

HARRIS SCHOOL WORKING PAPER
SERIES 06.05E

CHAPTER 4: THE S-DIAGRAM *

Howard Margolis

**This is a draft chapter from Cognition and Extended Rational Choice
(forthcoming Routledge 2007).*

Chapter 4 The S-diagram

We are better off with a government that collects taxes and provides public services than with no government and no services. But individually I would be even better off if the government collected taxes from everyone else, but somehow neglected to collect from me. The overall level of public services would look the same whether I paid or not (provided the rest of you continue to pay), but what I could spend on myself would be noticeably larger. So why pay?

In terms of what is still standard economic theory, the answer is not so obvious as might first be supposed. Of course, I can get into serious trouble by not paying. Yet the tax system depends on a very significant measure of "voluntary compliance" (Roth et al. 1988). More broadly, no society could survive if its citizens violated its rules whenever a favorable opportunity arose: that is, whenever the risk of punishment was small enough to make the gamble tempting as a matter of narrow self-interest. On the other hand, beyond small communities no society elicits such a high degree of voluntary compliance that it can get along without police and tax audits.

In general, understanding how societies function depends a great deal on how compliance with social rules is sustained and at what cost that compliance is secured. How that works will vary across sorts of rules and across conditions under which individuals might be tempted to violate the rules. Under some circumstances, as noticed earlier, normal levels of compliance can break down, although calamities (such as earthquakes) that produce social disorder in one community may elicit exceptionally cooperative behavior in another. There is always tension between forces favoring compliance with social norms and forces favoring narrowly self-interested behavior. Somehow a balance ordinarily prevails that allows a stable basis for social life. But that stability is not automatic or unchallengeable. Sometimes the balance is upset, and not always for the worse.

I want to construct an abstract model of how social coordination is usually sustained, but nevertheless is subject to change, and sometimes revolutionary change. In fact I will do that twice, interpreting the same geometry developed in the next section first (and briefly) in a way entirely independent of the NSNX argument, and then (in detail) in a way entirely contingent on that argument. The first reflects social contagion

effects which parallel those of epidemics, and which have been brought to especially wide popular attention through Malcolm Gladwell's *The Tipping Point*. But the principal concern here is with parallel dynamics that arise out of the NSNX tension between self-interested and group-interested motivation. So there are parallel processes, one whose striking effects are too visible to be doubted (contagion of epidemics and social contagion of fads and fashions) and the other not so easily seen (NSNX effects on social coordination and cooperation), since unlike fashions, motivation is not something on easily visible display. NSNX effects are unimportant for understanding cycles of fads and fashion, but the reverse holds for the effects of social contagion in NSNX dynamics. I will return to that point at the end of the chapter.

The convenient way to start is again by way of an analog to the trading equilibrium of a market. In the market context, the equilibrium that emerges at the social level (conspicuously, prices) is contingent on and interacts with the equilibrium personal spending choices of individuals. So the analysis of market equilibrium is built up from the properties of equilibrium individual choices. We want to consider an analogous dual equilibrium, with individuals allocating between social and self-interested use of their resources, yielding a social equilibrium contingent on but also interacting with those individual allocations.

The analytical device I will use is derived from Schelling's (1978) simple but very rich diagrammatic treatment of the relation between individual and social choice, but with various changes to suit the purpose here.

The Basic Diagram

Start from the standard model of strictly self-interested choice and from the stark case of strictly voluntary choices by these strictly self-interested individuals with respect to a pure public good.

FIGURES 4.1 - 4.6 SHOULD BE CLUSTERED & PROBABLY AS A FULL PAGE ABOUT HERE. SO THEY ARE REALLY ONE FIG. W/ 6 COMPONENTS WITHIN IT. I'VE PUT THEM TOGETHER INTO ONE FILE (N4.1)

The horizontal axis in figure 4.1 measures the extent of cooperation on a scale that runs from 0 (no one is making the cooperative choice) to 1 (everyone is). And the vertical axis indexes value to the *marginal chooser* at the corresponding point on the horizontal axis, contingent on that chooser's choice between free-riding (the FR curve) and cooperation (the C curve). In contrast to Schelling's original, where the chooser is everywhere an average chooser, here the marginal chooser would ordinarily be a different individual at each point. So (again in contrast to Schelling) it would be meaningless to interpret the curves as showing social value. The value to the marginal chooser is arbitrary. What counts is only the difference (Δ) between the FR and C curves. Ordinarily indeed, as the fraction cooperating grows value to each chooser also grows (the more cooperation, the better off for everyone). But that is not necessarily the case. Though we are starting here from a positive case, cooperation with a corrupt regime might be making (almost) everybody worse off.

The context could be one of seeking voluntary donations of money to some public function (say to a July 4 fireworks display or a political campaign), or seeking donations of time (say, to clean up a neighborhood), or sometimes the costs could be something subtler and in a dimension that involves neither time nor money, like offending old friends. The activity (taken to be beneficial but contrary cases are also important) could be anything that provides benefits to everyone in a group, whether or not they have helped provide the benefits. The relevant community (conspicuously so in a political campaign) might be only a segment of what in another context the chooser would see as his community. On almost any such matter there will be some temptation to go along for the ride, letting other people handle the costs. But if we all do that, nothing gets done, and we are all worse off.

If the group is small, and everyone pretty well knows what everyone else is doing, then the free-rider temptation may be small even if choice is strictly voluntary and self-interested. But as numbers grow large, anonymity and other advantages to self-interest make free-riding an increasingly tempting option.

As with markets, the analytics are simplest for cases with large numbers of actors, no single one of whom alone has a perceptible effect on the social outcome. But whatever the size of population, if there are N choosers each chooser represents a length

$1/N$ of the total space between 0 and 1 on the horizontal axis of figure 4.1. For large N , this width occupied by a single chooser would lie inside the thickness of any line we could draw. You can think of the lines in the figure as actually steps, but with the individual steps so tiny that it looks like just a smooth line.

Moving left to right in the Schelling diagrams as we will use them, we are always considering the person most easily tempted to change his or her choice (switch from cooperation to free-riding, or the reverse), given any particular level of cooperation by others. The propensity to cooperate = $C - FR = \Delta$ is a *net* value. If there are risks that go with free-riding or rewards to cooperation that a free-rider loses, that would shift the FR curve down, making Δ larger. The advantage a chooser would gain by free-riding is worth less if free-riding entails a risk of punishment or forgoes a reward. So Δ , if originally negative (favoring free-riding), may eventually reach 0 or change sign, prompting cooperation. We want to think about conditions under which that can happen, starting from the simplest possible case in fig. 4.1, so that Δ is everywhere negative.

In this large- N setting, individual choosers cannot see the individual social effect of their choice, as the same individuals cannot see the effect of their individual choices about private consumption on the prices and quantities of good in the market. But an individual can extrapolate what that effect might be from a judgment of the aggregate effect of "people like me" making the choice. And the effect on self-interest will be directly perceptible to the marginal chooser (if he should choose to contribute), namely the cost of his contribution, though perhaps adjusted because he gained some reward available to contributors or avoided some risk of punishment for not cooperating. But for the strictly voluntary case of figures 4.1 there is no prospect of reward or risk of punishment, so that the cooperate curve C is just the free-ride curve FR net of the cost of cooperating.

The only equilibrium is at the extreme left. As we move across the horizontal axis, the marginal chooser is changing. But whatever we might assume about the prevailing level of cooperation, that marginal chooser always does better to free-ride. Wherever we started, consequently, the prevailing level would move a bit to the left, where another person is the marginal chooser: and that person also will choose to free-ride. So wherever we are when the process starts, unraveling continues until we reach the

long-familiar result (Samuelson 1954, Olson 1965) that although everyone would be better off if the social choice were at 1 (100 percent cooperation), the only equilibrium outcome for strictly voluntary self-interested choice would be at the extreme left, where no one is cooperating.

But since actual communities commonly manage to avoid this dismal result, it follows that somehow there are commonly incentives present (beyond what can be accounted for in terms of strictly voluntary self-interest) whose aggregate effect is sufficient to offset the cost of cooperating.

Tipping Points and Equilibria

So next consider the simplest conceptually possible ways in which the curves of figure 4.1 might be shifted to allow a non-zero level of cooperation. Figures 4.2-4.6 show various possibilities, starting with the complete reversal of figure 4.3. This complete reversal could reflect, for example, some coercive threat sufficient to make every chooser see her self-interest as best-served by cooperating. Now the *cooperate* curve (C) dominates the free-rider curve. The unique equilibrium is at the far right, with 100 percent cooperation. But in realistic situations, what needs to be done to obtain increasing levels of cooperation almost always eventually become prohibitively difficult, so that the complete reversal (as in figure 4.2) will rarely be seen. But it is a logical possibility, and if the marginal cost of contributing is low it can even be observed, or at least closely approximated. Think of applause after a good concert performance. The applause is satisfying to the audience as well as the performer. It is a public good. You get to enjoy it even if you don't bother to applaud yourself. But although it is strictly voluntary, essentially everyone chooses to contribute, though not absolutely everyone, since there is a cost (getting caught in the crowd) unless you start to leave immediately.

Figure 4.3 shows a far more prevalent case in which the offsetting incentives are most effective when cooperation is low, with diminishing marginal effectiveness with increasing cooperation. The net advantage favors cooperation at the left of figure 4.3, but the advantage falls and eventually free-riding does better as the fraction induced to cooperate increases. Beyond the intersection the net incentives becomes increasingly negative (adverse to cooperation). Or, in terms of the relative advantage of cooperating rather than free-riding, Δ is positive at the left of the diagram but decreasing so that the C

and FR curves intersect. Since choosers are ordered by decreasing propensity to contribute, this situation easily emerges even if the cost of incentives favoring cooperation to the marginal chooser is constant.

Figure 4.4 shows the opposite case, in which the offsetting incentives are most effective when cooperation is high, so that the strong free-rider advantage when cooperation is low decreases as we move to the right, reaches 0 at the intersection, and thereafter the net advantage increasingly favors cooperation. This situation is not typical but also not terribly uncommon even when the cost of contributing is not trivial. For as fewer and fewer choose to free-ride, it becomes easier to identify severe shirkers, and more feasible to punish them. If N is not very large, this is a common situation, and it is approximately right even for many large N situations where the cost of cooperating is not very large but more than trivial.

The structure in figure 4.3 most easily arises when free-riding is not at issue. Using one of Schelling's examples, if there are two roads (A and B) connecting a pair of locations, drivers will distribute themselves to make travel time approximately equal either way, since as more and more drivers choose one road (increasing congestion) it becomes increasingly attractive to choose the other. Draw the situation as a Schelling diagram, with the declining line showing choices of road B and the increasing line showing choices of road A. With the horizontal axis showing the fraction choosing Road B, we get a version of figure 4.3, with an equilibrium near the middle. Here self-interest and socially effective choice are aligned, so there is no problem of how to reach a reasonable social outcome. What we want to understand is what happens, or might happen, to reach a reasonably good social outcome when they are not.

Figures 4.3 and 4.4 illustrate starkly different contexts for the evolution of the social result. In figure 4.3, where the cooperate curve (C) crosses the free-ride curve (FR) from above, the crossover identifies an equilibrium (Q). In figure 4.4, where the cooperate curve crosses the free-ride curve from below, the crossover identifies a *tipping point* (t), with two equilibria: one with zero cooperation, the other at 100 percent cooperation. To follow the balance of the argument, it is essential to be completely clear about why in figure 4.3 the crossover is an equilibrium, but in figure 4.4 it is a tipping point.

In Fig. 4.3, starting from any point along the horizontal axis, the social process would always move toward Q. For a chooser anywhere to the left of Q, the net advantage is always to cooperate. The C curve lies above the FR curve. The chooser would make herself better off if she chose to cooperate. So by construction, the marginal individual would always choose C, which moves the level of cooperation to the right in the figure (fraction cooperating is a bit larger than if this marginal chooser chose to free-ride) until we get to Q. But starting from any point to the right of Q, the opposite net advantage prevails, and again we would always move toward, never away from, the equilibrium at Q.

But for figure 4.4, the opposite holds. We would always move away from the crossover at t until we reached equilibrium, which here would be at the extremes. The only stable points are either 0 cooperation or 100 percent cooperation. In contrast to figure 4.3, where social evolution is always moving toward a unique equilibrium no matter where we start, the result for figure 4.4 would be fully contingent on events leading to the eventual equilibrium. Path-dependence would determine whether we ended up with no cooperation or full cooperation, though underlying individual preferences are identical for both Fig. 4.4 outcomes. Conceivably we could be exactly at the tipping point (t), like a pole balanced exactly on its end. That could happen, but the pole would not stay put very long, because the slightest perturbation would set things moving away from perfect balance, and the same would hold for the social analogue.

A simple example that approximates the figure 4.4 situation is driving on the left vs. driving on the right. It is socially best if everyone would adopt the same practice. But which possibility emerged would be path-dependent. If either way would work just as well there would be no social dilemma. But suppose that (as in Britain once the Chunnel made it much more common for British drivers to be on the continent and the reverse), it would really be better if the community switched from to driving on the left to driving on the right. How to get there would be a policy puzzle whose resolution (whether to try, how to do it) would be contingent how to manage what would necessarily be a path-dependent route to the alternative equilibrium. And devising a feasible path would be a puzzle with a free-rider component, which would come especially at the stage of enlisting political support for getting the change made, since a proposal to move would

not be immediately popular even if it was certain that within a few years essentially everyone would agree it was the smart thing to do. In the short run it would be unpopular, and ambitious politicians might not be willing to expend political capital on it, and especially so if they could not be sure their rivals would. I will return to this simple example in Chapter 5.

But until the change was made, a more exact diagram would be that in Fig. 4.5, where a considerable number of accidents occur because visitors to Britain sometimes forget the local practice until an oncoming car reminds them, sometimes too late. In Fig. 4.5 a visible fraction of people are not cooperating with the local practice (in the figure on the far right) even though it is in their own interest to do so. And of course there are always some people not conforming to local practice for nastier reasons. Although I have am using the label "free-riding" for non-cooperation that is to be understood (as the driving example requires) as just a label for not cooperating in some way, only sometimes, not necessarily, with the invidious connotation of free-riding.

Finally, we could get more than one crossover, and in particular (for the balance of the discussion) under conditions where cooperating or not cooperating with prevailing practice turns on conflicting motivation between self-interested and group-interested choice. A combination of decreasing marginal effectiveness for some incentives and increasing marginal effectiveness (up to some high level) for others could yield situations like that of figure 4.6, where there is a tipping point and two equilibria (as in figures 4.4 and 4.5) but now with neither equilibrium at the extreme of complete free-riding or complete cooperation. We can notice that this is an important configuration. For we can easily point to situations in which a practice had been rarely favored rather abruptly (relative to some usual timescale) turns it into the dominant practice in the community. But in real situations, there are usually some people making a choice when almost no one else is, and some who are not making that choice even when almost everyone else is. The high and low equilibria look like those in figure 4.6, not like those in Fig. 4.4 or 4.5.

Rewards and Punishment

So now consider characteristics of the requisite incentives to yield social coordination with respect to some choice if the ex ante situation was that of Fig. 4.1.

What might shift the curves of figure 4.1 to yield a high level of cooperation in that kind of context? Consider at first only incentives to individual choosers with given beliefs and strictly self-interested preferences. Assume there is no disagreement that the coordination at issue would benefit everyone, but the cost of participating is such that no one finds it worthwhile to bother in terms of their private self-interest. The possibilities for shifting the free-rider advantage then could come only from either imposing a risk of punishment on those who fail to cooperate or offering a promise of reward for those who do (Olson's "selective incentives"). We want to consider how the effectiveness of such negative or positive incentives might vary as we move from low toward high compliance. Or put another way, we want to consider typical variations in returns to scale of incentives.

Either rewards or penalties could work to push down the net advantage of free-riding. But workable incentives must yield more value than they cost to provide. Incentives to cooperate must provide something analogous to gains from trade. Some entity interested in influencing the result (call this entity the Agent: it might be the government, or an issue-oriented foundation, for example) will look for incentives which cost less to provide than the value to the Agent of the effect on cooperation they can secure. There are now three distinct "values" in play. The *value to the chooser* of the reward must be enough to offset the cost of cooperating. But this value has no necessary connection with either the *value to the Agent* of securing this chooser's cooperation or the cost to the Agent of providing the incentive. The value to the Agent must be enough to offset the *cost to the Agent* of providing the reward. But whatever that cost might be, we can expect to see typical patterns in how cost of incentives would usually scale with the level of cooperation they could be expected to achieve.

The outstanding example of something which costs Agent very little but might be of high value to a chooser is the honorary award. Honorary rewards – an invitation to a White House reception, a knighthood -- which cost essentially nothing to give may be highly valued by individuals, and indeed may have substantial economic as well as psychic value to some set of potential cooperators. But (conspicuously for honorary awards) the value of an incentive that costs little to give is tied to how rarely it is awarded. Positive incentives typically encounter either increasing costs to the agent

(since rewards whose values are contingent on their rarity are soon exhausted) or decreasing value to marginal compliers, or both. Awards with significant value to choosers even though widely distributed would rarely cost nothing to give. So marginal effectiveness in eliciting cooperation (the gain from the incremental cooperation vs. the cost of providing that incentive) for rewards will ordinarily decline as we move toward higher levels of cooperation. And, as noticed, this could be because the value of the incentive is contingent on its rarity, or because it is not cheap at all, hence can be provided only to some subset of potential cooperators judged to be of particularly high value, or because only a limited subset of potential cooperators are particularly enticed by this reward. Across the whole society, positive individual incentives with increasing or even constant net marginal effectiveness must be rare.

The more general situation is certainly that net value to an Agent of positive incentives diminishes as the needed level of cooperation increases. This implies that, considering positive incentives only, free-rider temptations overcome by positive incentives would have the general shape of figure 4.3, where the cooperate curve (C) is above the free-rider curve (FR) when cooperation is low but eventually falls below it. And for many contexts the limited opportunities for positive incentives to be workable will mean that the crossover point (Q) will be far to the left of the intersection shown in figure 4.3.

The situation is more complicated for negative incentives. With exceptions (such as honorary awards), which can be effective only if awarded to a few, providing positive incentives is costly, but the administrative costs are low. People who have earned a reward will not ordinarily conceal their behavior or evade attempts to supply the promised consequences. But of course the opposite applies for penalties. It is possible and sometimes actually happens that negative incentives are profitable for Agent. This is at least sometimes the case for fines, or denying ordinarily available benefits, even taking administrative costs into account. But that is not the usual situation. Indeed when negative incentives are cheap, or even profitable, something quite different from what is under discussion here is likely to be going on. What we would be seeing would mostly not be a matter akin to organizing efficient incentives to correct a market failure, but of predatory exploitation of vulnerable targets.

Much more usually, the various components of policing -- what it costs to detect noncompliance, identify noncompliers, and bring them to the point at which penalties have been imposed -- make negative incentives costly. But an effective negative incentive requires only some risk (not certainty) that not complying in any particular case will be detected and punished. Even if it looks like you could completely safely run a red light, so you see neither social nor private value in waiting, you are unlikely to do it. Sufficient risk to elicit compliance might be small enough that it could be reasonably cheap to create, but almost always only with the proviso that it does not need to be enforced very often. In the extreme case, if detection is easy and punishment reliable the negative incentive can be very cheap to provide, since it almost never actually needs to be imposed.

Just where negative incentives can be efficient (from the perspective of the Agent) depends on options that might be available and on penalties that can then be feasible. The agent can start by choosing particularly favorable targets (for example, people who are particularly easy to watch, or particularly risk-averse, particularly likely to be bound by habit if a habit of compliance can be established). But plainly a high level of cooperation cannot be gotten that way. Marginal costs then must rise as less and less vulnerable individuals must be covered. But negative incentives might work very well to maintain a high level of cooperation already in place, since now there would be few candidates for application of the punishment, and the costs of enforcement averaged across all cooperators might be very low. This would yield the Fig. 4.4 mirror-image analogue of the figure 4.3 situation. Now it is at high levels of cooperation, not at low levels, that free-riding no longer looks tempting relative to cooperating. In Fig. 4.3, rewards to the most easily tempted cooperators push up the Cooperate curve from the left. In Fig. 4.4 threats which become cost-effective only if compliance is high push down the free-rider curve on the right. But that opportunity comes into play only if somehow the situation has moved to the right of the tipping point. A high level of cooperation is sustainable, but only if Agent has somehow already gotten there.

Suppose Agent has invested in a fence, but many people take to scrambling over the fence. Unless Agent can start shooting -- there are cases in which Draconian measures are taken but more often that would be suicidal (politically and sometimes

physically) -- the incentives cannot be enforced. But when only a few people are trying to scale the fence -- when most people are cooperating (in the diagram when the status quo is well to the right) deterrence of all but the hardest cases might become cheap, given the initial investment in the fence.

And allowing for a fraction of especially hard cases at the right of the diagram, and starting from mainly easy to reward cases at the left, we could get a curve like figure 4.6, which exhibits the multiple crossovers mentioned at the end of the previous section. Mainly positive incentives elicit some modest level of cooperation, but decreasing marginal effectiveness limit what can be elicited. Then there is a range of cooperation over which it is too expensive to provide adequate positive incentives and impractical to impose effective negative incentives. But eventually, if participation were sufficiently high, negative incentives can become effective. So if we could somehow get beyond a tipping point, we would be in a region where potential free-riders are successfully deterred, where coordination could move up to a point where increasingly intense efforts to apprehend and punish increasingly reluctant non-cooperators exhausts what is practical.

Yet so long as we are dealing with strictly self-interested choice, this logical possibility is inadequate for the level of cooperation with social goals we see in actual societies. When there is an emphatic alignment of self-interest and social value (as, using again one of Schelling's examples, with everyone driving on the same side of the road) a satisfactory social outcome can emerge spontaneously. That happy result is in fact not uncommon, but it is far from reliable enough to set aside what remains. Societies on a large scale that work always rely on more than self-interest. In many contexts self-interested choice would be dominated by free-rider temptations (absent further incentives the situation would look like figure 4.1), or there is a visible positive equilibrium as in figure 4.3 but dangerously far to the left, or comfortably far to the right as in figure 4.5 or 4.6 but there is no way to get there. Then feasible strictly self-interested rewards and punishments might be catastrophically insufficient.

** All that is so even though, as mentioned at the outset, there is a familiar class of cases where self-interest alone is sufficient to get past a tipping point. For fads and

fashion (and not entirely differently, for the far more fundamental case of the evolution of language), all that is needed to get a version of Fig. 4.6 is that people vary in how far they are moved to do what nearly everyone else seems to be doing. Everyone has a tendency in that direction (most of the time even a determined non-conformist is in fact conforming), but with variation across issues and persons. For matters of mere fashion, the basic story requires only that there be some segment of the community that wants (on this matter) to be "ahead of the crowd", a much larger segment which want to keep up with the crowd, and a segment which wants to stay with what they liked in the past. We can then see the influence of the social contagion that is the focus of Gladwell's book, where the formal dynamics resemble those of epidemiology. A novelty gradually spreads but eventually dissipates as the conditions for contagion weaken (for fad or fashion, its novelty appeal ages with those most likely to be motivated to be ahead of the crowd).

But if the novelty grows sufficiently before starting to fade, and in particular if by chance or design (usually needing some help from chance) new centers of contagion are ignited, it may reach a critical mass or tipping point where the number of new cases accelerates and the novelty quickly spreads across a large fraction of the whole community. That happens when it is no longer just specially sensitive or specially located people who are exposed but essentially everyone in the community. Then all face multiple exposures and all but the most resistant become likely recruits. Gladwell is very good on this.

Sometimes ideas and passions, not just fads and diseases, can take hold in that way. We are interested in the contagion of ideas and passions, but especially in cases where the process in fact requires more than self-interested motivation, such as adopting or complying with social norms even when you could get away with not bothering, or joining a demonstration defying your government, even when you might get shot. What needs to be considered is how (for example) an Agent who sees how to sustain a low equilibrium (with rewards) and how he could sustain a high equilibrium (with punishments) might somehow get over the hump, reach a tipping point and then the high equilibrium. Even when self-interested motivation is not enough, an essential contribution is likely to come through the contagion process just described. But that part of the story is most easily seen when self-interest alone is sufficient. And even when

something more than strict self-interest is at issue, if the cost to an individual is low, the negative incentives needed to sustain high cooperation could be cheap. The tendency to want to be like other people, or merely to feel uncomfortable not being like other people, might be widespread enough on the issue at hand that if a tipping point were crossed the punishment for not cooperating might need to be only the embarrassment that deviants feel on being seen as deviants (for once cooperation is high, non-cooperators become deviants).

So we can give an account of how social equilibrium can shift where for a few people the spontaneous rewards of being ahead of the crowd generate the incentive to produce a seed for social change at Q^- , and then occasionally a sufficient number of ignitions to reach a critical mass might occur in some chancy way, so that once beyond the tipping point the spontaneous punishment of being left behind the crowd is sufficient to maintain a high equilibrium at Q^+ until a gradual erosion of conformity opens the way for another cycle to displace the current fashion. And (conspicuously in the case of the fashion industry) that can be speeded up and even routinized when agents with a serious interest in promoting the cycles come on the scene and resources become available to sweeten the bare incentive to be ahead of the crowd, or at least not behind the crowd.. When what is needed is only coordination that no more than mildly conflicts with self-interest, that may be as much of a story as is needed. But by itself this becomes increasingly inadequate to account for large scale cooperation when there is at least a phase where self-interest puts a substantial barrier in place. Nor would it usually be adequate to explain how cooperation is sustained past the tipping point in the face of inevitable transient shocks that can erode it.

We can see that societies can sometimes get there, and stay there. Indeed societies that survive somehow do get there on numerous aspects of social life, and even revolutions and other social movements facing strong opposition (so there are incentives, perhaps severe, to discourage participation) sometimes get there. To capture that possibility we need to consider motivation beyond self-interest. A second tier of what I will now call the S-diagram (because it is like a Schelling diagram and derived from it, but different enough to need a different label) comes into play. This arises from the possibility that people motivated to be "neither selfish nor exploited" might see sufficient

social value in the cooperation at issue, or (contrariwise) might see that cooperation as sufficiently socially perverse, that in addition to the social contagion effects that can generate a Schelling diagram like Fig. 4.6 a second set of factors come into play which also generate that structure.

Social motivation

Rewrite the NSNX equilibrium defined in Chapter 1 ($W = G'/S'$) as $WS' = G'$. We can then reinterpret the C and FR curves of figure 4.6 as reporting G' and WS' for the marginal chooser. Call this the *S-diagram*. See the enlarged copy is printed as Fig. 5.1 at the start of the next chapter. The marginal chooser will now contribute to whatever cause is at issue (donate to the fireworks, spend time on the neighborhood cleanup, join the protest march) if $G' > WS'$, and won't if $G' < WS'$. As before, the slopes of the curves are arbitrary. What is relevant is the vertical distance (Δ) between the lines, which scale with how much change in incentives would be needed to change choices. The discussion of Schelling diagrams to this point can be straightforwardly applied, with rewards and punishments (positive and negative incentives) still affecting choice directly by altering the S' factor as already described. But now we are also interested in possibly altering W and G' .

FIGURE 5.1 (used throughout ch 5) SHLD COME (PERHAPS FULL PAGE)
IMMEDIATELY FOLLOWING CH 4

Even prior to this section we were already concerned with more complicated situations than Schelling had occasion to explore. In creating the diagrams, Schelling dealt only with average or typical choosers. But for the fad-and-fashion models to yield striking effects heterogeneity across choosers is essential. There have to be some people who care about being ahead of the crowd, some people who can afford to be different, many people who mainly do not want to be left behind. Allowing for that gives up some features of Schelling's original. And extending the discussion to social choice shaped by NSNX motivation sharpens the departures.

We need to allow for preferences about the choice at issue that not only vary from one chooser to another but often conflict. Even with adjustments for possibly transient side-payment.(rewards and punishments), WS' and G' are not entities which can be the

basis for showing aggregated social or private values even of the marginal chooser (in particular they can be influenced by tactical considerations, transient public excitement, and so on). I show the curves with an upward trend, suggesting that more cooperation is better, as is usually the case. But the absolute positions of the two curves carries no information. What is significant is just the way Δ changes in relation to participation, which yields group-interest choice from the marginal chooser when $G' > WS'$ is positive and self-interested choice when $WS' > G'$. It is contention over social values and competition between social and private values that can be analyzed in terms of the S-diagram, not social value itself. Unsurprisingly, the S-diagram does not capture everything.

Contention over social value has played no role until now. In large-number contexts, that would be so for practical purposes (on the argument about voting reviewed in Chapter 2) even though a self-interested chooser might have a high valuation of cooperation. For a self-interested chooser could not be significantly motivated by the effect of her individual contribution on the cooperative endeavor. She might strongly believe one outcome over another (in an election, for example) would have great social value, but in a large numbers context, the marginal value for herself of her own choice, even in terms of even some notion of enlightened or enlarged self-interest, will still be inconsequential.

What makes the difference between the discussion here and a discussion in a more standard framework is not that for a self-interested chooser G' would be zero. Even narrow self-interest is not inconsistent with having a perception of the social importance of a policy or electoral choice, and as mentioned at the outset, current versions of maximizing a utility function often go beyond narrow self-interest. What makes G' inconsequential in a model of even broadened self-interest is not that it must be zero, but that in terms of what a consistent (rational) self-interested chooser would be willing to pay to have their preference included in an overall social weighting it might as well be zero. It is not zero, but in a large-number setting it is too small to make a difference big enough to motivate an individual to participate. But if NSNX holds, that would no longer be so, as discussed in Chapter 2. And we can observe that actual choosers who understand the rational choice argument that it cannot be rational to vote nevertheless

may find themselves voting anyway. This yields two inferences. Even if NSNX does not hold, people nevertheless seem to behave *as if* it held. So perhaps it does hold. And even if NSNX holds, that does not mean that choosers have any conscious access to access to the NSNX equilibrium condition. Feeling selfish or feeling exploited is accessible. Coming closer to $W = G'/S'$ is not, any more than in a Darwinian world agents are consciously maximizing fitness.

As before, we are lining up choosers such that the person most easily tempted (or coerced, as mentioned in Chapter 1) into cooperating is at each position on the horizontal axis. So the very first person ready to join in some effort is at the extreme left and the very last holdout is at the extreme right. With $\Delta = G' - WS'$ as the incentive to cooperate, as we move from left to right along the horizontal axis, the marginal chooser is always the person for whom the absolute value of Δ is the smallest. If a free-rider, it is the free-rider who can most easily be moved to cooperate; if a cooperator (Δ will now be > 0), it is the cooperator most easily tempted to defect.

If all choosers are identical, or the curves represent the values of the average chooser, then the aggregate value of a particular level of cooperation to the whole community could be calculated by multiplying the value to a cooperating chooser by the fraction of cooperating choosers, and adding to that the value to a free-rider multiplied by the fraction of non-cooperators. But here the values are increments as judged by the marginal individual at each point along the curve. The summation just described would be meaningless. I show the curves with an upward trend, suggesting that more cooperation is better, as is usually the case. But the slopes are really arbitrary. All that is meaningful is the way Δ changes.

Could the same individual turn up as the marginal chooser at more than once? That will sometimes happen. The most obvious case would be a contrarian who wants to do whatever is the opposite of what most people are doing. And although I will not explicitly deal with the point (and the diagram as drawn makes no attempt to show it), choosers near each other, and with nearly the same Δ 's, might be responding to very different combinations of G' and S' and in subtler ways even W . It is convenient and harmless for the discussion here to keep the diagram simple. But it is only Δ that should be thought of as changing by small enough increments (across marginal choosers) to look

continuous in the diagram. For equal values of Δ might come from adjacent choosers one of whom has a higher value of G' but also a higher value of S' . But dealing explicitly with these and various other complications I will set aside would not change anything substantial in the discussion here.

An S-diagram could even look like figure 4.1, where Δ is always negative. There now is no seed visible. Whatever motivation beyond self-interest is in play is everywhere too small to noticeably overcome the temptation to free-ride. This in fact is common. Endless possibilities for social coordination at least a few people favor which never get off the ground. Or the situation might be like that of figure 4.4, where until others are cooperating even those most ready to cooperate require a large incentive beyond what is at hand. But if the individual most easily moved to cooperate was indeed moved by some new incentive (a bribe or threat or maybe even a revelation) or by a change in circumstances, the next most ready person would (by construction) require a bit milder bribe or threat or revelation. But until one of those possibilities arises even the most easily moved chooser never in fact moves.

The Δ for a particular chooser will depend on that chooser's own sense of the marginal social value of the marginal choice at issue (G') and on the cost to that chooser of cooperating (S'), both net of adjustments already described. S' for the chooser would respond to both the direct cost of contributing and to chooser's individual vulnerability to any risk from free-riding and to this individual's valuation of any rewards from cooperating. G' would respond to the chooser's sense of the social value of chooser's own increment to cooperation on the matter at hand, allowing for any costs to the group in reaching that cooperation and for favorable or unfavorable effects on future cooperation. With or without adjustments, G' would be a very tiny increment on the scale of social value, but the individual cost to the chooser would also be very tiny on that scale. The relevant ratio, G'/S' , need not be small at all. See the discussion here in Chapter 2 and the more detailed discussion in Chapter 6 of my 1982 book. And Δ would depend also on the weight (W) given to self-interest in this choice, as governed by the NSNX equity rule. An Agent who wants to discourage, or an opposing Agent who wants to encourage the cooperation at issue will look for opportunities to shift the curves through effects on G' , or on S' , or on W , and usually on all three together.

For the fireworks example, as with many others, there may be no significant contest about what would be socially valuable. Essentially everyone agrees on what would be good, and rewards and punishments are only relevant when purely voluntary action will not produce enough of that good thing. But many choices are contested, which means that some actors are interested in encouraging cooperation and reducing free-riding, but others have an opposing interest in discouraging cooperation on just that matter. So actors on both sides will be interested in how to encourage potential supporters and discourage potential opponents.

Conflicting preferences among Agents can then yield competing promises of rewards and threats of punishment, possibly including threats or promises affecting G' rather than S' . Sometimes incentive effects on G' will have large consequences. A threat, after all, may target a group or a community not just an individual chooser contingent on just her choice. Or if the cooperation at issue is to pursue some reform, then an Agent opposed to that kind of cooperation might try to reduce G' by some gesture that seems to favor the reform. So the propensity to cooperation might be reduced by individual rewards to non-cooperators or punishment for cooperation (either, if effective, increasing net S' , hence reducing Δ). But Δ might also be reduced by punishment that targets the group, hence reduces G' indirectly by adding a component of loss to the group, or reduces G' directly by some partial reform that makes the object of cooperation seem less urgent. When we come to discuss concrete cases, examples will readily arise.

With that framework, start again from the case of purely voluntary action, considering what we might treat as the typical shape of a diagram reflecting the NSNX social effects when *no* private incentives due to an active agent are in play. As in the earlier discussion of purely self-interested social contagion, but now setting aside effects of social contagion, we are again led to the S-diagram.

What we are looking for is some "in general", or usual way in which the W , S' and G' would vary with respect to an interesting issue, conditional on the level of participation by other people. We would like to define a base case that will work across a wide range of contexts, and from which occasional departures from the "normal" configuration could be assessed. Within standard analysis of markets, the basic supply/demand diagram (with downward sloping demand curve and upward sloping

supply curve) provides a base case which continues to be useful as a point of departure for interpreting special contexts where the usual interactions are violated (by snob or bandwagon effects, for example). We want to establish a parallel to that.

Start from the situation of "most people", as distinguished from the small fraction of people who are joiners when virtually everyone else is a free-rider -- or not even a free-rider, since perhaps they are not yet even conscious of anything they might be free-riding on. At the other pole, there are always at least a few people who do not go along even when everyone else is. In this respect we are running parallel to the fads-and-fashions context, but here what makes the distinction is not that some want to be ahead of the crowd and others (on this matter) want to just conform with the crowd but that some are aroused to a social opportunity or risk while others are not. In between the poles of some ordinarily small fraction easily aroused and some also ordinarily small fraction irrevocably opposed, tautologically, are most people. In Fig. 5.1 this is the range along the horizontal axis some distance to the right of Q^- and across to some distance to the left of Q^+ . And for "most people", the W , G' and S' functions can ordinarily be expected to follow a usual pattern. Weight to self-interest (W) must go down as participation by others increases, and must go up when as chooser's own participation increases. Define g as the current ratio of group-interested to total use of resources for a chooser. And define g^* as the typical value of g for what that individual, in this context, recognizes as "people like me". For any given g^* , NSNX equity (Rule 2, p. --) requires that W increase as the chooser's own group-interested spending fraction, g , increases. But for any given individual, W becomes smaller as g^* grows larger. So for most people, as what "people like me" are doing grows more social, a chooser becomes less easily concerned with being exploited by contributing, and more easily concerned about being selfish by not contributing. Although NSNX mechanics are out of sight to a chooser (NSNX does not assume conscious calculations of the state of the equilibrium condition any more than standard modeling assumes conscious calculation of marginal utilities) these affective responses are not always tacit, and I doubt that any reader is unfamiliar with them.

So directly from Rule 2, weight to self-interest, W , will slope down as we move left to right across the "most people" range in the S -diagram, since exactly what that means is that more people "like me" are contributing. Recalling the remarks of Chapter

1, it is an affective, not a rational choice, issue to say what governs what a person treats as “people like me” in a particular context. But for this general discussion, we can just say that for "most people" the reference group in a situation is just about everyone facing the choice, or at least everyone who is not conspicuously *not* "like me". To the degree that most are cooperating, W declines and the converse.

On the other hand, basic S' , setting aside individual incentives or a conformity effect, is the cost of cooperating, which for the base case we are defining can be taken as constant across the range of participation that covers "most people". But Recall that a chooser's sense of S' must reflect three kinds of effect on private interests. There will be the direct (out-of-pocket) costs of contributing and there will also conformity effects. But participation often has private costs or benefits beyond these, most conspicuously so in violent situations where there are determined partisans with respect to the social cooperation at issue. A person taking part might be shot or otherwise harmed (his house destroyed, his job lost, and so on). Hence aside from conserving the direct value of resources that would be needed to participate, a chooser who free-rides will also gain whatever there is to be gained by avoiding the risks of participation, but lose whatever might be gained by rewards for that participation or by avoiding penalties that might be inflicted for not participating. Net S' might be complicated by competing incentives from opposed interests.

But Δ for "most people" must ordinarily decrease as participation by others increases even aside from the conformity effect. If tens of thousand are in the streets, the government will have to be more cautious about ordering that demonstrators be shot than if there are hundreds. And even if the order is given, and obeyed, the probability that any particular demonstrator is shot will be far less when there are tens of thousands than when there are a few hundred. And the risk that friends and neighbors will notice and resent and otherwise behave in ways that might prove costly accordingly grows larger when more are participating. Net S' , even aside from any conformity effects, will almost always grow smaller when larger numbers are participating.

When incentives conflict (the government is discouraging participation, the opposition is encouraging it), there may be penalties for not participating from one side and rewards from the other. Both affect choice, but unless an Agent who wants to

discourage participation is in a position to apply and willing to apply ruthless force, the net of these side-payments will usually favor doing what most others are doing. And indeed if participation becomes very prevalent, a different kind of reward component appears, as participation becomes something of a consumption good: not only would the mere conformity effect be high, but out there with the demonstrators can become "the place to be".

And since W , by Rule 2, slopes down, WS' for "most people" will be compounded of two downward sloping factors, and so certainly must ordinarily slope down. And when we come to consider NSNX incentive effects (as well as conformity effects) we will see that this inference inevitably becomes stronger.

G' , on the other hand, can be expected to increase for "most people" as participation by others increases. For as others join in, a typical individual will feel increasingly reassured that there is real social value to this effort, that social cooperation is succeeding, that "everyone knows" it is important that people support it. The combination of a downward slope for WS' and an upward slope for G' yields a *potential* tipping point in the central zone of figure 5.1, where "most people" are. To the left of t in the figure, cooperation will unravel towards the low equilibrium at Q^- , to the right it will cumulate toward the high equilibrium at Q^+ . And that there will be an actual not merely potential tipping point in a situation becomes more likely when we now consider incentive effects (rewards and punishments).

In sum, then, with respect to "most people" S' net of side-effects must ordinarily be declining between Q^- and Q^+ . And as already noticed, directly from the NSNX equity rule 2, W must decline. Hence odd cases aside, WS' must certainly tend to decline. But as discussed, G' will tend to rise with increasing participation and ordinarily be at least stable. Hence for most people, as participation increases $\Delta = G' - WS'$ must ordinarily be increasing (for free-riders, Δ is coming even less negative, for cooperators it is becoming more solidly positive). We get the geometry of the S-diagram.

But all this has been setting aside conformity effects which are apparent across the entire taxonomy of choice (fashion, language, and anything else that might be mentioned). Allowing for conformity effects, S' would be higher when the prevalence of cooperation is low, since there is some discomfort in behaving differently from others,

and S' would be lower when prevalence is high, and a person is not only comfortably choosing in what is the usual way but also has minimal decision costs since he can see that "everyone knows" what to do, so he does not have to think about it. How steep an allowance for mere conformity ought to be must depend on how costly cooperation is on the matter at hand. But on overwhelming evidence, it will rarely be inconsequential. And it reinforces the tendency of S' to decline across the range of "most people", reinforcing the point of the previous paragraph.

Movers and shakers

So now turn to the choosers who are not like "most people". In figure 5.1 they are near or to the left of Q^- and also sometimes of particular interest, near or to the right of Q^+ . Consider first the situation on the left of "most people" in the figure, where participation is very low. To elicit participation, people near this pole must have (relative to most people when participation is low) some combination of low weight to self-interest (W), high G' and low S' . But there need be -- indeed, since they must be atypical in some significant way, there only could be -- only a relatively few such people providing what I label the *seed*. Positive Δ ($WS' < G'$), favoring cooperation, could result from many different combinations of sufficiently high G' and sufficiently low W and S' . So it is the existence of a region near 0 participation where there are choosers with positive Δ that creates the seed, not some particular combination. Indeed, as will play an important role when we consider some exemplary applications in Chapter 11, similar contexts might have contrasting seeds, and even within a given situation there may be important heterogeneity within the seed. But all should be characterized by some combination of plausible values of W , G' and S' that distinguishing them from "most people".

High G' could reflect special knowledge, or special ability to influence outcomes, or longer than typical time horizons, or special experience. A person with special expertise who knows (or at least believes he knows) what is of special importance can have a strong sense of social value for (high G') for some project that "most people" would never think about, as could a person with powerful religious or political or social commitments: all these provide examples of individuals who might see G' for some social

activity as very much larger than a typical person. These are also people who would be more inclined to act in terms of where an initial commitment might eventually lead even if in the short run there is no visible social effect.

On the other hand, S' can be small for many of these same people. If you are rich, the private sacrifice in contributing money to help get something going may be not merely tiny but zero: a matter of allocating among social causes not of sacrificing anything that would otherwise be used for private needs. If you are a celebrity, even if not rich, all you may be contributing is your name, not involving any private sacrifice. A person with a sense of a calling -- a religious or political revolutionary -- may put very little value on what would seem enormous sacrifices of private values to another person, especially when there is a prospect of life chances if the effort succeeds that seem entirely out of scale with ordinary lives. (The counterpart in terms of self-interested motivation might be found, for example, among the 100 or so people a year who pay about \$70,000 to spend some days in severe discomfort climbing Mt. Everest, and do so knowing that almost every year some of them will die miserably along the way.) And sometimes there will be an interaction between S' and G' . A celebrity who can command media attention, or a rich person who can not only do a lot with her own resources but who is well-connected to other well-endowed people can sense a high G' for what (for them) is only a minimal cost. For they might see a multiplier effect of their own effort on engaging the efforts of others.

Or on a much less grand scale, there are a great many "local" cases of social action, where S' may be low because a person can do as well, or almost as well, for their private interest by the social commitment as they could by attending only to private interests. If you are an insurance agent in a small town, the private cost of working on town concerns, organizing a boy scout troop, and so on are plainly diminished by the obvious point that such activity can hardly be anything but good for business. So although the commitment will ordinarily be real, and go substantially beyond what would be warranted merely by being good for business, the effect on diminishing the opportunity cost of social behavior need not be trivial.

Further, for some people near the zero participation pole, even W can be atypically favorable (low weight to self-interest). W responds to how "people like me"

are choosing, which for people at the extremes of the social distribution may be quite different from how most people see that. Aristocrats, revolutionaries, or (more commonplace in our own society) leading professionals in any field commonly see themselves as special, hence come to see as a proper reference group how other special people -- people like themselves -- are behaving. How ordinary people are choosing must have a constraining effect on this, since the elite group as a whole must not have a shared sense of being exploited or (on the NSNX argument) they would shrink back from what would then come to seem an excessive commitment. But we can expect to see a constrained but still substantial special "people like me" effect on W near the low participation pole. Most people, for these people, are not "people like me".

Members of an elite group will see other members of that elite as "people like me", and participation will be easier to organize within such a grouping of people who often know each other and almost always know of each other, and if not easily find mutual connections. And (not always crucial but also rarely negligible) there will also be opportunities to lighten the burden by providing special benefits ("perks") to members of that elite, and special costs to someone who would not want to be ill-thought of within that elite, all of which reduce the opportunity cost of joining the effort (S'). So across the whole range of situations that turn on promoting social cooperation, from revolutions to small town civic improvement, it is not hard to see how a small core of initial support might arise, where some sufficient combination of G' , S' and W are atypically favorable to participation, yielding the region of early cooperation (the seed) at the left in figure 4.6. In a variety of ways, varying hugely across individuals and contexts, NSNX conditions can arise sufficient for the small number of people needed to "get something going" when almost no one else is attending to the matter.

Of course, on many matters of conceivable social cooperation there is essentially no seed. Nothing is going on, and until a seed grows to noticeable size, nothing is likely to happen. The S-diagram, we can say, is often degenerate. But when a seed is apparent, by construction as we move away from the left pole in fig. 5.1 we encounter choosers for whom some combination of W increasing (marginal choosers more readily see themselves as like those not yet active than like those who are active), S' increasing (the cost to self-interest is not so readily borne) or G' decreasing (the marginal chooser is less

likely to see effort on this issue as far more important than most people realize) yields a declining value of Δ . Eventually we must reach a point where the WS' and G' curves cross, beyond which the incentive to free-ride becomes stronger (Δ continues to decline even after it has already turned negative until the general tendency for Δ for "most people" to shrink in the opposite direction takes hold. This yields the low equilibrium at Q^- , and to the right increasingly reluctant marginal choosers until that tendency is overcome, once we are clearly into the region of "most people", by the characteristic decline in W and S' and increase in G' as participation increases.

And when conditions are sufficiently favorable, since as I have mentioned there is no guarantee, eventually this may yield a tipping point at t , where the curves again cross. But further to the right eventually we begin to encounter marginal choosers who are now no longer like "most people" so that again there is a shrinking of Δ to creating the high equilibrium at Q^+ as we increasingly encounter people who unlike "most people" exceptionally resist cooperation on this issue. To the right of Q^+ we then have the *holdouts*, forming a mirror-image of the *seed*. Here are people who will not be won over to cooperation even if almost everyone else has already joined in. These holdouts might be people alienated from the community, for whom G' is negligible, or even negative, even when almost everyone else is going the other way. And (partly the same people), there will be choosers in desperate positions (for whom S' is exceptionally high), and choosers who have opportunities to escape social sanctions that are not available to other people, because they have access to mountains or borders, or have criminal connections, or in some other way can more easily evade risks imposed on those who avoid what almost everyone else accepts (so although negative incentives can be intensely targeted they are hard to reach). And at that right-hand pole there will also sometimes be a group of special interest for us: people who are in fact the counterparts of the social activists at the opposite pole but with contrary views of what would be good for society. Individual choosers -- and more significantly, on some occasions a substantial fraction of all individuals -- may see social value in defying a policy, hence a social loss in compliance. We then see the atypical but familiar and important case of principled civil disobedience. Then treating noncompliance as just another form of free-riding would miss the point. But the analytics are just those already sketched. For these non-compliers, G' for the

cooperation with established authority is negative, so of course they would want to avoid cooperating: a point which came up in Chapter 1 (p. --).

The prospects for getting beyond the tipping point will be contingent on how far to the right participation must grow to reach a tipping point and by how large deficit Δ is over the interval from Q^- to t . Jointly these two elements determine the size of the *hump* that must be gotten over. But once beyond t , participation will spontaneously expand as far as Q^+ , providing a *cushion* that now protects the new status quo against reversal. But the process is reversible in principle and sometimes in practice, with the former Soviet Union providing a spectacular example.

Reviewing the labeling of features in the generic S-diagram: The WS' and G' curves typically cross at the three points already defined, $Q^- > 0$ and $Q^+ < 1$, with a tipping point (t) between. On the left, a low participation equilibrium at Q^- marks the participation level currently achieved by a group of early activists. Between the left pole and Q^- , the G' curve is above the WS' curve. Participation will grow (the marginal chooser will join in rather than stand aside) up to Q^- . But to the right of Q^- up to t , the reverse holds. As things stand, participation will not grow beyond Q^- . If some temporary shock shifts the curves -- for example, something prominent in the news shifts the G' curve upward for a few days -- that would move participation to the right. But unless it moves things so far right as to get beyond the tipping point, the increase in participation will be only temporary and participation may return to the equilibrium at Q^- . On the right, there is a high participation equilibrium at Q^+ , with the characteristic downward tendency of the WS' curve I have described in between Q^- and Q^+ . If the status quo is at Q^+ , then again, unless something changes sufficiently to push participation below t , we will stay there. High cooperation can be sustained even though in terms of self-interest free-riding temptations should erode it away.

But this means there is an intrinsic potential for instability built into the S-diagram. As things stand, the equilibrium will stay low, or stay high. But if things change, a radical shift can occur. That is appropriate, since we can observe that potential instability as a feature of the world. Even for quite mundane sorts of emergent social cooperation, we routinely talk -- and talk from experience in the world -- of reaching a

"critical mass", "getting the ball rolling" and so on. And for more dramatic situations, such as revolutions and outbreaks of ethnic violence, instability is even more strikingly observed -- not routinely, which it certainly is not -- but often enough to be a major feature of human sociality.

Specifically, if the curves go as in figure 5.1, and participation is high (the equilibrium is at Q^+), a disturbance that pushes compliance down to the left of t will lead to unraveling. Unless some further consideration intervenes (for example, effective imposition of martial law) compliance would fall all the way back to the inferior equilibrium at Q^- . So we want to think about the social processes that could affect the shape of the S-diagram and in particular yield movement in either direction across the tipping point. If the social equilibrium is at Q^- , we want to see how a society might reach the high compliance equilibrium at Q^+ , since we can observe that sometimes happens. And, if Q^+ is reached, we can observe that that favorable (or at least favorable for those in power) equilibrium could still be vulnerable to some shock with potentially radical consequences for the society. For sometimes there is a collapse of established authority.

And indeed, although compliance with the social order is ordinarily high, we do occasionally see striking shifts. Apparently something like the unraveling implicit in the existence of tipping points actually occurs. Sometimes (as in Iran at the time of the fall of the Shah, or more recently throughout Eastern Europe) these shifts in compliance with the social order develop suddenly enough to surprise knowledgeable observers inside and outside the country (Kuran 1989). Thus certain features of observed social processes suggest that diagrams for real social situations might often take the form of the S-diagram. A theory whose logic generates a structure which exhibits these features might yield insight into real cases.