1-2005

The Emperor's Dilemma: A Computational Model of Self-Enforcing Norms

Damon Centola
University of Pennsylvania, dcentola@asc.upenn.edu

Rob Willer

Michael Macy

The Emperor's Dilemma: A Computational Model of Self-Enforcing Norms

Abstract
The authors demonstrate the uses of agent-based computational models in an application to a social enigma they call the "emperor's dilemma," based on the Hans Christian Andersen fable. In this model, agents must decide whether to comply with and enforce a norm that is supported by a few fanatics and opposed by the vast majority. They find that cascades of self-reinforcing support for a highly unpopular norm cannot occur in a fully connected social network. However, if agents' horizons are limited to immediate neighbors, highly unpopular norms can emerge locally and then spread. One might expect these cascades to be more likely as the number of "true believers" increases, and bridge ties are created between otherwise distant actors. Surprisingly, the authors observed quite the opposite effects.

Disciplines
Communication | Social and Behavioral Sciences

The Emperor's Dilemma: A Computational Model of Self-Enforcing Norms[1]

Damon Centola, Robb Willer, and Michael Macy
Cornell University

The authors demonstrate the uses of agent-based computational models in an application to a social enigma they call the "emperor's dilemma," based on the Hans Christian Andersen fable. In this model, agents must decide whether to comply with and enforce a norm that is supported by a few fanatics and opposed by the vast majority. They find that cascades of self-reinforcing support for a highly unpopular norm cannot occur in a fully connected social network. However, if agents' horizons are limited to immediate neighbors, highly unpopular norms can emerge locally and then spread. One might expect these cascades to be more likely as the number of "true believers" increases, and bridge ties are created between otherwise distant actors. Surprisingly, the authors observed quite the opposite effects.

> Naturally, the best proof of the sincerity of your confession was your naming others whom you had seen in the Devil company.
> –Arthur Miller, 1996

THE POPULAR ENFORCEMENT OF UNPOPULAR NORMS

In "The Emperor's New Clothes" Hans Christian Andersen ([1837] 1998) tells the story of three rogues who sell a foolish monarch a nonexistent robe that they claim cannot be seen by those who are "unfit for office" or "incorrigibly stupid." Fear of exposure leads the emperor, and in turn, each of the citizens, to express admiration for the new clothes, which then

[1] The authors wish to express their gratitude to Andreas Flache for help with Voronoi diagrams and to the National Science Foundation for their support of Centola through an IGERT fellowship in nonlinear dynamics and Macy and Willer through grant SPS 0241657. The authors also thank the Fetzer Institute for support of Macy and the Javitz Foundation for support of Willer. Direct correspondence to Damon Centola, Department of Sociology, Cornell University, Ithaca, New York 14853. E-mail: dc288@cornell.edu

0002-9602/2005/11004-0005$10.00

reinforces the illusion of widespread support for the norm. The spell is broken when a child, innocent of the norm, laughs at the naked old man.

It is not hard to find everyday examples of this fable in the academic kingdom. We can all think of prestigious scholars who are widely proclaimed as having the most brilliant new ideas, yet privately, people find the work entirely incomprehensible. Some may worry that perhaps they are indeed inadequate—that those who cannot see these beautiful ideas must be "incorrigibly stupid." Others are quite certain that the emperor is naked but worry about being dismissed as an intellectual lightweight by enthusiasts who clearly seem to understand and appreciate every word. The safest course is to go along with the charade and admire the emperor—thereby reinforcing this same false belief among our colleagues.

The problem is not limited to faculty. Studies of campus attitudes toward drinking find that students anticipate negative social consequences for failing to participate in drinking rituals that celebrate intoxication as a symbol of group identity, especially in fraternities (Nagoshi et al. 1994; Perkins and Wechsler 1996; Baer 1994; for a review, see Borsari and Carey [2001]). Yet Prentice and Miller (1993) found that students were privately less comfortable with alcohol use than they (falsely) perceived other students to be. The study suggests that, contrary to campus legend, students are actually somewhat uncomfortable about excessive drinking, at least when they are sober.

According to Prentice and Miller (1993), students in their college drinking study are victims of "pluralistic ignorance," a term first coined by Allport (Katz and Allport 1931, p. 152). Pluralistic ignorance describes situations where a majority of group members privately reject a norm, but assume (incorrectly) that most others accept it (see Miller and McFarland [1991] and O'Gorman [1986] for reviews). It is, in Krech and Crutchfield's (1948, pp. 388–89) words, the situation where "no one believes, but everyone thinks that everyone believes." The illusion of support is validated when it motivates widespread public compliance.

Pluralistic ignorance has been documented not only among groups that indulge but also among those that abstain. For example, in Schank's (1932) classic investigation, the members of a religious community were observed publicly endorsing norms against gambling, smoking, and drinking that they violated in private. More recently, Kitts (2003) found that students in five vegetarian housing cooperatives overestimated public support for dietary norms that were publicly enforced but privately violated. Kitts tested relational explanations ("selective exposure" and "selective disclosure") against social psychological theories of cognitive bias. Consistent with theories of pluralistic ignorance, he found greater support for the relational effects of differential access to information about others' compliance.

Other examples are more disturbing. O'Gorman found that American whites grossly exaggerated other whites' support for segregation in the late 1960s (1975; O'Gorman and Garry 1976). A similar pattern can be found in other repressive regimes. In his book *Private Truths and Public Lies*, Timur Kuran (1995*b*) points to widespread but illusory support for the communist regime in the former Soviet Union, based in part on fear of denunciation for revealing private opposition to neighbors whose apparent enthusiasm for the regime was in fact equally a charade, and for the same reason.[2]

A similar dynamic is evident in witch hunts. As noted by Erikson (1966), witch hunts are caused not by an outbreak of deviance, but by an outbreak of enforcement. Witches are created by anxious neighbors seeking to affirm their status in the community by accusing others of deviance, thereby perpetuating the fear that fuels the need for affirmation. Those accused can then save themselves only by revealing the names of yet other neighbors. Perhaps no one in this population actually believes in the existence of witches. Yet a terrified public turns out to cheer at the executions, in public expiations of a collective anxiety that is of their own making. This self-reinforcing dynamic indeed casts a spell on the community as powerful as that of any witch.

We need not assume this dynamic is some historical relic of superstition. Witch hunts were highly publicized on both sides in the early years of the Cold War. Contemporary witches may also include gays assaulted by young thugs eager to affirm their manhood. A study by Adams, Wright, and Lohr (1996) found that homophobic men rated themselves as having lower levels of arousal than other men when shown videos of homosexual intercourse. However, physiological measures of sexual response were found at higher levels among the homophobic men. The results suggest that aggressive same-sex enforcement of heterosexual norms may be motivated by anxiety over the transparency of hidden deviation. Research on adolescent gangs (Willis 1977; MacLeod 1995) shows how homophobic humor is used to ridicule group members who lack the requisite toughness and to affirm the status and loyalty of those who might otherwise become suspect themselves.

The willingness to feign support for a public lie has also been demonstrated under laboratory conditions. In a classic study, Asch (1951) showed that participants would conform to a consensus judgment they

[2] Kuran (1991, 1995*a*; see also Bicchieri and Fukui 1999) also highlights the potential for rapid collapse, triggered when a few vanguards finally express their actual belief, which encourages others to overcome social pressure and express their actual belief, and so on.

knew to be false rather than risk social isolation as a deviant. When participants were assured anonymity, the false compliance disappeared.

It is not difficult to find other familiar examples of compliance with, and enforcement of, privately unpopular norms:

1. the exposure of the "politically incorrect" by the righteously indignant who thereby affirm their own moral integrity;
2. gossiping about a social *faux pas* by snobs anxious to affirm their own cultural sophistication;
3. public adoration of a bully by fearful schoolboys who do not want to become the next victim;
4. "luxury fever" (Frank 2000) among status seekers who purchase $50 cigars, $17,000 wristwatches, and $3 million bras, in an arms race of conspicuous consumption and one-upmanship that leaves the contestants no happier but perhaps a bit less affluent.

Naked emperors are easy to find but hard to explain. It is easy to explain why people comply with unpopular norms—they fear social sanctions. And it is easy to explain why people pressure others to behave the way they want them to behave. But why pressure others to do the opposite? Why would people publicly enforce a norm that they secretly wish would go away?

One hypothesis is that very few would actually enforce the norm, but no one knows this. If people estimate the willingness to enforce based on the willingness to comply, and they comply based on the false belief that others will enforce, they become trapped in pluralistic ignorance—an equilibrium in which few people would actually enforce the norm but no one realizes this. However, this equilibrium can be extremely fragile. As in the Andersen story, all that is needed is a single child to laugh at the emperor and the spell will be broken.

THE ILLUSION OF SINCERITY

A more robust explanation is that most people really will enforce the norm, and for the same reason that they comply—social pressure from others in the group, for whom mere compliance is not enough.[3] To the

[3] Norms mandating the enforcement of other norms are sometimes referred to as "metanorms" (Axelrod 1986; Horne 2001). Explicit obligations to enforce can be found in systems of collective sanctioning in which a group is made responsible for the compliance of its members. A good example is an honor society that obligates individuals to punish family members or fellow students who dishonor the group by cheating on a spouse or an exam (Vandello and Cohen 2004). Classmates who fail to report a student who cheated on an exam are also guilty, as are those who fail to report those who tolerated cheating. In "cultures of honor" (Nisbett and Cohen 1996), a daughter's

true believer, it is not sufficient that others go to the right art galleries, display the right body jewelry, purchase the right sports car, or support the right wing. They must do it for the right reason. Zealots believe that it is better not to comply at all than to do so simply to affirm social status (Kuran 1995*a*, p. 62). Such compliance lasts only so long as behavior can be monitored and social pressure is sufficient to induce acquiescence (Hechter 1987). Thus, true believers reserve special contempt for imposters. Those who comply for the wrong reason must worry about being exposed as counterfeit.

The hypothesized anxiety is supported by research on the "illusion of transparency" (Gilovich, Savitsky, and Medvec 1998). This refers to a tendency to overestimate the ability of others to monitor our internal states. Savitsky, Epley, and Gilovich (2001) found that individuals tend to overestimate how harshly others will judge them for a public mishap. Across four experimental studies, actors anticipated being more harshly evaluated than was actually the case.

Applied to the emperor's dilemma, the "illusion of transparency" suggests that those who admire the emperor out of a desire for social approval fear that their posturing will be apparent to others. They then look for some way to confirm their sincerity. Enforcing the norm provides a low-cost way to fake sincerity, to signal that one complies—not as an opportunist seeking approval—but as a true believer.

What better way to signal one's sincerity than to act in a way that encourages others to comply (Kuran 1995*b*, p. 61)? When one's moral, political, or professional "fitness for office" is challenged, people rarely turn the tables on their inquisitors. If conformity is sanctioned, while enforcement is not, conformists may be suspected of posturing in order to gain social approval, but those who enforce conformity appear to be the genuine article. This use of enforcement to signal sincerity explains the apparent fanaticism of "new recruits" who must prove their loyalty to the established members of a cult or gang, and it also raises the possibility that the thought police may actually be imposters themselves, a

sexual transgression dishonors the entire family until someone (e.g., the father or oldest brother) restores the family's social position by carrying out the obligatory "honor killing." Honor codes also obligate punishment of out-group members who violate strict rules regarding displays of respect (Bourdieu 1966; Elster 1990). Failure to carry out the vendetta leads to being labeled a coward and can result in further attacks on one's family (Gould 2000). Those who befriend a "coward," in turn, come to be tarnished with the same brush. Simply put, every violation of the code of honor becomes an acid test of everyone's loyalty to and standing within the community.

defensive tactic Freud (1894) called "reaction formation" (Baumeister, Dale, and Sommer 1998).[4]

In the college drinking example, insecure freshmen who worry about social acceptance may be tempted to drink—and to celebrate intoxication—in order to appear "cool." However, they must not appear to be motivated by this goal or they risk being scorned as a "poser." Thus, it is not enough to "party"—they must also express the belief that drinking is cool and act accordingly, thereby adding to the social pressure that leads others like them to join in.

Or consider those who pretend to appreciate some highly opaque scholar in order to affirm their erudition.[5] Privately, they have no clue what the writings mean (if anything), and they worry that true believers will see them as fakes (the illusion of transparency). The solution is easy: simply disparage those intellectually shallow scholars who fail to appreciate real genius. But when one does this, one adds yet another voice to the chorus of intimidation that induces the insecurity motivating the behavior in the first place. The norm becomes self-enforcing.

Can self-enforcing norms emerge in a reluctant population, in the absence of any top-down institutional repression, or does it require a powerful emperor to jump-start the process? Can unpopular enforcement be entirely self-organizing? How many true believers are required to trigger a cascade that pulls in the disbelievers as well? Do these cascades depend on the structure of social networks? How stable is an unpopular equilibrium, and how large a disturbance is required for it to collapse?

One way to find the answers to these questions is to systematically study norm enforcement in communities suspected of being trapped in an emperor's dilemma. Although it is not difficult to find empirical case studies, it may be nearly impossible to find convincing evidence that the supporters of the norm are actually imposters. It is not clear that people even really know their own beliefs (Nisbett and Wilson 1977), given the power of self-deception and the subconscious state of reaction formation, but even if we assume that people have perfect interior knowledge of mental states, they have no direct access to the mental states of others. If people are unable to distinguish true believers and imposters, this may also be true for social scientists, including participant observers.

These difficulties have led researchers to study unpopular norm com-

[4] Based on their review of empirical work on reaction formation, Baumeister et al. (1998) concluded that "people respond to the implication that they have some unacceptable trait by behaving in a way that would show them to have the opposite trait" (1998, p. 1,085).

[5] This anxiety is supported by research on "the imposter phenomenon" and refers to the feeling, common among academics and professionals, that one is an "intellectual fraud" (Clance 1985).

pliance through formal methods, such as game theory. It can be shown, for example, that if an entire population is enforcing a norm that compels intolerance of deviance by oneself and others, then enforcement of the norm is a Nash equilibrium. Even if everyone prefers that the norm would disappear, no one has an incentive to change strategy unilaterally—thereby becoming the lone deviant in a population of enforcers (Heckathorn 1990; Binmore 1998).

However, knowing that an equilibrium exists does not mean that this outcome is likely or even attainable. A growing interest in out-of-equilibrium dynamics has led game theorists to use evolutionary models to study the emergence and stability of equilibria from a variety of initial conditions (Skyrms 1996; Young 1998; Gintis 2000). However, these analytic models are often mathematically tractable only if populations are assumed to be fully connected or randomly connected. Yet research on cascades points to the decisive importance of the structure of local interactions (Watts 2002; Centola, Macy, and Eguiliz 2004).

Given these difficulties with traditional methods, computational modeling can be a useful approach for studying normative cascades. An empiricist might object that computer simulation is just as much a fairy tale as Andersen's story. We do not dispute that, but we would point out that fairy tales populated by computational agents are capable of attaining a much higher level of logical consistency than those populated by characters expressed in natural language.

On the other side, analytical game theorists criticize computational models as numerical rather than mathematical, but as Abbott (1998, pp. 176–77) has noted, "[analytical] game theory will not get us very far because it is ignorant, except in the most general terms, of a serious concern with structure and with complex temporal effects. But simulation may help us understand the limits and possibilities of certain kinds of interactional fields, and that would be profoundly sociological knowledge."

Multiagent models can be useful for studying processes such as informal social control that lack centralized coordination. These models focus on how simple and predictable local interactions generate familiar but often enigmatic global patterns, such as cascading enforcement of unpopular norms. By looking for ways to generate these cascades under controlled conditions in a population of computer agents, we may find some clues about the dynamics to look for when we try to model the natural world empirically.

UNPOPULAR NORMS: AN AGENT-BASED COMPUTATIONAL MODEL

We model a heterogeneous population of agents who differ in their beliefs and convictions. Each agent i has a binary private belief, B_i, which defines the agent as either a "true believer" ($B_i = 1$) or a disbeliever ($B_i = -1$). A small group of true believers (analogous to Andersen's three rogues) have such strong convictions that they always comply with the norm, regardless of social pressure not to comply. When dissatisfied with the level of compliance by others, they may also enforce the norm. We call this "true enforcement" because the agent is enforcing compliance with its true (privately held) belief.

The remainder of the population (analogous to Andersen's citizens) consists of skeptics ("disbelievers") who privately oppose the norm, but with less conviction compared to that of the true believers. This opposition can lead them to deviate from the norm and even to pressure others to deviate as well. This is also "true enforcement" because the disbeliever is enforcing its true belief, which happens to be in opposition to the norm. However, because their convictions are not as strong as those of true believers, disbelievers can also be pressured to support the norm publicly. This support includes not only compliance with the norm but can also include pressuring others to comply as well. We call this "false enforcement" because the agent is enforcing a behavior that does not conform with the agent's private beliefs. Thus, there can be three enforcement possibilities:

1. *true enforcement* by true *believers* (who truly support the norm);
2. *true enforcement* by true *disbelievers* (who truly oppose the norm);
3. *false enforcement* by false *disbelievers* (who privately oppose but publicly support the norm).[6]

At each iteration, each agent observes how many of its neighbors comply with the norm and how many deviate. Agents also observe how many neighbors are pressuring others to comply and how many are pressuring others to deviate. These numbers are used to ascertain the level of public support for and opposition to the norm. These distributions, in turn, influence the two decisions that the agent must make—whether to comply with or deviate from the norm, and whether to pressure others to comply or to deviate.

[6] In one experiment, we also allow for a fourth possibility: false believers who privately support but publicly oppose the norm.

Compliance

The compliance decision is based on the level and direction of social pressure relative to the strength S of an agent's convictions, where $0 < S \leq 1$. More precisely, we model agent i's decision to comply with the norm (C_i) as a binary choice, where $C_i = 1$ if i chooses to comply and $C_i = -1$ otherwise. Social pressure is defined as the sum of enforcement decisions by i's neighbors. Each neighbor j who enforces the norm ($E_j = 1$) increases the pressure on i to comply, and each neighbor who enforces deviance ($E_j = -1$) increases the pressure to deviate. A positive net value means that there is greater pressure in the direction of compliance, while a negative net value promotes deviance. A disbeliever complies with the norm if social pressure overcomes the agent's opposition, given the strength of the agent's conviction:

$$C_i = \begin{cases} -B_i & \text{if } \dfrac{-B_i}{N_i}\displaystyle\sum_{j=1}^{N_i} E_j > S_i \\ \\ B_i & \text{otherwise.} \end{cases} \tag{1}$$

Equation (1) states that an agent can be expected to violate its belief (such that a disbeliever complies and a believer deviates) if and only if the proportion of neighbors enforcing falsification minus the proportion enforcing the opposite is sufficient to overcome the strength of the agent's conviction.[7] Thus, in order for opponents of the norm ($B_i = -1$) to be pressured into compliance ($C_i = 1$), positive social pressure must exist, and for supporters of the norm ($B_i = 1$) to be pressured into deviance ($C_i = -1$), negative social pressure is necessary.

By default, we assume that true believers ($B_i = 1$) have maximal conviction ($S_i = 1$) and therefore always comply, even if all their neighbors enforce deviance. For disbelievers, however, convictions are not so strong and may be overcome if the social pressure is sufficiently positive.

The public compliance decision intersects with private beliefs to create four agent types: *true believers* and *true disbelievers* (whose public behavior conforms to their private beliefs), and *false believers* and *false disbelievers* (whose public behavior differs from their private beliefs, in

[7] Although the model is dynamic and decisions are iterated, the model is not indexed on time because we assume asynchronous updating, which means that agents decide to comply and then enforce based on the pressures that exist at the moment of the decision. Their decisions then alter the conditions on which other agents (choosing later) base their decisions. To avoid order effects, we randomize the sequence in which agents make decisions. To test robustness, we also replicated our experiments using synchronous updating (in which all agents update their decisions at the same time, based on conditions that existed at the end of the previous iteration) and found no qualitative differences in the results.

response to the effects of social pressure). By default, we assume that true believers cannot become false disbelievers, because their convictions are too strong to be overcome by any amount of pressure to deviate. However, we also explore a special case in which private beliefs of false believers ($B_i = -1$) can eventually conform to public behavior ($C_i = 1$), creating "converts" with $B_i = 1$. These newly converted believers have weaker convictions than those of the original true believers and could eventually be pressured to "flip-flop" on the issue. Equation (1) thus also allows for the possibility of a "false disbeliever" and shows the level of negative pressure required to flip these converts (now with $B_i = 1$) back to their original behavior ($C_i = -1$).

Enforcement

An agent's enforcement decision is informed by concerns that differ depending on whether the agent's enforcement is true or false. Since false believers are secretly opposed to the norm, they have no interest in pressuring others to comply with a norm that goes against their private convictions. Unlike true believers, who sanction to promote compliance, and true disbelievers, who sanction to promote deviance, the false believer sanctions to avoid exposure as an opportunistic imposter. The tendency to falsely enforce thus increases with increasing social pressure to support the norm and decreases with increased conviction, exactly as with the decision to comply. However, we assume that enforcement imposes an additional cost, beyond the costs associated with compliance.[8] Thus, the threshold for false enforcement is higher than the threshold for compliance by an amount corresponding to the additional cost K incurred by those who choose to enforce as well as comply, where $0 < K < 1$:

$$E_i = \begin{cases} -B_i & \text{if } \left(\dfrac{-B_i}{N_i}\displaystyle\sum_{j=1}^{N_i} E_j > S_i + K\right) \wedge (B_i \neq C_i) \\ +B_i & \text{if } (S_i W_i > K) \wedge (B_i = C_i) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The top line of equation (2) is identical to equation (1), except that greater

[8] To simplify the model, we omit cost from the compliance algorithm and treat the cost of enforcement as the added cost of also enforcing a norm with which one is in compliance. This added cost might be relatively small for sanctions based on social approval or quite large for those based on material rewards or any form of disapproval that carries a risk of inciting retaliation.

social pressure is required to induce false enforcement than to induce false compliance, by the amount K, corresponding to the added cost of enforcement. A false believer ($B_i \neq C_i$) can be expected to enforce a norm that the agent privately opposes if and only if social pressure is sufficiently positive (i.e., supporters of the norm outweigh the opponents) to overcome the agent's reluctance to enforce, given both the strength of the agent's private opposition to the norm and the cost of enforcement.

Only "posers" feel the need to affirm the sincerity of their false behavior by pressuring others to act likewise. For everyone else, the decision to enforce conformity to their belief is more straightforward. This decision is based on one's interest in having others conform with the way one wants them to act, given one's private belief (B_i) and the strength of one's conviction (S_i). Thus, the effect of conviction on true enforcement is the opposite of the effect on false enforcement. Conviction *inhibits* false enforcement (thus S appears on the right-hand side of line 1 in eq. [2]), while conviction *promotes* true enforcement (thus S appears on the left-hand side of line 2). We assume further that the effect of conviction depends on the need for enforcement. Agents only pay the cost of enforcement when it is needed to promote the desired behavior, and they stop enforcing when there is sufficient conformity with their beliefs that enforcement is no longer warranted. For example, if everyone is complying, true believers have no need to enforce, even though their convictions are maximally strong. The higher the level of deviance among their neighbors, the greater the need of true believers to invest effort in a campaign of social control. Similarly, the higher the level of compliance among a disbeliever's neighbors, the greater the need for a disbeliever to invest effort to pressure others into noncompliance with the norm.

The need for enforcement (W_i) is simply the proportion of i's neighbors whose behavior does not conform with i's beliefs (B_i), or

$$W_i = \frac{1 - (B_i/N_i) \sum_{j=1}^{N_i} C_j}{2}. \tag{3}$$

Equation (3) rescales the aggregation of C over i's N neighbors j (which ranges from -1 when all j deviate to $+1$ when all j comply), so that the result corresponds to the proportion of neighbors whose compliance behavior does not conform with i's belief B_i (which ranges from 0 when all j conform to i's belief to 1 when all j violate i's belief). Substituting for W_i in the middle line of equation (2), the algorithm implies that agents enforce conformity with their private beliefs when conformity falls below the level they are willing to tolerate, given their belief in the norm and the strength of their conviction.

We assume that agents can only enforce compliance if they have also complied, and they can only enforce deviance if they have also deviated.[9] However, given the cost of enforcement (K), it remains possible to comply with a norm but fail to enforce it and to deviate without enforcing deviance. Thus, unlike the complementary rates of compliance and deviance, the rates of enforcement for and against the norm need not sum to unity. Everyone must choose whether to comply with the norm, but it is possible that no one enforces anything ($E = 0$).

Equations (1)–(3) require start-up assumptions about the initial levels of compliance and enforcement. A conservative assumption in a test of false enforcement is that agents initially conform to their private convictions and no one enforces anything. Hence, there is initially no pressure to comply, nor any pressure to enforce falsely.

We initialized the population as either fully connected or clustered into small, ego-defined "neighborhoods" such that each agent has eight neighbors. We fixed the distribution of beliefs so that the norm could range in popularity from zero to nearly universal. True believers' convictions were fixed at 1.0, meaning that they were immune to social pressure. For disbelievers, convictions were uniformly distributed in the range $0 < S_i \leq .38$ (slightly above three out of eight neighbors), with a mean of 0.19. Although the mean conviction of true believers ($S_i = 1$) is over five times that of disbelievers, they are also outnumbered 100 to 1, making the norm highly unpopular overall. For a disbeliever with minimal conviction to be pressured into compliance, there must be at least one more neighbor enforcing compliance than enforcing deviance. A disbeliever with maximal conviction ($S = .38$) and eight neighbors requires at least four more neighbors enforcing compliance than enforcing deviance in order to be pressured into compliance. The threshold for false enforcement is higher than the threshold for compliance by the value of K, which we fix at .125. For agents with eight neighbors, this means that at least one additional neighbor must enforce before a false believer becomes a false enforcer as well.

These values of conviction and cost form a cumulative uniform distribution of false enforcement thresholds across the population of disbelievers, such that, for agents with eight neighbors, if two more neighbors are enforcing compliance than are enforcing deviance in every disbeliever's neighborhood, then about one-third of the disbelievers will falsely enforce. If three more neighbors enforce compliance than deviance, then about two-thirds of disbelievers will falsely enforce. And if four more

[9] While agents may hypocritically enforce a norm they privately question, they never enforce norms which they violate. We leave for future research the effect of hypocritical enforcement based on the rule "Do as I say and not as I do" (Heckathorn 1989).

neighbors enforce compliance than deviance, then just under 99% of disbelievers will falsely enforce.

We tested several variations of equations (1)–(3) to assess the robustness of the model, which uses a deterministic step function for the decisions to comply and enforce. The step function implies that agents are indifferent to changes in social influence at all levels of influence except for a single critical value. We relaxed this assumption by replacing the step function with a sigmoidal stochastic approximation (implemented as a cumulative logistic function), in which the probability of the decision increases with influence, such that the probability is 0.5 when the threshold is crossed and approaches the natural limits asymptotically. The results affirmed all conclusions based on the deterministic model. Moreover, with stochastic decisions, cascades of false enforcement can be randomly triggered in a population with no true believers, which is impossible in the deterministic specification. However, the latter has the didactic advantage of greater simplicity, and we therefore use the deterministic model for the presentation of our results.

EXPERIMENTAL DESIGN

We want to know if a small minority of fanatics can force a few vulnerable disbelievers ($S_i \approx 0$) to help enforce the norm, which will then increase the pressure on more stalwart disbelievers, triggering a cascade of false enforcement that sweeps through a skeptical population. We also want to see if this cascade process ultimately produces a stable enforcement equilibrium that would persist even if every true believer were to exit the population. To find out, we use a series of computational experiments to study the emergent dynamics as we manipulate a set of structural conditions:

1. access to information about the behavior of other agents (from global to local, where local access limits enforcement pressure to neighbors in a social network);
2. the frequency distribution and clustering of true believers (from less than 1% to more than 99% of the population and from highly clustered to randomly dispersed);
3. the network topology (from highly ordered networks to irregular grids to small world networks with a much shorter characteristic path length).

We explore network topologies using a two-dimensional cellular automata (von Neumann 1966) consisting of 1,000 agents located on a

40 × 25 torus, with every cell occupied by one agent.[10] In a regular lattice, each cell has a "neighbor" (or social tie) to the immediate right, left, top, and bottom. This elementary structure, called the von Neumann neighborhood, is characterized by the absence of transitivity (or triad closure, in which one's neighbors are also neighbors of one another). If we also include the four adjacent diagonal neighbors (creating a Moore neighborhood), we create ties among neighbors, a property of social networks that is supported by research in structural balance theory (Cartwright and Harary 1956).

Although transitive social ties are more plausible than von Neumann neighborhoods, the regularity of Moore neighborhoods remains highly stylized. We relaxed the regularity of the lattice in two ways. First, we created Voronoi diagrams or "irregular grids" (Flache and Hegselman 2001) that preserve the transitivity and overlap of Moore neighborhoods but allow agents to vary in degree (or number of neighbors). The results using irregular grids were identical to those using Moore neighborhoods, so only the results with the simpler regular structures are reported.

Second, following the procedure used by Watts (1999), we randomly rewired ties with probability P ($0 \leq P \leq 1$). With $P = 0$, the network is a regular lattice, and with $P = 1$ the network is completely random. An intermediate regime, called a small world network, is characterized by a high degree of clustering with a low characteristic path length and has been shown to be representative of a large number of social and biological populations (Watts and Strogatz 1998). Using the standard nine-cell Moore neighborhood as our baseline, we explore alternative network topologies and observe their effects on cascades of false enforcement.

RESULTS

To preview the main results, we report three notable findings on the effects of network topology and the distribution and clustering of true believers:

1. *Embeddedness*.—Although universal enforcement of a highly unpopular norm is an equilibrium state in a fully connected population, a dynamic model shows that this equilibrium cannot be reached from out of equilibrium for a highly unpopular norm, supported only by a very small number of true believers. Yet the equilibrium is easily reached if an identical population is embedded in a network that restricts interaction to small but overlapping neighborhoods.

[10] We also tested lattices with up to 5,000 cells. Increasing the size of the network had no effects on the results.

2. *Less is more*.—Under conditions in which even very large numbers of true believers are unable to quash deviance by disbelievers, a much smaller number of true believers can successfully ignite cascades that lead to near-universal compliance and enforcement of the norm. Simply put, we identified conditions in which five true believers can accomplish what 500 cannot. By extension, when we allow false enforcers to convert into true believers, it does not stabilize the high-compliance equilibrium, but causes it to collapse.
3. *Small worlds*.—Bridge ties between otherwise distant neighborhoods—such as the "weak ties" (Granovetter 1973) that have been shown to promote the diffusion of information, innovations, rumors, and disease—are shown to inhibit cascades of false enforcement.

Experiment 1: Effects of Embeddedness

We begin by testing the effects of network embeddedness on cascades of enforcement of a highly unpopular norm (less than 1% of the population are true believers). Experiment 1 manipulates the distribution (random vs. clustered) and embeddedness (global vs. local) of true believers and measures the effects on the proportion of disbelievers who falsely comply. Since clustering is meaningless in a fully connected population, these two manipulations yield three experimental conditions: global, local + clustered, and local + random.

Before turning to the results of experiment 1, it is useful to bear in mind that in a fully connected population, even with no true believers, universal enforcement of the norm is an equilibrium. That is, if we initialize the model such that everyone is falsely enforcing compliance with the norm, then no agent will unilaterally change strategy and become the lone deviant in a population of enforcers.

Nevertheless, the thin solid line in figure 1 shows that this equilibrium cannot be reached in an unembedded population that is overwhelmingly skeptical (less than 1% true believers) if the starting point is far from equilibrium. Figure 1 reports the proportion of disbelievers who comply with the norm as we increase the proportion of true believers from less than 1% to more than 99%. For now, we focus only on the far left side of figure 1 (expanded in the right-hand corner insert) as we compare compliance rates among disbelievers in an overwhelmingly skeptical population that differs only in embeddedness and clustering.

In a fully connected (unembedded) population (shown by the thin line in fig. 1), each agent surveys every other agent to determine the levels of compliance and enforcement of the norm. If the population is overwhelmingly deviant at the outset, true believers are readily willing to pay the

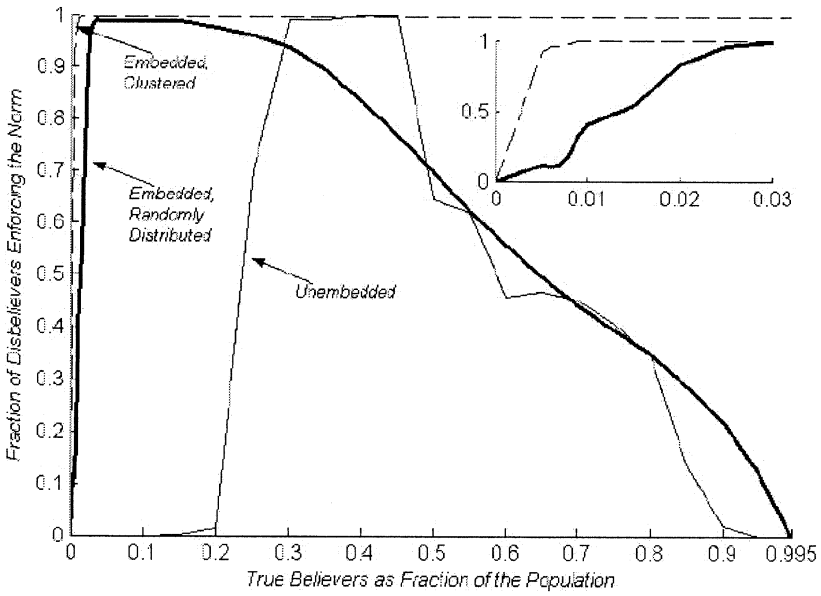

FIG. 1.—Effects of distribution and embeddedness of true believers (based on 1,000 agents and averaged over 100 replications at each parameter combination). The dark line indicates an embedded population in which the true believers are randomly distributed. The dashed line shows an embedded population with clustered true believers, and the thin line shows an unembedded population. The dashed line shows how a very small group of true believers must be clustered to produce sufficient pressure to ignite a cascade. As we increase their numbers, dispersed true believers can randomly cluster to form a critical mass (dark line). True believers have little effect on disbelievers in an unembedded population unless they comprise about one-third of the population (thin line). Above that level, their effectiveness diminishes due to high rates of compliance they observe among one another, which diminishes their zeal to enforce.

cost of enforcement ($K = .125$), but their numbers are too small to compel even the most spineless disbelievers to join their crusade. Had this been the Andersen fable, everyone but the three rogues would be laughing at the foolish emperor.

So how did it happen that the citizens did not laugh, but instead expressed fawning admiration? We hypothesize that the explanation centers on the assumption that social influence is not global but local.

To test this idea, we restricted agents' observations to their immediate neighbors and randomly distributed five true believers across a population of 995 skeptics. The result (indicated by the dark line in fig. 1) is similar to what we observe when information is global—everyone still laughs at the emperor. Once again, the true believers fail to ignite a cascade.

But now suppose that agents are clustered by beliefs, a network property called *homophily*, based on the principle that likes attract (Simmel

1955; Carley 1991; Mark 1998; McPherson, Smith-Lovin, and Cook 2001). When agents are clustered by belief rather than randomly distributed, a skeptic exposed to a true believer can now expect to have other true believers in its neighborhood. The results (shown by the dashed line in fig. 1) now change dramatically, and the outcome predicted by Andersen's fable becomes almost unavoidable. When true believers are clustered, even a very small number can spark a cascade of compliance.

To test the robustness of this result, we repeated the experiment, but this time we matched the true believers with an equal number of disbelievers who had equally strong convictions ($S = 1$). For the remaining disbelievers, S was uniformly distributed so that the average conviction of all disbelievers remained .19. Creating an equal number of "diehards" on both sides (while keeping true believers badly outnumbered) had no effect on the results. The fanatic disbelievers could not be converted and always enforced their true beliefs, yet they were powerless to stop cascades of false enforcement.

We also repeated the experiment using a stochastic version of the model, in which the probabilities of compliance and enforcement are cumulative logistic functions of the level of influence specified in equations (1) and (2), respectively. Remarkably, the results showed that *no* true believers are needed; cascades of false enforcement can be triggered "spontaneously" by the random waverings of uncommitted disbelievers.

To better understand the cascade dynamics, figure 2 gives a stylized illustration using a neighborhood with a cluster of three true believers (white cells 5, 8, and 11) surrounded by twelve disbelievers (light gray) at time 1. Each true believer is exposed to at least six deviants, which is sufficient to overcome the cost of enforcement ($K = .125$), and 5, 8, and 11 begin to pressure their neighbors at time 2 (turning dark gray). Meanwhile, two disbelievers (7 and 9) are likewise exposed to the three true believers, and this provokes them to fight back at time 2 by enforcing opposition to the norm (turning black).

At time 2, the true believers now observe not only deviance but also social pressure to deviate. However, given their strong convictions ($S = 1$), this pressure has no effect on their resolve. They continue to comply with the norm and to demand compliance of others. Conversely, agents 7 and 9 now observe not only compliance but also pressure to comply from three of eight neighbors (.375), causing them to falsify their beliefs at time 3 ($C_i \neq B_i$). With $K = .125$, we know from equation (2) that this pressure is sufficient to trigger false enforcement by any agent whose level of conviction is less than .25, which includes about two-thirds of the disbelievers. Suppose agents 7 and 9 succumb to the pressure. Now agents 4, 6, 10, and 12 will have three neighbors that are enforcing compliance with the norm and zero neighbors enforcing deviance. Hence,

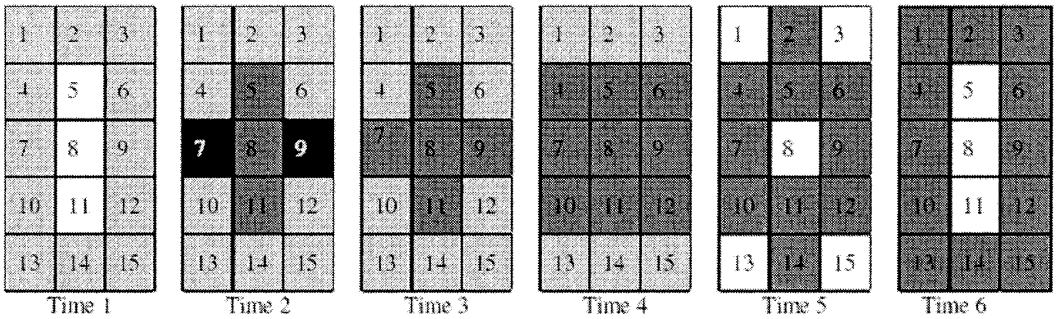

FIG. 2.—Cascade dynamics in a Moore neighborhood. The white cells are complying with the norm, and the light gray cells are deviating. The dark gray cells are also enforcing compliance, and the black cells are also enforcing deviance. Enforcement pressure created by the three true believers (cells 5, 8, and 11) compels neighboring disbelievers to falsely comply and falsely enforce. False enforcement adds to the pressure on other disbelievers, and soon not just their neighbors, but also their neighbors' neighbors are all enforcing the norm, allowing the true believers to stop enforcing and let disbelievers sustain the unpopular norm.

each of these agents will now be triggered to enforce falsely. By extension of this process, at time 4, the entire Moore neighborhood (agents 4–12) begins enforcing the norm (all dark gray).

At time 5, agents 1, 3, 13, and 15 only have two out of eight neighbors enforcing (.25). This pressure is likely to persuade them to comply with the norm but not to enforce it (since the enforcement threshold is higher by K). However, agents 2 and 14 now have three of eight neighbors enforcing the norm, and no pressure to deviate, so they will most likely join the cascade. Thus, at time 6, with the additional pressure from agents 2 and 14, agents 1, 3, 13, and 15 now have three neighbors enforcing compliance, and they too join in. The level of compliance is now sufficient that none of the true believers need to enforce any longer. Like the three swindlers in the Andersen fable, they can quietly leave the village, knowing the citizens will now do their work for them.

Experiment 2: Effects of the Number and Clustering of True Believers

Experiment 1 showed how even very small numbers of true believers can trigger a cascade of false enforcement in an embedded population, but only if they are sufficiently clustered in a neighborhood to generate sufficient peer pressure to trigger a chain reaction. Experiment 2 further explores what happens as we increase the proportion of true believers.

The thin line in figure 1 shows that this effect is U-shaped in a global population. As the number of true believers approaches one-third of the population, they can trigger a cascade in a fully connected population in the same way that three true believers can trigger a cascade in a local neighborhood of size 9. (Of course, one hesitates to refer to this as an "unpopular" norm, considering that one-third of the population now has an unwavering conviction in favor of norm compliance.)

However, as we increase the number of the true believers even further, their enforcement effectiveness begins to diminish. This happens because true believers have less need to enforce as the level of compliance increases, which means that the dynamics are self-limiting. In contrast, false believers feel more pressure to enforce as the cascade progresses—it is a self-reinforcing dynamic. Thus, the larger the proportion of true believers (who always comply), the lower the pressure on disbelievers. For example, when there is an overwhelming majority of true believers (e.g., 90% of the population), they observe that almost everyone is willing to comply without their having to pay the costs of enforcement. Thus, true believers do not bother to enforce their views, and the tiny minority of disbelievers can enjoy an atmosphere of tolerance (but only so long as their numbers do not attract too much attention from the majority).

We see a similar U-shaped effect in a population of embedded agents

with randomly distributed true believers (the dark line in fig. 1). However, the governing dynamics are somewhat different. As the number of randomly distributed true believers increases, there is a greater likelihood that a critical mass of true believers will end up in the same neighborhood. Moreover, the larger the number of such clusters, the greater the odds that a cluster will form in a neighborhood with disbelievers whose convictions are relatively low, such that a cascade can find a foothold. Thus, cascades of false enforcement are more robust as the number of true believers increases—but only up to a point. As in the unembedded case, there can also be too many true believers for their own good. True believers who are surrounded by compliance do not see the need to enforce, and this situation allows fringe pockets of deviance to persist in a population that is almost entirely compliant by conviction. Had there been fewer true believers, they would have pressured nearby disbelievers to join the cascade, who would have pressured other disbelievers, and so on, until these pockets of dissent were eliminated by the self-reinforcing dynamics of false enforcement.

Finally, we tested for the U-shaped effect in a population of agents clustered by belief. When true believers are tightly clustered, it takes only a very small number to trigger a cascade leading to universal enforcement, as already noted. As their numbers increase, we might expect them to become less effective, as observed for the global population and for a local but randomly distributed population, due to the self-limiting dynamics of true enforcement. However, when the true believers are densely clustered, we observe no reduction in effectiveness as their numbers increase.

This remarkable scalability is caused by the existence of an ordered boundary between the two populations. The ratio of the two groups along this boundary is constant, regardless of the relative sizes of the two populations. Even when the norm is very popular, the true believers along the border are exposed to high levels of deviance along the frontier, just as they would be if they were a minority in an isolated neighborhood. This contact motivates the true believers to enforce, and the cascade then fans out across the cluster of disbelievers, like a prairie fire. This boundary effect suggests a structural interpretation of Erikson's (1966, p. 19) emphasis on the importance in social control of "those we call deviants [who] patrol its boundaries."

As an extension of the experiment on the number of true believers, we also allowed for the possibility that agents might eventually change their private beliefs to conform to their public behavior.[11] We therefore added

[11] This change of belief to conform to behavior is hypothesized to occur under conditions where social pressure prevents agents from changing their behavior to conform to their

an algorithm that modifies agent i's private belief B_i based on the weighted cumulative experience of false enforcement, relative to the strength of conviction:

$$R_i^{t+1} = \begin{cases} R_i^t - \lambda E_i^t B_i^t & \text{if } E_i^t \neq B_i^t \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

$$B_i^{t+1} = \begin{cases} C_i^t & \text{if } (R_i^{t+1} > S_i) \wedge (B_i^t \neq C_i^t) \\ B_i^t & \text{otherwise.} \end{cases} \tag{5}$$

Equation (4) states that cognitive pressure to reconcile a dissonant belief increases by λ with each decision to falsely enforce ($E_i = -B_i$), is unaffected if the decision is not to enforce ($E_i = 0$), and is reset to zero should the agent truly enforce ($E_i = B_i$). Equation (5) states that private belief B_i is then reconciled with public behavior when the dissonance is sufficient to overcome the "stubbornness" of i's belief, given the strength of i's conviction (S_i). The learning parameter λ ($0 \leq \lambda \leq 1$) controls how quickly private beliefs change to conform to public behavior. We set λ at .0001 so that the cascade process would have sufficient time to complete before private beliefs begin to accommodate public behavior (with $\lambda = .0001$, nearly 3,800 updatings will be required before disbelievers with $S_i = 0.38$ will change beliefs).

Note that this experiment differs from the previous manipulation in which we increased the number of true believers *prior* to the cascade. Once the cascade has succeeded, one might expect increasing support for the norm to simply stabilize the equilibrium of norm compliance. Surprisingly, the effect is exactly the opposite. Figure 3 shows what happens as false enforcers begin to believe in what they are preaching—the unpopular norm destabilizes and eventually collapses.

The time series in figure 3 illustrates the emergence of an unpopular norm in an embedded population with clustered true believers, followed by the decay of false enforcement. Agents with weaker convictions are among the first to be converted to false enforcers. As equation (5) indicates, they also have a high susceptibility to eventually being converted into true believers. At some point after the population reaches an equilibrium of norm enforcement, weak-willed disbelievers, who have been falsely

beliefs. Significant research on cognitive dissonance and self-perception theories points to this assimilative effect of behavior on beliefs (Festinger 1957; Bem 1972). Prentice and Miller's research on unpopular drinking norms showed a tendency for male students' behaviors *and attitudes* to gravitate over time toward what they falsely thought to be the norm (1993). As William James said, "*We need only* ACT *as if the thing in question were real, and keep acting as if it were real, and it will infallibly end by growing into such a connection with our life that it will become real*" (James [1890] 1981, emphasis in original).

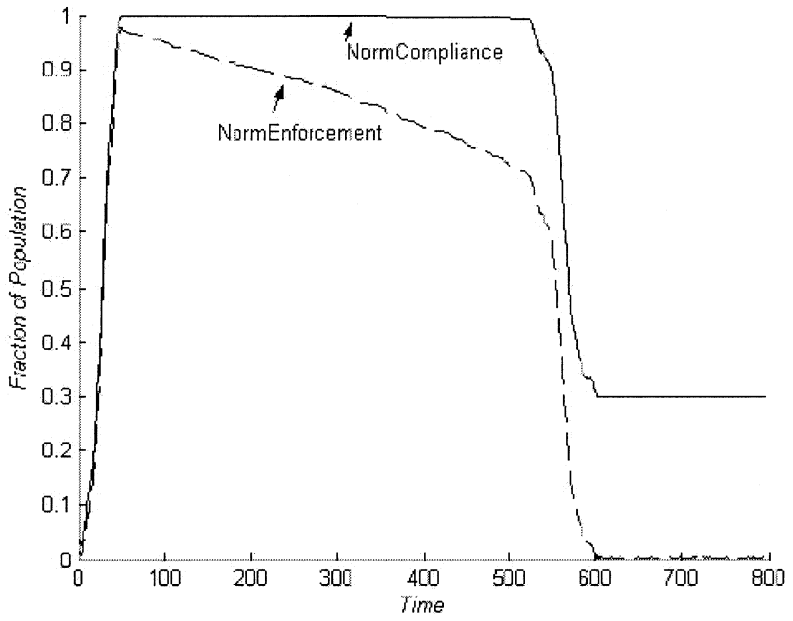

FIG. 3.—Effect of conversion of disbelievers (based on a representative time series for 1,000 agents). The solid line indicates norm compliance and the dashed line indicates norm enforcement. The norm quickly spreads, and by time 80 the entire population is enforcing the norm except the true believers, whose efforts are no longer needed. Disbelievers with low convictions eventually begin to change their private beliefs, becoming believers in the norm. With no need now to hide their posturing, they stop enforcing, which reduces pressure on other disbelievers. At time 560, there is a steep reverse cascade from false enforcement to widespread deviance, and by time 600, two-thirds of the population is refusing to comply.

enforcing for the longest time and are the least committed to their beliefs, begin to convert. As soon as disbelievers become true believers they are no longer complying because of social pressure and thus no longer feel the need to prove the sincerity of their compliance by publicly pressuring others. Nor is the weak-willed true believer willing to enforce out of conviction unless the compliance level approaches zero (eq. [2]). When these converted disbelievers stop enforcing the norm, there is less local pressure on the remaining disbelievers, which also allows those with stronger convictions to stop enforcing, further reducing local pressure. A new equilibrium then obtains, characterized by behavior that is largely voluntary, with minimal enforcement (limited to a few of the original high-conviction true believers, true disbelievers with high convictions located near them or their converts, and false disbelievers who enforce out of pressure from their true disbeliever neighbors).

Once support for the unpopular norm collapses, a new cascade of false

enforcement does not begin again. In order to start a new cascade, the original true believers need deviant neighbors with low convictions who can be pressured into false enforcement (to prove their sincerity). However, those with relatively low convictions have now been converted and are no longer deviant. Although the original true believers still have strong convictions, their newly converted but weak-willed neighbors now comply with the norm, but their convictions are not strong enough to motivate them also to enforce their views. The conversion of private beliefs has robbed the true believers of the spineless imposters that are needed to launch a successful cascade—the tinder needed for the fire. Paradoxically, although the norm is not as unpopular as before, the population has been inoculated against another cascade by the conversion of those with low convictions.

Note the contrast with the collapse in Andersen's fable. In the original story, the spell collapses when a child laughs at the naked emperor. Our model shows that the spell can also collapse for the opposite reason—when the adult sycophants actually begin to believe they can see the emperor's clothes—a counterintuitive but logical possibility that has gone largely unnoticed in the literature on social control.

Experiment 3: Effects of Network Topology

As a test of the robustness of our results across network topologies, we repeated experiments 1 and 2 using von Neumann neighborhoods as well as Moore neighborhoods of varying depth. As an additional robustness test, we used a Voronoi diagram that relaxes the spatial regularity of lattice networks and allows degree to vary over the population of agents, as illustrated in figure 4. Across all these network structures, the results were qualitatively the same, indicating that the governing dynamics do not depend on network regularity (which is removed in the Voronoi diagrams) or clustering (which is zero in von Neumann neighborhoods and increases with neighborhood depth in Moore geometries).

Further testing revealed that the spread of false enforcement does depend on network structure, but in a way that is quite surprising. A growing body of theoretical research demonstrates that increasing the proportion of random ties in a regular network dramatically increases the propagation rate of cascades (Newman 2000; Kleinberg 2002). We therefore tested the effect on cascades as we perturb network order using Watts's "rewiring" technique, replacing local ties with random ties with probability P. Watts shows how a very small number of random bridge ties between otherwise distant pairs is sufficient to drastically reduce the characteristic path length, with minimal reduction in clustering. The world remains "small"

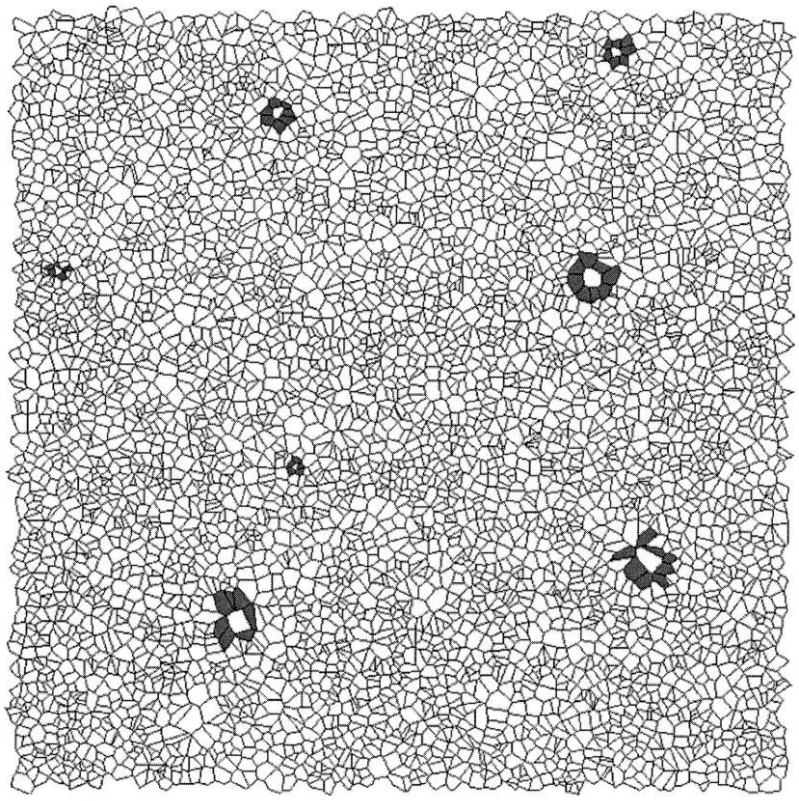

FIG. 4.—Voronoi diagram. The shaded cells are neighbors of the clear cell in the center of the neighborhood. The network retains the spatially ordered relations of grid structures but without their structural regularity.

(i.e., highly clustered) but with remarkably few "degrees of separation" between any two randomly chosen nodes.

One might then expect that random rewiring would also promote the spread of false enforcement of unpopular norms. The surprising result is that random rewiring not only failed to noticeably increase the rate of propagation, but it actually inhibited cascades as the level of randomness increased above a critical value of P. Figure 5 illustrates the effect of reducing network order on the fraction of disbelievers who enforce the norm. For a significant part of the small world regime ($P < .01$), cascades of false enforcement are as robust as in regular graphs ($P = 0$). However, for values of $P > .01$, there is a noticeable decline in the magnitude and frequency of cascades, and for $P > .1$ cascades are entirely precluded.

The unexpected inhibiting effect of random rewiring is due to the de-

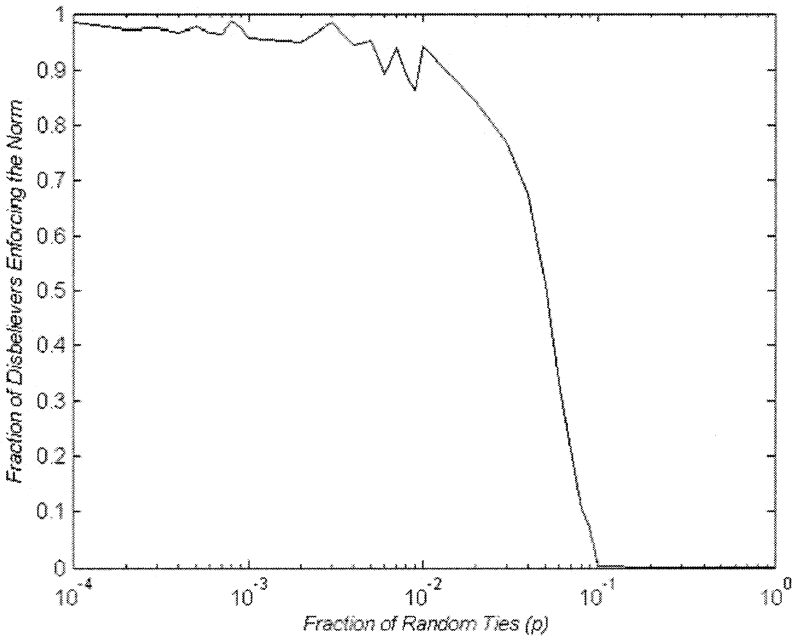

FIG. 5.—Effect of network perturbation (based on 1,000 agents, averaged over 100 replications at each setting). A logarithmic scale on the horizontal axis highlights the steep drop-off in the spread of false enforcements as P is increased. The sharp increase in the effect of P at about .01 indicates a phase transition. Below that level, random ties have little effect, and for $P > .1$ cascades do not occur at all.

creasing fraction of overlap between neighborhoods as network order decreases, as illustrated in figures 6 and 7. Suppose node A in figure 6 is activated, along with all of A's eight neighbors. With $P = 0$, nonadjacent agents A and B share three common neighbors, giving them an overlap of .375. So long as B's activation threshold is below .375, the cascade is guaranteed to move beyond A's neighborhood, even if none of B's other neighbors are activated.

Figure 7 shows how rewiring reduces the average overlap between neighborhoods. A is now connected to otherwise distant nodes, creating a shorter characteristic path length for the network. However, A and B now have only one neighbor in common instead of three. Pressure from A's neighborhood may be sufficient to induce B to comply with the norm, but with $K = .125$, it is not sufficient to induce B to enforce. More generally, as the network is perturbed, either by rewiring existing ties or by adding new ties, the average overlap between neighborhoods decreases, reducing the chance that a neighborhood of enforcers will have sufficient overlap with nearby neighborhoods to propagate the cascade.

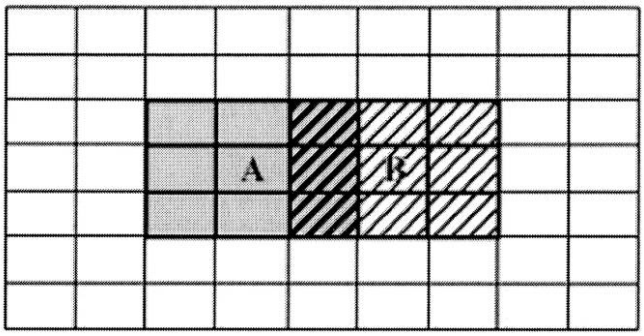

FIG. 6.—Regular neighborhoods. The shaded cells are A's neighborhood (enforcing the norm), the hashed cells are B's neighborhood (deviants), and the three shaded hashed cells are the common neighbors of A and B (enforcing the norm). With $K = .125$, if the strength of B's conviction is below .25, the pressure from these three neighbors will be sufficient to make B enforce as well.

DISCUSSION

The results we report, across all three experiments, tell a very similar story: cascades of false enforcement of an unpopular norm depend on the spread of misinformation about the distribution of support for the norm. Unpopular norms thrive on local misrepresentations of the underlying population distribution. Simply put, it is a sampling problem. This is more easily appreciated if we examine various conditions in which we found that cascades *fail*:

1. in an unembedded (fully connected) population;
2. in an embedded population with a small number of randomly dispersed true believers;
3. when random ties reduce the overlap between local neighborhoods.

In the first case, the cascade fails because disbelievers have an accurate estimate of the true distribution of public support for the norm. Similarly, in the second case, the cascade fails because the distribution of true believers in each neighborhood does not deviate very far from the underlying population distribution. Hence the information is not sufficiently inaccurate to persuade disbelievers that the norm is far more popular than it really is.

In the third case, random rewiring gives disbelievers access to information from outside their Moore neighborhoods. These random ties to otherwise distant nodes give disbelievers a more representative sample and thus a more accurate picture of the true state of the world (see Merton 1968; O'Gorman 1986). Although we discuss this effect as the *inhibition* of the diffusion of social pressure, we could also frame the effect as the

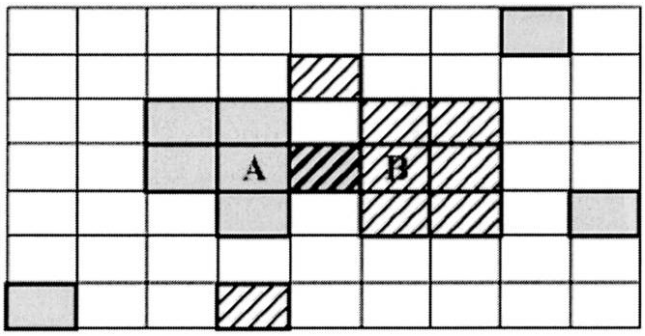

FIG. 7.—Perturbed neighborhoods. A and B now have only one neighbor in common out of eight (overlap is .125). With $K = .125$, B can no longer be pressured into enforcing the norm by members of A's neighborhood.

promotion of the diffusion of information. Random ties inhibit cascades of misinformation by providing access to a more representative sample of the true population.[12]

The results also converge around a second theme, one that calls into question the conventional wisdom among sociologists ranging from functionalists to utilitarians. Despite deep differences in their theoretical approaches, the functionalist and choice-theoretic accounts converge around the prevailing idea that norms are enforced because they are useful, either to society at large (in functionalist accounts) or to those who enforce them (in choice theory). A prominent social scientist sums up the consensus position very succinctly by noting that "norms of social behavior" can be viewed as "reactions of society to compensate for market failures." In effect, when markets create a mess of things, society creates norms with which to clean it up. Was this statement from a functionalist? On the contrary, it was the economist Kenneth Arrow (1971, p. 22), expressing a view recently echoed and underscored by game theorists Bowles and Gintis (2001, p. 6) and by Hechter and Opp (2001, p. xvi). "The view that norms are created to prevent negative externalities, or to promote positive ones," write Hecther and Opp, "is virtually canonical in the rational choice literature."

The emperor's dilemma tells a different story, one that invites us to revisit our sociological intuitions about how and why norms emerge and

[12] In a study of norm perception among five vegetarian communities, Kitts (2003) found support for the hypothesis that local interaction shapes the accuracy of norm perception. However, his research showed that structures encouraging local interactions led to more accurate norm perception because individuals were more likely to disclose their deviance with a friend, and close friendships tend to be clustered.

spread. Contrary to both functionalist and choice-theoretic accounts, norms do not necessarily solve social dilemmas, correct market failures, or promote social welfare. Norms can also create social dilemmas and undermine social welfare. When that happens, the real culprits are not the true believers, whose motivation for enforcement increases with the level of deviance. The key to the emergence of an unpopular norm is the cascade process centered on the self-reinforcing motivations of those who succumb to social pressure. Like a witch hunt, the process can quickly spiral up into a powerful (and dangerous) social movement.

A CAUTIONARY CONCLUSION

We have used a highly specific model of social influence to explore the propagation dynamics of unpopular norms. We do not mean to suggest that we believe this is the only possible specification of social influence. Rather, we used this model to show how these cascades are highly sensitive to the structural conditions in which they occur, and to show how computational models are useful tools for getting leverage on embedded dynamics that would otherwise be mathematically intractable. We believe it will prove useful to explore alternative models of the influence process, as a way to more precisely specify the loosely used concept of "peer pressure" as the motivation to enforce unpopular norms. Using agent-based models, three modes of influence can be compared:

1. *Conformist* influence, which is based on the false belief that others are correct, as in herd behavior (Banerjee 1992) and information cascades (Bikhchandani, Hirshleifer, and Welch 1992). This process applies to the spread of beliefs that are *uncertain* (Sherif 1936) rather than unpopular (Asch 1951).
2. *Imaginary enforcement pressure*, which is based on the false belief that those who comply will also enforce, when in fact they will not, as in the Andersen fable, in which a single violation (e.g., a child who laughs at the emperor) is sufficient to disturb a highly fragile equilibrium.
3. *Real enforcement pressure*, which is based on false compliance and false enforcement, as in the model used in this study.

Note that these elaborations of our model to include alternative specifications of the influence dynamics are not intended to make the model more realistic, nor do we know of any empirical studies that clearly favor one mechanism over another. Rather, we propose these alternative models as dynamic thought experiments in a program of systematic theoretical research that explores a series of *what if*s. The emperor's dilemma model

demonstrates the usefulness of this approach, especially for those interested in structural determinants of microsocial influence processes underlying puzzling population dynamics.

We have also limited ourselves to a computational model of these dynamics, and here again, we do not mean to suggest that other modeling techniques, such as game-theoretic approaches, would not provide additional insights that we may have missed. Nevertheless, we encourage readers to appreciate the particular strengths of this methodology. Agent-based models of dynamic social interaction are more tractable (but less generalizable) than mathematical modeling and more rigorous (but less nuanced) than natural language (Hanneman, Collins, and Mordt 1995). Expressed in natural language, a theory of enforcement of unpopular norms has the intuitive appeal of Andersen's fable, but we cannot know if the intuitions can be trusted, or which assumptions were necessary for the results. A mathematical model is useful for identifying possible equilibria in a fully connected or completely random network but not for studying the dynamics of propagation across a population embedded in a complex network structure. The latter turns out to be decisively important for the spread of unpopular norms.

REFERENCES

Abbott, Andrew. 1998. "The Causal Devolution." *Sociological Methods and Research* 27 (2): 148–81.

Adams, H. E., L. W. Wright, and B. A. Lohr. 1996. "Is Homophobia Associated With Homosexual Arousal?" *Journal of Abnormal Psychology* 105 (3): 440–45.

Andersen, Hans Christian. (1837) 1998. *The Emperor's New Clothes.* Orlando, Fla.: Harcourt Brace.

Arrow, Ken. 1971. "Political and Economic Evaluation of Social Effects and Externalities." Pp. 3–23 in *Frontiers of Quantitative Economics*, edited by M. Intriligator. Amsterdam: North-Holland.

Asch, Solomon E. 1951. "Effects of Group Pressure upon the Modification and Distortion of Judgments." Pp. 177–90 in *Groups, Leadership, and Men*, edited by H. Guetzkow. Pittsburgh: Carnegie Mellon University Press.

Axelrod, Robert. 1986. "An Evolutionary Approach to Norms." *American Political Science Review* 80 (4): 1095–1111.

Baer, J. S. 1994. "Effects of College Residence on Perceived Norms for Alcohol Consumption: An Examination of the First Year in College." *Psychology of Addictive Behaviors* 8 (1): 43–50.

Banerjee, Abhijit V. 1992. "A Simple Model of Herd Behavior." *Quarterly Journal of Economics* 107:797–817.

Baumeister, Roy F., Karen Dale, and Kristin L. Sommer. 1998. "Freudian Defense Mechanisms and Empirical Findings in Modern Social Psychology: Reaction Formation, Projection, Displacement, Undoing, Isolation, Sublimation, and Denial." *Journal of Personality* 66 (6): 1081–1124.

Bem, D. J. 1972. "Self-Perception Theory." Pp. 1–62 in *Advances in Experimental Social Psychology*, vol. 6, edited by L. Berkowitz. New York: Academic Press.

Bicchieri, Cristina, and Y. Fukui. 1999. "The Great Illusion: Ignorance, Informational

Cascades, and the Persistence of Unpopular Norms." *Business Ethics Quarterly* 9: 127–55.
Bikhchandani S., D. Hirshleifer, and I. Welch. 1992. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy* 100:992–1026.
Binmore, Ken. 1998. *Game Theory and the Social Contract*. Cambridge, Mass.: MIT Press.
Borsari, Brian, and Kate B. Carey. 2001. "Peer Influences on College Drinking: A Review of the Research." *Journal of Substance Abuse* 13 (4): 391–424.
Bourdieu, Pierre. 1966. "The Sentiment of Honour in Kabyle Society." Pp. 191–241 in *Honour and Shame: The Values of Mediterranean Society*, edited by J. G. Peristiany. Chicago: University of Chicago Press.
Bowles, S., and H. Gintis. 2001. "Social Capital and Community Governance." Working paper. Santa Fe Institute.
Carley, Kathleen. 1991. "A Theory of Group Solidarity." *American Sociological Review* 56 (3): 331–54.
Cartwright, D., and F. Harary. 1956. "Structural Balance: A Generalization of Heider's Theory." *Psychological Review* 63:277–93.
Centola, Damon, Michael Macy, and Victor Eguiliz. 2004. "Inhibitory Effects of Network Perturbation." Working paper. Cornell University.
Clance, P. R. 1985. *The Impostor Phenomenon: Overcoming the Fear that Haunts Your Success*. Atlanta: Peachtree Publishers.
Elster, Jon. 1990. "Norms of Revenge." *Ethics* 100 (4): 862–85.
Erikson, Kai T. 1966. *Wayward Puritans: A Study in the Sociology of Deviance*. New York: Wiley & Sons.
Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. White Plains, N.Y.: Row & Peterson.
Flache, Andreas, and Rainer Hegselmann. 2001. "Do Irregular Grids Make a Difference? Relaxing the Spatial Regularity Assumption in Cellular Models of Social Dynamics." *Journal of Artificial Societies and Social Simulation* 4:4.
Frank, Robert. 2000. *Luxury Fever: Money and Happiness in an Era of Excess*. Princeton, N.J.: Princeton University Press.
Freud, Sigmund. 1894. "The Neuro-Psychoses of Defense." Pp. 43–61 in *Complete Psychological Works*, vol 3. London: Hogarth.
Gilovich, Thomas, Kenneth Savitsky, and Victoria Husted Medvec. 1998. "The Illusion of Transparency: Biased Assessments of Others' Ability to Read One's Emotional States." *Journal of Personality and Social Psychology* 75 (2): 332–46.
Gintis, Herbert. 2000. *Game Theory Evolving.* Princeton, N.J.: Princeton University Press.
Gould, Roger V. 2000. "Revenge as Sanction and Solidarity Display: An Analysis of Vendettas in Nineteenth-Century Corsica." *American Sociological Review* 65 (5): 682–704.
Granovetter, Mark S. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78 (6): 1360–80.
Hanneman, Robert A., Randal Collins, and Gabriele Mordt. 1995. "Discovering Theory Dynamics by Computer Simulation: Experiments on State Legitimacy and Imperialist Capitalism." *Sociological Methodology* 25:1–46.
Hechter, Michael. 1987. *Principles of Group Solidarity*. Berkeley and Los Angeles: University of California Press.
Hechter, Michael, and Karl Dieter Opp. 2001. *Social Norms.* New York: Russell Sage.
Heckathorn, Douglas. 1989. "Collective Action and the Second-Order Free-Rider Problem." *Rationality and Society* 1:78–100.
———. 1990. "Collective Sanctions and Compliance Norms: A Formal Theory of Group-Mediated Control." *American Sociological Review* 55 (3): 366–84.

Horne, Christine. 2001. "The Enforcement of Norms: Group Cohesion and Meta-Norms." *Social Psychology Quarterly* 64 (3): 253–66.

James, William. (1890) 1981. *The Principles of Psychology*. Cambridge, Mass.: Harvard University Press.

Katz, Daniel, and Floyd H. Allport. 1931. *Student Attitudes*. Syracuse, N.Y.: Craftsman.

Kitts, James A. 2003. "Egocentric Bias or Information Management? Selective Disclosure and the Social Roots of Norm Misperception." *Social Psychology Quarterly* 66 (3): 222–37.

Kleinberg, Jon. 2002. "Small-World Phenomena and the Dynamics of Information." In *Advances in Neural Information Processing Systems* 14, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani. Cambridge, Mass.: MIT Press.

Krech, David, and Richard S. Crutchfield. 1948. *Theories and Problems of Social Psychology*. New York: McGraw-Hill.

Kuran, Timur. 1991. "Now Out of Never: The Element of Surprise in the East European Revolution of 1989." *World Politics* 44 (1): 7–48.

———. 1995*a*. "The Inevitability of Future Revolutionary Surprises." *American Journal of Sociology* 100 (6): 1528–51.

———. 1995*b*. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge, Mass.: Harvard University Press.

MacLeod, J. 1995. *Ain't No Makin' It*. Boulder, Colo.: Westview.

Mark, N. 1998. "Beyond Individual Differences: Social Differentiation from First Principles." *American Sociological Review* 63 (3): 309–30.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415–44.

Merton, Robert K. 1968. *Social Theory and Social Structure*. New York: Free Press.

Miller, Arthur. 1996. "Why I Wrote *The Crucible*." *New Yorker*, October 21.

Miller, D. T., and C. McFarland. 1991. "When Social Comparison Goes Awry: The Case of Pluralistic Ignorance." Pp. 287–313 in *Social Comparison: Contemporary Theory and Research*, edited by J. Suls and T. A. Wills. Hillsdale, N.J.: Erlbaum.

Nagoshi, C. T., M. D. Wood, C. C. Cote, and S. M. Abbit. 1994. "College Drinking Game Participation Within the Context of Other Predictors of Other Alcohol Use and Problems." *Psychology of Addictive Behaviors* 8:203–13.

Newman, Mark. 2000. "Models of the Small World." *Journal of Statistical Physics* 101:819–41.

Nisbett, Richard E., and Dov Cohen. 1996. *Culture of Honor: The Psychology of Violence in the South*. Boulder, Colo.: Westview.

Nisbett, Richard E., and Timothy D. Wilson. 1977. "Telling More than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84:231–59.

O'Gorman, Hubert. 1975. "Pluralistic Ignorance and White Estimates of White Support for Racial Segregation." *Public Opinion Quarterly* 39 (3): 313–30.

———. 1986. "The Discovery of Pluralistic Ignorance: An Ironic Lesson." *Journal of the Historic and Behavioral Sciences* 22:333–47.

O'Gorman, H. J., and S. L. Garry. 1976. "Pluralistic Ignorance: A Replication and Extension." *Public Opinion Quarterly* 40:449–58.

Perkins, H. W., and H. Wechsler. 1996. "Variation in Perceived College Drinking Norms and Its Impact on Alcohol Abuse: A Nationwide Study." *Journal of Drug Issues* 26 (4): 961–74.

Prentice, Deborah A., and Dale T. Miller. 1993. "Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm." *Journal of Personality and Social Psychology* 64 (2): 243–56.

Savitsky, Kenneth, Nicholas Epley, and Thomas Gilovich. 2001. "Do Others Judge Us as Harshly as We Think? Overestimating the Impact of Our Failures, Shortcomings, and Mishaps." *Journal of Personality and Social Psychology* 81 (1): 44–56.

Schank, R. L. 1932. "A Study of Community and Its Group Institutions Conceived of as Behavior of Individuals." *Psychological Monographs* 43 (2): 1–133.
Sherif, Muzafer. 1936. *The Psychology of Social Norms*. New York: Harper.
Simmel, Georg. 1955. *Conflict and the Web of Group Affiliations*. New York: Free Press.
Skyrms, Brian. 1996. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
Vandello, J. A., and D. Cohen. 2004. "When Believing Is Seeing: Sustaining Norms of Violence in Cultures of Honor." In *The Psychological Foundations of Culture*, edited by M. Schaller and C. Crandall. New York: Lawrence Erlbaum.
von Neumann, John. 1966. *The Theory of Self-Reproducing Automata*, edited and completed by Arthur W. Burks. Urbana: University of Illinois Press.
Watts, D. J. 1999. "Networks, Dynamics and the Small-World Phenomenon." *American Journal of Sociology* 105 (2): 493–527.
———. 2002. "A Simple Model of Global Cascades on Random Networks." *Proceedings of the National Academy of Arts and Sciences* 99:5766–71.
Watts, D. J., and Steven Strogatz. 1998. "Collective Dynamics of Small World Networks." *Nature* 393 (6): 440–42.
Willis, Paul. 1977. *Learning to Labor*. New York: Columbia University Press.
Young, H. Peyton. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Insitutions*. Princeton, N.J.: Princeton University Press.