# THE EVOLUTION OF TRUST AND COOPERATION BETWEEN STRANGERS: A COMPUTATIONAL MODEL*

**Michael W. Macy**
*Cornell University*

**John Skvoretz**
*University of South Carolina*

*Social and economic exchanges often occur between strangers who cannot rely on past behavior or the prospect of future interactions to establish mutual trust. Game theorists formalize this problem as a "one-shot prisoner's dilemma" and predict mutual noncooperation. Recent studies, however, challenge this conclusion. If the game provides an option to exit (or to refuse to play), strategies based on "projection" (of a player's intentions) and "detection" (of the intentions of a stranger) can confer a "cooperator's advantage." Yet previous research has not found a way for these strategies to evolve from a random start or to recover from invasion by aggressive strategies that feign trustworthiness. We use computer simulation to show how trust and cooperation between strangers can evolve without formal or informal social controls. The outcome decisively depends, however, on two structural conditions: the payoff for refusing to play, and the embeddedness of interaction. Effective norms for trusting strangers emerge locally, in exchanges between neighbors, and then diffuse through "weak ties" to outsiders.*

Social exchanges sometimes involve an unavoidable time lag between promise and delivery (Coleman 1990). In these situations, both sides can benefit from an honest exchange, yet one or both may be cheated. The quality of goods or services exchanged may turn out to be less than expected after it is too late to recover. The exchange partners are then caught in a "social trap" in which the pursuit of self-interest can lead to mutual harm (Platt 1973).

Game theorists formalize this problem as the prisoner's dilemma (or PD)—a game between two players in which both have two choices: to "cooperate" (exchange honestly) or to "defect" (cheat, misrepresent quality, renege on promises, and so on). These two choices intersect at four possible outcomes, and each outcome has a designated payoff. $R$ (reward) and $P$ (punishment) are the payoffs for mutual cooperation and defection, respectively, while $S$ (sucker) and $T$ (temptation) are the payoffs for cooperation by one player and defection by the other. By definition, $T > R > P > S$. This payoff schedule creates an interesting tension between individual and collective interests. Collectively, the players are better off cooperating (and receiving $R$) than defecting (and receiving $P$). Individually, however, each player is better off defecting in a single encounter, no matter what the partner chooses. If the partner cooperates, it is better to defect (and receive $T$) than to cooperate (and receive $R$). If the partner defects, it is better to defect (and receive $P$) than to cooperate (and receive $S$). The trap is that the optimal choice for each player leads to a suboptimal (and often the *least* optimal) collective outcome.

The standard solution to these social traps is to impose formal or informal social con-

* Direct all correspondence to Michael W. Macy, Department of Sociology, Uris Hall, Cornell University, Ithaca, NY 14853 (mwm14@cornell.edu). This research was supported by grants from the National Science Foundation to Michael W. Macy (SBR 95-11461) and John Skvoretz (SES-9223192). We thank Robert Frank, James Kitts, and members of the Sante Fe Institute Workshop on "Social Interactions and Economic Activity," and *ASR* referees for their comments on early drafts of this paper.

trols (Heckathorn 1993). Formal controls require the organized ability to monitor and sanction compliance with the terms of an agreement (Hechter 1987). However, formal control can be costly if monitoring is difficult or if sanctions require extensive litigation. Informal controls usually rest on "the shadow of the future" (Axelrod 1984)—the prospect of retaliation or loss of reputation can deter both sides from defecting (Coleman 1990).

Not all exchanges, however, involve familiar faces or third-party regulation. In the dark alleys of social life the future does not cast a shadow. There, strangers meet outside the watchful eye of a Leviathan capable of enforcing compliance with a negotiated agreement. In such unregulated exchanges, rapscallions can renege with impunity, without fear of future retaliation, loss of reputation, or prosecution by enforcement agencies.

These conditions pose the prisoner's dilemma in its purest form, as players are stripped of all institutional or structural protection against exploitation. There is no tougher test of the possibility of cooperation between self-interested actors, or more generally, of self-interest as the basis of social order. In the absence of institutional protections, it is not surprising that many businesses shun better deals in the open market in favor of established suppliers (Yamagishi and Yamagishi 1996). These parochial tendencies toward "protection" and "in-group favoritism" run counter to universalistic principles of "free trade" and cultural pluralism.

The problem of transient and anonymous exchange is not only a matter of considerable practical interest; it is also one of the most theoretically compelling social traps. While the incidence of cheating may be higher between strangers than between neighbors, it is obviously not universal. Not all strangers are dishonest, nor are all cultures reluctant to do business with "outsiders." Why not?

We report the results of simulation experiments that identify a possible solution to the enigma of cooperation between strangers, a puzzle that has perplexed game theorists for over 40 years. Our design builds on the work of Orbell and Dawes (1991) and Frank (1988), who have identified traits that might protect cooperators from exploitation in a one-shot PD game with an option to exit (that is, an option to refuse to play). Orbell and Dawes posit a "cooperator's advantage," insofar as cooperators may not be as suspicious as defectors and thus would be less prone to elect the exit option. Frank adds the idea that cooperators may read "telltale signs of character" in order to avoid exchanging with defectors. Both of these processes would increase the odds that exchange will occur between cooperators.

These studies advance our understanding of the conditions required for cooperation, but they also raise new questions. If optimism confers a "cooperator's advantage," what keeps defectors from learning that paranoia does not pay? And once a convention is established for identifying trustworthy strangers, why do defectors not learn to feign trustworthiness? The answer, we suggest, lies in the structural embeddedness of the dilemma. Frank (1988), like Orbell and Dawes (1991), assumes a world in which any two members have an equal probability of interacting.[1] In contrast, we assume that the probability of a given pairing varies with the social and geographical proximity of prospective exchange partners. We hypothesize that conventions for trusting strangers take root in "neighborhoods" or "cliques" characterized by relatively dense interactions, and that these rules then diffuse to other regions via weak ties to members of socially distant neighborhoods or cliques.

## A PRISON WITH AN OPTION TO EXIT

Orbell and Dawes (1993) point out that the prisoner's dilemma is rarely played by "prisoners." Most exchange partners are free to walk away. With this "exit option," players have three choices: to *cooperate* (exchange honestly), to *defect* (cheat), or to *exit* (refuse to play). The decision to exit may be informed by an estimate of a prospective partner's trustworthiness. Previous research has proposed two mechanisms for making this estimate: projection of one's own inten-

[1] By "interaction," we mean an encounter between two potential exchange partners, which may or may not be consummated, depending on whether the interactants trust one another. Frank (1988) assumes an equal probability of *interaction* but not of a *consummated* exchange.

tions onto the partner, or detection of the partner's intentions.

### *Projection Strategies*

Consider a large population of *D*s (who always defect) and *C*s (who always cooperate) playing a series of noniterated PD games, each with a randomly chosen partner. Suppose the game has an option to exit and players project their own intentions onto prospective partners. As compared with *C*s, *D*s will be suspicious and more likely to refuse to play. The players that participate in exchange will therefore have a higher percentage of *C*s than will the overall population, reducing *C*s' vulnerability to exploitation and conferring a "cooperator's advantage" on those players who expect others to cooperate (Orbell and Dawes 1991).

As Orbell and Dawes acknowledge (1991: 525), this advantage does not explain the *evolution* of a tendency to project one's intentions. Any distribution of strategies that gives blindly trusting *C*s a "cooperator's advantage" would give an even larger "defector's advantage" to "mutant" optimistic *D*s. Hence, projection requires additional assumptions about exogenous selection pressures that protect *C*s against the evolution of nonprojecting, optimistic *D*s.

### *Detection Strategies*

A second mechanism for estimating a partner's trustworthiness is related to what Orbell and Dawes (1991) call "translucence." Suppose players can detect and interpret behavioral cues that reveal the intentions of a prospective partner. This will allow players to exchange selectively with only those who will cooperate. Thus, while projection makes *C*s more likely to accept an offer of exchange, detection makes offers by *C*s more likely to be accepted. Conversely, projection makes *D*s more likely to avoid exchange, while detection makes them more likely to be rejected when they try to participate. Either way, the expected outcome is the same—a cooperator's advantage.

Frank (1988, 1993) proposes that "moral sentiments," such as sympathy, compassion, or remorse, provide "telltale signs" that make cooperation between strangers viable in human populations. He reviews strong evidence indicating physiological links between emotional states and involuntary nonverbal behavior, especially facial expressions and voice tone (1993:165). These involuntary behavioral expressions of emotional states serve to telegraph an individual's intentions: So, when trustworthy individuals play prisoner's dilemma, two things happen. First, they feel emotionally compelled to do the right thing, even though they might be able to cheat with impunity. Second, they display body language triggered by the associated emotion, and these behavioral cues provide assurance against being cheated in one-shot exchanges.

An effective detection strategy requires the evolution of two traits: the ability to signal trustworthiness in a way that cannot be easily faked, and the ability to correctly read others' signals. The evolution of each of these abilities depends on the existence of the other.[2] Frank (1988) thus concludes:

> [I]t is extremely improbable for the physical symptoms of moral sentiments to have originated *because* of their signaling effects. The first small mutation toward a given symptom in the bearer of a sentiment would not have created the impression of a general relationship between the symptom and the behavior. So it is difficult to see how natural selection could have favored the symptom on account of any signaling role. (P. 110)

Instead, Frank prefers "an explanation in which the sentiments began by serving some purpose unrelated to their physical symptoms," namely, resistance to shortsighted temptation in iterated PD (1988: 110). Moral sentiments originate in ongoing relationships where they help players resist the temptation to cheat so that they might secure the benefits of more farsighted behavior. Players with such moral sentiments thus outperform those who lack the "moral fiber" needed to overcome shortsighted

[2] See Dawkins (1982:145–53) for a longer discussion of the "green beard problem" in which a pleiotropic gene is postulated with two complementary phenotypes—a conspicuous marker (e.g., a green beard) and altruistic behavior toward those similarly marked. Dawkins regards this fortuity as "too good to be true" and "vulnerable to a mutant arising which conferred the label without the altruism" (1982:145).

temptation, causing the sentiments to spread across the population.

The physical symptoms associated with moral sentiments are not needed in iterated PD games because the players can simply use each other's past behavior as an indicator of character. Nevertheless, having fortuitously evolved, these behavioral cues (or "telltale signs") now make cooperation viable in one-shot games as well. Frank (1988) shows how these cues facilitate the evolution of cooperation in a PD game with an option to exit. If the costs of effective scrutiny and refusal to exchange are not too great, *C*s with a keen eye will enjoy improved prospects for interacting with others like themselves.

More precisely, if everyone can distinguish between *C*s and *D*s without error, *C*s will never exchange with *D*s, so *D*s will be forced either to "work alone" or to interact with other *D*s, each receiving *P*, the punishment for mutual defection. *C*s will only interact with *C*s, each receiving *R*, the reward for mutual cooperation. When a *C* meets a *D*, the *C* will exit. Each side then receives *X*, the "exit" payoff (where $X = P$ in Frank's argument). *C*s thus come to dominate the population, and *D*s are driven to extinction.

Of course, if *C*s and *D*s cannot be distinguished, *C*s are doomed if they choose to play. However, Frank (1988) argues that the most likely outcome falls somewhere between the two extremes because of "noisy" signals. *D*s may give signals that are difficult to distinguish from those of *C*s, or worse yet, *D*s may deliberately try to deceive naive partners by mimicking *C*s. The problem parallels the danger posed by projection strategies: "Telltale signs" confer an advantage to fakers, just as projection creates a niche for optimistic cheaters. Whatever the advantage to the genuinely trustworthy, an even larger advantage accrues to counterfeits.

Nevertheless, detection and projection strategies differ in important ways. If *D*s mutate optimism as a counter-strategy to projection, the *C*s' only recourse is to become cynical, which deprives them of the "cooperator's advantage." Thus, the *D* counter-strategy cannot be foiled. In contrast, when *C*s use detection strategies, the *D*s' counter-strategy, faking trustworthiness, can be defeated. *C*s have a recourse: They can learn to spot counterfeits. Indeed, as the number of fakers in the population increases, the advantage begins to shift away from deception and toward detection. Thus, Frank (1988:268–69) predicts that the proportion of *C*s will move to an equilibrium that balances deception and the *C*s' ability to detect it, given the payoffs and the error rate in reading the "telltale signs."

Moreover, in the evolutionary foot race between deception and detection, the latter enjoys an inherent advantage according to Frank (1988). Because *D*s do not actually experience benevolent moral sentiments, their signals are less likely to be interpreted as genuine. Furthermore, *C*s can improve their ability to identify counterfeit *D*s, at some cost, by investing in better means of detection. Thus, Frank concludes that the population will always contain some *C*s who cannot be fooled and whose signals cannot be faked by *D*s.

This assumption provides the starting point for Frank's (1988) evolutionary argument. The opportunity for *C*s to hold out for honest partners generates stable equilibrium proportions of *C*s and *D*s in a population. Neither group can drive the other out. Deception means that some *D*s will remain active in the population, while the ability to see through deception means that some *C*s can also survive.

Although Frank's account is highly plausible, it nonetheless fails to fully explain why the distinctive behaviors associated with moral sentiments persist once cooperation flourishes. What prevents fakers from now evolving identical behaviors and wiping out the correlation on which the signaling convention depends?

Frank (1988) answers that moral sentiments are difficult to fake: Emotional mimicry is both imperfect and costly. This assumption may be reasonable if our interest is restricted to the genetic evolution of involuntary behavioral cues like facial expressions. Cultural cues, however, may be much easier to mimic. If so, then we make things too easy for ourselves by simply assuming away the ability to fake. Clearly, if we assume, on the basis of strong empirical evidence, that effective signaling systems exist, it is not difficult to explain the evolution of trust between strangers (for which there is also strong empirical evidence). But we are

still left wondering how these signaling systems evolved in the first place and how they resisted invasion by fakers. How do cooperators manage to evolve a code that defectors cannot crack?

The problem is similar to that confronted by Orbell and Dawes (1991), who show that an association between trustworthiness and trust can give *C*s a "cooperator's advantage." Frank (1988) adds an intervening step: an association between trustworthiness and a behavioral marker, and between the marker and trust. Like Orbell and Dawes, Frank fixes these underlying associations at some arbitrarily high value and then calculates the expected payoffs to *D*s and *C*s as the proportion of *C*s increases from zero to one. This procedure reveals stable equilibrium proportions of *D*s and *C*s at the point where the payoff differential is zero.

Under either mechanism—projection or detection—thriving cooperation creates an ecological niche for mutants who violate the association on which cooperation depends. Simply put, an association between trust and trustworthiness seems ripe to become a victim of its own success. In the projection model, *D*s should learn that paranoia no longer pays. In the detection model, *D*s should learn to fake the "telltale sign." Either way, the *C*s are doomed. Yet this has not always happened. Why not? Cooperation with strangers remains an enigma.

## THE EVOLUTION OF TRUST IN EMBEDDED PRISONER'S DILEMMA

Previous efforts to unlock this paradox have failed, we believe, because one-shot PD experiments have overlooked the embeddedness of the game in social networks that limit the possibilities for interaction. Formal models of the cultural evolution of cooperation have overlooked the social structure in which interaction takes place, perhaps because it is assumed that social ties are not relevant in games that are played in a world of one-night stands. This random-pairing assumption, however, obscures what may be important properties of social interaction—in particular, the density of interaction may differ within and between groups.

Recent modeling work in biology and economics underscores the need to explicitly theorize about the structures governing encounters between PD players. Enquist and Leimar (1993) show that the prospects for cooperation are adversely affected when free-riders move through the population, exploiting cooperators in series of one-shot games and switching partners before they can be sanctioned. Suspiciousness of strangers lengthens search times to find new victims, and this, in turn, enhances the prospects for cooperation. In addition, as Wilson, Pollock, and Dugatkin (1992) show, some interaction beyond the local group is necessary for cooperation to spread once it develops in a local population, given that players benefit from the actions of nearby cooperators in direct proportion to the number of cooperators in the population. In economics, Ellison's (1993) analysis shows that the rate at which coordination games converge to "risk dominant" equilibria depends greatly on the structure of interaction among players. He proves that in large populations random interaction within the entire population produces very long convergence times, while interactions confined to nearest neighbors dramatically abbreviate the time to convergence.

Although Frank (1988) allows only one-shot games in his formal model, he assumes that the initial correlation between "telltale signs" and strategies has already evolved from ongoing relationships. We build on Frank's insight by pushing the starting point of the model back in evolutionary time, to the emergence of moral sentiments in repeated games. We hypothesize that embedded social ties, as might be found in tightly knit congregations or neighborhoods, facilitate the coordination of effective trust conventions, while nonembedded encounters between random strangers diffuse these conventions across the population.

To test this theory, we use a genetic algorithm to implement a dynamic model based on the static models used by Frank (1988) and by Orbell and Dawes (1991). We then simulate the evolution of trust and cooperation in a computational ecology, but with two important differences. First, in his analytical model, Frank assumes that everyone is a stranger. We assume, however, that exchanges are embedded in a social network, such that all players participate in two types of relationships—with "neighbors" and with "strangers." Inter-

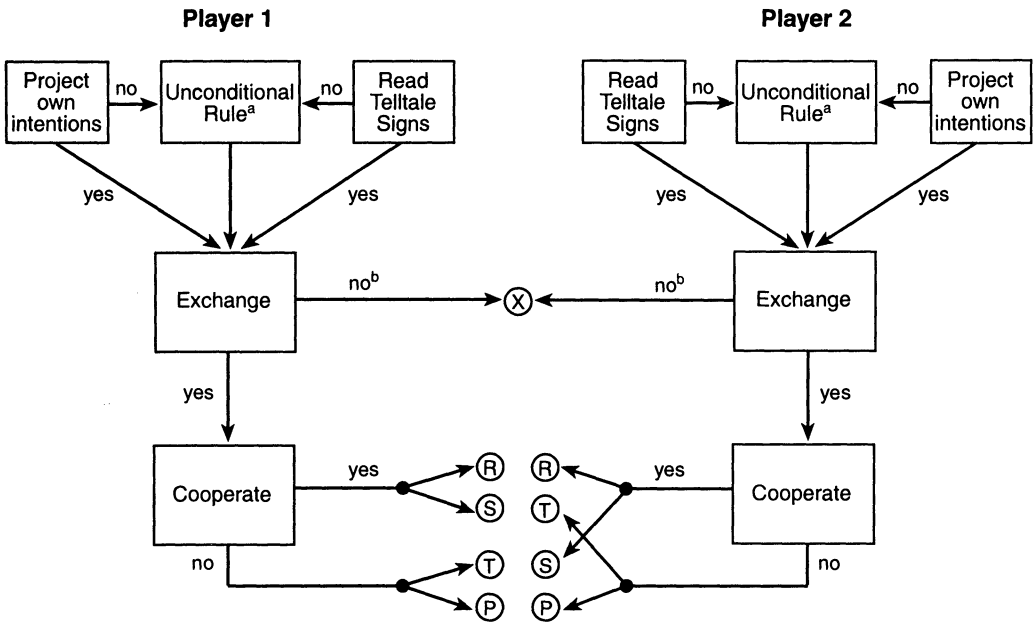


**Figure 1. Decision Tree for Prisoner's Dilemma with Option to Exit**

[a] "Always exchange" or "always exit."

[b] If either player exits, both receive payoff = *X*.

actions with neighbors exhibit a high degree of embeddedness, insofar as the partners a player encounters are likely to (a) encounter each other, and (b) re-encounter the player. Interactions with strangers, which are the focus of our concern, are minimally embedded, insofar as strangers are unlikely to re-encounter one another or to encounter their partners' associates.[3]

Second, previous researchers assume a fixed, exogenous correlation between cooperation and either trust (Orbell and Dawes 1991) or telltale behavioral cues (Frank 1988). In our dynamic model, these correlations must evolve. There is no guarantee that *C*s will be more likely to exchange with strangers or to display the appropriate marker. Nor do we assume that the display cannot be mimicked by *D*s, or that *C*s will know how to read the marker correctly. We then test to see if effective rules for trusting others can evolve in neighborhood interactions and spread across the population through contact with strangers.

### *A Genetic Learning Model*

Our model of the evolution of trust and cooperation between strangers assumes that players make several interrelated decisions when they contemplate interacting with a partner. These decisions are diagrammed in Figure 1. First, a player may choose to read or ignore several possible clues to the partner's character (detection), or may simply assume that the partner is just like oneself (projection). Based on whatever information (if any) is used, a player must then decide either to avoid the partner ("distrust") or to participate in an exchange, that is, to play a round of PD ("trust"). Players who trust their partners and elect to exchange expose themselves to the risk of being cheated.[4] Finally, if the decision is to trust,

[3] For convenience, we refer to the embedded PD as an "iterated" game and the unembedded PD as a "one-shot" game. However, "embedded" and "iterated" are not strictly equivalent. More precisely, the embedded PD carries a very high probability that the partner will be revisited, while the unembedded PD carries a very low probability.

[4] "Trust" generally implies "trustworthiness," which precludes the possibility that defectors can trust their partners. However, we follow Yamagishi and Yamagishi (1996) who use a behavioral

then the player must decide to cooperate or defect. These decisions are based on strategies or rules that evolve over time in response to successes and failures.

To model the evolution of trust and cooperation between strangers we use a genetic algorithm, a simple and elegant way to write game-playing strategies that can progressively improve on performance by building on partial solutions.[5] Each strategy in a population consists of a string of symbols that code behavioral instructions. These symbols are binary digits (or "bits") with values of 0 or 1. A string of symbols is analogous to a chromosome containing multiple genes. A set of one or more bits that contains a specific instruction is analogous to a gene. The values of the bits and bit-combinations are analogous to the alleles of the gene. A one-bit gene has two alleles (0 and 1), a two-bit gene has four alleles (00, 01, 10, and 11), and so on. The number of bits in a gene depends on the complexity of the instruction. For example, an instruction to cooperate/defect requires only a single bit. However, an instruction to cooperate, defect, or refuse to play requires two bits. The first tells whether to play, and if so, the second tells whether to play without cheating.

A strategy's instructions, when followed, produce an outcome (or payoff) that affects the player's reproductive fitness relative to other players in the computational ecology. In our model, fitness is a function of cumulative payoffs in repeated PD games. Relative fitness determines the probability that each strategy will propagate. Propagation occurs when two mated strategies recombine. If two different rules are both effective, but in different ways, recombination allows them to create an entirely new strategy that may integrate the best abilities of each "parent," making the new strategy superior to either contributor. If so, then the new rule may go on to eventually displace both parent rules in the population of strategies. In addition, the new strings may contain random copying errors. These mutations continually refresh the heterogeneity of the population, counteracting selection pressures that tend to reduce it.

### *Chromosome Structure*

We use a "chromosome" of 15 "genes" to code relevant behavioral instructions (described in Table 1). Gene 1 defines a player's "character" in exchanges with strangers. *D*s ($G_1 = 0$) will always cheat a stranger (defect), while *C*s ($G_1 = 1$) will always exchange honestly (cooperate). A *C*'s only protection against an unfamiliar *D* is to exit, based on strategies of conditional exchange, such as projection and detection.

Projection is triggered by the player's internal state without regard to any information about a prospective partner. Detection strategies, however, require effective cues. Genes 2 and 3 control the expression of these signs. Gene 2 controls a cultural or physiological marker (for instance, religious affiliation or facial expression) whose display is linked to character. If $G_2 = 0$, then *D*s (i.e., players with $G_1 = 0$) display the marker and *C*s (i.e., players with $G_1 = 1$) do not; whereas if $G_2 = 1$, then *C*s display the marker and *D*s do not.[6] Gene 3 controls a behavioral marker (or "greeting") that is unrelated to character; it is either displayed ($G_3 = 1$) or hidden ($G_3 = 0$). In order for detection strategies to be effective, players must learn to see past the "greeting" and focus on the marker tied directly to character by $G_2$.[7]

The remaining genes control a player's decision to trust a prospective exchange partner. The behavioral indicator of trust is the decision to exchange (to play the PD game) and thus to expose oneself to being cheated. Five cues may be used to decide whether to trust a prospective partner: (1) one's own in-

definition of trust: the willingness to participate in a risky venture. Hence, trust can apply to defectors as well as cooperators. For example, a dishonest shopkeeper might trust an unfamiliar customer and accept a bad check for shoddy goods.

[5] Recent work in sociology using genetic algorithms includes Egidi and Marengo (1995), Macy (1996), McCain (1995), Skvoretz and Fararo (1995), and Treuil (1995).

[6] Note that it is possible for both *D*s and *C*s in a population to display the marker only if there is a correlation between the first two genes such that the first two values are either 00 or 11.

[7] Genes 1 through 3 also can also be modeled as a single pleiotropic three-bit gene with eight alleles, ranging from marked *D*s who do not greet (000) to marked *C*s who greet you with a smile (111).

**Table 1. "Chromosome Structure" for the Genetic Algorithm**

| Gene | Allele | Rule |
|---|---|---|
| 1 | 0 | *D* (*defector*, always defects). |
| | 1 | *C* (*cooperator*, always cooperates). |
| 2 | 0 | Display marker if you are a defector ($G_1 = 0$). |
| | 1 | Display marker if you are a cooperator ($G_1 = 1$). |
| 3 | 0 | Do not greet your partner. |
| | 1 | Greet your partner. |
| 4 | 0 | Ignore your own intentions. |
| | 1 | Base trust on your own intentions, given $G_9$. |
| 5 | 0 | Ignore partner's $G_1$ marker. |
| | 1 | Attend to partner's $G_1$ marker, given $G_{10}$. |
| 6 | 0 | Ignore partner's greeting. |
| | 1 | Attend to partner's greeting, given $G_{11}$. |
| 7 | 0 | Ignore partner's group membership. |
| | 1 | Attend to partner's group membership, given $G_{12}$. |
| 8 | 0 | Ignore partner's relative success. |
| | 1 | Attend to partner's relative success, given $G_{13}$. |
| 9 | 0 | Assume others are the opposite of your $G_1$. |
| | 1 | Assume others are the same as your $G_1$ (project). |
| 10 | 0 | Distrust those who display the $G_1$ marker, trust those who do not. |
| | 1 | Trust those who display the $G_1$ marker, distrust those who do not. |
| 11 | 0 | Distrust those who greet you, trust those who do not. |
| | 1 | Trust those who greet you, distrust those who do not. |
| 12 | 0 | Distrust neighbors, trust strangers. |
| | 1 | Trust neighbors, distrust strangers. |
| 13 | 0 | Distrust those who are relatively successful, trust those who are not. |
| | 1 | Trust those who are relatively successful, distrust those who are not. |
| 14 | 0 | Distrust everybody. |
| | 1 | Trust everybody. |
| 15 | 0 | Cooperate or defect unconditionally, based on $G_1$. |
| | 1 | Override $G_1$ and cooperate if the partner is a neighbor. |

tentions ($G_4$); (2) the partner's $G_2$-controlled marker ($G_5$); (3) the partner's "greeting" ($G_6$); (4) whether the partner is a neighbor or stranger ($G_7$); and (5) the partner's relative success ($G_8$).

Genes 9 through 13 specify how these cues are used. For example, if $G_5 = 1$, then the player judges character on the basis of the marker controlled by the partner's $G_2$. If $G_{10} = 0$, then those who do not exhibit the marker are trusted; but if $G_{10} = 1$, then those who exhibit the marker are trusted. If $G_7 = 1$, the player evaluates character based on whether the partner is a neighbor or a stranger. If $G_{12} = 0$, the player trusts strangers and distrusts neighbors; but the opposite occurs if $G_{12} = 1$. If a player attends to none of the five cues, that is, genes 4 through 8 all equal 0, then gene 14 determines whether to trust a prospective partner, based on a strategy of unconditional exchange ($G_{14} = 1$) or unconditional exit ($G_{14} = 0$).

If more than one bit of information is used, each bit counts equally. Thus, a partner who emits "mixed signals" elicits trust with a corresponding probability. For example, if a player uses three bits of information to assess whether to exchange, and two of the three suggest trust while the third indicates distrust, then the player can be expected to trust such a partner two times out of every three such encounters.[8]

Finally, gene 15 allows *D*s to evolve a strategy to cooperate only with neighbors ($G_{15} = 1$), instead of cheating everyone ($G_{15} = 0$). ($G_{15}$ thus has no effect on *C*s.) We want to see if a cooperative strategy of conditional exchange, based on projection or detection, can compete successfully against a "parochial" strategy of conditional cooperation ("Cooperate with your neighbors and cheat everyone else").

[8] Alternatively, we could have equipped players with the ability to search for all possible patterns consisting of zero to five bits of information. We could also greatly expand the number of sensory inputs, or simply assign greater weight to ineffective cues, in order to find the minimum signal-to-noise ratio for players' ability to locate effective cues. However, the purpose of this study is to establish the possibility of cooperation between strangers in one-shot PD, not to define the limits of that possibility. Thus, we leave the latter for future research.

To illustrate this design, consider the following 15-bit string, extracted from a keen-eyed $C$ in one of the simulation runs:

110010001101110.

Reading from left to right, this player is a $C$ ($G_1 = 1$) who is marked ($G_2 = 1$) but who does not offer an arbitrary greeting ($G_3 = 0$). For this player, $G_4$ through $G_8$ reads 01000, which means that the player only uses the behavioral marker for judging prospective partners ($G_5 = 1$), and trusts those who display this sign ($G_{10} = 1$). The player does not project intentions onto others ($G_4 = 0$), is not fooled by the partner's greeting ($G_6 = 0$), and does not care if the partner is a neighbor ($G_7 = 0$) or is relatively successful ($G_8 = 0$).

Why these five "trust genes" and not others? As we noted above, previous research has suggested that projection and detection are important partner-selection strategies, so we created a genetic structure in which such strategies might evolve. The embeddedness of the PD game in a social network clearly mandates that we allow for the evolution of strategies for trust and cooperation exclusively with neighbors. Given our assumption that selection pressures are based on fitness (cumulative payoffs), we must allow for the possibility that success might trigger trust and cooperation as well as the selection of a role model. Finally, we include a voluntary signal ("greeting") because we want to allow for the possibility that detection strategies must cope with a marker that is not related to character.

Obviously, this player architecture determines the results by restricting the opportunities available to the evolutionary process. We preclude other abilities, such as empathic sensitivity or a gift for remembering faces, that might make "telltale signs" entirely superfluous. This makes the "chromosome" structure nonrandom, as in all models based on genetic programming. The random start refers to the contents of the chromosome (the random allocation of alleles), not to its structure. The key question is not whether we can obtain the same results with other structures, but whether the structure we have adopted guarantees (or "wires in") our results. It does not. Unlike Orbell and Dawes (1991), we do not build in the tendency for cooperators to project—only the possibility. A correlation between cooperation and trust may or may not evolve, depending on the effectiveness of this strategy in any given population of strategies. Unlike Frank (1988), we do not preclude the possibility that defectors could learn to perfectly mimic cooperators. They can, and if they do they will undermine the effectiveness of "telltale signs." In short, we have designed a computational experiment to test the evolutionary plausibility of the assumptions that were the starting points in Orbell and Dawes's and Frank's studies of cooperation between strangers.

### *Experimental Design*

Our experiment involves a series of PD games, some embedded and some not, with the bilateral payoffs $R = 3$ for mutual cooperation and $P = 1$ for mutual defection, and the unilateral payoffs $S = 0$ for cooperation and $T = 4$ for defection. We also test $T = 5$, the value assumed by Frank (1988). Evolution begins from a random start (randomly selected 15-bit strings). Reproductive fitness is determined by a weighted moving average of cumulative payoffs in consecutive games, each with a partner chosen from a population of 1,000. The formula is:

$$F_{it} = \frac{\sum_{t'=1}^{t} .999^{t-t'} \mathbf{O}_{t'}}{\sum_{t'=1}^{t} .999^{t-t'}}, \qquad (1)$$

where $F_{it}$ is the fitness score for player $i$ at time $t$, $\mathbf{O}_{t'}$ is the vector of previous payoffs from $t' = 1$ to $t$, and $.999^{t-t'}$ is the weighting factor that assigns greater weight to more recent outcomes. The weighted moving average minimizes sensitivity to the start parameters while allowing sufficient continuity for selection pressures to test emerging solutions.

The experiment involves three structural manipulations: (1) the relative cost of exit ($X = P$ or $P < X < R$), (2) neighborhood size, and (3) embeddedness of interaction. By "embeddedness of interaction" we mean the extent to which interaction with prospective partners is limited to those in a player's neighborhood. If embeddedness equals 1, its maximum, all interactions are limited to neighbors. If embeddedness equals 0, all in-

teractions are with those outside the neighborhood. If embeddedness equals .5, then a player is just as likely to interact with a neighbor as with a stranger. We assume that players are more likely to encounter neighbors than strangers, and that neighborhoods are small, relative to the size of the general population. We experiment with neighborhoods ranging in size from 10 to 50 members, with embeddedness ranging from .5 to .9 (i.e., 50 percent to 90 percent of all encounters are with neighbors). If player $i$ is to be assigned a partner from $i$'s neighborhood, then $i$'s choice of partner is a random selection from $i$'s $S - 1$ neighbors, where $S$ is the size of $i$'s neighborhood and $i \in [1, S]$ such that $i$ has not already been selected as a partner by someone else. The pairing algorithm for assigning $i$ to neighbor $A_i$ at time $t$ is as follows:

$$A_{it} = (i - 2 + \text{ran}\{2..S\}) \bmod S + 1. \quad (2)$$

If player $i$ is to be assigned a stranger, $i$ first randomly selects a neighborhood, $z_{A'}$, from the other $Z - 1$ neighborhoods, and then randomly selects a partner, $A'_i$, from any of that neighborhood's $S$ residents. The pairing algorithm for assigning $i$ to a stranger $A'_i$ (where $i$ has not already been chosen as a partner by anyone else) is as follows:

$$A'_{it} = \text{ran}\{1..S\} + ([z_i - 2 + \text{ran}\{2..Z\}] \bmod Z)\, S. \quad (3)$$

### *Reproduction and Propagation*

We assume that strategies propagate through social contact. Hence, a relatively successful player can influence only those he or she encounters as potential game partners.[9] We also explore the idea that influence is limited to a narrower circle of contacts, based on the intuition that influence might be compromised by the failure to consummate an exchange. We thus test two specifications: one in which influence is limited to the population that a player encounters, and a more restrictive condition in which influence is limited to just those partners with whom an exchange is consummated.

Although the genetic algorithm borrows its logic from natural selection, it is not limited to models of biological evolution. Rather, ours is a generic search algorithm based on an abstract model of recombinant replication in an artificial world. Accordingly, *"genes" in our model refer to strings of binary instructions, not to pieces of DNA*. We are not trying to predict the consequences of sexual reproduction or to explain the biological or cultural evolution of traits that might be observed in natural settings. Rather, our objective is to test the possibility of emergent social order under conditions previously believed to preclude it. This is a computational "thought experiment" whose results have greater plausibility and relevance if the model is highly abstract and elementary, rather than closely tied to a particular selection mechanism, whether biological or cultural.

Specifically, selection pressures in our design operate through direct transfer of genetic instructions from those with higher fitness to those with lower fitness. When two players are paired, the less fit partner randomly replaces selected portions of its chromosome with the corresponding parts of the more fit partner's chromosome. Unlike sociobiological specifications of the genetic algorithm, our design does not use crossover between paired chromosomes. Instead, we allow each bit on the chromosome to be replaced independently of any other. This eliminates any effect of an arbitrary placement and sequencing of genes (Goldberg 1989; Holland 1975). For each of the 15 bits, the allele for the less fit partner is switched to that of the more fit partner with a probability of .5.[10] Thus, a $C$

[9] We leave for future research a more complex social structure in which players participate in two independent networks—one that limits potential exchanges and another that constrains the spread of social influence.

[10] In this experiment, influence is not exchanged and mutation does not occur between equally fit individuals. Future research could explore constraints on the propagation of successful rules, such as making the probability of gene-replication a function of the difference in fitness of the two partners (rather than a constant probability of .5), or even assigning some cost to the adoption of a new rule. These changes would slow the evolutionary process and increase the step size in the gradient search. Our interest here, however, is only to demonstrate the evolutionary *possibility* of trust and cooperation between strangers. Future researchers may want to inves-

could become a $D$, or a marked $C$ could become an unmarked $C$, and so on.

Finally, genes are copied with an error rate of .01 (on average, 1 out of every 100 genes is copied with its allele switched). This rate of mutation ensures sufficient heterogeneity to keep evolution vigorous and provides a conservative test of the stability of emergent equilibria.[11]

## EXPERIMENTAL RESULTS

We report the results of our simulation experiments in three parts. The first part compares two populations that differ only in the payoff for refusing to exchange (exit). Neighborhood size and interaction embeddedness are held constant. In the first population, the exit payoff equals the payoff from mutual defection; in the second population, the exit payoff is preferred to the payoff from mutual defection but is less attractive than the payoff from mutual cooperation. The second part explores the basis for trust and cooperation between strangers in the second population. In the third part, we vary neighborhood size and embeddedness of interaction and test their effects on trust and cooperation between strangers. In these populations, the exit payoff is preferred to the payoff from mutual defections, but the payoff from mutual cooperation is preferred to the exit payoff.

### *Exit Payoffs and Trust and Cooperation between Strangers*

Our first experiment tests the hypothesis that the evolution of trust depends on the exit payoff (for refusing to play) relative to the payoff for mutual defection (playing but both choosing to defect). The experiment uses two populations, each consisting of 1,000 players, who play PD with 9 neighbors and 990 strangers. Interaction embeddedness is set at .67—two-thirds of all encounters are with a player's 9 neighbors. We allow evolution to proceed for 100,000 iterations and run 10 replications of each condition. The two populations differ only in the payoff for refusal to exchange. The first population plays a game based on Frank's (1988) assumption that the exit payoff equals the cost of mutual defection ($X = P = 1$). The second population uses Orbell and Dawes' (1991) assumption that the exit payoff is preferable to the payoff from mutual defection but is less attractive than the payoff from mutual cooperation ($P < X < R$). With $P = 1$ and $R = 3$, we let $X = 2$. All other conditions remain unchanged.

Table 2 compares the mean levels of trust and cooperation (averaged over 10 replications) between neighbors and between strangers for the two populations. Trust (the decision to play) and mutual cooperation (the choice of both sides to play and cooperate) refer to the phenotypical behaviors of the players (i.e., what they actually do in the encounters). These phenotypes are disaggregated for encounters with neighbors and strangers. Table 2 also reports the distribution of genotypes for conditional and unconditional cooperation ($G_{15}$ and $G_1$) as well as the difference between the means for the two groups of replications ($X = 1$ and $X = 2$).

The results show that when the exit option is relatively costly ($X = 1$), players generally elect to trust one another, whether the partner is a neighbor (.87) or stranger (.79). With exit as costly as mutual defection, the players have little choice but to trust. However, while players often cooperate with neighbors (.53), they tend to defect with strangers (the rate of cooperation is only .04). The chance to interact frequently with a small group of familiar faces leads to cooperation among neighbors. But neighborhood cooperation does not "jump start" cooperation between strangers. This result is consistent with the elementary game-theory principle that cooperation is only viable in PD games when there is a high probability of re-encountering

tigate the robustness of trust and cooperation under various assumptions about the efficiency of the search mechanism.

[11] Note that gene-copying, and hence mutation, occurs immediately after a given interaction. Note also that our formulation of the mutation process means that mutation is no more likely to occur if two mated strategies differ in nearly all alleles or only in one. Bergin and Lipman (1996) recently have shown that selection of a strategic equilibrium follows from the assumption that mutation rates depend on system states, but their results do not apply to our model because our mutation rate does not depend on the particulars of the strategies involved.

**Table 2. The Effect of Exit Payoff on Trust and Cooperation between Strangers: Means and Standard Errors over 10 Replications in Each Payoff Condition**

| Variable | Exit Payoff | | Difference between Means |
|---|---|---|---|
| | $X = 1$ | $X = 2$ | |
| "Always cooperate" (Gene 1) | .09 (.01) | .57 (.03) | .48** |
| "Cooperate with neighbors" (Gene 15) | .76 (.003) | .48 (.07) | −.28* |
| Trust in neighbors [a] | .87 (.001) | .60 (.02) | −.27** |
| Trust in strangers [a] | .79 (.005) | .49 (.03) | −.30** |
| Cooperation with neighbors [b] | .53 (.016) | .42 (.03) | −.11* |
| Cooperation with strangers [b] | .04 (.005) | .29 (.04) | .25** |

*Note*: Numbers in parentheses are standard errors.

[a] Observed (phenotypic) rate of exchange.

[b] Observed (phenotypic) rate of cooperation when choosing to exchange.

*$p < .05$ **$p < .01$ (two-tailed tests)

others. Although the players also play an iterated PD, defection remains the dominant strategy in the one-shot game, and this is reflected in the low level of mutual cooperation with strangers. Inspection of the resulting genetic distribution shows that the population is dominated by a strategy of conditional cooperation based on social or geographical distance. The relative frequency of $G_{15} = 1$, averaged over 10 replications of 100,000 iterations, is .76, compared to only .09 for $G_1 = 1$.

### *Less Costly Exit: The Evolution of Trust and Cooperation between Strangers*

In the second population, with the exit payoff raised to $X = 2$, Table 2 shows a strikingly different result: a sharp increase in the level of cooperation between strangers. While the opportunity to exit caused the level of trust in strangers to drop significantly (from .79 to .49), the rate of mutual cooperation between strangers increased significantly (from .04 to .29). This reflects the decline in the relative frequency of $G_{15} = 1$, from .76 to .48, and the increase in $G_1 = 1$, from .09 to .57.

Figure 2 offers a dynamic look at this second population in which the exit payoff is less costly than mutual defection ($X = 2$). The figure is based on a representative simulation from the 10 replications used in Table 2 and reports the genotypes for conditional and unconditional cooperation ($G_{15}$ and $G_1$) and phenotypes for trust and mutual cooperation with a neighbor or stranger. At about $t = 5{,}000$, the proportion of the population that cooperates unconditionally ($G_1 = 1$) increases from nearly 0 to about .9 at $t = 7{,}000$. The increase occurs while the level of trust (or willingness to exchange) between strangers is about .2, giving a conditional probability of exchange of about .04. Trust in neighbors emerges very quickly, but trust between strangers is not established until about $t = 20{,}000$. This lag in willingness to exchange with strangers indicates that trust conventions form in local, embedded social relations and then spread through contact to distant others.

We do not need simulations to predict the consequences if the exit payoff exceeds that of mutual cooperation ($X > R$). No one will exchange and no one will learn how to exchange safely with strangers. Trust conventions are superfluous in a world where relative payoffs favor exiting, or refusing to exchange. Thus, if it is better to avoid exchange, the $C$ population cannot learn how to interact successfully, making participation in exchange more difficult and dangerous. But if refusing to exchange is too costly, honesty has no place to hide in a predatory world. Trust between strangers can emerge only within a defined range in the relative cost of exit, $P < X < R$.[12]

[12] This result held up even when we imposed more restrictive assumptions about the propagation of strategies, namely, that players who exit cannot influence or be influenced by a distrusted partner. Players who never exchange are then immune to social influence, creating an evolutionary trap, a stable equilibrium of universal distrust, with no possibility of further evolution. To avoid this, we assumed that the game did not always have an option to exit. Five percent of encounters required both sides to exchange. We found that this rate of forced interaction was sufficient to allow evolution to search for a solution that might

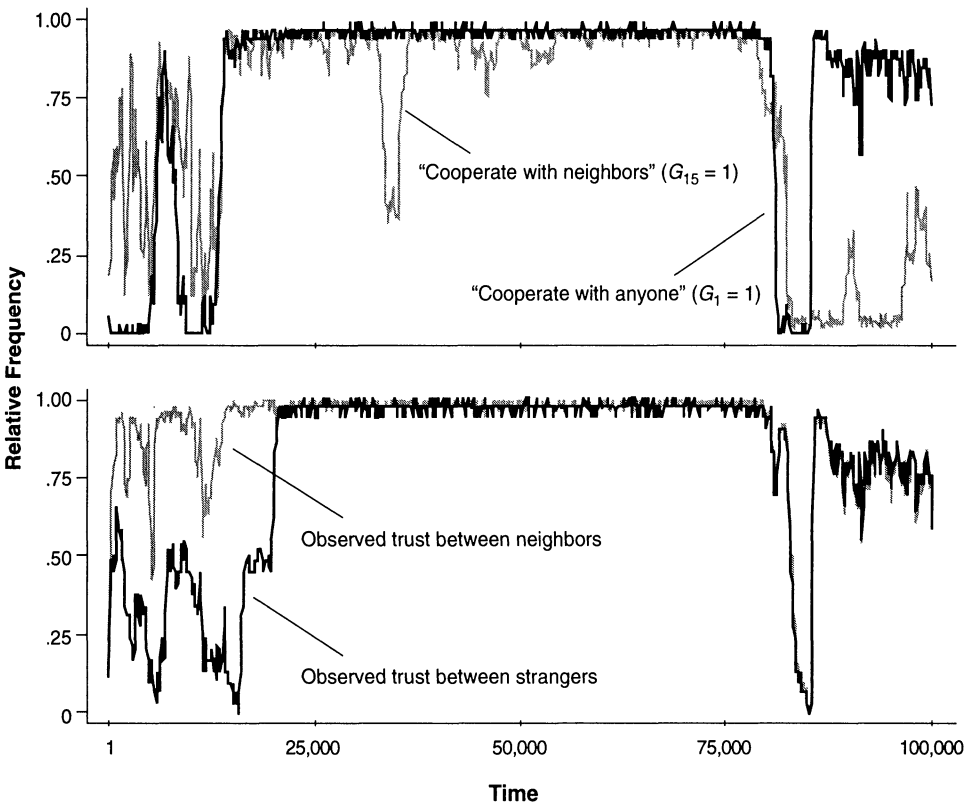


**Figure 2. The Evolution of Trust and Cooperation between Neighbors and between Strangers**

*Note*: $N$ = 1,000; $T$ = 4, $R$ = 3, $X$ = 2, $P$ = 1, $S$ = 0; neighborhood size = 10; embeddedness = .67.

While it is useful to identify scope conditions for the evolution of trust between strangers, what is most remarkable about these results is that such cooperation is not impossible. Although previous researchers (Frank 1988; Orbell and Dawes 1991) have demonstrated the possibility of cooperation between strangers, they assumed that cooperators had innate abilities to project and to signal other cooperators in ways that could not be effectively faked by defectors. Skeptics might rightly question whether these abilities could even evolve in the first place. Our experiment shows that they can. In certain conditions, cooperation between strangers can evolve from a random start, that is, from a state in which the abilities to project and to spot fakers are not already present at the outset.[13]

A caveat is in order, however. We do not claim that this central finding is robust across other specifications of the chromosome structure or network structure in this experiment. On the contrary, the evolution of cooperation between strangers is exceedingly fragile. It is trivial to introduce changes in the player architecture and social structure

be superior to universal distrust. The results were parallel to those observed with less restrictive assumptions about influence but had slightly lower levels of cooperation. This was due to the toll that occasional forced interactions imposed on the $C$ population. We also tested the effect of raising the temptation to exploit to $T$ = 5, the level assumed by Frank (1988). As expected, this made it much easier for $D$s to learn how to take advantage of emergent norms, but it did not prevent effective conventions from reemerging.

[13] "Random start" refers only to the initial randomness in the associations between cooperation and trust, between cooperation and marker, and between marker and trust. We do not begin with a completely unstructured interaction, but unlike previous researchers, we do not wire in the associations that are necessary for cooperation with strangers to evolve.

that will preclude cooperation outside the neighborhood, regardless of the exit payoff. Our cumulative research experience shows overwhelmingly how extraordinarily difficult it is for cooperation between strangers to survive the relentless evolutionary pressure to "hit and run." The difficulty of evolving cooperation in a one-shot game makes this demonstration of a logically feasible evolutionary path all the more noteworthy.

## HOW TRUST BETWEEN STRANGERS EMERGES: THE FORMATION OF TRUST CONVENTIONS IN LOCAL INTERACTION

The population data reported in Figure 2 show that effective trust conventions can evolve, but these data do not provide a micro-level description of the evolutionary sequence at the neighborhood level. The graphs in Figure 3 track the genetic basis (or genotypes) for the behavioral changes reported in Figure 2. The graphs identify which social cues were most responsible for the viability of cooperation between strangers. Figure 3 also shows, as a benchmark, the rate of trust between strangers.

### *Stages in the Emergence of Trust*

The emergence of trust between strangers involves a series of stages, beginning with a state of anomie and ending with effective trust conventions. The sequence unfolds as follows:

***Widespread distrust.*** Following the random start, all members of the population generally distrust neighbors and strangers alike. In Figure 2, this is evident when $t < 300$. The level of trust is about .15 for both neighbors and strangers, giving an expected rate of consummated exchange of about .02.

***Drift toward trustworthiness.*** With widespread distrust, selection pressures weakly distinguish between alleles for defection and cooperation, allowing both $G_1 = 1$ and $G_{15} = 1$ to diffuse by evolutionary drift. Between $t = 0$ and $t = 300$, the number of cooperators among neighbors increases from 0 to .38, based on $G_{15}$ (see Figure 2).

***Local trust.*** The spread of cooperation is uneven, with some neighborhoods becoming more cooperative than others. One of the more cooperative neighborhoods becomes the site where trust first gains a foothold. Figure 3 shows that the earliest trust rule is based on social distance—trust neighbors but not outsiders ($G_7 = 1$). This rule emerges at around $t = 300$ and leads to a sharp increase in the rate of local cooperation.

***Local cooperation.*** Whenever two like-minded $C$ neighbors meet, they trigger the strategy to "trust and cooperate" and receive $R$ instead of $X$, thereby gaining in fitness relative to their neighbors. Because of this fitness advantage, neighbors are more likely to imitate them. Their local influence grows and provides them with additional neighbors with whom they can safely cooperate. The $C$s now flourish within this neighborhood at the $D$s' expense.

***Diffusion via contact with strangers.*** Members of the vanguard neighborhood encounter suspicious strangers from "backward" neighborhoods (where trust has not yet evolved), but they do not consummate exchange due to their partners' distrust. However, these vanguard players still "infect" their contacts with their rules for cooperation and trust. If infected contacts are more influential than their neighbors, the rule will spread throughout the infected neighborhood, and from there to new neighborhoods, causing the $C$ population to rapidly increase. In Figure 2, after several "false starts," the population with $G_{15} = 1$ ("cooperate with neighbors") stabilizes at about .9 by $t = 16{,}000$. Distrust among strangers is still rampant at this point, allowing $G_1 = 1$ to coevolve as a rule for neighborhood cooperation. By $t = 16{,}000$, $G_1 = 1$ is as widespread as $G_{15} = 1$.[14]

***Universal trust.*** If the vanguard neighborhood is originally organized around a rule to trust only neighbors, the spread of cooperative strategies now creates an opening for a rule to trust strangers. Two rules are possible:

[14] In a population that cooperates only with neighbors ($G_{15} = 1$), a rule to trust only neighbors ($G_7 = G_{12} = 1$) is also viable and often emerges. However, these rules allow the population to drift toward $G_1 = 1$, which opens up a niche for more courageous rules (such as projection or detection) that not only reap the rewards of local interaction, but in addition learn to take advantage of what would otherwise be missed opportunities for exchange with trustworthy strangers.

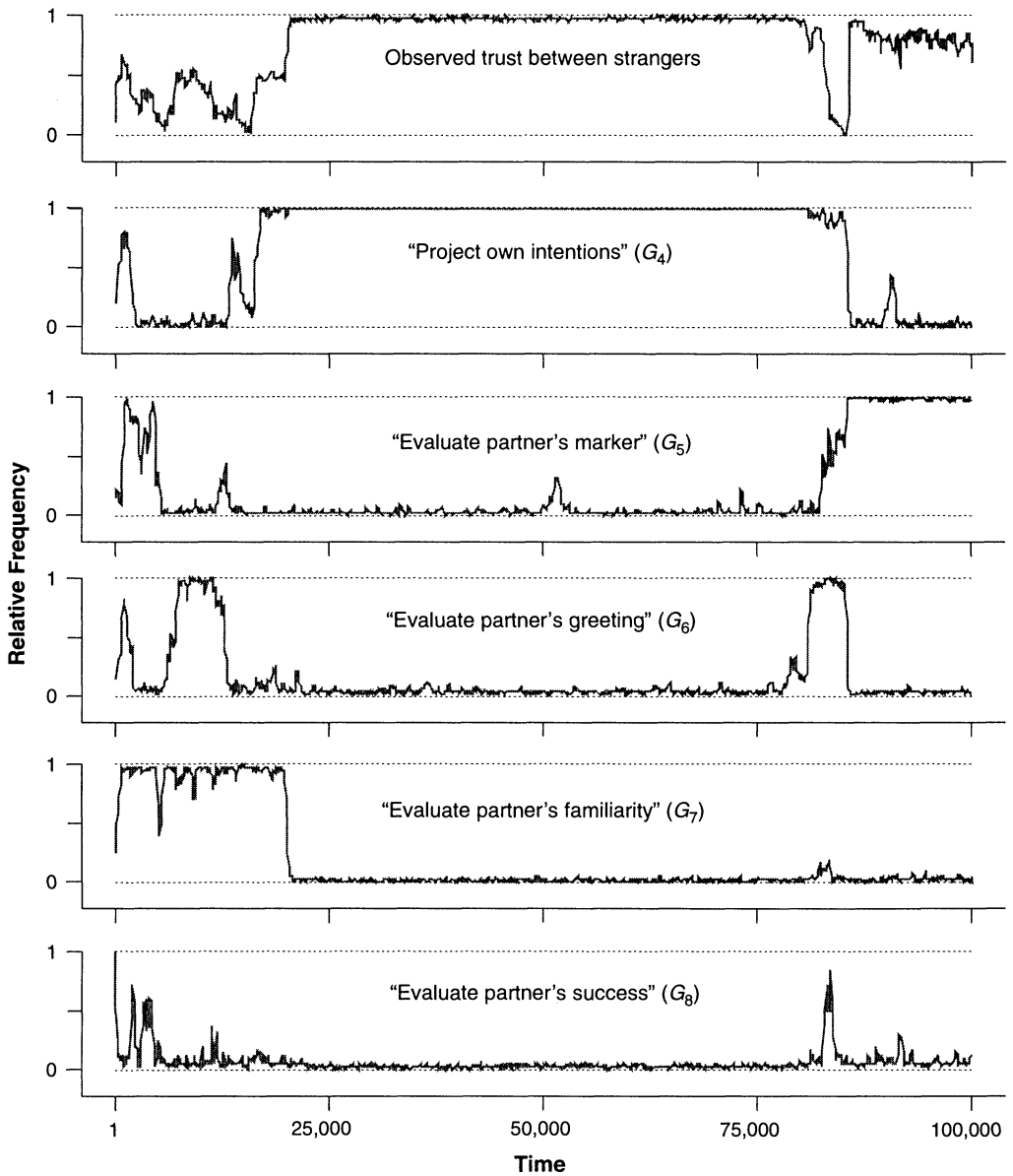


**Figure 3. Strategies for Conditional Exchange**

*Note*: $N = 1{,}000$; $T = 4$, $R = 3$, $X = 2$, $P = 1$, $S = 0$; neighborhood size = 10; embeddedness = .67.

one based on projection and one based on detection. In Figure 3, the first period of sustained cooperation among strangers ($16{,}000 < t < 80{,}000$) is based on a gene for projection ($G_4$), but following the collapse of trust, its reestablishment (at around $t = 85{,}000$) occurs on the basis of an effective rule for judging character ($G_5$). In a benign population, inadvertently created by the parochial rule to trust only neighbors, these two rules outperform and quickly displace parochial strategies that avoid all strangers. Note how the parochial rule to trust only neighbors collapses at around $t = 20{,}000$, shortly after the emergence of projection.

This successful evolution of trust conventions poses two questions: (1) How can projection and detection rules eventually win out over rules for trusting only neighbors, and (2) with widespread trust, why don't *C*s quickly learn to defect, causing trust conventions to collapse?

**Table 3. Correlations between Five Trust Strategies and the Viability of Trust and Cooperation between Strangers**

| Rule (Gene) | Cooperation | *C*-Payoff | Projection | Marker | Greeting | Neighbor |
|---|---|---|---|---|---|---|
| Projection ($G_4$) | .591** | .292* | — | — | — | — |
| | (.129) | (.116) | | | | |
| Marker ($G_5$) | .583** | .426** | .173 | — | — | — |
| | (.152) | (.110) | (.214) | | | |
| Greeting ($G_6$) | .052 | .055 | .077 | .071 | — | — |
| | (.150) | (.114) | (.140) | (.136) | | |
| Neighbor ($G_7$) | −.411** | .010 | −.378** | −.239* | .192 | — |
| | (.071) | (.104) | (.101) | (.097) | (.094) | |
| Success ($G_8$) | −.268** | −.186 | −.104 | −.162 | −.041 | −.035 |
| | (.072) | (.091) | (.090) | (.105) | (.103) | (.072) |

*Note*: Reported values are means over 10 replications. Numbers in parentheses are standard errors.
*$p < .05$ **$p < .01$ (two-tailed tests)

### *The Path to Parochialism*

Compared to more universalistic rules, a rule for in-group favoritism (based on $G_7$) is less vulnerable to exploitation by strangers, which suggests that parochial strategies might be selected over rules that take chances with "outsiders." Both "parochialism" (rules that favor neighbors) and "universalism" (rules that apply to neighbors and strangers alike) will produce in-group trust and cooperation. However, universal strategies also allow cooperation between strangers once a sufficient number of neighborhoods have adopted a compatible rule. Clearly, both patterns are evident in natural settings that exhibit cultural variations in ethnic intolerance, economic protectionism, nativism, and xenophobia.[15] Why do parochial strategies not prevail in this simulation?

Figure 3 reports a simulation in which universalism and parochialism vie for control, but universalism ultimately dominates. For a relatively brief period, players used a strategy of conditional exchange, triggered by the interactant's social distance (neighbor versus stranger). Note how trust and cooperation flourish only after exchange rules become less parochial. Once universal rules have spread to other neighborhoods, parochial strategies cannot compete with those that have learned to trust strangers. Universalism "locks out" parochialism by creating a world in which some strangers can be trusted.

Replications of this experiment show that a universalistic strategy that cooperates with everyone does not always prevail over a parochial rule that favors neighbors.[16] Over 10 replications of the experiment shown in Figure 3, parochialism was the dominant strategy only twice and universalism seven times. In one case, no rules for cooperation took hold within the 100,000 iterations.

Table 3 gives a more detailed summary of the results of these 10 replications. The table reports the mean serial correlations (over 10 trials) between each of the five cues for conditional exchange and two outcome measures—the rate of mutual cooperation among strangers, and the payoff performance for *C*s compared to *D*s in interactions between strangers. Correlations were aggregated over all 10 replications of the experiment; each replication was treated as an independent event giving a sample size of 10. Table 3

[15] These differences have been attributed to changes in the social division of labor (Durkheim [1893] 1960), the progressive rationalization of social life (Weber [1904] 1958), and the emergence of markets (Smith [1776] 1937).

[16] Of course, this is not meant to suggest that universalism is more common in natural settings. Again, the purpose of these experiments was not to replicate human cultural evolution but to demonstrate the theoretical possibility that trust and cooperation between strangers can evolve from a random start, a possibility that most game theorists have generally rejected in the absence of a preexisting system of social control.

shows that projection and detection strategies share the credit for cooperation about equally. They have nearly identical effects on the rate of mutual cooperation and there is a slightly stronger effect of detection on the payoff difference between *C*s and *D*s (but the difference is not statistically significant). Greetings did not hurt or help ($r = .052$, $p = .736$), while exchange rules tied to the partner's familiarity ($r = -.411$, $p < .001$) or success ($r = -.268$, $p = .005$) tended to undermine cooperation between strangers.

### *Invasion by Defectors*

With widespread trust established, what prevents mutations from *C*s to *D*s from causing trust conventions to collapse? Although the simulation in Figures 2 and 3 ends with cooperation dominant, it is only a matter of time before a *C* mutates into a *D* and so begins to defect. However, the embeddedness of interaction makes collapse of trust among neighbors highly self-limiting. If a mutant cheats its neighbors, the neighborhood quickly goes downhill and becomes vulnerable to the influence of strangers from neighborhoods where trust remains intact.[17]

Figure 3 shows the resilience of strategies for cooperation with strangers. The spikes in defection at around $t = 7{,}000$ and $t = 15{,}000$ correspond to successful penetration of *C* defenses (which at this time are based on projection, a relatively naive safeguard). These invasions are to be expected. As the proportion of *C*s approaches unity, selection pressures for vigilance become attenuated, creating an opening for *D* subterfuge. With one exception, after each assault, *C*s managed to quickly regroup and repel the invasion.

The one exception occurs around $t = 80{,}000$, when the collapse spreads from the point of mutation to surrounding neighborhoods. Note how the collapse at $t = 80{,}000$ allows the basis of trust to switch from projection to detection, a more vigilant strategy. Although it does not happen in this simulation, a collapse can also lead to the restoration of parochialism. If a strong epidemic of defection wipes out every enclave of trust, and local trust happens to reemerge based on $G_{15}$ and $G_7$, cooperation might be restored in its parochial form. This process can also restore universalism following the collapse of a parochial culture. And in some cases, neither trust pattern is restored and the world remains in an extended state of anomie. The expected pattern, therefore, is a punctuated equilibrium, with abrupt transitions between highly cooperative and highly distrustful regimes and between universal and parochial regimes.

The most sinister strategy exchanges with everyone ($G_{14} = 1$), but cooperates only with neighbors ($G_1 = 0$ and $G_{15} = 1$) and fakes the "telltale sign" ($G_2 = 0$). This strategy outperforms neighboring strategies that do not cheat strangers. Consequently, the strategy spreads to its neighbors, making the neighborhood more successful than others. The rule will now be adopted by anyone who comes into contact with this neighborhood. The ensuing epidemic quickly wipes out cooperation between strangers. The epidemic runs its course when the population learns to avoid strangers rather than cheat them (a lesson that can only be learned if $X > P$). The disappearance of trust between strangers, in turn, allows *C*s to drift back in (now that selection pressures can no longer distinguish between *C*s who avoid strangers and strategies that cheat strangers but cooperate with neighbors). By chance, some neighborhoods will contain only *C*s. The probability of this occurrence increases with the number of neighborhoods and decreases with their size (as elaborated below). These neighborhoods then become the sites where rules for detection and projection are rediscovered. Trust between strangers is restored when *C*s learn to recognize one another, first locally and then globally.

Over 10 replications of the simulation reported in Figures 2 and 3 (1 million iterations), cooperation between strangers was viable about one-third of the time, and distrust (nonexchange) between strangers prevailed about two-thirds of the time—a level of cooperation that is consistent with experimental evidence on cooperation in one-shot

[17] The spread of prosocial norms from successful to unsuccessful neighborhoods closely resembles "group selectionist" models of the evolution of cooperation (Boyd and Richerson 1990; Soltis, Boyd, and Richerson 1995). However, selection pressures in our model act only on individuals, not groups. Trust conventions spread via contact with relatively fit individuals from successful neighborhoods.

games (Sally 1995). This should not be regarded as a failure of cooperation. On the contrary, any cooperation at all is more than might be expected from a conventional game-theory analysis of a PD game played by strangers.

To sum up, neighborhoods compete for global influence. The neighborhood that gains an early lead becomes the "model neighborhood" for the larger population. Neighborhoods self-organize along two evolutionary tracks, "parochialism" and "universalism." Parochial trust is based on $G_{15}$ and $G_7$, a rule-set that instructs players to trust and cooperate only with neighbors and avoid everyone else. The universal model instructs players to cooperate with everyone ($G_1 = 1$), and to assume that others have the same intention ($G_4 = 1$), and/or to look for a telltale sign indicating character ($G_5 = 1$). Compared to parochial trust, projection/detection strategies are more vulnerable to exploitation by strangers. However, if a vanguard neighborhood becomes successfully organized using a rule of universalism, the rule will spread to other neighborhoods through contact with strangers. When sufficiently widespread, the rule can then dominate parochial strategies by taking advantage of opportunities for cooperation with strangers.

### *Neighborhood Size and Embeddedness of Interaction*

The simulation results reported thus far apply only to a world of small neighborhoods whose members meet one another twice as often as they do strangers. We hypothesize that this structural constraint on interaction makes possible the evolution of trust and cooperation between strangers. To test this idea, we removed the constraint by assigning all 1,000 players an identical cultural marker and placing them in a single 1,000-member neighborhood. In a second test, we allowed two 500-member neighborhoods and gave everyone a .5 probability of interacting with a neighbor. (In other words, at each iteration, every player had an equal probability of interacting with a neighbor or a stranger.) As we expected, trust and cooperation failed to evolve under either condition—an undifferentiated population or a minimally differentiated population.

We then increased the embeddedness of interaction until trust and cooperation with strangers became viable. Figure 4 reports the effects of network structure. In this experiment we manipulated both the size of the neighborhood (from 10 to 50 members, in steps of 10 members) and the embeddedness of interaction with neighbors (from .5 to .9, in steps of .1). Neighborhood size and embeddedness of interaction combine to determine the probability of exchanging with any given neighbor, including the immediately preceding partner. For example, with 20-person neighborhoods and 50-percent local interaction, the chance of interacting with a particular neighbor is .025, about the same as in a 30-person neighborhood with 80-percent local interaction (.026).

Figure 4 shows that the effect of network structure on trust and cooperation is not due to changes in the prospects for re-encountering others (that is, the conditional probability of immediately revisiting a given neighbor). In 20-person neighborhoods with .5 embeddedness, the rate of cooperation with strangers is .16 (averaged over 10 replications). In 30-person neighborhoods with .8 embeddedness, the conditional probability of re-encounter is equivalent, yet the rate of cooperation is much higher, at .46 ($p < .01$, $F[1,16] = 104.4$). So it is not the prospect of re-encounter that helps conventions for trusting strangers to evolve.

Rather, the outcomes depend on the ability to coordinate effective trust conventions in local interaction. Coordination complexity, in turn, is a function of neighborhood size and the embeddedness of interaction. Below .5 embeddedness, cooperation between strangers is unlikely to evolve, even in populations distributed across large numbers of small (10-person) neighborhoods. Above .5, cooperation increases as a linear function of the embeddedness of interaction, regardless of neighborhood size. At .9 embeddedness, the rate of cooperation between strangers is highest (.62) when neighborhoods contain 10 members and declines as a linear function of neighborhood size, to .45 with 50-member neighborhoods.

Even though smaller neighborhoods reduce coordination complexity in the evolution of signaling conventions, the diffusion of rules is impeded by interactions that are not suffi-

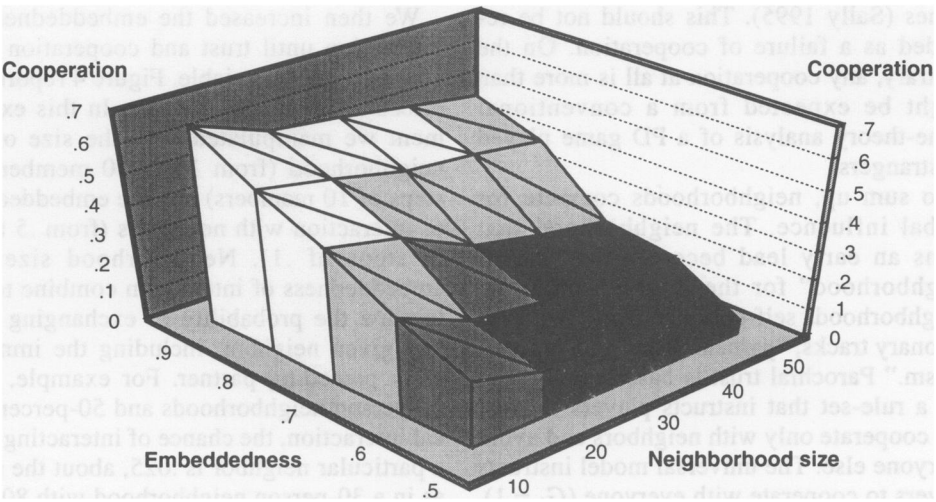


**Figure 4. Effect of Neighborhood Size and Embeddedness on Cooperation between Strangers**

*Note*: Mean cooperation rates are for 10 replications of 100,000 iterations each; $N = 1{,}000$. Embeddedness is reported as the proportion of interactions with neighbors.

ciently embedded in the neighborhood. The more that neighbors interact with one another (and not with strangers), the greater the evolutionary advantage of successful local coordination (due to the relative increase in $R$ payoffs). That advantage in turn allows a local convention to spread, as long as there is contact with at least one outsider. If the convention becomes universal, it can make cooperation with strangers viable as well.

## DISCUSSION: UNIVERSAL AND PAROCHIAL SOLIDARITY

According to Weber ([1904] 1958), Protestantism thrived in America as a highly successful adaptation to social conditions that created high levels of uncertainty in business transactions. Common membership in a Protestant sect provided an effective cue that a potential exchange partner could be trusted. This facilitated transactions between strangers with the same religious marker.

Boudon (1987) formalizes and elaborates Weber's argument as follows:

> In the United States it is more difficult for two persons, say A and B, who wish to conduct business, to know whether they can have confidence in their potential partner. First, because of the greater mobility in the United States, they are less likely to know each other; previous familiarity is less likely to provide indicators of the degree of trustworthiness of the partner. Second, because of the weaker visibility of status symbols, the latter will not be as easily usable as in France or Germany. [This] creates a demand for symbols, for signals whereby A and B would know whether they can have a sufficient degree of confidence in each other before embarking on a business relationship that will, given the complexity of the economic system, in many cases include effects delayed over time. (Pp. 53–54)

Both Weber and Boudon show why trust conventions were needed, but neither shows how the need for reliable social cues led to their evolution. In effect, they assume that Protestant trustworthiness is an equilibrium. The problem is that a successful convention invites fallen souls to infiltrate the church.

Frank's (1988) formal model presents a similar problem. He assumes an exogenous correlation between trustworthiness and a behavioral cue that is difficult to counterfeit. Our dynamic specification relaxes this assumption. We find that reliable cues do not evolve if refusal to exchange is as costly as mutual defection (the condition assumed by Frank). However, we also identify conditions in which telltale behavioral signs can evolve from a random start and repel most (but not all) invasions by fakers. First, the exit option must be less costly than mutual defection but less profitable than mutual cooperation

($P < X < R$), and second, interaction must be concentrated in small neighborhoods, with only occasional contact with outsiders. This outcome is semistable, producing a punctuated equilibrium similar to that observed in evolutionary experiments using the iterated, no-exit PD game (Axelrod 1984; Nowak and Sigmund 1993).

Our simulations reveal two relatively stable strategies: One is triggered by social proximity (stranger versus neighbor) which produces a pattern of "parochial solidarity" based on in-group altruism and out-group defection ("Cooperate with your neighbors and cheat or avoid everyone else"). The other strategy is most effectively triggered by a player's own intention to cooperate (projection) or by the partner's telltale behavioral sign (detection). It produces a pattern of "universalistic solidarity" characterized by openness to strangers ("Judge strangers the same way you judge your neighbors"). As noted by Weber ([1904] 1958), universalism makes possible the reorganization of social life around unembedded open-market transactions that require high levels of trust.

When neighborhoods compete for cultural hegemony in the creation of norms of exchange, players in parochial neighborhoods have a decisive evolutionary edge over universalistic competitors. The effectiveness of a parochial strategy depends on the rate of adoption within one's neighborhood but is independent of the norms adopted elsewhere. In contrast, the success of universalism depends not only on local agreement but also on the probability of meeting a like-minded stranger. The upside potential of universalism is greater than parochialism, but only in the long run after most neighborhoods have converted to the emergent norm.

We find that parochialism tends to dominate populations in which exits are blocked by the high cost of refusing to exchange. Conversely, the ability to avoid exchange with strangers protects emergent *C*s from predation and allows them to spread by evolutionary drift until they are sufficiently prevalent that "xenophobes" cannot compete with discriminating "universalists." Of course, if exit is too easy (more rewarding than mutual cooperation), then no one will learn to trust or cooperate, even with neighbors. The nonlinear effect of the exit payoff suggests a resonant paradox that fits liberal interpretations of historical experience: Universalism thrives on the capacity for "rugged self-reliance" ($P < X$), tempered by a preference for mutual cooperation ($X < R$).

The evolution of universalistic strategies also depends decisively on social structure. Trust conventions congeal in locally embedded social ties and then diffuse from neighbors to strangers. A large number of small neighborhoods confers two advantages. First, small neighborhood size reduces coordination complexity in the standardization of signaling rules and thus facilitates the emergence of effective trust conventions. The embeddedness of neighborhood interaction (or probability of interacting with a neighbor) increases the evolutionary advantage of successful local coordination. Occasional contact with strangers, in turn, allows a successful local convention to become universal, making cooperation with strangers viable as well.

Second, a population divided into a large number of small neighborhoods facilitates the restoration of order following an epidemic of distrust. Once trust and cooperation are thriving outside the neighborhoods, it is only a matter of time before a mutant learns how to cheat strangers, thereby increasing its relative influence. The infection first spreads to the mutant's neighborhood and then to other areas via contact with strangers.

These epidemics are of two types. Some mutants ($G_1 = G_{15} = 0$) cheat neighbors as well as strangers, causing local cooperation to quickly collapse. These strains flourish briefly but then succumb to competition from "healthy" neighborhoods where local trust and cooperation continue to flourish. The smaller the neighborhoods, the faster the infected neighborhoods decline in fitness. The larger the number of neighborhoods, the better the odds that a few will remain healthy until fitness in infected neighborhoods falls below the level required to spread the infection.

More devastating epidemics are caused by mutations that only cheat strangers ($G_1 = 0$, $G_{15} = 1$). These strains do not undermine local cooperation and are therefore much more virulent. With $X > P$, the epidemic ends when the population learns to avoid strangers. *C*s can then safely reappear and spread by evolutionary drift. Again, population structure

affects the recovery. Holding overall population size constant, the larger the number of neighborhoods and the smaller their size, the greater the chances that a few neighborhoods will contain only *C*s. If all neighbors are *C*s, universalist rules (based on detection or projection) will earn *R* in local exchanges and *X* or *R* in interactions with strangers (depending on whether the stranger is like-minded). Parochial rules (that cooperate only with neighbors and avoid all strangers) also earn *R* in local exchanges, but they can only earn *X* outside the neighborhood. With $R > X$, universalists have a potential advantage, depending on their relative numbers in the population. The more of them there are, the better they all perform, and the better they perform, the greater their influence.

### *The Evolution of Protestant Sects Revisited*

We referred above to Weber's ([1904] 1958) theory of Protestantism and trust in an emerging market society, and to his tacit assumption that the need to discern character led Americans to rely on church membership as a telltale sign. Our study formalizes and tests the internal validity of the evolutionary argument that underlies Weber's explanation. Our model is highly abstract, and any application to specific historical circumstances carries considerable risk. Nevertheless, it may be instructive to consider the implications of these simulation experiments for Weber's theory. Our results suggest that Protestant church membership became associated with trustworthiness in local congregations characterized by highly embedded interactions. The convention then spread through occasional contact with strangers. But what protected the emergent norm of trust from subversion? The solution, our analysis suggests, may have been partly structural: the proliferation of small churches.[18]

[18] One *ASR* reviewer wondered why the local church was focal, ". . . rather than the structurally similar local saloon?" We see two possible reasons. First, in addition to structural factors, the Protestant creed itself may have promoted trustworthy behavior. Second, interaction in saloons may be less embedded, depending on the relative frequency of "regulars" and the ability to remember social interactions the next morning.

## DIRECTIONS FOR FUTURE RESEARCH

We restricted our study to evolutionary learning at the population level via the recombination and diffusion of successful strategies. Learning also occurs at the individual level via reinforcement, of course, and we expect this fact may make the coordination of conventions more difficult. More work is needed before we can know if the results reported here hold up when adaptation operates at both the macro and micro levels.

Our study was also limited to a simple diffusion process by which more-fit strategies replace less-fit strategies over time. Therefore we did not include other types of influence, such as conformity, peer pressure, or social impact (Latané 1981). Nor did we allow for more sophisticated strategies of conditional cooperation with neighbors, such as "Tit-For-Tat" (Axelrod,1984) or "Pavlov" (Macy 1996; Nowak and Sigmund 1993). Further research is needed to see if sanctioning and reciprocity promote or block the discovery of rules that allow cooperation to emerge outside the "shadow of the future" (Axelrod 1984).

We also made the degree of embeddedness exogenous to the model and assumed identical payoffs in games with neighbors and strangers. It might be instructive to allow players to decide whether to leave the "safety" of their village in pursuit of higher payoffs available in the "open market," a problem recently addressed by Yamagishi and Yamagishi (1996) in their comparative study of trust and cooperation in the United States and Japan. Suppose most neighbors refuse to interact with outsiders and with those neighbors who break ranks (and have thereby been exposed to outside influence). These exclusionary practices would preclude the diffusion of trust conventions through contact with strangers. All else being equal, this highly parochial culture would then be unlikely to evolve universalistic solutions to the problem of uncertainty in social exchange. However, if the opportunity cost of avoiding the open market is sufficiently great, a trickle of "defections" may eventually trigger a cascade toward the adoption of universalistic rules. New computational experiments are needed to systematically ex-

plore the dynamic properties of collectively self-imposed structural constraints on the choice of exchange partners.

More broadly, our simulations suggest the potential for building formal models of symbolic interaction (also see Kollock and O'Brien 1992). Interactionist approaches to the analysis of emergent norms have characteristically avoided formal theorizing (Blumer 1969). Although our model can be outlined in sentential logic, the dynamics cannot be rigorously tested without formal specification. Nevertheless, we believe our study demonstrates the rich possibilities for using computer simulation to model the dynamic properties of emergent social conventions that are structured by the coevolution of protocols for social exchange.

***Michael W. Macy*** *is Professor of Sociology at Cornell University. He studies the emergence and diffusion of social norms using computational models of adaptive agents and laboratory experiments with human participants. His article on "Social Learning and the Structure of Collective Action" recently received the ASA Theory Section Best Paper Award. He is an Associate Editor of* Advances in Group Processes *and serves on the editorial board of* Social Psychology Quarterly.

***John Skvoretz*** *is a Carolina Distinguished Professor of Sociology at the University of South Carolina. Current research projects concern formal models of social networks, status and participation in discussion groups, power in exchange networks, and theoretical studies of action structures. He is a member of the Sociological Research Association, is an Associate Editor of the* Journal of Mathematical Sociology, *and serves on the editorial board of* Social Psychology Quarterly.

## REFERENCES

Axelrod, Robert. 1984. *The Evolution of Cooperation.* New York: Basic Books.

Bergin, James and Barton Lipman. 1996. "Evolution with State-Dependent Mutations." *Econometrica* 64:943–56.

Blumer, Herbert. 1969. *Symbolic Interactionism.* NJ: Prentice-Hall.

Boudon, Raymond. 1987. "The Individualistic Tradition in Sociology." Pp. 45–70 in *The Micro-Macro Link*, edited by J. C. Alexander, B. Geisen, R. Münch, and N. J. Smelser. Berkeley, CA: University of California Press.

Boyd, Robert and Peter J. Richerson. 1990. "Group Selection among Alternative Evolutionarily Stable Strategies." *Journal of Theoretical Biology* 145:331–42.

Coleman, James S. 1990. *Foundations of Social Theory.* Cambridge, MA: Harvard University Press.

Dawkins, Richard. 1982. *The Extended Phenotype.* Oxford, England: Oxford University Press.

Durkheim, Emile. [1893] 1960. *The Division of Labor in Society.* New York: Free Press.

Egidi, Mason and Luigi Marengo. 1995. "Division of Labour and Social Co-ordination Modes: A Simple Simulation Model." Pp. 40–58 in *Artificial Societies: The Computer Simulation of Social Life,* edited by N. Gilbert and R. Conte. London, England: University College London Press.

Ellison, Glenn. 1993. "Learning, Local Interaction, and Coordination." *Econometrica* 61: 1047–71.

Enquist, Magnus and Olaf Leimar. 1993. "The Evolution of Cooperation in Mobile Organisms." *Animal Behaviour* 45: 747–57.

Frank, Robert. 1988. *Passions within Reason: The Strategic Role of the Emotions.* New York: Norton.

———. 1993. "The Strategic Role of Emotions: Reconciling Over- and Undersocialized Accounts of Behavior." *Rationality and Society* 5:160–84

Goldberg, David. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning.* New York: Addison-Wesley.

Hechter, Michael. 1987. *Principles of Group Solidarity.* Berkeley, CA: University of California Press.

Heckathorn, Douglas. 1993. "Collective Action and Group Heterogeneity: Voluntary Provision Versus Selective Incentives." *American Sociological Review* 58:329–50

Holland, John. 1975. *Adaptation in Natural and Artificial Systems.* Ann Arbor, MI: University of Michigan Press.

Kollock, Peter and John O'Brien. 1992. "The Social Construction of Exchange." Pp. 89–112 in *Advances in Group Processes,* vol. 9, edited by E. Lawler et al. Greenwich, CT: JAI Press.

Latané, Bibb. "The Psychology of Social Impact." *American Psychologist* 36:343–56.

Macy, Michael. 1996. "Natural Selection and Social Learning in Prisoner's Dilemma: Co-Adaptation with Genetic Algorithms and Artificial Neural Networks." *Sociological Methods and Research* 25:103–37.

McCain, Roger. 1995. "Genetic Algorithms, Teleological Conservatism, and the Emergence of Optimal Demand Relations: The Case of Learning by Consuming." Pp. 126–42 in *Artificial Societies: The Computer Simulation of Social Life,* edited by N. Gilbert and R. Conte.

London, England: University College London Press.

Nowak, Martin and Karl Sigmund. 1993. "A Strategy of Win-Stay, Lose-Shift that Outperforms Tit-for-Tat in the Prisoner's Dilemma Game." *Nature* 364:56–58.

Orbell, John and Robin Dawes. 1991. "A 'Cognitive Miser' Theory of Cooperators' Advantage." *American Political Science Review* 85: 515–28.

———. 1993. "Social Welfare, Cooperators' Advantage, and the Option of Not Playing the Game." *American Sociological Review* 58: 787–800.

Platt, John. 1973. "Social Trap." *American Psychologist* 28:641–51.

Sally, David. 1995. "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992." *Rationality and Society* 7:58–92.

Skvoretz, John and Thomas Fararo. 1995. "The Evolution of Systems of Social Interaction." *Current Perspectives in Social Theory* 15:275–99.

Soltis, J., R. Boyd, and P. J. Richerson. 1995. "Can Group Functional Behaviors Evolve by Cultural Group Selection? An Empirical Test." *Current Anthropology* 36:473–94.

Smith, Adam. [1776] 1937. *An Inquiry into the Nature and Causes of the Wealth of Nations.* New York: Modern Library.

Treuil, Jean. 1995. "Emergence of Kinship Structures: A Multi-Agent Approach." Pp. 59–85 in *Artificial Societies: The Computer Simulation of Social Life*, edited by N. Gilbert and R. Conte. London, England: University College London Press.

Weber, Max. [1904] 1958. "The Protestant Sects and the Spirit of Capitalism." Pp. 302–22 in *From Max Weber*, edited by H. Gerth and C. Mills. New York: Oxford University Press.

Wilson, D. S., G. B. Pollock, and L. A. Dugatkin. 1992. "Can Altruism Evolve in Purely Viscous Populations?" *Evolutionary Ecology* 6:331–41.

Yamagishi, Toshio and M. Yamagishi. 1996. "Trust and Commitment in the United States and Japan." *Motivation and Emotion* 18:129–66.