# Follow the Leader: Local Interactions with Influence Neighborhoods*

Peter Vanderschraaf and J. McKenzie Alexander†‡

We introduce a dynamic model for evolutionary games played on a network where strategy changes are correlated according to degree of influence between players. Unlike the notion of stochastic stability (Foster and Young, 1990), which assumes mutations are stochastically independent and identically distributed, our framework allows for the possibility that agents correlate their strategies with the strategies of those they trust, or those who have influence over them. We show that the dynamical properties of evolutionary games, where such *influence neighborhoods* appear, differ dramatically from those where all mutations are stochastically independent, and establish some elementary convergence results relevant for the evolution of social institutions.

**1. Introduction.** Game theorists analyze the strategic aspects of interactions. Social network theorists analyze the structures that determine who interacts with whom. Game theory and social network theory meet when those who are connected via a network play a game. An emerging literature explores how players engaged in such network games can gradually settle into an equilibrium. In this literature, the network game is modeled as a dynamical system of players who interact with their neighbors[1] and who adjust their strategies over time. The attracting points of certain dynamical adjustment processes are often Nash equilibria of the network game. Some Nash equilibria are also stochastically stable (Foster and

*Received May 2003; revised May 2004.

†To contact the authors, please write to: Peter Vanderschraaf, Department of Philosophy, College of Humanities and Social Sciences, Baker Hall 135, Carnegie Mellon University, Pittsburgh, PA, 15213; email: peterv@andrew.cmu.edu, or to J. McKenzie Alexander, Department of Philosophy, Logic, and Scientific Method, London School of Economics, Houghton Street, London, United Kingdom, WC2A 2AE; email: jalex@lse.ac.uk.

‡We would like to thank Brian Skyrms, Cristina Bicchieri, and two anonymous referees for their helpful comments and suggestions.

1. Player $i$ is said to be a *neighbor* of another Player $j$ if $i$ and $j$ are connected by an edge in the social network.

Philosophy of Science, 72 (January 2005) pp. 86–113. 0031-8248/2005/7201-0005$10.00

Young 1990), in the sense that these equilibria emerge and persist when the dynamical system is perturbed with independent random changes in strategy or mutations.[2] Game theorists have proved some powerful convergence theorems for network games that evolve according to dynamics perturbed with independent mutations (Ellison 1993, 2000; Young 1998, Morris 2000). Some argue that these theorems are important in explaining the evolution of social institutions (Young 1998; Binmore 1998).

In this paper we introduce a dynamical adjustment process for network games with correlated mutations. This is significant because previous convergence results for network games invoke implausibly strong assumptions. In particular, these results assume that random mutations, which perturb the network game, are stochastically independent and identically distributed. But stochastic independence clearly fails in many settings. People often imitate others, even in experimental situations, which prevents mutations from being stochastically independent. In this paper, we present a model that allows players to imitate the behavior of another player who mutates, thereby introducing correlated mutations. This kind of correlated mutation has a natural interpretation: If a player in the network experiments and can signal her intent to some of the other players, those who receive the signal might imitate this signaler if she has sufficient influence over them. The players in this influence neighborhood who imitate the signaler 'follow the leader'. We show that the dynamical properties of evolutionary games allowing for influence neighborhoods can differ dramatically from ones where all mutations are stochastically independent. We also argue that this dynamics more closely mirrors the process by which societies reform, rather than the dynamics of stochastically independent mutation.

In Section 1 we review some basic notions of network games, using the Assurance game to develop motivating examples. Section 2 discusses how inductive best-response dynamics are applied to network games, giving an example where best-response dynamics with independent random mutation never reaches the stochastically stable equilibrium in a reasonable time frame. We argue that this result casts doubt upon the explanatory power of models that assume stochastically independent mutations. In Section 3 we relax the independence assumption by introducing influence neighborhoods. We show how influence neighborhoods can greatly accelerate the transition from a suboptimal equilibrium to an optimal and stochastically stable equilibrium. We also show how influence neighborhoods can drive a network out of a stochastically stable equilibrium, and even converge to an optimal equilibrium that is not stochastically stable.

2. In line with the other literature on network games, in this paper, a mutation is a random change in strategy, not a biological mutation.

| | | Player $j$: $s_1$ | Player $j$: $s_2$ |
|---|---|---|---|
| Player $i$ | $s_1$ | $(x,x)$ | $(0,y)$ |
| | $s_2$ | $(y,0)$ | $(z,z)$ |

$x > y \geq z > 0$

Figure 1. The Assurance game.

We argue that this influence neighborhood model can be applied to test the likelihood of cooperation in social dilemmas and Hardin's (1995) dual coordination account of the persistence and fall of political regimes. In Section 4 we give formal definitions of influence neighborhoods and best-response dynamics with correlated mutations, together with some elementary convergence results.

**2. The Assurance Game Played with Neighbors.** Figure 1 summarizes the symmetric 2-player Assurance game.[3] The Assurance game plays an important role in moral and political philosophy. Philosophers use the Assurance game to represent collective action problems ranging from cooperation in the Hobbesian State of Nature to pollution control to political revolutions.[4] The Assurance game also illustrates some of the challenges of accounting for equilibrium selection in games. In the Figure 1 game, ($s_1$, $s_1$) and ($s_2$, $s_2$) are both coordination equilibria (Lewis 1969) with the property that neither player's payoff is improved if one of them deviates from either ($s_1$, $s_1$) or ($s_2$, $s_2$). The equilibrium ($s_1$, $s_1$) is Pareto optimal and yields each player his highest possible payoff. However, each player is certain to gain a positive payoff only if he follows $s_2$. Should rational players contribute to an optimal outcome or play it safe?

The classical game theory of von Neumann and Morgenstern (1944) and Nash (1950, 1951a, [1951b] 1996) gives no determinate answer to this question. Harsanyi and Selten (1988) tried to answer this question by introducing a refinement of the Nash equilibrium concept called risk dominance. A strategy $s$ is a player's best response to a strategy profile of the other players or a probability distribution over these profiles, when $s$

3. Following standard conventions, Player $i$'s (Player $j$'s) payoff at each outcome of the game is the first (second) coordinate of the payoff vector in the cell of Figure 1 that characterizes this outcome. For instance, if $i$ chooses $s_1$ and $j$ chooses $s_2$ then $i$'s payoff is 0 and $j$'s payoff is $y$.

4. See especially Taylor and Ward 1982; Kavka 1986; Hampton 1986; Taylor 1987; Jiborn 1999; Skyrms 2001, 2004.

maximizes the player's payoff given this profile or distribution. If the players in a symmetric 2 × 2 game each assign a uniform probability distribution over the other's pure strategies and $s^*$ is the unique best response for both, then ($s^*$, $s^*$) is the risk dominant equilibrium. In the game of Figure 1, ($s_1$, $s_1$) is risk dominant if $x > y + z$ and ($s_2$, $s_2$) is risk dominant if $y + z > x$. Harsanyi and Selten argue that a player should follow her part of a risk dominant equilibrium, because this strategy is the best response over the larger share of possible probabilities with which the other player follows his pure strategies (Harsanyi and Selten 1988, 82–83). Risk dominance is an important concept in game theory, but it raises obvious questions: Why *shouldn't* a player's probabilities over her opponent's strategies lie outside the range that makes her end of the risk dominant equilibrium her best response? Why shouldn't a player optimistically assign a high probability to her counterpart choosing $s_1$, even if ($s_2$, $s_2$) is risk dominant, or pessimistically assign a high probability to her counterpart choosing $s_2$, even if ($s_1$, $s_1$) is risk dominant? In the end, there really is no determinate solution to the Assurance game. Given appropriate probabilities reflecting a player's beliefs about what the other player will do, either pure strategy can be a best response. Rational players might fail to follow an equilibrium at all, even if they have common knowledge of their rationality.[5]

Now suppose that, in a population of players, everyone plays the Assurance game with a subset of the population, known as her 'neighbors'. At a given time, each player follows one strategy in her interactions with all her neighbors.[6] Explicitly identifying the neighbors with whom each player interacts embeds the Assurance game in a local interaction structure or network. Formally, a network is an undirected graph in which the nodes represent the players. Player $j$ is Player $i$'s *neighbor* if the nodes representing $i$ and $j$ are linked with an edge. If $n_i(s_1)$ of $i$'s neighbors follow $s_1$ and $n_i(s_2)$ of $i$'s neighbors follow $s_2$, then $s_1$ is a best response for $i$ if

$$n_i(s_1)x \geq n_i(s_1)y + n_i(s_2)z. \tag{1}$$

5. David Lewis (1969, 56–57) presented the first analysis of common knowledge. A proposition $A$ is Lewis-common knowledge among a group of agents if each agent knows that all know $A$ and knows that all can infer the consequences of this mutual knowledge. Lewis-common knowledge implies the following better known analysis of common knowledge: $A$ is common knowledge for a group of agents if each agent knows $A$, each agent knows that each agent knows $A$, and so on, ad infinitum.

6. To motivate this assumption, common throughout the network game literature, one can suppose that each player interacts with all her neighbors simultaneously, or that she cannot keep track of which neighbors follow any particular strategy, so she must adopt a single strategy for interacting with them all.

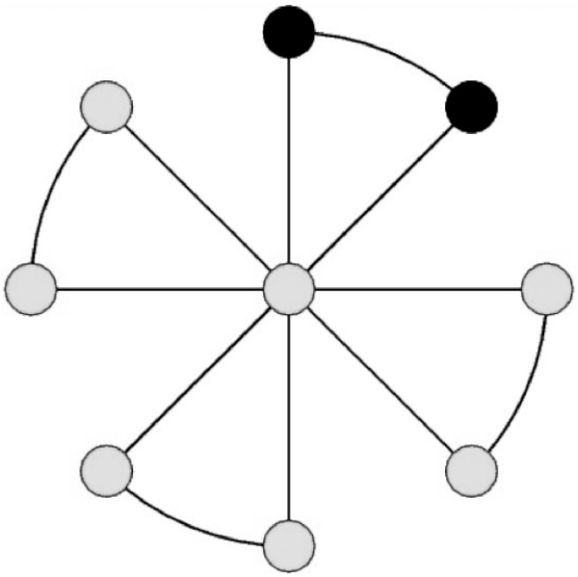

Figure 2. Black players follow $s_2$. Light players follow $s_1$.

$s_2$ is a best response if the reverse inequality is satisfied. In the special case where $y = z$, (1) is equivalent to

$$x \cdot \frac{n_i(s_1)}{n_i(s_1) + n_i(s_2)} \geq z. \tag{2}$$

That is, $s_1$ is a best response if the weighted average of payoffs, which $i$ receives from her neighbors who follow $s_1$, exceeds the guaranteed payoff of following $s_2$. To illustrate, Figure 2 depicts a 'propeller' graph, where eight outer players are each linked with the same central player and one outer player. The outer players form the central player's Moore-8 neighborhood. The Figure 2 graph and Figure 1 game together define a *network game*. If, for instance, $x = 9$ and $y = z = 5$, then by (2), $s_1$ is a best response for the central Player $i$ if $9 \cdot [n_i(s_1)/8] \geq 5$ or $n_i(s_1) \geq 40/9 > 4$.

A priori analysis cannot predict what players in a network game will do any more than classical game theory can predict what a pair of players who meet in the Assurance game will do. Indeed, local interaction structures complicate the equilibrium selection problem. If the players in a network play the Assurance game with their neighbors, then this system is at one equilibrium if all follow $s_1$, and another if all follow $s_2$. In addition, there are polymorphic equilibria where some players follow $s_1$, while others follow $s_2$. If the players in the Figure 2 graph play the Assurance game

with $x = 9$ and $y = z = 5$ for each Player $i$, then along with the all-$s_1$ and all-$s_2$ equilibria, any state, where the central player follows $s_1$ and exactly two of the outer players linked with each other follow $s_2$, is an equilibrium. Figure 2 depicts one of these polymorphic equilibria. Which, if any, of all these equilibria will the players in a local interaction structure adopt?

*2.1. Best-Response Dynamics, and an Apparent Anomaly.* In recent years, game theorists have made increasing use of dynamical adjustment processes to analyze equilibrium selection. This approach explores how individuals test and revise their strategies over time until, gradually, they converge to an equilibrium of a game. The formal model of the process by which players update their behavior characterizes a dynamical system. The popularity of this dynamical systems approach is recent, but the underlying idea appears early in the history of game theory. John Nash included a dynamical updating method for equilibrium selection in his original presentation of the Nash equilibrium concept (Nash 1951b).[7] Strikingly, David Hume's analysis of convention in *A Treatise of Human Nature* foreshadows both the Nash equilibrium concept and a dynamical explanation of equilibrium selection (Hume [1740] 1976, 490).[8]

Over the past decade, several authors (Young 1993, 1998; Kandori, Mailath and Rob 1993; Ellison 1993, 2000; Morris 2000) have proved a set of results that establish important connections between risk dominant equilibria in a wide class of games and the stochastically stable equilibria (Foster and Young 1990) of a variety of adaptive dynamics. One can perturb an adaptive dynamic so that each player occasionally mutates by following a new strategy chosen at random. Informally, an equilibrium is stochastically stable if it is robust against a low but steady 'bombardment' of stochastically independent random mutations in the dynamics. If a game has a stochastically stable equilibrium of an adaptive dynamic, then over an infinite sequence of plays, players, who update according to this dynamic perturbed with independent random mutations, will gravitate to this equilibrium a nonnegligible part of the time. If the game has a unique stochastically stable equilibrium, then over infinitely many plays, the players gravitate to this equilibrium for all but a negligible amount of time.

With network games, game theorists standardly investigate the prop-

7. Nash's dynamical model foreshadows the *fictitious play processes* (Brown 1951; Fudenberg and Levine 1998) that have become a staple tool for analyzing equilibrium selection in games.

8. For discussion of Hume's informal game-theoretic insights, see Lewis 1969 and Vanderschraaf 1998.

erties of the inductive best-response dynamic with random perturbations. According to the best-response dynamic, a player follows a strategy that yields the highest payoff against the strategies her neighbors have just followed. This dynamic tacitly assumes that players react myopically to their situation. If the players in a network play a game with a risk dominant equilibrium, the strategy of this equilibrium characterizes the unique stochastically stable equilibrium of the system for the best-response dynamic with independent random mutation (Ellison 1993; Young 1998). So, we evidently have a dynamical account of the emergence of risk dominant equilibrium play between neighbors.

The relationship between risk dominance, a static concept from rational choice game theory, and stochastic stability, a dynamical concept, is of fundamental theoretical importance. Nevertheless, it is not clear how far stochastic stability results go in explaining how players in the real world might interact more successfully. The following example illustrates this point.

*2.2. Example 1. Assurance Game Played on a Torus with Independent Mutations.* Let $m > 1$ be an integer and let $N = \{1, \ldots, n\}$ where $n = m^2$. Define a bijective function $\iota : N \rightarrow \{1, \ldots, n\} \times \{1, \ldots, n\}$ that assigns to each Player $i$ a unique index $\iota(i) = (\iota_1(i), \iota_2(i))$. The graph

$$N = \{\{i, j\} : |\iota_1(i) - \iota_1(j)| = 1 \bmod m \text{ and/or}$$
$$|\iota_2(i) - \iota_2(j)| = 1 \bmod m\}$$

consists of links between each $i$ and the 8 neighbors that immediately surround $i$. These links define $i$'s Moore-8 neighborhood. This 2-dimensional graph is topologically a torus, and can be mapped onto a square whose edges 'wrap around'. A number of authors use this graph to model various local interactions because it roughly approximates the interactions of agents who neighbor each other in a geographic region.[9] We set $m = 100$, so the entire network contains 10,000 players.

Next, we augment the local interaction structure with strategies and payoffs. Each player in the network plays the Figure 3 Assurance game with each of his Moore-8 neighbors, choosing a single strategy for interaction. The risk dominant equilibrium of the Figure 3 game is $(s_1, s_1)$. So if players in this system update according to the best-response dynamic with independent random mutations, then the stochastically stable equilibrium of this system is the equilibrium where all follow $s_1$. The all-$s_1$ equilibrium is the unique stable attractor of this dynamic for any positive

9. See, for instance, Nowak and May 1992; Nowak, Bonhoeffer and May 1994; Grim, Mar, and St. Denis 1998; Alexander 2000; Alexander and Skyrms 1999.

| | | Player $j$: $s_1$ | Player $j$: $s_2$ |
|---|---|---|---|
| Player $i$ | $s_1$ | (6,6) | (0,3) |
| | $s_2$ | (3,0) | (2,2) |

Figure 3. Assurance game with ($s_1$, $s_1$) risk dominant.

rate of mutation, no matter how small (Young 1998). In particular, if the system starts in the suboptimal equilibrium with all players following $s_2$, best-response dynamics with random mutation eventually move the entire population to the optimal all-$s_1$ equilibrium.

One should wonder how long it takes this movement to occur. To test the speed of this convergence, we ran a computer simulation of this system.[10] All 10,000 players were initially assigned the strategy $s_2$, starting the system at the suboptimal all-$s_2$ equilibrium. At each time period, every player played the Figure 3 Assurance game with her Moore-8 neighbors, updating her strategy according to a perturbed best-response dynamic. Stochastically independent mutants appeared at a rate of .05. Each mutant chose one of the pure strategies, $s_1$ or $s_2$, at random with equal probability. We deliberately chose this rather high mutation rate so as to bias the dynamics against the initial all-$s_2$ equilibrium.

While the all-$s_2$ equilibrium is not stochastically stable, it proves surprisingly robust in the face of independent random mutations. The system was allowed to evolve for 100 million periods.[11] Figure 4 depicts the state of this 100 × 100 lattice at the final stage of this simulation. Even though the mutation rate was relatively high, so that at any stage an average of 5% of the players mutated, the $s_1$-mutants were consistently overwhelmed and could not establish a permanent foothold. Hence, the $s_1$-strategy never overthrew the incumbent $s_2$-equilibrium. Indeed, in this simulation the suboptimal all-$s_2$ equilibrium gave the appearance of being stochastically stable!

One might object that the test of the attracting power of the all-$s_1$

10. All of the simulation experiments summarized in this paper were run using the *Evolutionary Modeling Lab*, developed by Alexander. The Evolutionary Modeling Lab is accessible at http://evolve.lse.ac.uk/eml/, and the specific programs run to perform the simulations of this paper are available upon request from the authors.

11. The pseudo-random number generator that the *Evolutionary Modeling Lab* uses is the Mersenne twister algorithm known as MT19937, which has a provable period of $2^{19937} - 1$.

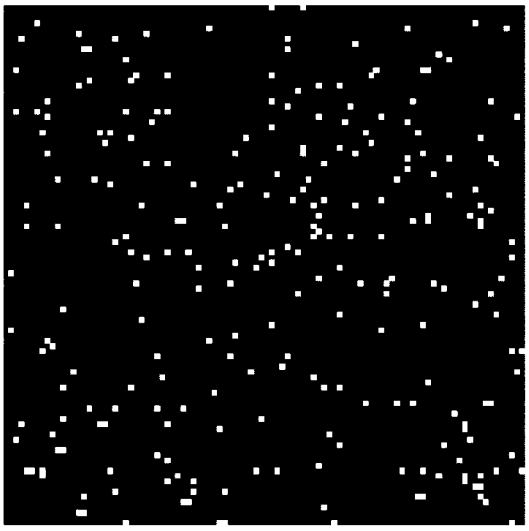

Figure 4. State of the model after 100,000,000 iterations. Black players follow $s_2$. Light players follow $s_1$.

equilibrium in Example 1 is too severe. Perhaps rational agents in such a network would seldom, if ever, all begin by following $s_2$. In fact, we did relax the severity of the test, and found that the perturbed best-response dynamic with .05 mutation rate can converge to and never overthrow the suboptimal all-$s_2$ equilibrium over 100 million rounds of play, even if the system is initially set with as many as 20% of the network players following the strategy $s_1$. Still, we think the conditions of Example 1 are not so farfetched. Social dilemmas occur when individuals are reluctant to contribute towards a common good, even when they realize that all are better off if all contribute. A network Assurance game models a social dilemma where a player contributes to the common good by following $s_1$ and withholds his contribution by following $s_2$. Suppose initially that the benefit of the common good is small compared against the security of not contributing, so that all tend to follow $s_2$ so as to avoid the costs of contribution. Then conditions change, making the relative benefit of the common good significantly greater. The Example 1 network corresponds to such a situation, since the ($s_1$, $s_1$) equilibrium of the Figure 3 game is both optimal and risk dominant. However, by (1) at least half of any player's neighbors must change from $s_2$ to $s_1$ before $s_1$ becomes this player's best response. Example 1 shows that players who best respond to their neighbors' previous strategies can have great difficulty making the transition from consistently following $s_2$ to consistently following $s_1$, even when

the network is continually 'bombarded' by independent random mutations appearing at a high rate. The initial all-$s_2$ equilibrium of Example 1 models a state that seems ripe for reform. But the dynamical behavior of this system reflects the fact that the road to social reform can be a long one.

**3. Influence Neighborhoods.** Theory tells us that random mutations will lead players to converge to stochastically stable equilibria almost surely in the long run. But we have just seen that players in a network who mutate independently at a high rate can fail to reach the stochastically stable equilibrium in over 100 million rounds of play. We believe Example 1 casts doubts upon the explanatory power of stochastic stability theorems applied to network games played by humans. Most, if not all, human interaction networks change and even dissolve long before the people in the network approach a 100 millionth consecutive round of interactions, yet a network of players who mutate independently can fail to approach its long run limit over 100 million rounds. If stochastic stability theorems require extraordinarily long waiting times—as Example 1 clearly shows—how can such theorems be relevant for explaining why actual people behave the way they do? No person changes his belief 100 million times in the course of his life, much less within a single repeated game. While we agree that 100 million rounds of interaction constitutes a short period of time from the point of view of the ergodic theory underlying stochastic stability theorems, one must appreciate a crucial difference between physical and social systems. Ergodic theory provides useful analyses of physical phenomena simply because, according to the time scale of many physical events, each elementary component (i.e., atom, molecule, etc.) can be involved in an extraordinarily large number of interactions in a relatively short period of time. The same is not true for social systems. Social and physical systems fail to be analogous precisely where required if ergodic theory is to be explanatorily relevant.

What if mutations in the dynamics can be *correlated*? The following examples show that the evolution of behavior in a network of best-response updaters can change dramatically if we relax the assumption that all the mutations are stochastically independent.

*3.1. Example 2. Assurance Game Played on a Torus with Influence Neighborhoods.* We revisit the network game of Example 1, with players (arranged on a 100 × 100 torus) who play the Figure 3 Assurance game with Moore-8 neighbors. Again, each player updates his strategy according to a perturbed best-response dynamic. However, now we allow for correlation in the mutations. If a given Player $i$ spontaneously mutates at stage $t$, then each of $i$'s Moore-8 neighbors and each of *their* Moore-

8 neighbors imitate $i$'s stage $t$ strategy with probability $\lambda_i(t)$.[12] The 24 players whose stage $t$ strategies are correlated with $i$'s mutant strategy are $i$'s Moore-24 neighborhood in the torus. Each player in this Moore-24 neighborhood follows the mixed strategy of playing $i$'s mutant strategy with probability $\lambda_i(t)$. $\lambda_i(t)$ is a value sampled from the uniform distribution over [0,1], that is, $\lambda_i(t)$ is a probability for imitation drawn at random from [0,1], for each $i$.[13] We set the spontaneous mutation rate for a 'leader' at a low .0001, so that an average of only one 'leader' appears in the entire network per period. A 'leader' spontaneously mutates to $s_1$ with probability .5 and to $s_2$ with probability .5.[14]

As in Example 1, in every simulation we ran of this dynamic we started the network game at the suboptimal all-$s_2$ equilibrium. In each of these simulations, in fewer than 800 generations, the $s_1$-followers had spread throughout the entire system of players so that all followed $s_1$ except for occasional areas of $s_2$-followers that emerged due to this correlated mutation.[15] These occasional $s_2$-following clusters were quickly overwhelmed and converted back to $s_1$-following. Figure 5 depicts the state of this 100 × 100 graph at the 100th, 300th, 500th and 700th generations of one of our simulations.

Note that the system converged rapidly to the all $s_1$-equilibrium even though at any given stage the overall mutation rate was bounded from

12. Why use the Moore (24) neighborhood rather than just the Moore (8) neighborhood? Quite often one's social influence spreads beyond one's immediate neighbors or acquaintances. It is not uncommon for the following situation to occur: $A$ knows $B$, $B$ knows $C$, and $A$ does not know $C$. Nevertheless, $A$ exerts influence upon $C$ through $B$, because $B$ tells $C$ that $A$ believes something or did something. The Moore (24) neighborhood is a crude first approximation at capturing this phenomenon. Clearly other influence neighborhoods are worthy of examination. Such a study, though, lies beyond the scope of this paper.

13. One might consider the use of randomly chosen probabilities as an extreme case. However, this strikes us as a not implausible assumption. For example, I may have an extremely skeptical neighbor, yet he may have a neighbor who is capable of being easily influenced. In such a case, I may have little influence over my immediate neighbor yet have considerable influence over my neighbor's neighbor. Lifting the assumption of randomly chosen probabilities requires making further assumptions about the specific way influence is exercised and implemented in the social system, assumptions we do not wish to make at this time.

14. One can also allow independent random mutations to appear alongside the mutations correlated with the 'leaders'. In this simulation experiment, the independent mutation rate was set to .0 so that the 'leader' players who might be followed by some of their Moore-24 neighbors received no additional 'help' from independent random mutants.

15. We achieved similar results when we changed the parameters of the dynamics in various ways, such as setting $\lambda_i(t)$ to be constant over the Moore-24 neighborhood or varying the sizes of the neighborhoods of correlated mutations.

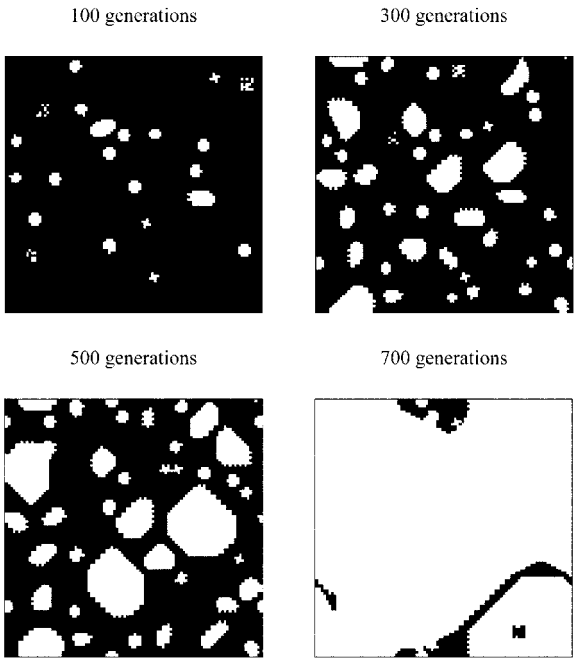


Figure 5. Black players follow $s_2$. Light players follow $s_1$.

above by $25 \cdot (1/10{,}000) = .0025$, the overall expected mutation rate if *all* of a 'leader' player's Moore-24 neighbors imitated the 'leader's' strategy.

The correlation in mutations described in Example 2 is a correlation over a 'leader's' influence neighborhood. In this example if Player $i$ is a leader mutant at period $t$, then his influence neighborhood $I_i(t)$ is the set of his Moore-24 neighbors. At period $t$, each $j \in I_i(t)$ imitates $i$ with probability $\lambda_i(t)$. A natural to way to justify this sort of correlation in strategies is to allow for the possibility of costless communication, or what game theorists call 'cheap talk'. If players can communicate, then they can correlate their strategies with the leader players whose messages they receive. When $i$ is a leader at period $t$, $i$ mutates to strategy $s_i(t)$ and communicates this fact to each $j \in I_i(t)$. $\lambda_i(t)$ measures the strength of $i$'s influence over those in the neighborhood $I_i(t)$. Those in $I_i(t)$ who imitate $i$'s strategy $s_i(t)$ at period $t$ 'follow their leader'. In Example 2, the correlated mutation of influence neighborhoods moves the network game from the suboptimal to the optimal equilibrium, even though the influence neighborhoods rarely appear. The road to reform in this example is shortened considerably by the introduction of influence neighborhoods.

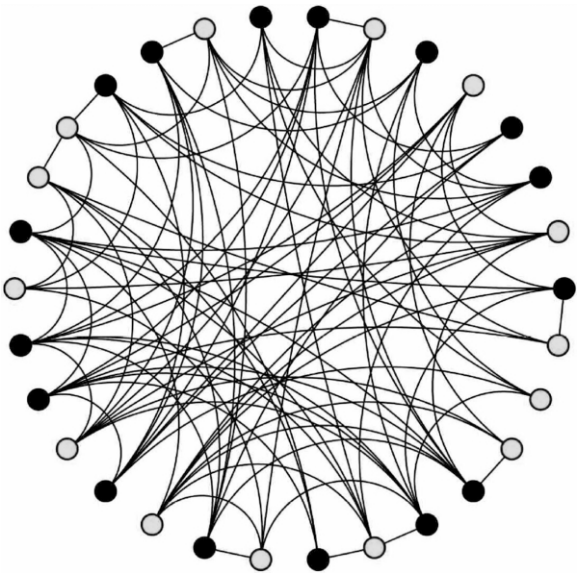

Figure 6. Black players follow $s_2$. Light players follow $s_1$.

Example 2 shows that risk dominant play can overtake an interaction network fairly rapidly when some players' strategies are correlated. This example suggests a framework for modeling the 'bandwagon' effect that drives social reformation. Moreover, the next examples show that when correlated influence neighborhood mutations are possible, players who update according to a perturbed best-response dynamic need not converge to risk dominant play. This suggests a framework for modeling the stability of political regimes and the process of revolt.

*3.2. Example 3. Assurance Game Played on a Bounded Degree Network with Influence Neighborhoods.* Figure 6 depicts a bounded degree network, where each node is linked with at least 4, and at most 8, others. Again, the nodes represent players and the edges define each player's neighbors. Each player plays the Figure 7 Assurance game with each of her neighbors in the network. In Figure 7, the suboptimal ($s_2$, $s_2$) equilibrium is risk dominant. Consequently, in this network game the all-$s_2$ equilibrium is the unique stochastically stable equilibrium of the best-response dynamic. In computer simulations, we found that when updating occurred according to the best-response dynamic with independent random mutations, the network converged to the all-$s_2$ equilibrium even if it was initially set at the all-$s_1$ equilibrium. Moreover, these random mutations never gen-

| | | Player $j$ | |
|---|---|---|---|
| | | $s_1$ | $s_2$ |
| Player $i$ | $s_1$ | (9,9) | (0,5) |
| | $s_2$ | (5,0) | (5,5) |

Figure 7. Assurance game with $(s_2, s_2)$ risk dominant.

erated a permanent foothold of $s_1$-followers in the network, even when the system was 'bombarded' by a high mutation rate of .10 for 100,000 periods. These results were not surprising, given that only all-$s_2$ is stochastically stable.

However, the all-$s_2$ equilibrium does not retain its high attracting power when mutations can be correlated via influence neighborhoods. In a second set of computer simulations, the spontaneous mutation rate was set at .001, and again a spontaneous mutant followed $s_1$ or $s_2$ at random with equal probability. If a leader Player $i$ spontaneously mutated to $s_i$, then $i$'s neighbors together with their neighbors each followed $s_i$ with probability $\lambda_i(t)$ selected at random from [0,1]. In these simulations, even when the network was initially set at the stochastically stable all-$s_2$ equilibrium, it oscillated between all-$s_2$ and all-$s_1$. Figure 8 summarizes the evolution of strategies over this network during 5,000 periods of best-response updating perturbed with these influence neighborhoods. In this network, no equilibrium is stable with respect to these correlated mutations.

In Examples 2 and 3, influence neighborhoods appear in the network at a fixed rate and a fixed size across pure strategies. In the next example, influence neighborhoods appear at different rates and in different sizes across the pure strategies.

*3.3. Example 4. Assurance Game Played on a Bounded Degree Network with Differential Influence Neighborhoods.* We consider another bounded degree network game, where each of 50 players is linked with at least 4, and at most 8, other players. Each player plays the Figure 7 game with her neighbors. As in the network game of Example 3, in this case the all-$s_2$ equilibrium is the unique stochastically stable equilibrium of the best-response dynamic. In computer simulations, when the system was set at the all-$s_2$ equilibrium the best-response dynamic with independent random mutations could never establish a stable foothold of $s_1$-followers over 100,000 periods even with a mutation rate as high as .10. Additionally,

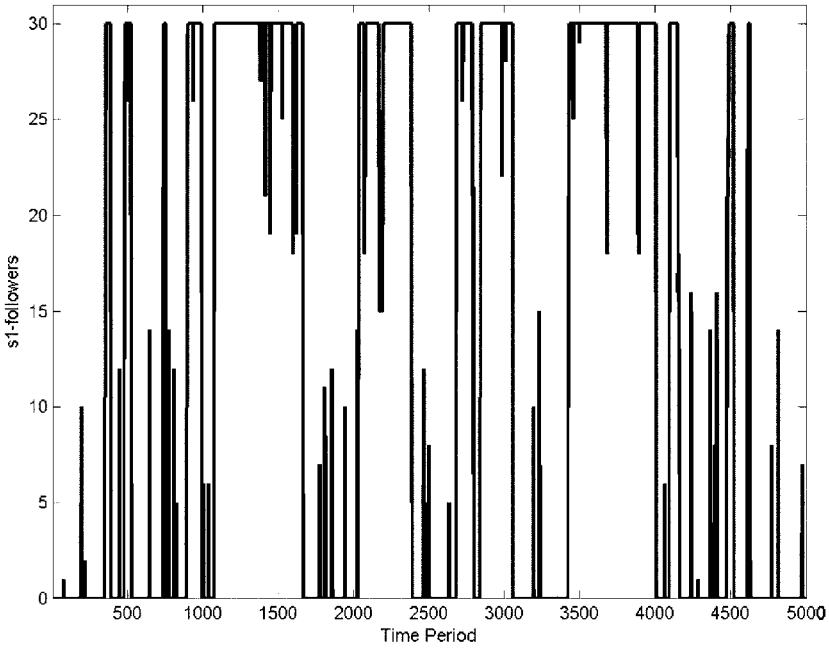


Figure 8. Frequency of $s_1$-followers in the 30-player bounded degree network game who update according to a best-response dynamic with influence neighborhoods.

the all-$s_2$ equilibrium is not robust against the introduction of influence neighborhoods of the sort applied to the network game of Example 3.

We next examined the following perturbed best-response dynamics: At each time step, independent mutants of $s_2$-followers appear with probability .1, and $s_1$-mutants appear spontaneously with probability .001. When an $s_1$-mutant appears spontaneously, her neighbors together with their neighbors each follow $s_1$ with probability $\lambda_i(t)$ chosen at random from [0,1]. This dynamic always converged to the optimal all-$s_1$ equilibrium, even though all-$s_2$ is stochastically stable, and at any time period, an average of 10% of the players spontaneously mutated to $s_2$. Figure 9 summarizes the results of one computer simulation over 5000 periods where the system was initially set at the all-$s_2$ equilibrium.

In this network game, all-$s_1$ is the unique stable attractor of the best-response dynamics perturbed with these influence neighborhoods. This result is especially striking because $s_2$-following mutants appear 100 times as often as leader $s_1$-following mutants appear, and even when an $s_1$-following leader $i$ appears at period $t$, the correlation in her influence neighborhood might be weak depending upon $\lambda_i(t)$. The high influx of

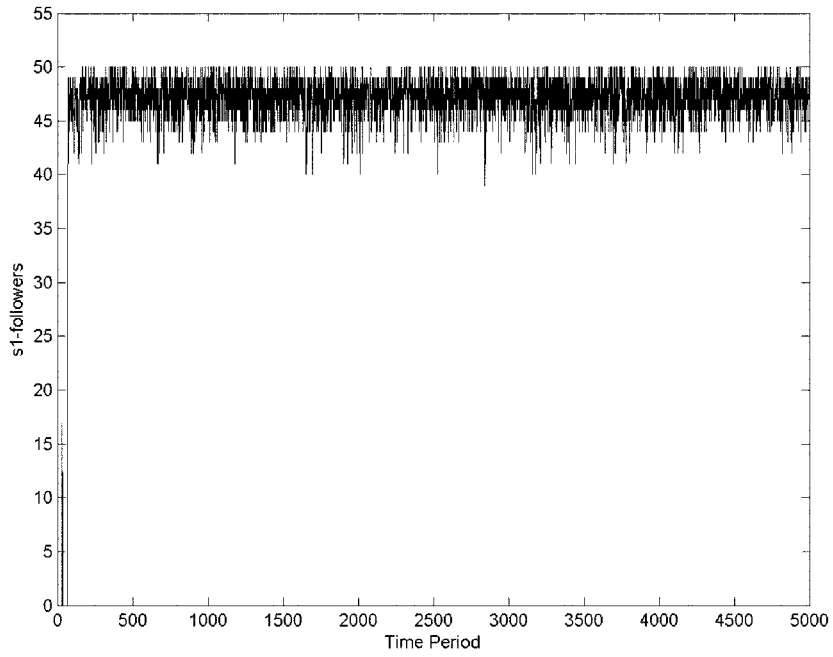


Figure 9. Frequency of $s_1$-followers in the 50-player bounded degree network who update according to a best-response dynamic with influence neighborhoods.

$s_2$-mutants cannot prevent the overthrow of the all-$s_2$ equilibrium because the $s_1$-mutations are correlated. $s_1$-mutant leaders appear seldom in the network game, but the coordinated play across their influence neighborhoods enables the $s_1$-followers to conquer the network and to suppress the high influx of $s_2$-mutants over time.

Example 4 shows that influence neighborhoods appearing at random in a network game can drive this game to an optimal equilibrium, robust against a high rate of independent mutations, even when the suboptimal equilibrium of the 2-player base game is risk dominant. The stability of the optimal equilibrium with respect to this dynamic depends upon the $s_1$-mutants being correlated, while the $s_2$-mutants remain independent. To explain this asymmetry, one can allow for differences in ability to communicate across individuals. That is, leader $s_1$-following mutants may have access to some communication channel used to send messages to those in their influence neighborhoods, while $s_2$-mutants have no reliable means of communicating. So even though leader $s_1$-mutants appear seldom in the network, their ability to signal their plans to others enables those they have influence over to coordinate more effectively. On the other hand,

even though $s_2$-mutants appear at a much higher rate, they are unable to communicate and consequently cannot coordinate their activity. So $s_1$-followers can overthrow an incumbent all-$s_2$ equilibrium and even fight off a continual high influx of new $s_2$-followers.

This suggests a method for testing Hardin's (1995) dual coordination explanation of the duration of regimes and successful revolt. Hardin maintains that a generally despised regime will remain in power so long as agents of the regime can simultaneously coordinate their activity and prevent those under their jurisdiction from coordinating. This explains why repressive regimes suppress communication. Hardin argues that if dissidents become able to communicate and thereby coordinate, while the regime's agents lose these abilities, the regime becomes vulnerable. One can interpret the network game and influence neighborhoods of Example 4 as follows: To follow $s_2$ is to obey a regime all dislike. Suppose, though, that dissident $s_1$-followers establish an underground broadcasting network enabling them to send messages to others while jamming the attempts of reinforcement sent by the $s_2$-following regime. Given these conditions, the $s_1$-followers stage a successful revolt.

**4. The Formal Model.** Let $N = \{1, \ldots, n\}$ denote the set of *players* and let $ij$ denote the subset $\{i, j\} \subseteq N$. Each $ij \subseteq N$ with $i \neq j$ is an undirected *link* (or edge) for $N$. We use $K_n$ to denote the complete undirected graph over $N$. A subset $\mathcal{N} \subseteq K_n$ defines an *interaction network*. If $ij \in \mathcal{N}$, then $i$ and $j$ are *neighbors* and are said to be *linked*. The set of all players that a given Player $i$ is linked to is called the neighborhood of $i$ and is denoted $\mathcal{N}_i$. Since each player in the network interacts with at least one other player, $\mathcal{N}_i \neq \varnothing$ for all $i$.

All players in the network play a symmetric, noncooperative 2-player game $\Gamma$ with pure strategy set $S$ and payoff matrix $u : S \times S \rightarrow \Re^2$. If $i$ follows $s_i$ and $j$ follows $s_j$, the payoff to $i$ is $u(s_i, s_j)$ and the payoff to $j$ is $u(s_j, s_i)$. $\mathcal{N}$ and $\Gamma$ characterize the network game $N_\Gamma$.

A state of a network game is a vector $\vec{s} = (s_{k_1}, \ldots, s_{k_n})$, specifying a strategy for each player in the network. In each round of play, each player plays the game with all of her neighbors, receiving a score equal to the sum of the payoffs. A strategy $s_{k_i}$ is a best response for Player $i$ to $\vec{s}_{-i}$[16] if $s_{k_i}$ maximizes $i$'s payoff, that is,

$$\sum_{j \in N_i} u_i(s_{k_i}, s_{k_j}) \geq \sum_{j \in N_i} u_i(s'_{k_i}, s_{k_j}) \quad \text{for each } s'_{k_i} \in S. \tag{3}$$

16. The subscript '$-i$' indicates the result of removing the $i$th component of an $n$-tuple. In particular, $\vec{s}_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ denotes the $n - 1$ tuple of strategies that Player $i$'s opponents follow when they all follow $\vec{s} = (s_1, \ldots, s_n)$.

$S_i^*(\vec{s}_{-i})$ denotes the set of Player $i$'s best responses to $\vec{s}_{-i}$. A state $\vec{s}^* = (s_{k_1}^*, \ldots, s_{k_n}^*)$ is a Nash equilibrium of $N_\Gamma$,

$$s_{k_i}^* \in S_i^*(\vec{s}_{-i}) \quad \text{for each } i \in N, \tag{4}$$

and $\vec{s}^*$ is *strict* if exactly one strategy satisfies (4) for each $i \in N$.

Since players may change strategies over time, we need to add explicit time-indexes to some of the above definitions. Consequently, let $s_i(t)$ denote the strategy $i$ follows at time $t$. Likewise, let $\vec{s}(t) = (s_1(t), \ldots, s_n(t))$ denote the state of the network at time $t$.

The (inductive) best-response dynamic (*BR*-dynamic) specifies how individuals change strategies as follows: For each $i \in N$, let $f_i$ be a (possibly random) choice function from subsets of strategies to single strategies, that is, $f_i : \mathcal{P}(S) - \varnothing \to S$. At each period $t > 0$,

$$BR_i(t) = f_i(\{s_k \in S : s_k \in S_i^*(\vec{s}_{-k}(t-1))\}). \tag{5}$$

In words, at time $t$ each Player $i$ adopts a strategy that is a best response to the strategies $i$'s neighbors followed at period $t-1$. If $S_i^*$ contains more than one pure strategy, then $i$ selects one of these best responses according to his choice function.

Notice that $BR_i(t) = BR_i(t-1)$ for each $i \in N$, only if $\vec{s}(t) = \vec{s}(t-1)$ is a Nash equilibrium of $N_\Gamma$, that is, the fixed points of the *BR*-dynamic are Nash equilibria of the network game. Also note that the converse need not hold, for if $\vec{s}(t-1)$ is a Nash equilibrium but not strict, then at time $t$ some players might choose best responses other than their respective parts of $\vec{s}(t-1)$. However, any strict Nash equilibrium is a fixed point of the *BR*-dynamic, since by definition each player's part of such an equilibrium is her unique best response to the others' strategies.

Let $\sigma_i$ denote a completely mixed strategy[17] for Player $i$ and $A_{\varepsilon_1}^t, \ldots, A_{\varepsilon_n}^t$ denote stochastically independent propositions such that $\Pr(A_{\varepsilon_i}^t) = \varepsilon_i$. Then the *$\overline{BR}$-dynamic with independent random mutation* is defined by

$$\overline{BR}_i(t, \varepsilon_i, \sigma_i) = (1 - 1_{A_{\varepsilon_i}^t}) \cdot BR_i(t) + 1_{A_{\varepsilon_i}^t} \cdot \sigma_i, \tag{6}$$

where $1_A$ denotes the indicator of a proposition $A$.[18] That is, at each stage

17. A player follows a completely mixed strategy by pegging his pure strategies on a random experiment such that each pure strategy has a positive probability of being followed according to the outcome of the experiment (von Neumann and Morgenstern 1944; Nash 1951a).

18. That is,

$$1_A = \begin{cases} 1 & \text{if } A \text{ obtains} \\ 0 & \text{otherwise.} \end{cases}$$

$t$, $i$ best responds with probability $1 - \varepsilon_i$, and with probability $\varepsilon_i$ chooses a pure strategy at random. $\varepsilon_i$ is $i$'s *mutation rate*. One may interpret a mutation $\sigma_i$ as $i$ experimenting or making an error, or as one individual being replaced by a fresh individual unfamiliar with the history of play.

Our conception of mutations as being dependent upon the influence neighborhood of a player may be formally defined as follows:

> **Definition.** Given a network game $N_\Gamma$, at each period $t$ there is a matrix
>
> $$\lambda_N(t) = \begin{pmatrix} \lambda_{11}(t) & \cdots & \lambda_{1n}(t) \\ \vdots & \ddots & \vdots \\ \lambda_{n1}(t) & \cdots & \lambda_{nn}(t) \end{pmatrix},$$
>
> where $(\lambda_{i1}(t), \ldots, \lambda_{in}(t)) = \vec{\lambda}_i(t)$ is a probability distribution. Player $i$'s influence neighborhood ($I$-neighborhood) at time $t$ is the set $I_i(t) = \{j \in N : \lambda_{ji}(t) > 0\}$. The *size* of an $I$-neighborhood is the cardinality of this set.

The influence neighborhood probabilities can vary over time periods, while the graph that defines the network remains fixed. The underlying intuition here is that changing one's neighbors is prohibitively costly, but cost-free communication with nearby players might at times be possible. So the players' interaction neighborhoods remain fixed, but their influence neighborhoods can change rapidly. Note that for a given $ij \in \mathcal{N}$, we can have $\lambda_{ij}(t) \neq \lambda_{ji}(t)$. This reflects the idea that influence need not be a symmetric relation between players. The weights can vary across players in a $I$-neighborhood so that influence might vary across players as well as across time.

The precise manner by which players correlate their strategies is defined by a variant of the best-response dynamic that incorporates influence neighborhoods:

> **Definition.** Let $A^t_{i1}, \ldots, A^t_{in}$ be mutually exclusive propositions such that $\Pr(A^t_{ij}) = \lambda_{ij}(t)$. Then the $BR^*$-dynamic with influence neighborhoods $(\lambda_N(t))$ is defined as follows: For $i \in N$,
>
> $$BR^*_i(t, \vec{\lambda}_i(t)) = \sum_{j \in N} s_j(t) \cdot 1_{A^t_{ij}}, \tag{7}$$
>
> where $s_i(t) = \overline{BR}_i(t, \varepsilon_i, \sigma_i)$. In words, Player $i$ imitates the strategy of Player $j$ with probability $\lambda_{ij}(t)$ if $i$ falls in $j$'s influence neighborhood.

If at time $t$ we have $\lambda_{ii}(t) = 1$ for all $i \in N$, then no player has influence over other players. For this special case, the $BR^*$-dynamic with influence neighborhoods reduces to the $\overline{BR}$-dynamic with independent random mu-

tation. At another extreme, if a Player $i$ has a number of other players falling in her $I$-neighborhood at period $t$ and $\lambda_{ji}(t) = 1$ for each $j \in I_i(t)$, then all of the players in $i$'s influence neighborhood are certain to correlate their strategies with $i$ at period $t$. One might think of this case as a 'perfectly disciplined' influence neighborhood whose members all follow their leader's command.

In Examples 2, 3, and 4, a leader Player $i$ with an influence neighborhood $I_i(t) \neq \{i\}$ appears at random in the network game, and $i$'s influence probability is constant over the influence neighborhood. This is why in these examples it makes sense to write $\lambda_{ji}(t) = \lambda_i(t)$ for each $j \in I_i(t)$. In these examples, a leader's influence over those in his influence neighborhood is set at random and lasts only for a single time period, save for the unlikely event that this leader spontaneously mutates over consecutive time periods. Of course, many other configurations of influence neighborhoods are possible. A leader player might have a fixed influence neighborhood over part of the network game and over an indefinite number of consecutive periods. Such fixed influence neighborhood leaders have their analogous counterparts in real life, such as political leaders and military commanders. If a single Player $i$ is such that $\lambda_{ji}(t) = 1$ for all time and all $j \in N$, then the entire network is $i$'s perfectly disciplined influence neighborhood. In this case, $i$ plays a role analogous to Hobbes' absolute sovereign. As noted above, perfect discipline is an extreme case. In many actual situations, a leader's 'clout' varies across the population. Varying influence probabilities reflect a leader's uneven sway over those who receive his messages. One virtue of the framework described here is its flexibility for treating such cases.

When correlated mutations are possible, a variety of long-term outcomes can emerge, depending upon the payoff structure and the topology of the interaction network and the influence neighborhoods. If influence neighborhoods remain fixed and perfectly disciplined over a stretch of time, the network can remain at a polymorphism of strategies where those in the influence neighborhoods follow their leaders and the rest converge to the strategy that defines the stochastically stable equilibrium. Examples 3 and 4 show that stochastically stable equilibria need not be robust against influence neighborhoods even when these neighborhoods appear momentarily at random at very low rates. These examples show that there is *no* universal convergence property of the best-response dynamic perturbed with influence neighborhoods analogous to that of stochastic stability when all mutations are stochastically independent.

However, it is possible to define convergence concepts for the $BR^*$-dynamic and to identify some sufficient conditions for convergence. Example 2 shows that influence neighborhoods can greatly accelerate the

convergence of the network to the strategy of the risk dominant equilibrium. Examples 2 and 4 suggest the following:

**Definition.** A state $\vec{s} = (s_1^*, \ldots, s_n^*)$ of a network game is an *attractor* of the *BR**-dynamic with influence neighborhoods $\vec{\lambda}_N(t)$ if for some state $\vec{s}' \neq \vec{s}^*$, when $\vec{s}(0) = \vec{s}'$, then

$$\Pr\left(\lim_{t\to\infty} s_i(t) = (1 - 1_{A^t_{\varepsilon_i}} s_i^* + \sigma_i \cdot 1_{A^t_{\varepsilon_i}})\right) = 1 \quad \text{for each } i \in N. \quad (8)$$

If (8) is satisfied for every state $\vec{s} \neq \vec{s}^*$, then $\vec{s}^*$ is the *global attractor* of the *BR**-dynamic.

This definition says that $\vec{s}^*$ is an attractor of a *BR**-dynamic if, with probability one, from an initial state $\vec{s}(0)$, players who update according to this dynamic all eventually follow $\vec{s}^*$, except when they mutate. In Example 2, $(s_1^*, \ldots, s_n^*)$ is the global attractor of the *BR**-dynamic, where each *I*-neighborhood is a Moore-24 neighborhood of varying discipline. For any initial state of this network, under this *BR**-dynamic all the players eventually follow $s_1$ except for the occasional *I*-neighborhood of $s_2$-followers that appears and is then eliminated.

**Proposition 1.** If $\vec{s}^* = (s_1^*, \ldots, s_n^*)$ is an attractor of the *BR**-dynamic with influence neighborhoods $\vec{\lambda}_N(t)$, then $\vec{s}^*$ is a Nash equilibrium.

**Proof.** By hypothesis, given some $\vec{s}(0) = \vec{s} \neq \vec{s}^*$ this *BR**-dynamic satisfies (8). Hence as $t \to \infty$, with probability one, each Player $i \in N$ follows $s_i^*$, unless $i$ mutates spontaneously or imitates the strategy of a leader in case $i$ falls in this leader's *I*-neighborhood. But then $s_i^*$ must be a best response for each $i \in N$ under the unperturbed *BR*-dynamic, and so (4) is satisfied. QED

Example 2 shows that a system of *BR**-updaters can converge to an optimal Nash equilibrium even when the *I*-neighborhoods appear randomly in the network at a very low rate, and the initial state is a suboptimal but strict equilibrium. In the remainder of this section, we show why this is the case and at the same time establish some convergence conditions for *BR**-dynamics. First, we define the notion of *BR*-stability for the unperturbed best-response dynamic.

**Definition.** Given the network game, a set $B \subseteq N$ is *BR-stable* with respect to $s \in S$ if given $s_i(t) = s$ for each $i \in B$, $BR_i(t + 1) = s$. If $B$ is *BR*-stable with respect to $s$, we say that $s$ is *BR-stable over B*.

Intuitively, a set of the players is *B*-stable with respect to the pure strategy $s$ if when all in $B$ start to follow $s$, the *BR*-dynamic cannot 'erode'

the $s$-following throughout $B$ even when all the rest of the players in $N_\Gamma$ do not follow $s$.

The following proposition establishes a convergence result for the $BR^*$-dynamic provided certain $BR$-stable sets exist.

**Proposition 2.** In a network game, let $I$-neighborhoods of bounded size $b < n$ appear with probability $\varepsilon$. Also assume that, for each $I$-neighborhood, the probability that the leader adopts the pure strategy $s \in S$ is $1/|S|$. For each $I$-neighborhood that appears, let $q > 0$ be a lower bound on the probability that this neighborhood is of size $b$, and let $p > 0$ be a lower bound on the probability everyone in this neighborhood follows the strategy of the leader. If for each $I$-neighborhood of maximum size $b$, $\vec{s}^*$ is the unique $BR$-stable strategy of some subset of that $I$-neighborhood, then $\vec{s}^* = (s^*, \ldots, s^*)$ is the global attractor of this $BR^*$-dynamic.

**Proof.** Let $(I_u)$ denote the sequence of $I$-neighborhoods that appears in the network lexically ordered according to time. With probability one, a perfectly disciplined $I$-neighborhood of maximum size $b$ whose players follow $s^*$ appears infinitely often in the sequence of plays. Let $(I_{u_k})$ denote the subsequence of $(I_u)$ such that $I_{u_k}$ is of size $b$ and each $i \in I_{u_k}$ follows $s^*$, and let $(B_{u_k})$ denote the sequence of $BR$-stable sets of $s^*$-followers that appear in $N_\Gamma$ as a result. We claim that $s^*$ satisfies (8), that is, $s^*$ overtakes the network game with probability one. Each $B_{u_k}$ introduces a number of $s^*$-followers that remain in the network over time until $B_{u_k}$ is disrupted by some influence neighborhood whose players follow some strategy other than $s^*$. Thus the $B_{u_k}$'s gradually increase the number of $s^*$-followers in the network until all but mutants follow $s^*$ unless all but some fixed finite number of the $B_{u_k}$'s are disrupted by 'counter' $I$-neighborhoods whose 'leaders' follow strategies other than $s^*$ that appear and overlap the $B_{u_k}$'s. But for this containment of the $B_{u_k}$'s to occur, a sequence $(A_{u_k})$ of $I$-neighborhoods synchronized with the $I_{u_k}$'s must appear in $N_\Gamma$ such that all but a fixed number of the $A_{u_k}$'s satisfy the following properties: (i) the leader Player $i_{u_k}$ of each $A_{u_k}$ follows some strategy other than $s^*$, (ii) $i_{u_k}$ appears in a part of the network where $A_{u_k}$ overlaps $I_{u_k}$, and (iii) enough players in $A_{u_k}$ imitate $i_{u_k}$'s strategy to disrupt the $s^*$-stability of $B_{u_k}$ so that the players in $B_{u_k}$ do not continue to follow $s^*$. (If these conditions are not met, then a subsequence of the $B_{u_k}$'s is not contained by the $A_{u_k}$'s and this subsequence then overtakes the whole network. But if $A'_{u_k}$ denotes the proposition that for a given $B_{u_k}$ a matching $A_{u_k}$ appears satisfying (i), (ii), and (iii), then $\Pr(A'_{u_k})$ is some value $\eta_{u_k} < 1$. For note that the probability that (i) occurs is fixed by hypothesis. The probability that (ii) occurs is some

fixed number, for there are only so many ways a $I$-neighborhood of size $b$ can overlap one of the $B_{u_k}$'s. The probability that (iii) occurs is bounded from above, since the 'best case' scenario for the 'disrupters' is if $A_{u_k}$ overlaps perfectly with $B_{u_k}$ and then sufficiently many players in $A_{u_k}$ imitate $i_{u_k}$ to destabilize the $s^*$ strategy in $B_{u_k}$. So the $\eta_{u_k}$'s are bounded from above by some $\eta < 1$. Hence if $A'$ denotes the event that the necessary sequence of $A_{u_k}$'s appears to contain the $B_{u_k}$'s, then

$$\begin{aligned} \Pr(A') &= \Pr(A'_{u_k} \text{ for all but finitely many } u_k) \\ &= \lim_{m\to\infty} \eta_{u_{k_1}} \cdots \eta_{u_{k_m}} \\ &\leq \lim_{m\to\infty} \eta^m = 0. \end{aligned}$$

QED

The key idea behind Proposition 2 is that $BR$-stable sets of $s^*$-followers appear in the network game and tend to persist, even when they do not appear in consecutive time periods and are not contiguous in the network. So with probability one, the appearance of these $BR$-stable sets together with the forces of the $BR^*$-dynamic results in $s^*$ overtaking the entire network game. This argument differs considerably from the proofs of the stochastic stability results for independent random mutations in works such as Young (1993, 1998), Ellison (1993, 2000), and Morris (2000), which consider the behavior of a network game in the rare event that sufficiently many independent mutations occur consecutively so as to drive the system out of and away from an equilibrium. Note also that the premises of Proposition 2 do not bias the $BR^*$-dynamic to make the influence neighborhoods of any one strategy more likely to appear or to persist over time than another. Finally, note that the proof of Proposition 2 does not depend upon specific values of $\varepsilon$, $q$, or $p$ as stated in the hypotheses. So the $BR^*$-dynamics that satisfy these hypotheses ultimately overtake the entire network game no matter how infrequently $BR$-stable sets of $s^*$-followers appear, as long as they appear with some positive probability at each period.

We can now identify certain network structures where a $BR^*$-dynamic that is not biased in favor of any pure strategy will converge to risk dominant equilibrium play. A network $\mathcal{N}$ is *c-regular* if each Player is linked with exactly $c$ other players.

**Corollary 3.** Let $\mathcal{N}$ be $c$-regular, and let $(s^*, s^*)$ denote the risk dominant equilibrium of the base game $\Gamma$. Let $I$-neighborhoods of bounded size $b < n$ appear with probability $\varepsilon$ where, for each $I$-neigh-

borhood, the probability that the leader adopts the pure strategy $s \in S$ is $1/|S|$. For each $I$-neighborhood that appears, let $q > 0$ be a lower bound on the probability that this neighborhood is of size $b$, and let $p > 0$ be a lower bound on the probability everyone in this neighborhood follows the strategy of the leader. If for each $I$-neighborhood $I_i(t)$ with $b$ members, a nonempty subset $B_i$ of $I_i(t)$ is such that each player in $B_i$ is linked with at least $c/2$ players in $I_i(t)$, then $\vec{s} = (s^*, \ldots, s^*)$ is the global attractor of this $BR^*$-dynamic.

**Proof.** As in the proof of Proposition 2, let $(I_u)$ denote the sequence of $I$-neighborhoods that appears in the network lexically ordered according to time periods. With probability one, a subsequence $(I_{u_k})$, where $I_{u_k}$ is of size $b$ and each $i \in I_{u_k}$ follows $s^*$ appears in the sequence of plays. By hypothesis, each $I_{u_k}$ contains a nonempty subset $B_{u_k}$ whose member nodes are each linked with at least $c/2$ players in $I_{u_k}$. Since $(s^*, s^*)$ is risk dominant, $s^*$ is the unique best response for each $i \in B_{u_k}$ at subsequent time periods, because at least half of $i$'s $N$-neighbors followed $s^*$. Hence $s^*$ is the unique $BR$-stable strategy for each $B_{u_k}$ in the sequence $(B_{u_k})$, so all of the hypotheses of Proposition 2 are satisfied. QED

Corollary 3 establishes that when the base game has a risk dominant equilibrium, a large class of $BR^*$-dynamics will converge to risk dominant equilibrium play in the special case where the interaction network is uniformly linked, as are the one-dimensional circular network games analyzed by Ellison (1993) and the two-dimensional lattice of Examples 2 and 3. Moreover, we can now explain why random mutations failed to overthrow the suboptimal $(s_2, \ldots, s_2)$ equilibrium of the lattice network of Assurance games in Example 1 while in Example 2 influence neighborhoods that entered the same network game rapidly moved the system to the $(s_1, \ldots, s_1)$ equilibrium. In order for any $BR$-stable set of $s_1$-followers to appear in the network, at least 12 players in a 'cross' configuration must simultaneously mutate to $s_1$. Given stochastic independence with $\varepsilon_i = .05$, the probability that even one such group of 12 appears in the network over 100 million generations is bounded from above by $2^{-24}$.[19] So it is not surprising that in Example 1 the independent $s_1$-mutants failed to establish a stable bridgehead in the network game over a million generations even though they appeared at such a high rate.

19. If $S(t)$ denotes the event that at least one $BR$-stable set appears at round $t$, then

$$\Pr(S(t)) \leq 10^4 \varepsilon_i^{12} \left(\frac{1}{2}\right)^{12} = 10^4 \left(\frac{1}{20}\right)^{12} \left(\frac{1}{2}\right)^{12} = 10^{-8} \cdot 2^{24},$$

so $\Pr(S(1) \vee \cdots \vee S(10^8)) \leq 2^{-24}$.

In Example 2, even though *I*-neighborhoods appeared at a rate of only .001, some of the *I*-neighborhoods of $s_1$-followers that appeared introduced *BR*-stable sets. Each *I*-neighborhood was a Moore-24 neighborhood, and since the network game was 8-uniformly linked, all but the four 'corner' players of a given *I*-neighborhood were linked with at least four players in the same *I*-neighborhood. So when a perfectly disciplined *I*-neighborhood of $s_1$-followers appeared in the network game surrounded by $s_2$-followers, the corner players converted to $s_2$ on the subsequent round of play but the remaining 20 players formed a *BR*-stable set of $s_1$-followers that persisted in the game. While these *BR*-stable sets appeared seldom in the network due to the very low leader mutation rate, they started a steady contagion of $s_1$-followers that rapidly overtook the network.

Proposition 2 and Corollary 3 are fundamental convergence results for *BR**-dynamics. They show that for certain classes of network games, influence neighborhoods large enough to introduce *BR*-stable sets will ultimately drive a network game to a unique Nash equilibrium, no matter how infrequently leader mutants appear. However, these results cannot be generalized to all network games. Example 3 shows that if the interaction network is not uniformly linked there might be no global attractor, or even a stable equilibrium, of a *BR**-dynamic that introduces influence neighborhoods following each pure strategy at equal rates. Example 4 shows that under a *BR**-dynamic that introduces influence neighborhoods at rates and of sizes that vary across pure strategies, a nonuniformly linked network game can converge to an equilibrium of nonrisk dominant play that is robust against a high rate of spontaneous mutation. Plainly, the impact of correlated influence neighborhood mutation varies according to the network structure and the specifics of the influence neighborhoods.

**5. Conclusion.** We have shown that correlated mutations profoundly influence the evolution of strategies across local interaction structures. Previous work established that when the base game of any network game has a risk dominant equilibrium, risk dominant play characterizes the unique stochastically stable state of the best-response dynamic (Ellison 1993; Young 1998). The generality of this result suggests that the payoffs of the base game alone determine the long-term limits of dynamical updating. However, the examples in this paper show that the tight connection between risk dominant play and dynamic stability dissolves when one relaxes the assumption that all mutations are stochastically independent. Network structure *does* play a role in determining where the players end up when the correlated mutations of influence neighborhoods can appear. Correlation via influence neighborhoods can help drive a network of players to a stable equilibrium of risk dominant play, or to some other stable equilibrium. And it is possible that no state is stable when influence neigh-

borhoods enter into the network game, even when this game has a unique stochastically stable equilibrium of risk dominant play.

Correlation via influence neighborhoods also dramatically accelerates the evolution of equilibria in some network games. We have seen that when only independent mutations are possible, the players in a network game can find themselves trapped at a suboptimal equilibrium that is not stochastically stable for a very long time. While according to theory, independent mutations will ultimately drive the network game to the stochastically stable equilibrium, this process may take so long that stochastic stability cannot be the basis for any realistic explanation of the emergence of a new optimal social equilibrium in human communities. Communities of people do occasionally reform their practices, and the process does not typically occur as the result of independent aberrations in behavior over millions of consecutive interactions. Successful reform requires coordinated departures from incumbent practice. Typically, such coordination requires planning, communication and leadership. Such coordination also succeeds by generating a 'bandwagon' effect that spreads quickly through society. Influence neighborhoods prove useful for modeling this coordination. Moreover, an optimal equilibrium that independent mutation never produces over millions of periods of interaction can emerge quite rapidly under the correlated mutations of influence neighborhoods. We believe that influence neighborhoods can be a valuable tool for analyzing social change.

Most of the literature on network games, including the stochastic stability literature, develops quite general convergence results from powerful assumptions that are mathematically convenient but not really well founded. In this paper, we have explored some of the consequences of relaxing one of these assumptions, namely, that all mutations are stochastically independent. Not surprisingly, we do not get convergence theorems for influence neighborhood dynamics as general as those of the stochastic stability literature, but we do get what we think is a more realistic model of how strategies develop over local interaction structures. Future work should investigate the consequences of relaxing some of the other robust assumptions common in the network game literature, in conjunction with relaxing the stochastic independence assumption. Players might not be so myopic as the literature assumes. Updating rules more sophisticated than the best-response dynamic should be explored. Players might not always interact with the same neighbors. Some authors have already proposed models where the interaction network itself evolves over time (Skyrms and Pemantle 2000; Goyal and Vega-Redondo 2000; Watts 2001, Jackson and Watts 2001a, 2001b). Combining different learning rules and evolving network structures with influence neighborhood mu-

tation may produce a theory of network games that has much greater explanatory power than the existing theory.

REFERENCES

Alexander, J. McKenzie (2000), "Evolutionary Explanations of Distributive Justice", *Philosophy of Science* 67: 490–516.
Alexander, J. McKenzie, and Brian Skyrms (1999), "Bargaining with Neighbors: Is Justice Contagious?", *Journal of Philosophy* 96: 588–598.
Binmore, Ken (1998), *Just Playing*. Cambridge, MA: MIT Press.
Brown, George W. (1951), "Iterative Solutions of Games by Fictitious Play", in T. C. Koopmans (ed.), *Activity Analysis of Production and Allocation*. New York: Wiley, 374–376.
Ellison, Glenn (1993), "Learning, Local Interaction and Coordination", *Econometrica* 61: 1047–1071.
——— (2000), "Basins of Attraction and Long Run Equilibria", *Review of Economic Studies* 67: 17–45.
Foster, Dean, and H. Peyton Young (1990), "Stochastic Evolutionary Dynamics", *Journal of Theoretical Biology* 38: 219–232.
Fudenberg, Drew, and David Levine (1998), *The Theory of Learning in Games*. Cambridge, MA: MIT Press.
Goyal, Sanjeev, and Fernando Vega-Redondo (2000), "Learning, Network Formation and Coordination", Working paper. Rotterdam: Erasmus University.
Grim, Patrick, Gary Mar, and Paul St. Denis (1998), *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling*. Cambridge, MA: MIT Press.
Hampton, Jean (1986), *Hobbes and the Social Contract Tradition*. Cambridge: Cambridge University Press.
Hardin, Russell (1995), *One for All: The Logic of Group Conflict*. Princeton, NJ: Princeton University Press.
Harsanyi, John, and Reinhard Selten (1988), *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.
Hume, David ([1740] 1976), *A Treatise of Human Nature*. Oxford: Clarendon Press.
Jackson, Matthew, and Alison Watts (2001a), "The Evolution of Social and Economic Networks", *Journal of Economic Theory* 106: 265–295.
——— (2001b), "On the Formation of Interaction Networks in Social Coordination Games", *Games and Economic Behavior* 41: 265–291.
Jiborn, Magnus (1999), *Voluntary Coercion*. Ph.D. Dissertation. Lund: Lund University.
Kandori, Michihiro, George Mailath, and Rafael Rob (1993), "Learning, Mutation, and Long-Run Equilibria in Games", *Econometrica* 61: 29–56.
Kavka, Gregory (1986), *Hobbesian Moral and Political Theory*. Princeton, NJ: Princeton University Press.
Lewis, David (1969), *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.
Morris, Stephen (2000), "Contagion", *Review of Economic Studies* 67: 57–78.
Nash, John (1950), "Equilibrium Points in N-Person Games", *Proceedings of the National Academy of Sciences of the United States* 36: 48–49.
——— (1951a), "Non-Cooperative Games", *Annals of Mathematics* 54: 286–295.
——— ([1951b] 1996), *Essays on Game Theory*. Cheltenham, UK: Edward Elgar Publishing.
Nowak, Martin, Sebastian Bonhoeffer, and Robert May (1994), "Spatial Games and the Maintenance of Cooperation", *Proceedings of the National Academy of Sciences of the USA* 91: 4877–4881.
Nowak, Martin, and Robert May (1992), "Evolutionary Games and Spatial Chaos", *Nature* 359: 826–829.
Skyrms, Brian (2001), "The Stag Hunt", *American Philosophical Association* (Proceedings) 75: 31–41.
——— (2004), *The Stag Hunt: Evolution of Social Structure*. Cambridge: Cambridge University Press.

Skyrms, Brian, and Robin Pemantle (2000), "A Dynamic Model of Social Network Formation", *National Academy of Sciences* (Proceedings) 97: 9340–9346.
Taylor, Michael (1987), *The Possibility of Cooperation*. Cambridge: Cambridge University Press.
Taylor, Michael, and Hugh Ward (1982), "Chickens, Whales and Lumpy Goods", *Political Studies* 20: 350–370.
Vanderschraaf, Peter (1998), "The Informal Game Theory in Hume's Account of Convention", *Economics and Philosophy* 14: 215–247.
von Neumann, John, and Oskar Morgenstern (1944), *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.
Watts, Alison (2001), "A Dynamic Model of Network Formation", *Games and Economic Behavior* 34: 331–341.
Young, H. Peyton (1993), "The Evolution of Conventions", *Econometrica* 61: 57–84.
——— (1998), *Individual Strategy and Social Structure*. Princeton, NJ: Princeton University Press.