Local Interactions with Influence Neighborhoods

Peter Vanderschraaf and J. McKenzie Alexander

May 12, 2003

Technical Report No. CMU-PHIL-142

Philosophy

Methodology

Logic

Carnegie Mellon

Pittsburgh, Pennsylvania 15213

Local Interactions with Influence Neighborhoods

Peter Vanderschraaf and J. McKenzie Alexander

Introduction

Game theorists analyze the strategic aspects of interactions. Social network theorists focus on 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.[2] The attracting points of certain dynamical adjustment processes are Nash equilibria of the network game. Some Nash equilibria are also *stochastically stable*, in the sense that these equilibria emerge and persist when the dynamical system is perturbed with independent random changes in strategy or *mutations*.[3] 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.

In this essay we introduce a dynamical adjustment process for network games where mutations can be *correlated.* Previous convergence results for the dynamics of network games rely upon strong assumptions. For example, it is often assumed that the random mutations that perturb the network game are stochastically independent and identically distributed. This assumption is clearly false in many cases of interest. People often imitate others, even in experimental situations, which prevents random mutations from being stochastically independent. In this paper, we present a model that relaxes this assumption by allowing some players in the network game to imitate the behavior of a single player who mutates spontaneously. This introduces a certain number of correlated

mutations. This kind of correlated mutation has a natural interpretation: If a given player in the network game plans to experiment and can signal her plan 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 do imitate the signaler "follow the leader". We show that the dynamical properties of evolutionary games in which influence neighborhoods can appear differ dramatically from ones where all mutations are stochastically independent. We also argue that this dynamics mirrors the process by which societies sometimes reform more closely than the dynamics of stochastically independent mutation.

This essay is organized as follows: In §1 we review some of the basic notions of games played over networks. We use the familiar Assurance game to develop motivating examples. In §2 we discuss how *inductive best-response dynamics* are applied to network games, and give an example of a network game where best-response dynamics perturbed with independent random mutation never reaches the stochastically stable equilibrium in a feasible amount of time. We argue that this result casts doubt upon the explanatory power of models that assume stochastically independent mutations. In §3 we relax the independence assumption by introducing influence neighborhoods. We show how influence neighborhoods can greatly accelerate the transition of a network game from a suboptimal equilibrium to an optimal and stochastically stable equilibrium. We also show how influence neighborhoods can drive a network game out of a stochastically stable equilibrium, and even converge to an optimal equilibrium that is not stochastically stable. In §4 we give the formal definitions of influence neighborhoods and best-response dynamics with correlated mutations, together with some basic convergence results.

§1. The Assurance Game Played With Neighbors

Figure 1 summarizes the symmetric 2-player *Assurance game*.[4]

Figure 1. The Assurance Game

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

$$x > y \geq z > 0$$

The Assurance game plays an important role in moral and political philosophy and illustrates some of the fundamental challenges of accounting for equilibrium selection in games. Philosophers use Assurance games to represent collective action problems ranging from cooperation in the Hobbesian State of Nature to pollution control to political revolutions.[5] As one can see, the structure of Assurance games contains an apparent conflict between optimality and risk. In the game of Figure 1, (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, 1951*a*,*b*) gives no determinate answer to this question. Harsanyi and Selten (1988) attempted to answer this question by introducing a refinement of the Nash equilibrium concept they dubbed *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 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 shown in 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 give the following rationale for claiming that a risk dominant equilibrium is the players' correct choice: One should follow one's part of a risk dominant equilibrium because this is the least risky strategy, in the sense that it is the best response over the larger share of possible probabilities with which the other player follows his pure strategies (Harsanyi and Selten 1988, pp. 82-83). Risk dominance is an important concept in rational choice game theory, but it raises obvious pointed 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 ascribe a high probability to her counterpart choosing s_1 even if (s_2, s_2) is risk dominant, or pessimistically ascribe 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.[6]

Now suppose that, in a population of players, each player plays the Assurance game with each of a given set of the others, who are her "neighbors". At a given time of play, each player follows one strategy in her interactions with all her neighbors.[7] Explicitly identifying the neighbors with whom each player interacts embeds their Assurance game in a *local interaction structure* or *network*. Formally, a network is a graph in which the nodes represent the players. Player j is Player i's *interaction neighbor* if the nodes representing Player i and Player j are linked with an edge. If $n_i(s_1)$ of Player

i's neighbors follow s_1 and $n_{s_2}(i)$ of Player i's neighbors follow s_2, then s_1 is a best response for Player i if

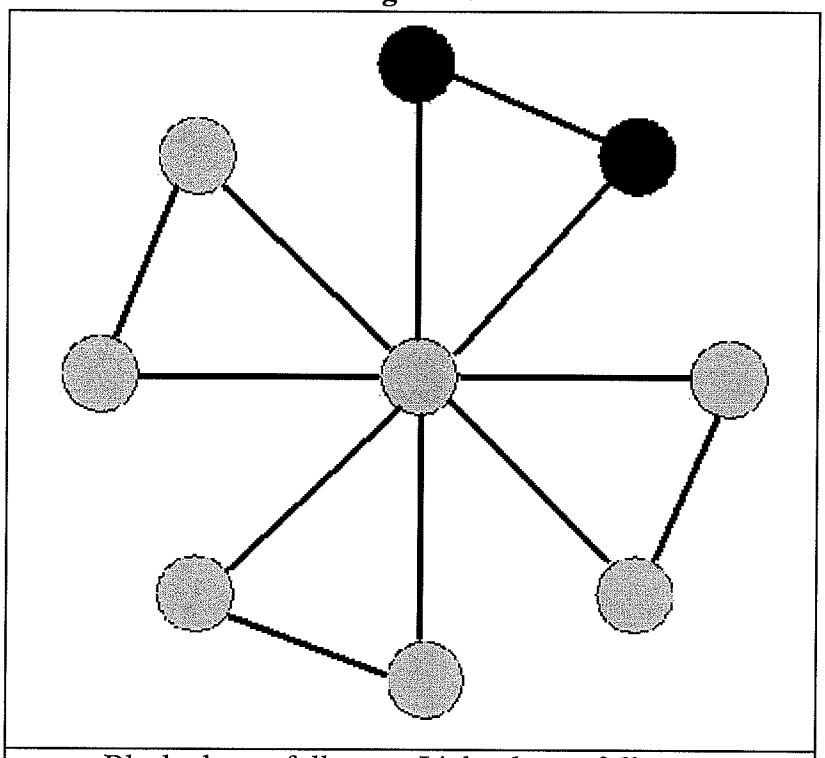
$$n_i(s_1)x \geq n_i(s_1)y + n_i(s_2)z \; . \tag{1.1}$$

and s_2 is a best response if the reverse inequality is satisfied. In the special case where $y = z$, (1.1) is equivalent to

$$x \cdot \frac{n_i(s_1)}{n_i(s_1) + n_i(s_2)} \geq z \tag{1.2}$$

that is, s_1 is a best response if the weighted average of payoffs Player i receives from her neighbors who follow s_1 exceeds the guaranteed payoff of following s_2. To illustrate, Figure 2 depicts a "propellor" graph with 9 players, 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 together with the Figure 1 game define a *network game* in which each player plays the Assurance game with every interaction neighbor.

Figure 2.



Black players follow s_2, Light players follow s_1

If, for instance, $x = 9$ and $y = z = 5$, then by (1.2) s_1 is a best response for the central Player i if $9 \cdot \frac{n_i(s_1)}{8} \geq 5$ or $n_i(s_1) \geq \frac{40}{9} > 4$, so at least 5 of Player i's neighbors must follow s_1 in order for s_1 to be Player i's best response.

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 a set of players play the Assurance game with their interaction 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 Assurance 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 follow?

§2. 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 to equilibrium selection 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 1951*b*).[8] Strikingly, David Hume's analysis of convention in *A Treatise of Human Nature* foreshadows both the Nash equilibrium concept and a dynamical approach to equilibrium selection (Hume 1740, p. 490).[9]

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, Young 1998) 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 part of the time.

With network games, game theorists standardly investigate the properties of the *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 usually react myopically to their situation. If the players in a local interaction structure 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 interaction 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 so clear how far these analytic stochastic stability results can go in explaining how players in the real world might interact more successfully. The following example illustrates this point.

Example 1. Assurance Played on a Torus with Independent Mutations

Let $m > 1$ be an integer and let $N = \{1, ..., n\}$ where $n = m^2$. Define a bijective function $\iota : N \to \{1, ..., m\} \times \{1, ..., m\}$. ι assigns to each Player i a unique index $\iota(i) = (\iota_1(i), \iota_2(i))$. The graph

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

consists of links between each Player i and the 8 neighbors that immediately surround Player i. These links define Player i's Moore-8 neighborhood. This 2-dimensional graph is topologically equivalent to 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 or even around the globe.[10] We set $m = 100$, so that 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, and must choose a single strategy for interaction.

Figure 3. Assurance with (s_1, s_1) Risk dominant

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

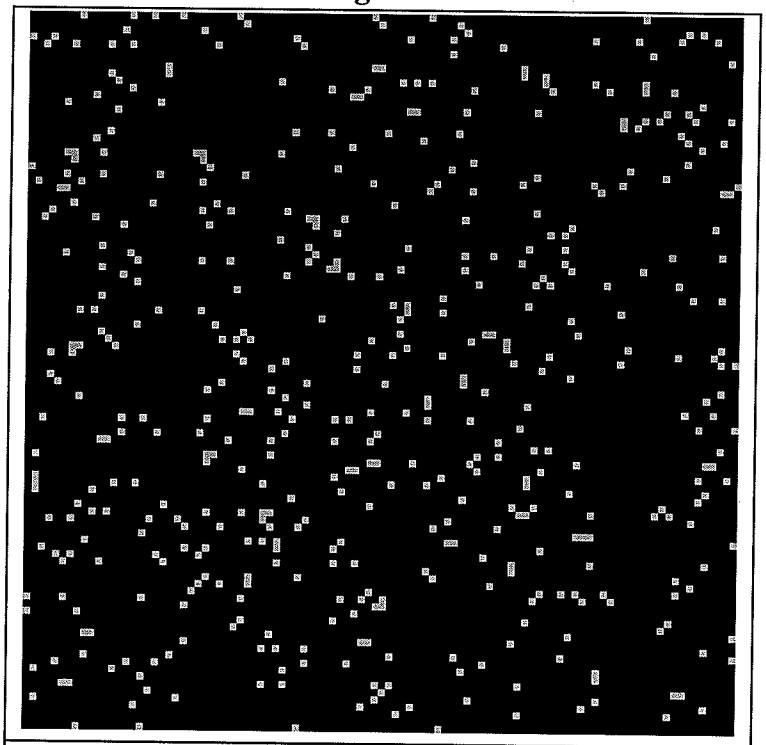
(s_1, s_1) is the risk dominant equilibrium of the Figure 3 game. So if the players in this system update according to the best-response dynamic with independent random mutations, then the stochastically stable equilibrium of this system is the equilibrium where all follow s_1. The all-s_1 equilibrium is the unique stable attractor of this dynamic for any positive 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 will eventually move the entire population to the optimal all-s_1 equilibrium.

One might well wonder how long it takes for this movement to the all-s_1 equilibrium to occur. To test the speed of this convergence, we ran a computer simulation of this system.[11] 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 0.10. 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 1,000,000 periods.[12] Figure 4 depicts the state of this 100×100 graph, mapped onto a square, at the final stage of this simulation.

Figure 4.



Black players follow s_2, Light players follow s_1

Even though the mutation rate was set relatively high, so that at any stage an average of 10% of the players mutated, the s_1-mutants were consistently overwhelmed and rendered unable to establish a permanent foothold in the network. So the s_1-strategy never started

to overthrow 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 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 a .10 mutation rate can converge to and never overthrow the suboptimal all-s_2 equilibrium over 1,000,000 rounds of play even if the system is initially set with as many as 20% of the network players following the s_1-strategy. Still, we think the conditions of the Example 1 test are not so far-fetched. Social dilemmas are those situations where 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 inequality (1.1) 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. However, Example 1 shows that this "long run" can be very long, indeed. One would not expect many people in the actual world to interact exclusively within a fixed neighborhood structure over a million consecutive rounds of play. The network of interactions between people might well change, and perhaps dissolve, over so many stages. But we have just seen that players who do interact this often within a fixed neighbor structure and who mutate independently can fail to replace a Pareto-inferior equilibrium with an equilibrium that is both Pareto-optimal and risk dominant.

Yet the apparent failure of our experiment suggests an interesting possibility. Over a lengthy time frame, best-response dynamics with stochastically independent mutations can fail to converge to the equilibrium that is the ultimate limit of the dynamics. 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.

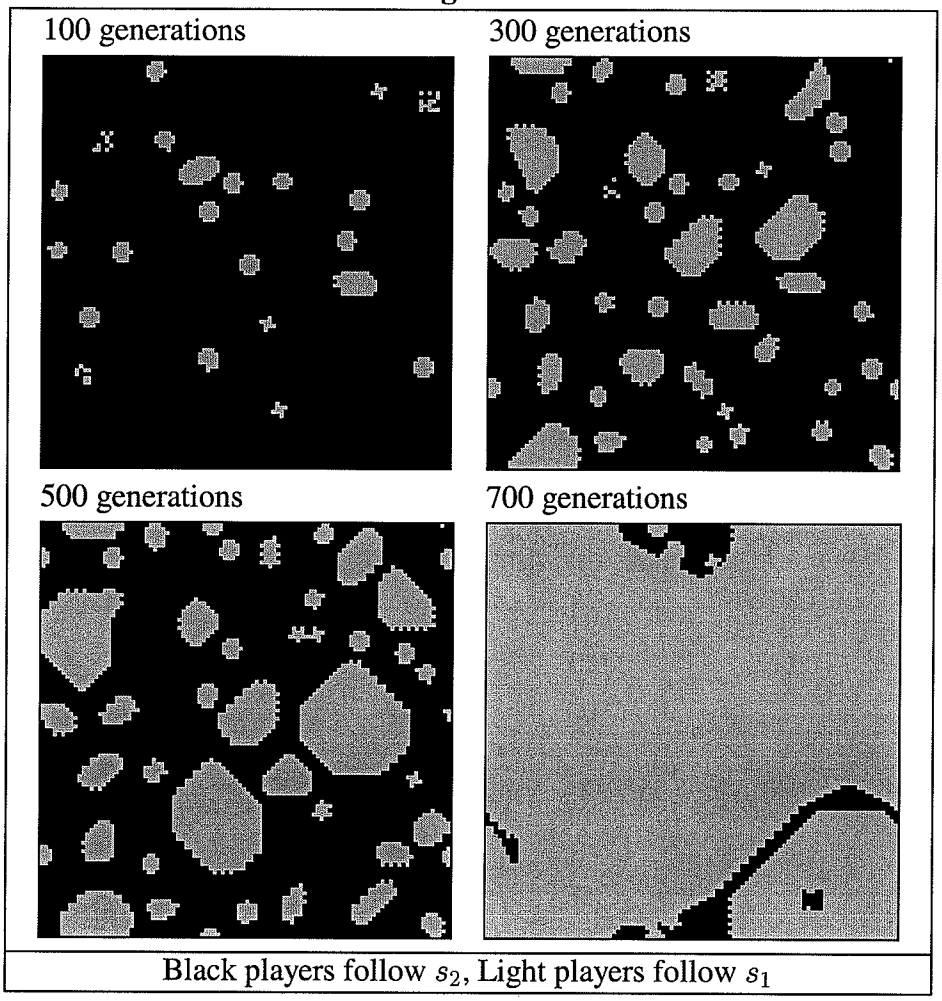
Example 2. Assurance Played on a Torus with Influence Neighborhoods

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 correlation in the mutations across certain individuals. If a given Player i spontaneously mutates at stage t, then each of this player's Moore-8 neighbors and each of *their* Moore-8 neighbors also mutate by imitating Player i's stage t strategy with probability $\lambda_i(t)$. The probability $\lambda_i(t)$ is sampled from the uniform distribution over $[0, 1]$. The 24 players whose stage t strategies are now correlated with Player i's

experiment are Player i's *Moore-24 neighborhood* in the torus. We set the spontaneous mutation rate for a "leader" player at a low 0.0001, so that an average of only one "leader" player per period appears in the entire network. A "leader" spontaneously mutates to s_1 with probability $\frac{1}{2}$ and to s_2 with probability $\frac{1}{2}$.[13]

We tested the properties of this dynamic with a series of computer simulations. 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. 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.

Figure 5.



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 above by $25 \cdot \frac{1}{10,000} = 0.0025$, 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 $\mathcal{I}_i(t)$ is the set of his Moore-24 neighbors. At period t, each $j \in \mathcal{I}_i(t)$ imitates Player 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,

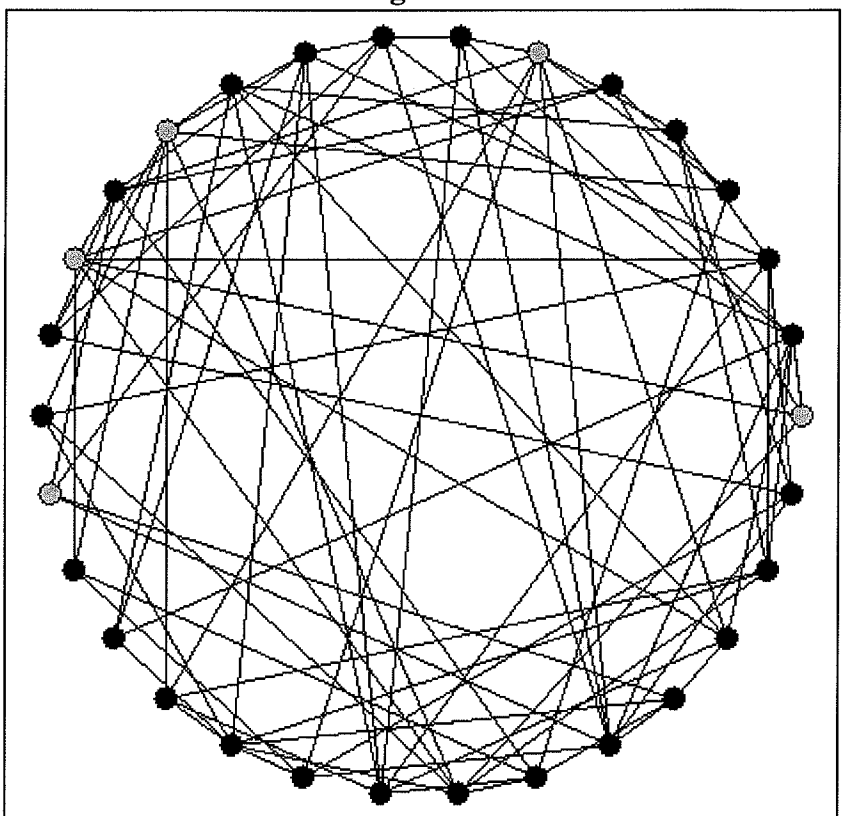
or what game theorists call *cheap talk*. If communication is effectively suppressed throughout the network, then no player has an influence neighborhood including any other player, and players cannot correlate their mutations. But if communication is possible, then players can correlate their strategies with the leader players whose messages they receive. When Player i is a leader at period t, Player i mutates to strategy $s_i(t)$ and communicates this fact to each player $j \in \mathcal{I}_i(t)$. $\lambda_i(t)$ is a measure of the strength of Player i's influence over those in the neighborhood $\mathcal{I}_i(t)$. Those in $\mathcal{I}_i(t)$ who imitate Player i's strategy $s_i(t)$ at period t "follow their leader". In Example 2, the correlated mutation of influence neighborhoods steadily moves the network game from the suboptimal to the optimal equilibrium, even though the influence neighborhoods appear at a low rate. The road to reform in this example is considerably shortened by the introduction of influence neighborhoods.

Example 2 shows that risk dominant play can overtake an interaction network fairly rapidly when some players' strategies are correlated via influence neighborhoods. The next examples show that players who update according to a perturbed best-response dynamic are by no means guaranteed to converge to risk dominant play when correlated influence neighborhood mutations are possible.

Example 3. Assurance Played on a Bounded Degree Network with Influence Neighborhoods

Figure 6 depicts a *bounded degree network*, where each of the 30 nodes is linked with at least 4 and at most 8 other nodes. Again the nodes in the graph represent players, and the edges define each player's interaction neighbors.

Figure 6.



Black players follow s_2, Light players follow s_1

Each player plays the Figure 7 Assurance game with each of her interaction neighbors in the network.

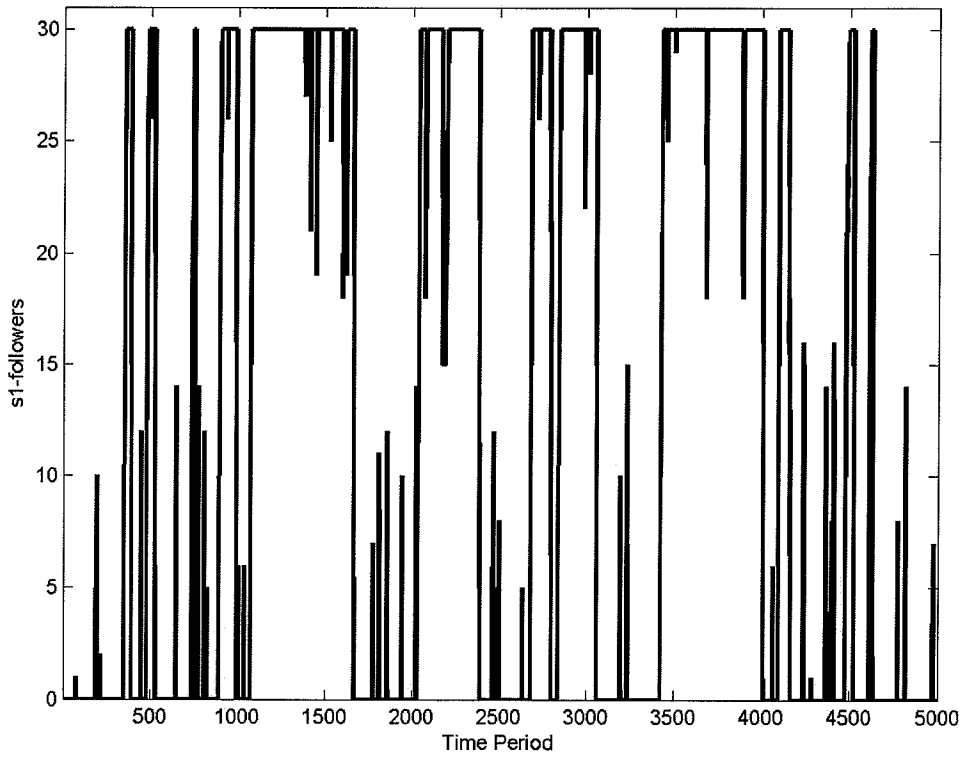
Figure 7. Assurance with (s_2, s_2) Risk dominant

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

In the Figure 7 game, the suboptimal (s_2, s_2) equilibrium is risk dominant. Consequently, the all-s_2 equilibrium is the unique stochastically stable equilibrium of the best-response dynamic over this network game. 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 generated a permanent foothold of s_1-followers in the network, even when the system was "bombarded" by a high mutation rate of 0.10 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 0.001, 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 Player i's interaction neighbors together with their interaction neighbors each followed s_i with probability $\lambda_i(t)$ sampled from the uniform distribution over $[0, 1]$. 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.

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

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.

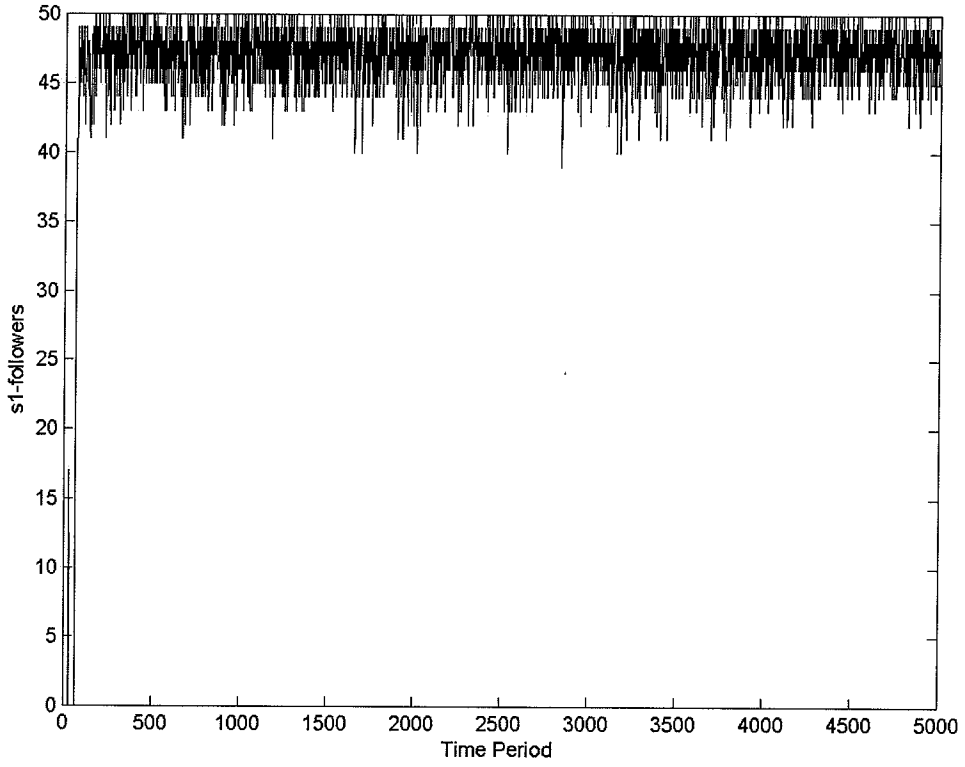
Example 4. Assurance Played on a Bounded Degree Network with Differential Influence Neighborhoods

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 interaction neighbors. As in the bounded degree network game of Example 3, in

this network game the all-s_2 equilibrium is the unique stochastically stable equilibrium of the best-response dynamic. As expected, we found in computer simulations that 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 0.10. We also found that the all-s_2 equilibrium is not robust against the introduction of influence neighborhoods of the sort we applied to the network game of Example 3.

We next tested the following variant of best-response dynamics: At each time period, independent mutants of s_2-followers appear with probability 0.1, and s_1-mutants appear spontaneously at a rate of 0.001. When an s_1-mutant appears spontaneously, her interaction neighbors together with their interaction neighbors each mutate with probability $\lambda_i(t)$ sampled from the uniform distribution over $[0, 1]$. We found that this dynamic always led the network game to converge 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. The results of one of our computer simulations over 5000 periods of updating are summarized in Figure 9.

Figure 9.



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

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 of course even when an s_1-following leader Player i appears at period t, the correlation in behavior over her influence neighborhood might be weak depending upon $\lambda_i(t)$. The reason the high influx of s_2-mutants cannot prevent the overthrow of the all-s_2 equilibrium or later destabilize the new all-s_1 equilibrium is precisely because the s_1-following mutations are correlated. s_1-following leader mutants appear seldom in the network game, but the coordinated play of players in their influence neighborhoods

enables the s_1-followers to conquer the network game and to suppress the high influx of s_2-mutants over time. ■

Example 4 shows that influence neighborhoods that appear at random in a network game can drive this game to an optimal equilibrium that is 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 mutants. That is, leader s_1-following mutants might have access to some communication channel they use to send messages to others in their influence neighborhoods, while s_2-mutants have no reliable means of communicating with others. So even though leader s_1-mutants appear seldom in the interaction network, their relatively effective ability to signal their plans to others enables those in their influence neighborhoods to coordinate their activity to a certain extent. On the other hand, even though s_2-mutants appear at a much higher rate than s_1-mutants, they are unable to communicate with others 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.

Our examples suggest 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 the agents of the regime can simultaneously coordinate their activity and prevent those under their jurisdiction from coordinating. This would explain why repressive regimes suppress communication. On the other hand, Hardin argues that if dissidents suddenly gain the ability to communicate and thereby coordinate while the regime's agents lose these abilities, the regime becomes vulnerable. One can interpret the network game and the influence neighborhood structure of Example 4 as follows: To follow s_2 is to obey a regime all dislike. Dissident s_1-

followers establish an underground broadcasting network that enables them to send messages to others and also to jam the attempts of new s_2-followers sent by the regime to communicate with others. Given these conditions, the s_1-followers stage a successful revolt.

§4. The Formal Model

We first review some basic notions of network games to establish notation. $N = \{1, ..., n\}$ is the set of *players*. ij denotes the subset $\{i, j\} \subseteq N$. Each $ij \subseteq N$, $i \neq j$, is an undirected *link* over N. $\mathcal{N}^N$ denotes the complete graph over N, that is, the set of all links over N. A subset $\mathcal{N} \subseteq \mathcal{N}^N$ defines an *interaction network*, or $\mathcal{N}$*-network*. If $ij \in \mathcal{N}$, then i and j are *interaction neighbors*, or $\mathcal{N}$*-neighbors*, and are said to be $\mathcal{N}$*-linked*. The set $\mathcal{N}_i = \{j \in N : ij \in \mathcal{N}\}$ is Player i's $\mathcal{N}$*-neighborhood*. $\mathcal{N}_i \neq \emptyset$ for each $i \in N$. This guarantees that each player interacts with at least one other player, that is, each player "gets to play". The *base game* Γ is a symmetric noncooperative 2-player game with pure strategy set S and payoff matrix $\boldsymbol{u}$:$S \times S \to \mathbb{R}^2$. With each of her $\mathcal{N}$-neighbors, Player i plays Γ and receives a payoff $u_i(s_{k_i}, s_{k_j}) = I_1 \circ \boldsymbol{u}(s_{k_i}, s_{k_j})$, where $I_1(\boldsymbol{x})$ projects $\boldsymbol{x} \in \mathbb{R}^2$ onto its 1st component, for each Player $j \in \mathcal{N}_i$. $\mathcal{N}_\Gamma = (\mathcal{N}, \Gamma)$ is the *network game* characterized by the network $\mathcal{N}$ and the base game Γ.

A *state* of a network game $\mathcal{N}_\Gamma$ is a vector $\boldsymbol{s} = (s_{k_1}, ..., s_{k_n})$, $s_{k_i} \in S$. Player i's *payoff* at state $\boldsymbol{s}$ is the sum of the payoffs Player i receives from playing with each of her $\mathcal{N}$-neighbors. A strategy s_{k_i} is a *best response* for Player i to $\boldsymbol{s}_{-i}$[14] if s_{k_i} maximizes Player i's payoff, that is,

$$\sum_{j \in \mathcal{N}_i} u_i(s_{k_i}, s_{k_j}) \geq \sum_{j \in \mathcal{N}_i} u_i(s'_{k_i}, s_{k_j}) \text{ for each } s'_{k_i} \in S\,. \tag{3.1}$$

$S_i^{\bigstar}(\boldsymbol{s}_{-i})$ denotes the set of Player i's best responses to $\boldsymbol{s}_{-i}$. A state $\boldsymbol{s}^* = (s^*_{k_1}, ..., s^*_{k_n})$ is a *Nash equilibrium* of $\mathcal{N}_\Gamma$ if

$$s^*_{k_i} \in S_i^{\bigstar}(\boldsymbol{s}^*_{-i}) \text{ for each } i \in N. \tag{3.2}$$

and $\boldsymbol{s}^*$ is *strict* if exactly one strategy satisfies (3.2) for each $i \in N$.

We need a few additional bits of notation to define the states of a network game when players interact with their $\mathcal{N}$-neighbors over time. At each period t, each Player i follows a strategy $s_i(t) \in S$. The state of the network game $\mathcal{N}_\Gamma$ at t is the vector $\boldsymbol{s}(t) = (s_1(t), ..., s_n(t))$. Clearly, $\boldsymbol{s}(t)$ is a Nash equilibrium if $\boldsymbol{s}(t) = \boldsymbol{s}^*$ satisfies (3.2).

In the sequel, 1_A is the *indicator* of a proposition $A \subseteq \Omega$, that is, $1_A = 1$ if A is the case and $1_A = 0$ otherwise. $\mu(A)$ denotes the probability that $A \subseteq \Omega$ obtains.[15]

Next we define the dynamics of our examples. Each dynamic presupposes an initial state $\boldsymbol{s}(0)$. $\boldsymbol{s}(0)$ is the boundary condition of the dynamic. The *(inductive) best-response-dynamic* (*BR-dynamic*) on a network game $\mathcal{N}_\Gamma$ is defined as follows: For each $i \in N$, let $f_i : 2^S - \emptyset \to S$ be a (possibly random) choice function.[16] At each period $t > 0$,

$$BR_i(t) = f_i\big(\{s_k \in S : s_k \in S_i^\bigstar(\boldsymbol{s}_{-k}(t-1))\}\big)\,. \tag{3.3}$$

In words, at period t each Player i adopts a strategy that is a best response to the strategies Player i's neighbors followed at period $t-1$. If $S_i^\bigstar$ contains more than one pure strategy, then Player i selects one of these best responses according to the choice function. Clearly, $BR_i(t) = BR_i(t-1)$ for each $i \in N$ only if $\boldsymbol{s}(t) = \boldsymbol{s}(t-1)$ is a Nash equilibrium of $\mathcal{N}_\Gamma$, that is, the fixed points of the *BR*-dynamic are Nash equilibria of the network game. Of course, the converse need not hold, for if $\boldsymbol{s}(t-1)$ is a Nash equilibrium but not strict, then at stage t some players might choose best responses other than their respective parts of $\boldsymbol{s}(t-1)$. 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 σ_i denote a completely mixed strategy[17] for Player i and $A^t_{\epsilon_1}, ..., A^t_{\epsilon_n}$ denote stochastically independent propositions such that $\mu\big(A^t_{\epsilon_i}\big) = \epsilon_i$. Then the *$\overline{BR}$-dynamic with independent random mutation* is defined by

$$\overline{BR_i}(t, \epsilon_i, \sigma_i) = \left(1 - 1_{A^t_{\epsilon_i}}\right) \cdot BR_i(t) + 1_{A^t_{\epsilon_i}} \cdot \sigma_i \,. \tag{3.4}$$

That is, at each stage t, Player i best responds with probability $1 - \epsilon_i$, and with probability ϵ_i chooses a pure strategy at random. ϵ_i is Player i's *mutation rate*. One may interpret a mutation σ_i as Player 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 $\mathcal{N}_\Gamma$, at each period t and for each $i \in N$ there is an associated probability distribution $\boldsymbol{\lambda}_i(t) = (\lambda_{i1}(t), ..., \lambda_{in}(t))$ over the players in N. Player i's *influence neighborhood* ($\mathcal{I}$*-neighborhood*) *at time* t is the set $\mathcal{I}_i(t) = \{j \in N : \lambda_{ji}(t) > 0\}$. The *size* of a $\mathcal{I}$-neighborhood is $|\mathcal{I}_i(t)|$, the number of players in $\mathcal{I}_i(t)$. If $j \in \mathcal{I}_i(t)$, $\lambda_{ji}(t)$ is Player i's *influence probability* over Player j. All of the influence neighborhoods over $\mathcal{N}_\Gamma$ at period t are specified by the matrix $\boldsymbol{\lambda}_N(t) = (\boldsymbol{\lambda}_1(t); ...; \boldsymbol{\lambda}_n(t))$. ■

The probabilities that characterize the influence neighborhoods can vary over time periods, while the graph that defines the local interaction structure of the network game remains fixed. The underlying intuition here is that changing one's interaction neighbors is prohibitively costly, but cost-free communication with nearby players might at times be possible. So the players' interaction subnetworks, the $\mathcal{N}$-neighborhoods, remain fixed. But their *communication subnetworks,* the $\mathcal{I}$-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 $\mathcal{I}$-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 $A_{i1}^t, ..., A_{in}^t$ be mutually exclusive propositions such that $\mu(A_{ij}^t) = \lambda_{ij}(t)$. Then the *BR*-dynamic with influence neighborhoods* $(\boldsymbol{\lambda}_N(t))$ is defined as follows: For $i \in N$,

$$BR_i^*(t, \boldsymbol{\lambda}_i(t)) = \sum_{j \in N} s_j(t) 1_{A_{ij}^t} \tag{3.5}$$

where $s_i(t) = \overline{BR_i}(t, \epsilon_i, \sigma_i)$. ■

In words, Player i imitates the strategy of Player j with probability $\lambda_{ij}(t)$ if Player i falls in Player j's influence neighborhood. Frequently, if not predominantly, a Player i will not fall in any other player's influence neighborhood. In this case, $\boldsymbol{\lambda}_i(t)$ is characterized by $\lambda_{ii}(t) = 1$, and Player i simply follows the best-response dynamic with independent random mutations. If at period t we have $\lambda_{ii}(t) = 1$ for all $i \in N$, then every influence neighborhood is a singleton, that is, $\mathcal{I}_i(t) = \{i\}$ for each $i \in N$. Intuitively, in this case no player communicates with any other player. Also for this case, at this period the *BR**-dynamic with influence neighborhoods reduces to the $\overline{BR}$-dynamic with independent random mutations. At another extreme, if a Player i has a number of other players falling in her $\mathcal{I}$-neighborhood at period t and $\lambda_{ji}(t) = 1$ for each $j \in \mathcal{I}_i(t)$, then all of the players in Player j's influence neighborhood are certain to correlate their strategies with Player j at period t. One might think of this case as a "perfectly disciplined" influence neighborhood whose members all follow the command of their leader.

In Examples 2, 3, and 4, a leader Player i with an influence neighborhood $\mathcal{I}_i(t) \neq \{i\}$ appears at random in the network game, and this leader'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 Player $j \in \mathcal{I}_i(t)$. In these examples, a leader's influence over his $\mathcal{I}$-neighbors 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 t and all $j \in N$, then the entire network is Player i's perfectly disciplined influence neighborhood. In this special case, $\mathcal{I}_i(t) = N$, $\mathcal{I}_j(t) = \emptyset$ for $i \neq j$, and Player i plays a role analogous to Hobbes' absolute sovereign. As noted above, perfect discipline is an extreme case. One would expect that in many actual situations, a leader's "clout" varies over those falling within his sphere of influence. Varying influence probabilities over the $\mathcal{I}$-neighborhood reflect a leader's uneven sway over those who receive his messages. And in the real world, actual influence over others at particular times is likely to be neither completely fixed nor purely random.

When correlated mutations are possible, a variety of long term outcomes can emerge in a network game, depending upon the payoff structure of the base game, the configurations of the interaction network and the influence neighborhoods. If influence neighborhoods remain fixed and perfectly disciplined over a stretch of time periods, then the network can remain at a polymorphism of strategies where those in the influence neighborhoods follow their leaders and the rest follow 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 that is the analog of stochastic stability when all mutations are stochastically independent.

On the other hand, 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 sometimes greatly accelerate the convergence of the

network to the strategy of the risk dominant equilibrium in the base game. This example and Example 4 suggest the following

Definition. A state $\boldsymbol{s}^* = (s_1^*, ..., s_n^*)$ of a network game $\mathcal{N}_\Gamma$ is an *attractor of the BR*-dynamic with influence neighborhoods* $(\boldsymbol{\lambda}_N(t))$ if for some state $\boldsymbol{s}' \neq \boldsymbol{s}^*$, when $\boldsymbol{s}(0) = \boldsymbol{s}'$ then $BR_i^*(t, \boldsymbol{\lambda}_i(t))$ in (3.5) is such that

$$\mu\Big(\lim_{t\to\infty} s_i(t) = \Big(1 - 1_{A_{\epsilon_i}^t}\Big)s_i^* + \sigma_i 1_{A_{\epsilon_i}^t}\Big) = 1 \text{ for each } i \in N. \tag{3.6}$$

If (3.6) is satisfied for every state $\boldsymbol{s} \neq \boldsymbol{s}^*$, then $\boldsymbol{s}^*$ is the *global attractor of the BR*-dynamic.* ■

This definition says that $\boldsymbol{s}^*$ is an attractor of the *BR**-dynamic with the $\mathcal{I}$-neighborhoods $(\boldsymbol{\lambda}_N(t))$ if, with probability one, from $\boldsymbol{s}(0)$ players who update according to this dynamic all eventually follow $\boldsymbol{s}^*$ except when they mutate, either spontaneously or by imitating the leader of a $\mathcal{I}$-neighborhood. In Example 2, $(s_1, ..., s_1)$ is the global attractor of the *BR**-dynamic where each $\mathcal{I}$-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 $\mathcal{I}$-neighborhood of s_2-followers that appears and is then eliminated. We have the following elementary result:

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

PROOF. By hypothesis, given some $\boldsymbol{s}(0) = \boldsymbol{s} \neq \boldsymbol{s}^*$ this *BR**-dynamic satisfies (3.6). Hence as $t \to \infty$, with probability one each Player $i \in N$ follows s_i^* unless Player i mutates spontaneously or imitates the strategy of a leader in case Player i falls in this leader's $\mathcal{I}$-neighborhood. But then s_i^* must be a best response for each Player $i \in N$ under the unperturbed *BR*-dynamic, and so (3.2) is satisfied. □

Example 2 shows that a system of *BR**-updaters can converge to an optimal Nash equilibrium even when the $\mathcal{I}$-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 $\mathcal{N}_\Gamma$, a set $B \subseteq N$ is *BR-stable* with respect to $s \in S$ if, given $s_i(t) = s$ for each $i \in B$, $BR_i(t+1) = s$ for each $i \in B$. If B is *BR*-stable with respect to s, we say that s is *BR-stable over* B. ■

Intuitively, a set B of the players is *BR*-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 $\mathcal{N}_\Gamma$ do not follow s.

In the following proposition, we show that if the influence neighborhoods of a *BR**-dynamic introduce *BR*-stable sets, this dynamic can converge to a Nash equilibrium from any initial state no matter how infrequently these influence neighborhoods appear.

Proposition 2. Let the network game $\mathcal{N}_\Gamma$ be given. Let $\mathcal{I}$-neighborhoods of bounded size $b < n$ appear in $\mathcal{N}_\Gamma$ with probability $\mu(\mathcal{I}_i(t)) = \epsilon$ where for $\mathcal{I}_i(t)$, $\mu(s_i(t) = s_k)$ is uniformly distributed across pure strategies. Let $\mu(|\mathcal{I}_i(t)| = b) \geq q > 0$ and $\mu(s_j(t) = s_k \text{ for all } j \in \mathcal{I}_i(t)) \geq p > 0$ for each $\mathcal{I}_i(t)$ that appears in $\mathcal{N}_\Gamma$. If for each $\mathcal{I}_i(t)$ of maximum size b, s^* is the unique *BR*-stable strategy of some subset $B_i \subseteq \mathcal{I}_i(t)$, then $\boldsymbol{s^*} = (s^*, ..., s^*)$ is the global attractor of this *BR**-dynamic.

PROOF. Let $(\mathcal{I}_u)$ denote the sequence of $\mathcal{I}$-neighborhoods that appears in the network lexically ordered according to time periods of play. With probability one, a perfectly disciplined $\mathcal{I}$-neighborhood of maximum size b whose players follow s^* appears infinitely often in the sequence of plays. Let $(\mathcal{I}_{u_k})$ denote the subsequence of $(\mathcal{I}_u)$ such that $\mathcal{I}_{u_k}$ is of size b and each $i \in \mathcal{I}_{u_k}$ follows s^*, and let (B_{u_k}) denote the sequence of *BR*-stable sets of s^*-followers that appear in $\mathcal{N}_\Gamma$ as a result. We claim that s^* satisfies (3.6), that is, s^* overtakes the network game with probability one. For 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^*. So 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" $\mathcal{I}$-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 $(\mathcal{A}_{u_k})$ of $\mathcal{I}$-neighborhoods synchronized with the $\mathcal{I}_{u_k}$'s must appear in $\mathcal{N}_\Gamma$ such that all but a fixed number of the $\mathcal{A}_{u_k}$'s satisfy the following properties: (*i*) the leader Player i_{u_k} of each $\mathcal{A}_{u_k}$ follows some strategy other than s^*, (*ii*) Player i_{u_k} appears in a part of the network where $\mathcal{A}_{u_k}$ overlaps $\mathcal{I}_{u_k}$, and (*iii*) enough players in $\mathcal{A}_{u_k}$ imitate Player 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 $\mathcal{A}_{u_k}$'s and this subsequence then overtakes the whole network.) But if $\mathcal{A}'_{u_k}$ denotes the proposition that for a given B_{u_k} a matching $\mathcal{A}_{u_k}$ appears satisfying (*i*), (*ii*) and (*iii*), then $\mu(\mathcal{A}'_{u_k})$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 $\mathcal{I}$-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 $\mathcal{A}_{u_k}$ overlaps perfectly with B_{u_k} and then sufficiently many players in $\mathcal{A}_{u_k}$ imitate Player i_{u_k} to destabilize the s^* strategy in B_{u_k}. So the η_{u_k}'s are bounded from above by some $\eta < 1$. Hence if $\mathcal{A}'$ denotes the event that the necessary sequence of $\mathcal{A}_{u_k}$'s appears to contain the B_{u_k}'s, then

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

The key idea behind Proposition 2 is that *BR*-stable sets of s^*-followers appear in the network game and 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 is quite different 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 so that the influence neighborhoods of any particular strategy are more likely to appear or to persist over time periods. Finally, note that the proof of Proposition 2 does not depend upon specific values of ϵ, 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, so 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. An interaction network $\mathcal{N}$ is *c-uniformly linked* if each $i \in N$, $|\mathcal{N}_i| = c$, that is, each Player i is $\mathcal{N}$-linked with exactly c other players.

Corollary 3. Let $\mathcal{N}_\Gamma$ be such that $\mathcal{N}$ is c-uniformly linked and (s^*, s^*) is the risk dominant equilibrium of Γ. Let $\mathcal{I}$-neighborhoods of bounded size $b < n$ appear in $\mathcal{N}_\Gamma$ where $\mu(\mathcal{I}_i(t)) = \epsilon$ and where for $\mathcal{I}_i(t)$, $\mu(s_i(t) = s_k) = \frac{1}{|S|}$. Let $\mu(|\mathcal{I}_i(t)| = b) \geq q > 0$ and $\mu(s_j(t) = s_k \text{ for all } j \in \mathcal{I}_i(t)) \geq p > 0$ for each $\mathcal{I}_i(t)$. If for each $\mathcal{I}_i(t)$ where $|\mathcal{I}_i(t)| = b$, a nonempty subset $B_i \subseteq \mathcal{I}_i(t)$ is such that each Player i is $\mathcal{N}$-linked with at least $\frac{c}{2}$ players in $\mathcal{I}_i(t)$, then $\boldsymbol{s}^* = (s^*, ..., s^*)$ is the global attractor of this *BR**-dynamic.

PROOF. As in the proof of Proposition 2, let $(\mathcal{I}_u)$ denote the sequence of $\mathcal{I}$-neighborhoods that appears in the network lexically ordered according to time periods. With probability one, a subsequence $(\mathcal{I}_{u_k})$ where $\mathcal{I}_{u_k}$ is of size b and each $i \in \mathcal{I}_{u_k}$ follows s^* appears in the sequence of plays. By hypothesis, each $\mathcal{I}_{u_k}$ contains a nonempty subset B_{u_k} whose member nodes are each $\mathcal{N}$-linked with at least $\frac{c}{2}$ players in $\mathcal{I}_{u_k}$. Since (s^*, s^*) is risk dominant, by (1.1) s^* is the unique best response for each Player $i \in B_{u_k}$ at subsequent time periods because at least half of Player i's $\mathcal{N}$-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. □

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 1-dimensional circular network games analyzed by Ellison (1993) and the 2-dimensional lattice network game of Examples 2 and 3. Moreover, we can now explain why random mutations failed to overthrow the suboptimal $(s_2, ..., s_2)$ 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 $(s_1, ..., s_1)$ equilibrium. Even though $(s_1, ..., s_1)$ is the unique stochastically stable equilibrium of this network game, over a million generations no set of the singleton $\mathcal{I}$-neighborhoods of s_1-followers introduced by independent random mutation appeared together in a *BR*-stable configuration. While they appeared at a high rate, the s_1-following mutants were relatively isolated from each other and consequently were not *BR*-stable with respect to s_1. Hence, 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 the high 0.1 rate.

In Example 2, even though $\mathcal{I}$-neighborhoods appeared at a rate of only 0.001, some of the $\mathcal{I}$-neighborhoods of s_1-followers that appeared introduced *BR*-stable sets.

Each $\mathcal{I}$-neighborhood was a Moore-24 neighborhood, and since the network game was 8-uniformly linked, all but the four "corner" players of a given $\mathcal{I}$-neighborhood were $\mathcal{N}$-linked with at least four players in the same $\mathcal{I}$-neighborhood. So when a perfectly disciplined $\mathcal{I}$-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 clearly do not generalize 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 non-risk 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 correlation in mutations can profoundly impact 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 payoff structure alone determines the long term limits of dynamical updating. If so, then of course network structure plays no role in determining the state to which the players ultimately converge, though it certainly influences how quickly they approach that state. But we believe it would be a mistake to draw this moral from the stochastic stability literature. 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 in the network game. 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 neighborhoods 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 optimal 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 actual human communities. Communities of people do occasionally reform their practices, and the process does not typically occur as the result of independent aberrations in the behavior of individuals 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. Independent random mutation cannot adequately model this kind of coordination. But influence neighborhoods are tailor made for modeling this coordination. And an optimal equilibrium that independent mutation never produces over a million 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 in which the interaction network itself evolves over time (Skyrms and Pemantle 2000, Goyal and Vega-Redondo 2000, Watts 2001, Jackson and Watts 2001*a*, 2001*b*). Combining different learning rules and evolving network structures with influence neighborhood mutation may produce a theory of network games that has much greater explanatory power than the existing theory.

REFERENCES

Alexander, Jason and Skyrms, Brian. 1999. 'Bargaining with Neighbors: Is Justice Contagious?' *Journal of Philosophy* 96: 588-598.

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.

Danielson, Peter. 1992. *Artificial Morality.* London: Routledge.

Ellison, Glenn. 1993. 'Learning, Local Interaction and Coordination.' *Econometrica,* 61, 1047-1071.

Ellison, Glenn. 2000. 'Basins of Attraction and Long Run Equilibria.' *Review of Economic Studies* 67, 17-45.

Foster, Dean and Young, H. Peyton. 1990. 'Stochastic evolutionary dynamics.' *Journal of Theoretical Biology*, 38, 219-232.

Fudenberg, Drew and Levine, David. 1998. *The Theory of Learning in Games.* Cambridge, Massachusetts: MIT Press.

Goyal, Sanjeev. 2002. 'Learning in Networks.' Unpublished manuscript.

Goyal, Sanjeev and Vega-Redondo, Fernando. 2000. 'Learning, Network Formation and Coordination.' Unpublished manuscript.

Grim, Patrick, Mar, Gary and St. Denis, Paul. 1998. *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling.* Cambridge: 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: Princeton University Press.

Harsanyi, John and Selten, Reinhard. 1988. *A General Theory of Equilibrium Selection in Games.* Cambridge, Massachusetts: MIT Press.

Hume, David. (1740, 1888) 1976. *A Treatise of Human Nature*, ed. L. A. Selby-Bigge. rev. 2nd. ed., ed. P. H. Nidditch. Oxford: Clarendon Press.

Jackson, Matthew and Watts, Alison. 2001*a*. 'The Evolution of Social and Economic Networks.' Unpublished manuscript.

Jackson, Matthew and Watts, Alison. 2001*b*. 'On the Formation of Interaction Networks in Social Coordination Networks.' Unpublished manuscript.

Jiborn, Magnus. 1999. *Voluntary Coercion*. Lund: Lund University.

Kandori, Michihiro, Mailath, George and Rob, Rafael. 1993. 'Learning, Mutation, and Long-Run Equilibria in Games.' *Econometrica* 61: 29-56.

Kavka, Gregory. 1986. *Hobbesian Moral and Political Theory*. Princeton: Princeton University Press.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge, Massachusetts: Harvard University Press.

Morris, Stephen. 2000. 'Contagion.' *Review of Economics Studies* 67, 57-78.

Nash, J. 1950. 'Equilibrium points in n-person games.' *Proceedings of the National Academy of Sciences of the United States* 36: 48-49.

Nash, John. 1951*a*. 'Non-Cooperative Games.' *Annals of Mathematics* 54: 286-295.

Nash, John. (1951*b*) 1996. 'Appendix: Motivation and Interpretation'. Reprinted in *Essays on Game Theory* by John Nash. Cheltenham, United Kingdom: Edward Elgar. 32-33.

Nowak, M. A. and May, R. M. 1992. 'Evolutionary Games and Spatial Chaos.' *Nature* 359, 826-829.

Nowak, M. A., Bonhoeffer, S. and May, R. M. 1994. 'Spatial Games and the Maintenance of Cooperation.' *Proceedings of the National Academy of Sciences of the USA*, 91, 4877-4881.

Sen, Amartya. 1967. 'Isolation, Assurance and the Social Rate of Discount.' *Quarterly Journal of Economics*, 81, 112-124.

Skyrms, Brian. 2002. 'The Stag Hunt.' *Proceedings and Addresses of the American Philosophical Association*, 75, 31-41.

Skyrms, Brian. forthcoming. *The Stag Hunt: Evolution of Social Structure.* Cambridge University Press.

Skyrms, Brian and Pemantle, Robin. 2000. 'A Dynamic Model of Social Network Formation.' *Proceedings of the National Academy of Sciences of the USA* 97: 9340-9346.

Sugden, Robert. 1986. *The Economics of Rights, Co-operation and Welfare.* Oxford: Basil Blackwell, Inc.

Vanderschraaf, Peter. 1998. "The Informal Game Theory in Hume's Account of Convention" *Economics and Philosophy* 14, 251-247.

Taylor, Michael and Ward, Hugh. 1982. 'Chickens, Whales and Lumpy Goods.' *Political Studies* 20, 350-370.

Taylor, Michael. 1987. *The Possibility of Cooperation.* Cambridge: Cambridge University Press.

Von Neumann, John and Morgenstern, Oskar. 1944. *Theory of Games and Economic Behavior*, Princeton: 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.

Young, H. Peyton. 1998. *Individual Strategy and Social Structure.* Princeton: Princeton University Press.

---

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

[2]Goyal (2002) surveys much of the progress achieved in the dynamical analysis of network games so far. Interestingly, some of the most original contributions in this new

literature come from philosophers (Danielson 1992, Grim, Mar and St. Denis 1998, Alexander and Skyrms 1999, Skyrms 2002, forthcoming) and computer scientists (Nowak and May 1992, Nowak, Bonhoeffer and May 1994).

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

[4]Following standard conventions, Player 1's (Player 2'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 Player i chooses s_1 and Player j chooses s_2, then Player i's payoff is 0 and player j's payoff is y.

Many authors have argued that versions of the Figure 1 game capture the logic of Rousseau's celebrated example of the stag hunt, given in Part II, paragraphs 8 and 9 of *Discourse on the Origin and Foundations of Inequality in Men.* The Assurance game gets its name from Sen (1967), the first author we know of who analyzed this game for $y > z$.

[5]See especially Taylor and Ward (1982), Kavka (1986), Hampton (1986), Taylor (1987), Jiborn (1999) and Skyrms (2001, forthcoming).

[6]David Lewis (1969) 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 1969, pp. 56-57). 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.*

[7]To motivate this assumption, which is 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 of her neighbors follows any particular strategy, so that she must adopt a single strategy for interacting with them all.

[8]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.

[9]For extended discussion of Hume's informal game-theoretic insights, see Lewis (1969), Sugden (1986) and Vanderschraaf (1998).

[10]See, for instance, Nowak and May (1992), Nowak, Bonhoeffer and May (1994) and Grim, Mar and St. Denis (1998).

[11]All of the simulation experiments summarized in this paper were run using *The Evolutionary Modeling Lab* developed by Alexander.

[12]The pseudorandom number generator that *The Evolutionary Modeling Lab* employs to introduce the mutations in our simulation experiments implements the Mersenne twister algorithm known as MT19937, which has a provable period of $2^{19937} - 1$.

[13]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 0.0 so that "leader" players who might be "followed" by some of their Moore-24 neighbors received no additional "help" from independent random mutants.

[14]The subscript '$-i$' indicates the result of removing the ith component of an n-tuple or an n-fold Cartesian product. In particular,

$$\boldsymbol{s}_{-i} = (s_1, ..., s_{i-1}, s_{i+1}, ..., s_n)$$

denotes the $n-1$-tuple of pure strategies that Player i's opponents follow when they all follow the state $\boldsymbol{s} = (s_1, ..., s_n)$.

[15]In much of the game-theoretic literature, probabilities are players' subjective beliefs, which need not agree over all propositions. In our model, the relevant probabilities are probabilities of mutations, which may be viewed from the perspective of

any player or an external observer. Hence in our model it makes sense to use a single probability measure rather than an entire system of subjective probability measures.

[16] $2^B - \emptyset$ denotes the set of all nonempty subsets of a nonempty set B. For a finite set $B = \{x_1, ..., x_n\}$, a choice function $f : 2^B - \emptyset \rightarrow B$ takes as an argument any nonempty subset $\{x_{k_1}, ..., x_{k_m}\} \subseteq B$ and returns $f(\{x_{k_1}, ..., x_{k_m}\}) = x_{k_j} \in \{x_{k_1}, ..., x_{k_m}\}$, a single element of $\{x_{k_1}, ..., x_{k_m}\}$. In particular, for a singleton $\{x\} \subseteq B$, $f(\{x\}) = x$.

[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 1951*a*).