Working Paper

Social Coordination and Social Change

*H. Peyton Young**

WP-96-32
June 1996

IIASA
International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 807 □ Fax: +43 2236 71313 □ E-Mail: info@iiasa.ac.at

Social Coordination and Social Change

*H. Peyton Young**

WP-96-32
June 1996

*Department of Economics, Johns Hopkins University
Baltimore, MD 21218-2685
and
International Institute for Applied Systems Analysis
Laxenburg, Austria

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 807 □ Fax: +43 2236 71313 □ E-Mail: info@iiasa.ac.at

Abstract

Social and economic institutions govern how people interact with each other -- they define the "rules of the game." Choosing the rules is at bottom a pure coordination problem, since people must agree on the rules in order to play. We posit that these rules evolve endogenously through the repeated interactions of individuals. They choose best replies to their environment subject to some inertia and error. Over the long run, such a process selects institutions (rules) that are efficient, and fair in the sense that the expected payoffs are centrally located on the Pareto frontier of the payoff possibility set.

Social Coordination and Social Change[1]

1. Games and social norms

Games govern how people interact with one another. They provide the framework of rules that constrain what agents do and what they expect others to do. When everyone knows what the game is, game theory tells us how rational people will play it and what the outcome will be. But how do people come to know what the game is? Most interactions are too complex to be understood strategically if we allow for all possible strategies the agents might use. Instead, interactions tend to follow stylized patterns that the players have come to expect in particular situations. It is this delimited and constrained form of interaction that we call a "game." A game is a *gestalt* that frames the strategic situation for the players, and allows them to coordinate their behaviors and expectations.

Consider, for example, the ways in which people can write a contract. In practice they do not negotiate over all possible forms of contracts; they take a standard contract for that situation and negotiate how to fill in the blanks. Real estate sales contracts have one format, construction contracts another, manufacturing employment contracts a third. Their terms depend not only on the economic activity at hand, but on the time and place -- on the social context. Over time, contracts evolve into a specific format through the cumulative experience of many individuals trying out different forms.

A second, considerably fuzzier, example, concerns the operation of households. The roles of men and women, children and adults, are governed by unwritten rules and common understandings in any society. There are, of course, many variations among households in such particulars as who does the dishes, who goes shopping, and who weeds the garden. But these are details that are negotiated within the broader conventions and understandings of the way in which households are supposed to operate within a given culture. These common understandings define the spheres within which the various parties have primary responsibilities in rearing children, earning income, caring for the

[1] I am indebted to Kenneth Arrow for suggesting this general line of inquiry, and to John Roemer and Mancur Olson for constructive comments on an earlier draft.

elderly, and so forth. They define the implicit game within which the relationships are played out. Though the rules of the game are quite complex, and one would be hard pressed to write them down explicitly, most people within a given culture know how to play the household game.

My point is that every sphere of interaction has its customary rules -- the rules of the game. There is no one way to structure an interaction as a game, any more than there is one way to run a household or one way to write a contract. Rather, the rules of the game need to be understood as social conventions that are the product of social evolution (North, 1981; Sugden, 1986; Binmore, 1994; Arrow, 1994). The question is whether some games tend to be favored by the evolutionary process. In this paper I analyze this issue using the methodology of evolutionary game theory (Foster and Young, 1990, Kandori, Mailath, and Rob, 1993, and Young, 1993a).

We start with the idea that alternative ways of structuring an interaction can be represented as a finite class of different games. All parties to a transaction must agree on what the rules of the game are, otherwise they cannot play. This amounts to solving a pure coordination problem. Over time, alternative conceptions of which game is *relevant* compete for acceptance in people's heads. Once everyone agrees on the relevant game for a given sphere of interaction, it is as if a species had successfully invaded a social niche. An institutional form has evolved that solves a social coordination problem.

Evolution does not stop there however: new ideas, like mutant species, keep coming along and eventually displace established ones. What is the source of these new ideas? One force for change comes from philosophers, preachers, social theorists, and assorted "radicals" who argue that the world should operate differently than it now does. If they are persuasive and find many converts, they may cause established convention to change. A second, more subtle, and probably more important source of new ideas is individual variation. People do not always follow established convention; sometimes they do things differently. If there are enough of them, and they push in the same direction, they can tip society into a new convention even if their actions are not consciously coordinated (Schelling, 1971, 1978).

We show that these forces select games (institutional forms) that are *efficient*: there is no other way of structuring the interaction so that all parties get higher expected payoffs. Second, it favors games that are centrally located on the Pareto frontier of the feasible payoff set, instead of near the boundaries. At a purely technical level this can be regarded as a central tendency theorem for a class of stochastic processes. Interpreted in the context of economic and social institutions, it says that the most stable institutional arrangements are those in which all sides enjoy substantial gains from cooperation within the set of feasible payoff opportunities.

2. A model of social coordination

Let society consist of n disjoint populations or *classes* of individuals $C_1, C_2, \ldots.,$ C_n whose members interact from time to time. The structure of their interaction can be represented by a finite set of n-person games $G_1, G_2, \ldots, G_m$, each of which embodies a different set of "rules" that have payoff implications for the participants. To keep the model simple, we shall assume that each way of specifying the rules leads to a unique equilibrium, and that each member of the population correctly anticipates the expected payoff to himself in that equilibrium.

At the beginning of each period, people are matched in groups, where a group consists of one person from each of the n social classes. A strategy is to name a game. If everyone in a given group names the *same* game i, they play it in the current period. If they fail to name the same game, they do not play (they are unattached for the current period). This assumption reflects the idea that social interactions are purely voluntary, and that there is no social decision rule to fall back on if they fail to coordinate. (Any such rule would have to be justified as a social institution that is also subject to evolution.)

Altogether, therefore, there are m + 1 "states" that an individual can occupy at any given time: play one of the m games, or be unattached. Let a_{ij} be the *expected utility* to a member of class j from playing game i ($1 \leq i \leq k$) and let a_{0j} be the utility from being unattached. For simplicity we shall assume that all games are worth playing, that is, $a_{ij} > a_{0j}$ for all j and all $i \neq 0$. There is no loss of generality in normalizing each person's utility function so that for every class j, $a_{0j} = 0$ and

$\max_i a_{ij} = 1$. We shall assume throughout that the utilities have been fixed in this fashion.

Over time, then, society is engaged in a meta-game that has the structure of a pure coordination game. Indeed, our results apply to *any* pure coordination game, whether or not it is motivated in this way. There are, however, two reasons for thinking of the coordination problem as one of choosing the rules of the game. First, choosing the rules is an especially natural example of a situation where *all* parties must agree if there are to be gains from cooperation. If they fail to agree, they simply do not play. (In particular, they do not have to *decide* whether the meta-game is a pure coordination game; it has this structure by default.) Second, the development of rules governing social and interactions is a long-term process in which evolutionary forces surely play a major role. Moreover, the time scale is so long, and the number of people is so large, that no one individual can expect to have much effect on the course of events. In such a setting, myopic optimization is a reasonable assumption about how agents respond to their environment.

Consider then a coordination game with payoffs

$$\begin{bmatrix} (a_{11}, \ldots, a_{1n}) & & & \\ & (a_{21}, \ldots, a_{2n}) & & \\ & & \cdot & 0, \ldots, 0 \\ & 0, \ldots, 0 & \cdot & \\ & & \cdot & \\ & & & (a_{m1}, \ldots, a_{mn}) \end{bmatrix}$$

Let each class consist of k persons, where k is a positive integer. The process evolves in discrete time intervals $t = 1, 2, \ldots$. The *state* at the end of period t is an m x n matrix $x^t = (x^t_{.1}, x^t_{.2}, \ldots, x^t_{.n})$, where each column vector $x^t{.j} = (x^t_{1j}, x^t_{2j}, \ldots, x^t_{mj})$ lists the number of agents in class j who proposed playing each of the m games in period t.

A *complete matching* is a collection of k elements from the product set $\prod C_i$ such that each member of the population occurs in one and only one element. Let M be the set of complete matchings. At the beginning of period t + 1, everyone in the population is matched by a random draw from M according to some conditional probability distribution $\mu(\cdot \mid x^t)$, which depends on the state but not on the time period (except insofar as the time period determines the state). The probability of a matching reflects the geographical and social proximity of individuals, which affects their probability of meeting. Unlike some models in the literature, however, we do not assume that individuals interact exclusively with their neighbors (Ellison, 1993, Blume, 1993, An and Kiefer, 1993). Instead, we posit that all matchings occur with positive probability.

Once a matching has been chosen, each person in each matched group names a game. If all n members of the group name the *same* game i, they play it in period t + 1 and receive expected payoffs $a_{i1}, a_{i2}, \ldots, a_{in}$. If the members of a given group fail to name the same game (i.e., they cannot agree on the rules of the game), they are unattached for the period and their payoffs are zero. An individual in class j decides which game to name by consulting the frequency distribution of demands x^t that were made in the previous period. From these he infers the probability that different games will be named by the people against whom he is matched now. In particular, people do not condition their demands on the identities of others; each individual is viewed as an anonymous representative of his or her class, and past behavior is taken as a predictor of present behavior. These assumptions obviously sacrifice some degree of realism, but still have considerable justification in a large-population setting.

Each representative agent from class j estimates, therefore, that the probability of everyone else in his group naming game i is $\prod_{j' \neq j} (x^t_{ij'}/k)$. The best reply is to choose a game i that maximizes $a_{ij} \prod_{j' \neq j} (x^t_{ij'}/k)$. If there are ties in the maximum, the j-agent chooses among them according to some probability distribution π_j that has full support. These tie-breaking rules are fixed throughout and will not be listed explicitly as parameters of the system. Individuals can deviate from best reply in two ways: through inertia and random error (which we can also interpret as experimentation). Let $\nu \in (0, 1)$ be an inertia probability and let $\varepsilon \in [0, 1)$ be an error probability. In each period, each individual sticks to his previous choice with probability ν, chooses a best reply with probability $(1 - \nu)(1 - \varepsilon)$, and chooses a game at random with probability $(1 - \nu)\varepsilon$.

These rules define a Markov process $P^{k,v,\varepsilon}$ on the finite state space X. When ε is positive, there is a positive probability of moving from any state x to any other state x' in one period, because everyone could make an error (i.e., choose a strategy at random). Hence the process is irreducible. It is a standard result that a finite, stationary, irreducible Markov process has a unique stationary distribution $\mu^{k,v,\varepsilon}(x)$, which represents the relative frequency with which the process visits state x during the first T periods as T becomes arbitrarily large.

A *social norm* is a state in which all matched players name (and play) the same game G_i. Such a state has the form $z_i = (ke_i, ke_i, \ldots, ke_i)$, where $e_i \in R^m$ is the unit column vector with 1 in position i and 0's elsewhere. When the probability of making errors is zero, every norm is an absorbing state. We claim that, in fact, the norms are the *only* absorbing states when $\varepsilon = 0$. Suppose indeed that x is absorbing. If i is a game such that $x_{ij} > 0$ for some j, then i must be a strict best reply by each j-player to the frequency distribution implied by x, for otherwise there is a positive probability that j will not play i in the next period (recall that the tie-breaking rule has full support). By assumption, the state in the next period is x with probability one. It follows that everyone in class j named i in x, that is, $x_{ij} = k$. Since the choice of i by class j must be a best reply to the choice of every other class, all classes must be choosing i, which means that $x = z_i$.

It can be shown that, when the disturbance term ε is small but positive, the stationary distribution $\mu^{k,v,\varepsilon}(\cdot)$ puts almost all of the probability on one or more norms z_i. (This follows from the proof of Theorem 1 given in the Appendix.) To be precise, there is a unique, nonempty set of norms $Z^* \subseteq \{z_1, z_2, \ldots, z_m\}$ such that

$$\lim_{\varepsilon \to 0} \mu^{k,v,\varepsilon}(x) > 0 \text{ if and only if } x \in Z^*. \tag{1}$$

Equivalently, Z^* is the minimal set of states such that, given any $p < 1$ there exists an ε_p such that $\sum_{x \in Z^*} \mu^{k,v,\varepsilon}(x) \geq p$ whenever $0 < \varepsilon \leq \varepsilon_p$. The states in Z^* are said to be *stochastically stable* (Foster and Young, 1990; see also Kandori, Mailath, and Rob, 1993, and Young, 1993a). The process is very likely to be in one of the stochastically stable states when the perturbation probability ε is small. In concrete terms, this says that when conventional forms of social interaction are occasionally but persistently challenged by innovations, the stable one(s) are much more likely to be observed than the others over the long run.

Of course, since the state space X is determined by the class size k, it is conceivable that stochastic stability also depends on k. Fortunately this is not an issue when k is large: for generic coordination games there is a unique coordination equilibrium i such that the corresponding norm z_i is stochastically stable for all sufficiently large k. As we show in the proof of theorem 1, this equilibrium is characterized as the unique minimum of a certain potential function. Even in nongeneric coordination games, the same potential function characterizes those norms that are stochastically stable for an infinite number of values of k. To simplify the statement of results, we shall therefore say that the ith coordination equilibrium of a pure coordination game is *stable*, and that the evolutionary process *selects* that equilibrium, if for every $\nu \in (0, 1)$ there are infinitely many values of k such that $\lim_{\varepsilon \to 0^+} \mu^{k,\nu,\varepsilon}(z_i) > 0$. For notational simplicity we shall henceforth suppress k and ν, and simply write μ^ε for the stationary distribution.

3. Welfare analysis.

By assumption, $0 < a_{ij} \leq 1$ is the expected payoff to each player from class j in coordination equilibrium i (which in keeping with our earlier interpretation we shall call "game i"). On the one hand, we may interpret a_{ij} as the expected gains from playing game i compared to being unattached, which has utility zero. On the other hand, we may think of $1 - a_{ij}$ as the level of dissatisfaction of class j, that is, the difference between what it gets in game i and what it could get from changing the rules of the game. Let us therefore say that a_{ij} is the *level of satisfaction* of class j in game i. The *satisfaction index* of game i is the level of satisfaction of the least satisfied class:

$$w_i = \min_j a_{ij}. \qquad (2)$$

Let $w^+ = \max_i w_i$. A *maximin game* i^* is one that maximizes the satisfaction index:

$$w_{i^*} = w^+ = \max_i w_i . \qquad (3)$$

While this criterion is reminiscent of Rawls's difference principle, it must be stressed that Rawls's principle is based on an index of "primary goods," not on comparisons of von Neumann Morgenstern utility (Rawls, 1971).[2] But why are agents permitted to compare von Neumann Morgenstern utilities in our model? The answer is that they do not: individuals maximize their private utilities given their expectations about others' behavior. They are classical myopic optimizers and make no interpersonal comparisons of utility whatsoever. In the evolutionary selection process, however, differences in utility functions translate into different speeds of adjustment by the different classes, which means that the process ends up making comparisons implicitly.

Game i is *efficient* if there exists no other game i' (in the class of feasible games) such that $a_{ij} < a_{i'j}$ for every class j. We are going to show that the evolutionary process selects (puts high probability on) norms that are efficient and approximately maximin. Before we state this result, however, a further remark is in order. Eking out the maximum possible gain for any one group usually means imposing a cost on other groups due to substitution possibilities in the design of the rules of the game. In other words, a game in which one particular class is very satisfied will, in most situations, be a game in which some other class is very dissatisfied. The extent to which such tradeoffs are possible has a bearing on the solution to the problem at hand. Roughly speaking, the greater the substitution possibilities, the closer is the evolutionary outcome to the maximin solution.

To state this result precisely, let w_j^- be the lowest level of satisfaction among all games in which class j is perfectly satisfied:

$$w_j^- = \min_i \{w_i : a_{ij} = 1\}. \qquad (4)$$

Further, define

$$w^- = \max_{1 \leq j \leq n} w_j^- \qquad (5)$$

[2] Gauthier (1986) advances a minimax principle of justice that is closer to ours, but his justification of it is quite different.

These ideas are illustrated graphically in Figure 1 for a population consisting of two classes. The payoffs (a_{i1}, a_{i2}) from each game are represented by a point in the nonnegative orthant. Each pair above the diagonal is associated with a vertical line, and each pair below the diagonal is associated with a horizontal line. The intersections of these lines with the diagonal coincide with the satisfaction indices of the various games.

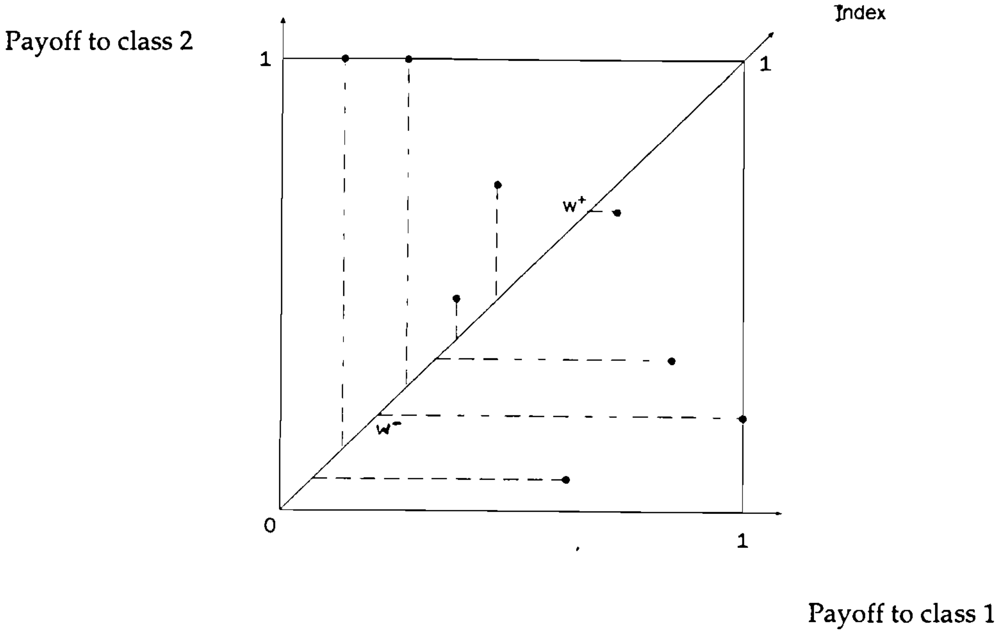


Figure 1. Satisfaction indices for a collection of two-person games.

Theorem 1. *In a pure n-person coordination game, every stable coordination equilibrium i is efficient and approximately maximin:*

$$1 \geq w_i / w^+ \geq \frac{[(1 - \rho/(w^+)^\eta)]^{1/\eta}}{(1 + \rho)} \qquad (6)$$

$$\textit{where} \quad \rho = \frac{(w^-)^\eta (1 - (w^+)^{2\eta})}{(1 + (w^-)^\eta)\,((w^+)^\eta + (w^-)^\eta)} \quad \textit{and} \quad \eta = 1/(n - 1). \qquad (7)$$

The parameter ρ measures the possible distortion from the maximin game. When w^- is small (whenever one class is perfectly satisfied, another is very dissatisfied), or when w^+ is close to one (all classes can simultaneously achieve a high level of satisfaction), ρ is close to zero and $w_i \approx w^+$.

4. Examples.

The detailed proof of theorem 1 is given in the Appendix. Here we shall describe a general method for computing the stable equilibria that can be used to analyze particular cases.

Consider any family of m strictly positive payoff vectors $a_1, a_2, \ldots, a_m \in R^n$. Construct a graph having m vertices, one for each vector, and for every ordered pair of vertices (i, i'), $i \neq i'$, draw a directed edge from vertex i to vertex i' and give it weight

$$r_{ii'} = \min_j \{(a_{ij})^\eta / ((a_{ij})^\eta + (a_{i'j})^\eta)\}, \text{ where } \eta = 1/(n-1). \qquad (8)$$

Define an *i-tree* to be a subset of directed edges in the graph such that every vertex i' other than i has exactly one exiting edge, and there is a unique directed path from i' to i (Freidlin and Wentzell, 1984). The vertex i is called the *root* of the tree. The *resistance* of the tree is the sum of the weights on its edges. The *stochastic potential* of payoff vector i is the smallest resistance among all i-trees. It can be shown (see the appendix) that the stable payoff vectors are those with minimum stochastic potential.[3]

To illustrate these ideas, consider norms regarding the control of property in a marriage. In many societies, men and women have substantially different expectations about the extent to which they can own and inherit property. (In eighteenth century Britain, for example, wives were expected to cede control over their property to their husbands, and inheritance by the eldest son was the norm.) To cast this into the form of a coordination game, imagine that each partner in a (prospective) marriage can propose one of three arrangements: the man is sole owner of the property, the woman is sole owner of the property, or they own it jointly.

[3] This follows from Young (1993a, Theorem 4). Kandori, Mailath, and Rob (1993) use similar combinatorial methods that are based on the work of Freidlin and Wentzell (1984).

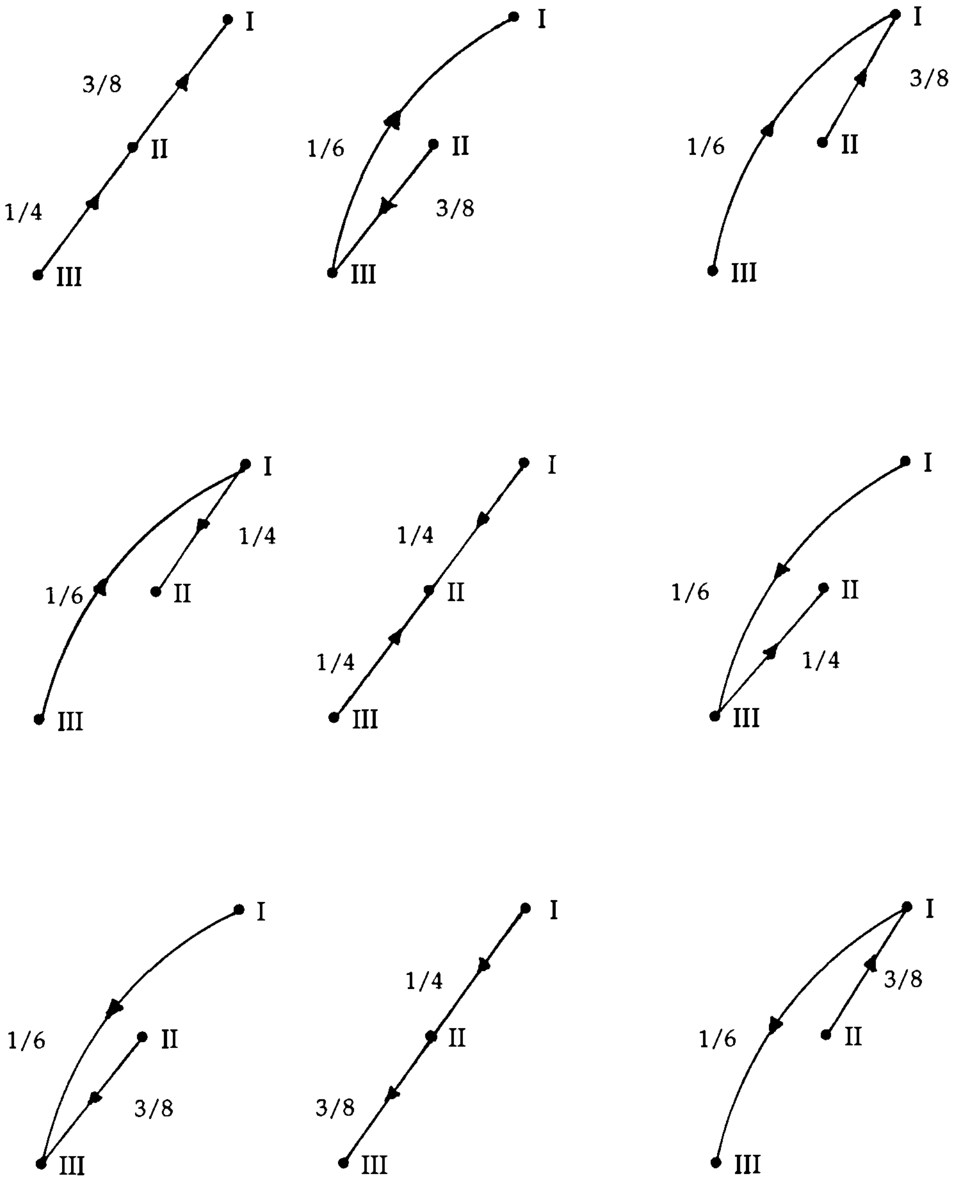
Let the payoffs be as follows:

		men		
		sole owner	joint owner	non-owner
women	sole owner	0, 0	0, 0	5, 1
	joint owner	0, 0	3, 3	0, 0
	non-owner	1, 5	0, 0	0, 0

The computations are illustrated in figure 2. There are nine rooted trees, and the two with minimum resistance (1/6 + 1/4) have the joint ownership norm as the root. Thus joint ownership is stochastically stable and will be observed with higher probability, over the long run, than either of the other arrangements.[4]

Of course we should not take this result too literally, as the model is highly simplified. Nevertheless, the *explanation* for the result is robust and makes intuitive sense. Norms with extreme payoffs are relatively easy to dislodge because it does not take many stochastic shocks to induce the most dissatisfied group to try something different. By contrast, norms with payoffs that are centrally located in the feasible payoff set are relatively hard to dislodge, because it takes a larger accumulation of stochastic shocks to induce any group to want to change. The net effect, over the long run, is to push society away from the boundaries of the feasible payoff set and toward the middle of the efficiency frontier. (The proof of the theorem requires a considerably more complex argument, but this is the essence of the matter.)

[4] Joint ownerhsip is stable even if the expected payoffs to sharing are lower than 3, say due to higher transactions costs incurred by joint management of the property. Any payoff greater than $\sqrt{5}$ will work.



I = women are sole owners
II = joint ownership
III = men are sole owners

Figure 2. Resistances of the nine rooted trees for the property game.

Our next two examples show why the theorem cannot be substantially strengthened. First we exhibit a situation where the stable norm is efficient but not strictly so. In Figure 3, norm 1 weakly Pareto dominates norm 2, but the least-resistant 1-tree (Figure 3a) and the least-resistant 2-tree (Figure 3b) are congruent and have the same resistance. Hence both norms are stable.

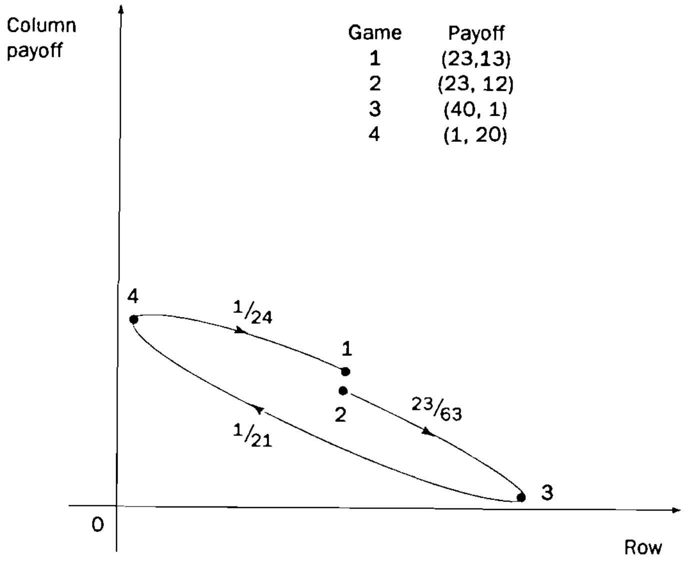


Figure 3a. Least resistant 1-tree.

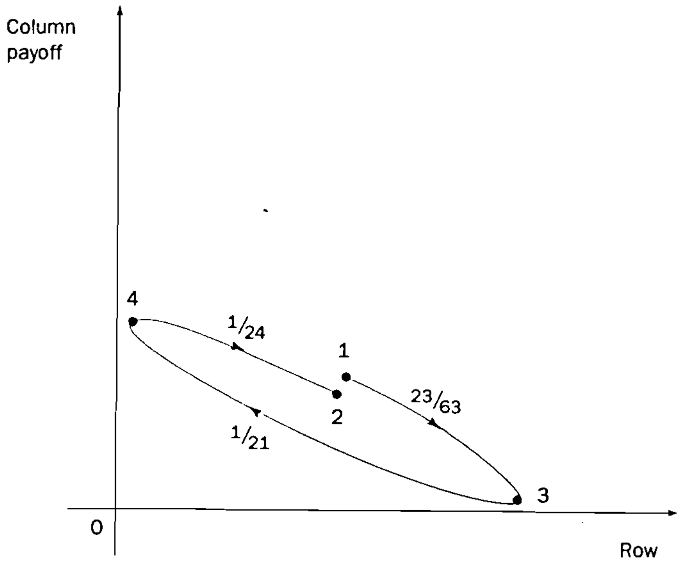


Figure 3b. Least resistant 2-tree.

Next we show that the lower bound in (6) is tight when there are at least four games. Choose real numbers w^-, w, and w^+ such that

$$0 < w^- < w < w^+ < 1.$$

Construct four two-person games with the payoffs shown in figure 4. Then w^- and w^+ have the meanings defined in (3) and (5). The least-resistant tree rooted at (w^+, w^+) is shown in figure 4a, the least-resistant tree rooted at $(w, 1)$ is in figure 4b. The payoffs (w^+, w^+) correspond to the maximin game, but the game with payoffs $(w, 1)$ has the same or lower stochastic potential if

$$\frac{w^-}{1+w^-} + \frac{w^-}{1+w^-} + \frac{w}{1+w} \geq \frac{w^-}{1+w^-} + \frac{w^+}{1+w^+} + \frac{w^-}{w^-+w^+}. \tag{9}$$

It can be verified that this reduces to $w/w^+ \geq (1 - \rho/w^+)/(1 + \rho)$, where ρ is defined as in (7) and $\eta = 1$. Hence the bound in (6) is best possible for two-person games, and a similar construction shows that it is best possible for n-person games. This example suggests that departures from maximin arise only in cases where there are relatively large gains to some group, and relatively small losses to the other groups, from moving away from the maximin equilibrium. Indeed theorem 1 gives a precise bound on how far below the minimax level the worst-off group can be pushed as a consequence of gains to some other group.

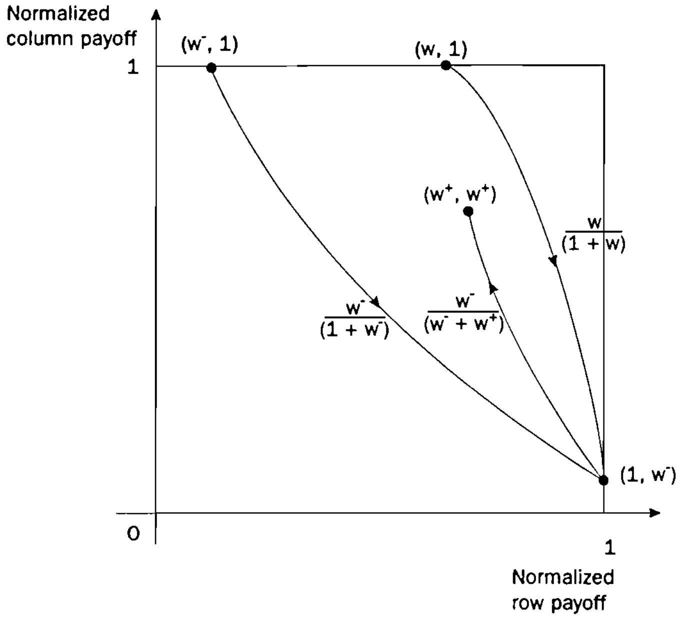


Figure 4a. Least-resistant (w^+, w^+) tree.

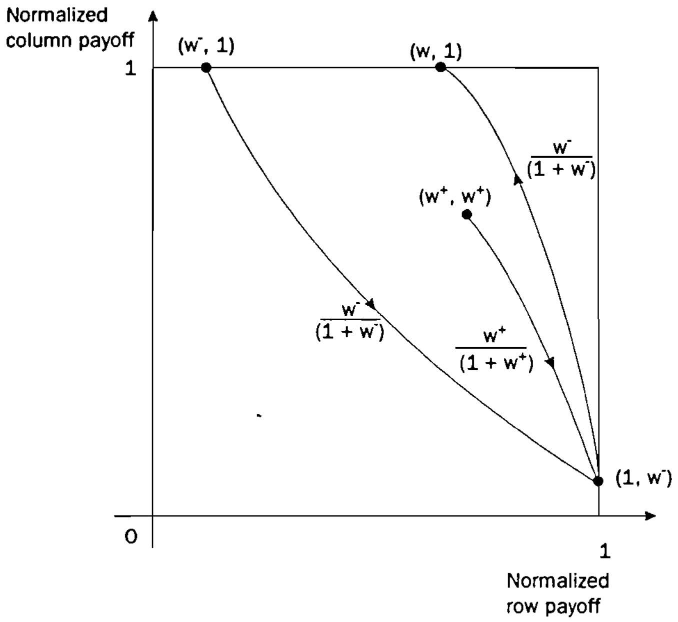


Figure 4b. Least-resistant $(w, 1)$-tree.

5. Specializations of the main result: 2 x 2 games.

In this section and the next we shall discuss several cases where the conclusions of theorem 1 can be sharpened. Consider first the situation where the competition is between one pair of two-person games with payoffs (a_1, b_1), (a_2, b_2). In this case we have an evolutionary selection process on a 2 x 2 coordination game of form

$$\begin{array}{cc} a_1, b_1 & 0,\ 0 \\ 0,\ 0 & a_2, b_2 \end{array}$$

where a_1, b_1, a_2, $b_2 > 0$. This situation has been treated (with slightly different assumptions about the adaptive process) by Kandori, Mailath, and Rob (1993) in the symmetric case, and by Young (1993a) in the asymmetric case. Namely, a coordination equilibrium is stable in a 2 x 2 game if and only if it is risk dominant. When the off-diagonal payoffs are zero, the risk dominant equilibrium is the one that maximizes the product of the agents' utilities.

In general, however, risk dominance and stochastic stability are not equivalent even in pure coordination games. As an example, consider the game

$$\begin{array}{cccc} 1,10 & 0,0 & 0,0 & 0,0 \\ 0,0 & 4,10 & 0,0 & 0,0 \\ 0,0 & 0,0 & 6,6 & 0,0 \\ 0,0 & 0,0 & 0,0 & 10,1 \end{array}$$

A straightforward calculation using the above methods shows that (6, 6) has the lowest stochastic potential, though (4, 10) is risk dominant.[5]

6. Symmetric games

Another special case arises when the games are all symmetric, that is, the payoffs are the same no matter what class a person is in. Let $a_{ij} = a_i$ be the expected payoff to every player in game i. Then the games can be ordered according to their welfare -- say $a_1 \geq a_2 \geq \ldots \geq a_m$. This case has already been treated in the literature (Kandori and Rob, 1995), and the stable games are precisely the games that maximize welfare. This can be established quite readily using the above methods. Choose any $i > 1$. From (8) it follows that the minimum-weight edge exiting from vertex i is directed toward vertex 1, and its weight is at most 1/2. On the other hand, every edge exiting from 1 has weight at least 1/2. It follows that, among all rooted trees, the tree rooted at vertex 1 and consisting of the directed edges $\{(i, 1): i = 2, 3, \ldots, m\}$ is the one with least resistance. Hence game 1 (or indeed any game with highest payoff) has minimum stochastic potential and is stochastically stable.

[5] Young (1993a) gives an example of a coordination game with non-zero off-diagonals in which risk dominance and stochastic stability differ.

7. Selecting from a continuum of payoffs: the Kalai-Smorodinsky solution.

Let us return to the asymmetric case and suppose that there is a continuum of games whose payoffs form a convex set. A plausible example would be the expected payoffs associated with all possible forms of *contracts* between n agents. In general, let B be a compact, convex, comprehensive, full-dimensional subset of R^n_+, that is, a *bargaining set*. Assume further that the utilities are normalized so that for each i,

$$x^+_i = \max \{x_i: x \in B\} = 1.$$

The *Kalai-Smorodinsky solution* is the unique vector $x^* \in B$ such that

$$x^* = \underset{x \in B}{\text{argmax}} \min_{1 \leq j \leq n} x_j$$

Discretize B as follows: for each positive integer N, let B^N consist of all payoff vectors $x \in B$ such that $Nx_1, Nx_2, \ldots, Nx_n$ are positive integers. B^N is nonempty for all sufficiently large N, and $B^N \to B$ in the Hausdorff topology as $N \to \infty$. Define $w^+(N)$ and $w^-(N)$ as in (3) and (5) respectively. As $N \to \infty$, $w^+(N) \to \min_j x^*_j$ and $w^-(N) \to 0$. It follows from (7) that $\lim_{N\to\infty} \rho(N) = 0$, so by theorem 1, the stochastically stable norm(s) of B^N converge to the Kalai-Smorodinsky solution of B.

We may state this result somewhat more generally as follows. For any finite set $X \subset R^n_+$, let cch(X) denote the convex, comprehensive hull of X.

Corollary 1.1. *Let* B *be a bargaining set in* R^n_+ *with Kalai-Smorodinsky solution* x^*. *Let* $B^1, B^2, \ldots B^h, \ldots$ *be a sequence of finite subsets of* B *such that* $\text{cch}(B^h) \to B$ *in the Hausdorff topology as* $h \to \infty$. *If* x^h *is stochastically stable in the set* B^h, *then* $x^h \to x^*$ as $h \to \infty$.

There is an interesting contrast between this result and the evolutionary bargaining model proposed in Young (1993b). In the latter model, two classes of agents play the Nash demand game: each demands a share of a fixed pie of size 1. They get what they ask for if the demands sum to one or less; otherwise they get nothing. The strategy space is discretized to keep the state space finite, and there are random perturbations in the agents' choices. Let all agents in class 1 have utility function $u_1(x)$ and all agents in class 2 have utility function $u_2(x)$, where $x \in [0, 1]$ is the agent's share. Then the evolutionary process selects a norm that is arbitrarily close to the Nash bargaining solution when the discretization of the state space is sufficiently fine. In other words, any stable division of the pie $(x, 1 - x)$ is close to the unique division $(x^*, 1 - x^*)$ that maximizes

$$[u_1(x) - u_1(0)]\,[u_2(1 - x) - u_2(0)].$$

The crucial difference between that model and the present one is the following. In the Nash demand game, agents get their demands *even when they miscoordinate* by asking for amounts that sum to less than unity, that is, when $x_1 + x_2 < 1$. Modelled as a pure coordination game, the agents would only get their demands if they are fully consistent, that is, $x_1 + x_2 = 1$. In this case, the evolutionary model selects the Kalai-Smorodinsky solution. This underscores the sensitivity of the outcome to the precise way in which we model the one-shot game. However, it is also consistent with the idea that evolutionary forces tend to select outcomes that are more or less centrally located on the efficiency frontier of the bargaining set. It is beyond the scope of this paper to establish this central tendency hypothesis in its most general form, but we conjecture that it holds for a wide variety of stochastic adaptive models of the type described here.

8. Variations in the model

The model described above, like all models, presents a very simplified picture of reality. In this section we shall suggest several ways in which it could be made more realistic. We shall then argue that these embellishments do not change the bottom line very much: essentially the conclusions of theorem 1 continue to hold.

One unrealistic assumption in our model is that individuals react to the *entire distribution* of actions by other agents in the preceding period. This is clearly absurd when the populations are large; people do not have that much information. A more plausible scenario is that agents know, through hearsay and personal experience, what *a few* other people in the other classes have been doing. In this sense information is partial, and it has a random component that depends on how the agent happened to hear about a given precedent. A natural way of incorporating this idea into the model is to suppose that each agent knows a random sample of size s drawn from each of the other populations, where $s \leq k$ (the number of agents in each class) and the draws are independent among populations. The agent then follows the same decision rule as before, where the outcome of the random sample forms his base of information. It is straightforward to show that the long-run behavior of this process is essentially the same as in the full information model: the absorbing states are the norms, and the resistance to transiting between norms is the same except that it depends on the size of s rather than the size of k. Thus the conclusions of theorem 1 continue to hold when the sample size is large.

A second unrealistic assumption is that changes in expectations arise solely from the cumulative impact of many *independent*, idiosyncratic choices by individuals. In reality, changes in expectations are often correlated, because individuals are reacting to a common event -- a news item (Rosa Parks refuses to sit in the back of a bus), a speech ("I Have a Dream"), or a new theory (*Das Kapital*). Without trying to minimize the variety and complexity of such influences, we can nevertheless say that correlated changes in behavior do not *by themselves* change the substance of the argument, as long as they represent sufficiently small fractions of the total population.

To see why, assume for simplicity that we are discussing two-person interactions and suppose that the current norm is z_i. To trip the process into the basin of attraction of some other norm $z_{i'}$ requires at least $R_{ii'}$ people in some class to switch from game i to game i'. When these changes arise from uncorrelated "mutations" in behavior, each having probability ε, the probability of this event is on the order of $\varepsilon^{R}ii'$. Suppose instead that individuals do not change idiosyncratically, but in groups: ideas fall from the blue and strike a certain number of people p at the same time. We can think of this as a Poisson type of process in which the probability that i ideas arrive in a given period is proportional to ε^{i}. Assume that all classes are equally likely to be hit, but that when an idea strikes it falls entirely on one class. Each idea is "labelled" with one of the m available games, $1 \leq i \leq m$. All those who are struck by an idea labelled "i" insist on playing game i in the next period. A given idea is equally likely to have any one of the labels $1 \leq i \leq m$.

For each pair of norms z_i and $z_{i'}$, the number of *individuals* who must change their minds in order to trip the process from z_i to $z_{i'}$ is the same as before, namely, $R_{ii'}$. In the correlated model, it therefore takes a succession of at least $R_{ii'}/p$ ideas labelled i' to trip the process from z_i to $z_{i'}$. Assuming that $p << R_{ii'}$, this event has probability on the order of $\varepsilon^{R_{ii'}/p}$. Therefore the *relative difficulty* of getting from any norm to any other is the same as in the uncorrelated model. Hence the conclusions of theorem 1 continue to hold provided that the correlations involve small enough fractions of the whole population.

9. Conclusion.

Of course, the forces that produce social change are far more complex than anything we can hope to capture in a simple model. Nevertheless, we would argue that the model described above does contain many of the key elements that shape the evolution of social norms. These include: the salience of precedent, boundedly rational responses by individuals to their environment, and random variation in expectations. The ways in which these features are formally modeled as a stochastic process may alter the conclusions to some degree, but it seems reasonable to conjecture that the central tendency property will hold under a wide variety of assumptions. The intuitive reason is that norms with payoffs near the boundary of the feasible payoff set tend to be unstable. They imply that some group is dissatisfied, and the more dissatisfied a group is, the more easily it is seduced by new ideas that give them hope of getting more. Social change, in other words, is driven by those who have the most to gain from change. Over the long run, this tends to favor institutions that are efficient, and that offer each group in society a fairly large share of the potential benefits they can realize from cooperation.

Appendix

Proof of Theorem 1. Let G be a pure n-person coordination game with m strategies for each player. In the text we interpreted a strategy as naming a particular game out of a given class of m games, but this interpretation is not necessary for the argument set forth here. Let $a_{ij} > 0$ be the payoff for members of class j when everyone chooses strategy i. We shall assume the payoffs have been normalized so that $\max_i a_{ij} = 1$ for every j, $1 \leq j \leq n$. Let $P^{k,\nu,\varepsilon}$ be the Markov process on the finite state space X defined in section 2. For notational convenience we shall temporarily fix $k \geq 1$ and $\nu \in (0, 1)$, and let $P^{\varepsilon} = P^{k,\nu,\varepsilon}$.

Consider first the situation in which $\varepsilon = 0$. An *ergodic class (recurrent communication class)* of P^0 is a subset of states Y in X such that the process never leaves Y once it is in Y, and every state in Y can be reached with positive probability from every other state in Y. Every norm $\{z_i\}$ constitutes a singleton ergodic class (i.e., an absorbing state) of P^0, because the probability of exiting z_i is zero once the process is in it.

We claim that the norms $\{z_i\}$ are, in fact, the *only* ergodic classes of P^0. To see why, consider any state $x \in X$ at the end of some time period t. There is a positive probability that, in period t + 1, all players will choose a best reply to the empirical distributions implied by x. (If there are ties in best reply, the full support of the tie-breaking rule implies that there is a positive probability that everyone chooses the *same* best reply.) Thus at the end of period t + 1 there is a positive probability of moving to a state x' in which everyone in the same class j plays the same strategy, say i_j.

Moving to the next period, there is a positive probability that everyone in classes 1 to n - 1 sticks with his previous choice out of inertia, and that everyone in class n chooses a best reply to the empirical distributions in x'. The games named by the members of classes 1 through n - 1 in x' are $i_1, i_2, \ldots, i_{n-1}$. A best reply to any probability distribution over members of this set is again in the set. Thus, barring ties in best replies, everyone in class n will play the same strategy, namely, one of the elements in $\{i_1, i_2, \ldots, i_{n-1}\}$. Even if there are ties, each will be chosen with positive (stationary) probability, so again there is a positive probability that *everyone* in class n will choose an element in $\{i_1, i_2, \ldots, i_{n-1}\}$. At this stage we have reached a state x" where everyone in at least two classes is coordinated on the same game.

In the next period, assume that everyone in these classes stays fixed out of inertia, while everyone else chooses a best reply to x". By now it is clear how, in a finite number of periods, the process can reach a state in which everyone names the same game. This proves that the only ergodic classes of P^0 are the n norms $\{z_i\}$.

Now consider the situation where $\varepsilon > 0$. Given a state x, define a *quirky* choice by some player to be a choice in the next period that is neither inertial nor a best reply to the state x. For every two states x and x', there is some number of quirky choices that transforms x to x' in one period. Let r(x, x') be the *least* such number. (Possibly r(x, x') = 0.) Then the process P^ε has the following properties

i) P^ε is aperiodic and irreducible whenever $\varepsilon > 0$,
ii) $\lim_{\varepsilon \to 0} P^\varepsilon_{xx'} = P^0_{xx'}$ for all $x, x' \in X$,
iii) $0 < \lim_{\varepsilon \to 0} \varepsilon^{-R(x, x')} P^\varepsilon_{xx'} < \infty$ for all $x, x' \in X$.

Such a process P^ε is said to be a *regular perturbation* of P^0 (Young, 1993a). It follows from (by now) standard arguments that P^ε has a unique stationary distribution μ^ε, and $\lim_{\varepsilon \to 0} \mu^\varepsilon(x) = \mu^0(x)$ is a stationary distribution of the process P^0.

A state x is *stochastically stable* if $\mu^0(x) > 0$. Since $\mu^0(x)$ is a stationary distribution of P^0, and the only ergodic classes of P^0 are the norms $\{z_i\}$, it follows that if $\mu^0(x) > 0$, then $x = z_i$ for some i. In other words, every stochastically stable state must be a norm. The proof of Theorem 1 proceeds by characterizing the stochastically stable norms analytically, drawing on the methods of Freidlin and Wentzell (1984) and Young (1993a).

Construct a complete, directed graph Γ having X as its set of vertices. Let the *resistance* of the directed edge $x \to x'$ be r(x, x'). Consider any two norms z_i and z_j. A *path* from z_i to $z_{i'}$ in Γ is a sequence of directed edges that begins at z_i and ends at $z_{i'}$. The *resistance* along this path is the sum of the resistances over its edges. For every two norms z_i and $z_{i'}$, define $R_{ii'}$ to be the *minimum resistance* over all directed paths from z_i to $z_{i'}$.

The numbers $R_{ii'}$ are computed as follows. In state z_i everyone names game i. To get to the norm $z_{i'}$ with as few quirky choices as possible, it is useless to name games other than i'. Thus, to compute $R_{ii'}$, we need to find the *smallest* number of individuals (summed over all classes) who must name game i' in order for the norm to tip from z_i into a state x from which the process can evolve, with no further quirky choices, to $z_{i'}$. That is, we need to find the smallest number of switches from i to i' that results in some state $x \in B^0(z_{i'})$, where $B^0(z_{i'})$ is the set of all states from which $z_{i'}$ is reached with positive probability under P^0. $B^0(z_{i'})$ is called the *basin of attraction* of $z_{i'}$ under P^0.

Fix a class j*, and consider a state x such that, in each class $j \neq j^*$, k_j individuals name i', and $k - k_j$ name i. (Recall that each class contains k individuals.) Then i' is a best reply for every member of class j* provided that

$$a_{i'j^*} \prod_{j \neq j^*} k_j/k \geq a_{ij^*} \prod_{j \neq j^*} (1 - k_j/k). \qquad \text{(A1)}$$

If this holds, then the state x is in the basin of attraction of $z_{i'}$. Let $m_{j^*} = \min \sum_{j \neq j^*} k_j$ over all sets of integers $(k_j)_{j \neq j^*}$ such that $0 \leq k_j \leq k$ and (A1) holds. Thus $R_{ii'} = \min_{j^*} m_{j^*}$.

Sidestepping (for the moment) the complications that arise from integer values, let $p_j = k_j/k$ and consider the solution to the system

$$\min \sum_{j \neq j^*} p_j \qquad \text{(A2)}$$

subject to

$$0 \leq p_j \leq 1 \text{ for all } j \neq j^*,$$

and

$$a_{i'j^*} \prod_{j \neq j^*} p_j \geq a_{ij^*} \prod_{j \neq j^*} (1 - p_j).$$

The minimum of $\sum_{j \neq j^*} p_j$ occurs when all p_j are equal, that is,

$$(p_j/(1 - p_j))^{n-1} = a_{ij^*}/a_{i'j^*} \text{ for all } j \neq j^*.$$

This is equivalent to

$$p_j = a_{ij^*}{}^{\eta}/(a_{ij^*}{}^{\eta} + a_{i'j^*}{}^{\eta}) \text{ for all } j \neq j^*, \text{ where } \eta = 1/(n - 1).$$

In general, let $\lceil y \rceil$ denote the least integer greater than or equal to a real number y. It follows from the above that

$$m_{j^*} = (n - 1) \lceil k a_{ij^*}{}^{\eta}/(a_{ij^*}{}^{\eta} + ai'j^* {}^{\eta}) \rceil.$$

Since $R_{ii'} = \min_{j^*} m_{j^*}$, and j* is arbitrary, it follows that

$$R_{ii'} = (n - 1) \lceil k \min_j \{a_{ij}{}^{\eta}/(a_{ij}{}^{\eta} + a_{i'j}{}^{\eta})\} \rceil.$$

For the most part we shall do calculations in terms of

$$r_{ii'} = \min_j \{a_{ij}{}^{\eta}/(a_{ij}{}^{\eta} + a_{i'j}{}^{\eta})\}, \qquad \text{(A3)}$$

noting that $R_{ii'} = (n - 1)\lceil k r_{ii'} \rceil$.

Now construct another directed graph Γ^*. This one has n vertices, one for each of the coordination equilibria $1 \leq i \leq n$. The weight (or resistance) on the edge $i \to i'$ is $R_{ii'}. = (n - 1)\lceil k r_{ii'} \rceil$. The *stochastic potential* $\phi_k(i)$ of equilibrium i is the resistance of the least-resistant i-tree. It can be shown that *the stochastically stable norms* z_i *are precisely those* i *that minimize the stochastic potential function* $\phi_k(i)$ (Young, 1993a, Theorem 4). Note that this function is independent of the inertia rate ν. It is also essentially independent of k in the following sense. Say that a coordination equilibrium i is *stable* if the corresponding norm z_i is stochastically stable for infinitely many values of k. Let $\phi(i)$ be the resistance of the least resistant i-tree when each weight $R_{i'i''}$ is replaced by $r_{i'i''}$. Among all those i that minimize ϕ, there is at least one that minimizes ϕ_k for infinitely many k, so it is stable. Thus stable equilibria exist. On the other hand, if i does not minimize ϕ, then it does not minimize ϕ_k for all sufficiently large k, hence it is not stable. Thus every stable equilibrium minimizes ϕ. To verify the claims of the theorem, it suffices to show that every i that minimizes ϕ is efficient and approximately maximin.

To prove the latter statement, we need to show that if i minimizes ϕ, then w_i/w^+ is bounded below as in expression (6). This is obviously true if $w_i = w^+$, so let us assume that $w_i < w^+$. Let $i^* \neq i$ be a maximin equilibrium: $w_{i^*} = w^+$. Let T_i be an i-tree having minimum resistance among all rooted trees, that is, $r(T_i) = \phi(i)$. There is a unique edge e_1 in T_i exiting from i^*. Let h be the least-satisfied class in game i, that is, $a_{ih} = w_i$. Let i' be a game in which h gets its maximum ($a_{i'h} = 1$). Among all such games i' choose one for which the welfare index $w_{i'}$ is lowest. Thus

$$w_{i'} \leq w^-. \tag{A4}$$

Assume for the moment that $i' \neq i, i^*$. (We shall dispose of the cases $i' = i$ and $i' = i^*$ in due course.) Let e_2 be the unique edge in T_i exiting from i. Construct an i^*-tree T_{i^*} by deleting the edges e_1 and e_2 from T_i and adding the edges $e'_1 = (i \rightarrow i')$ and $e'_2 = (i' \rightarrow i^*)$. See figure 5.

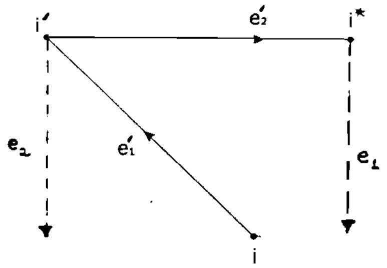


Figure 5.

By assumption T_i is a least resistant tree, hence $r(T_i) \leq r(T_{i^*})$. From this it follows that

$$r(e_1) + r(e_2) \leq r(e'_1) + r(e'_2), \tag{A5}$$

(Here r(e) denotes the resistance of edge e.) Let us evaluate the four resistances in expression (A5). We know that

$$r(e'_1) = r_{ii'} = \min_j \{a_{ij}^\eta/(a_{ij}^\eta + a_{i'j}^\eta)\}.$$

By choice of i', the minimum occurs when j = h, that is,

$$r(e'_1) = a_{ih}^\eta/(a_{ih}^\eta + a_{i'h}^\eta) = w_i^\eta/(w_i^\eta + 1). \qquad (A6)$$

The minimum resistance from i* to any other vertex is

$$\min_{i \neq i^*} \min_j a_{i^*j}^\eta/(a_{i^*j}^\eta + a_{ij}^\eta) = w_{i^*}^\eta/(w_{i^*}^\eta + 1) =(w^+)^\eta/((w^+)^\eta + 1).$$

Since e_1 exits from i* it follows that

$$r(e_1) \geq (w^+)^\eta/((w^+)^\eta + 1). \qquad (A7)$$

Similarly,

$$r(e_2) \geq w_{i'}^\eta/(w_{i'}^\eta + 1). \qquad (A8)$$

Finally, let us note that

$$r(e'_2) = \min_j \{a_{i'j}^\eta/(a_{i'j}^\eta + a_{i^*j}^\eta)\}. \qquad (A9)$$

By definition, $w^+ = w_{i^*} = \min_j a_{i^*j}$, so $a_{i^*j} \geq w^+$ for all j. From this and (A9) it follows that

$$r(e'_2) \leq \min_j \{a_{i'j}^\eta /(a_{i'j}^\eta + (w^+)^\eta)\} = w_{i'}^\eta/(w_{i'}^\eta + (w^+)^\eta). \qquad (A10)$$

Combining (A5) - (A10) results in the inequality

$$\frac{(w^+)^\eta}{(w^+)^\eta + 1} + \frac{w_{i'}^\eta}{w_{i'}^\eta + 1} \leq \frac{w_i^\eta}{w_i^\eta + 1} + \frac{w_{i'}^\eta}{w_{i'}^\eta + (w^+)^\eta} \quad .$$

After some algebraic manipulation, we obtain the equivalent expression

$$w_i^{\eta} \geq \frac{(w^+)^{\eta} - \rho(w_{i'})}{1 + \rho(w_{i'})} \quad \text{where } \rho(w_{i'}) = \frac{w_{i'}^{\eta}(1 - (w^+)^{2\eta})}{(1 + w_{i'}^{\eta})((w^+)^{\eta} + w_{i'}^{\eta})}$$

By choice of i', $w^- \geq w_{i'}$. It is straightforward to show that $\rho(w_{i'})$ is increasing in $w_{i'}$ provided that $(w_{i'})^2 \leq w^+$, which holds because $w_{i'} \leq w^+ \leq 1$. Letting $\rho = \rho(w^-)$ we therefore obtain

$$w_i^{\eta} \geq \frac{(w^+)^{\eta} - \rho}{1 + \rho},$$

which implies the inequality claimed in the theorem.

It remains to dispose of the cases i' = i or i' = i*. We claim that in fact neither is possible. Suppose on the one hand that i' = i. By definition, i' is an equilibrium in which the least satisfied class gets its maximum (namely 1). If i' = i, it follows that *all* classes get their maximum in equilibrium i, which contradicts our assumption that $w_i < w^+$.

Suppose on the other hand that i' = i*. Change the tree T_i by deleting the unique edge e_1 that exits from i*, and adding the edge $e'_1 = (i \rightarrow i')$. This results in an i'-tree, say $T_{i'}$. Since i is stable but i' is not, we must have $r(T_i) < r(T_{i'})$, that is,

$$r(e'_1) > r(e_1). \tag{A11}$$

Since the least satisfied class in equilibrium i gets its maximum in i', it follows that

$$r(e'_1) \leq w_i^{\eta}/(w_i^{\eta} + 1).$$

But (A7) implies that $r(e_1) \geq (w^+)^{\eta}/((w^+)^{\eta} + 1)$. Thus

$$r(e_1) \geq (w^+)^{\eta}/((w^+)^{\eta} + 1) \geq w_i^{\eta}/(w_i^{\eta} + 1) \geq r(e'_1),$$

which contradicts (A11).

Finally, we need to establish Pareto optimality. Suppose, by way of contradiction, that i is a stable equilibrium and that the payoffs from i are strictly dominated by the payoffs from i', where i' is not stable:

$$a_{ij} < a_{i'j} \text{ for all j, } 1 \leq j \leq n. \tag{A12}$$

Let T_i be an i-tree of least resistance. Since i is stable, none of the i'-trees can have a resistance *strictly smaller* than $r(T_i)$. We shall show, however, that there is such an i'-tree.

If (i' → i) is an edge in T_i, replace it by the opposite edge (i → i'). This creates an i'-tree, and its resistance is less than $r(T_i)$ because (A3) and (A12) imply that $r_{ii'} < r_{i'i}$. This contradiction shows that (i' → i) is not an edge in T_i.

Consider any edge (h → i) in T_i that points toward i (there is at least one such edge because i is the root). Replace it with the edge (h → i'), and do this for all edges that point toward i. In T_i there is a unique edge exiting from i', say (i' → h'), where h' ≠ i by the above. Replace it with the edge (i → h'). It is easy to check that this new object is an i'-tree. It has strictly lower resistance than T_i , because (A3) and (A12) imply that for all i, i', h, and h',

$$r_{hi'} = \min_j \{a_{hj}{}^{\eta}/(a_{hj}{}^{\eta} + a_{i'j}{}^{\eta}) < \min_j \{a_{hj}{}^{\eta}/(a_{hj}{}^{\eta} \ a_{ij}{}^{\eta}\} = r_{hi}.$$

and

$$r_{ih'} = \min_j \{a_{ij}{}^{\eta}/(a_{ij}{}^{\eta} + a_{h'j}{}^{\eta}\} < \min_j \{a_{i'j}{}^{\eta}/(a_{i'j}{}^{\eta} + a_{h'j}{}^{\eta}\} = r_{i'h'}.$$

This contradiction shows that no stable equilibrium is strictly Pareto dominated, and completes the proof of the theorem.

References

An, Mark Y. and Nicholas M. Kiefer (1993), "Evolution and Equilibrium Selection in Repeated Lattice Games," Preprint, Cornell University.

Arrow, Kenneth, (1994): "Methodological Individualism and Social Knowledge," *American Economic Review*, 84, 1-9.

Binmore, Ken (1994): *Game Theory and the Social Contract I: Playing Fair*. Cambridge, MA: MIT Press.

Blume, Larry (1993): "The Statistical Mechanics of Strategic Interaction," *Games and Economic Behavior*, 4, 387-424.

Ellison, Glenn (1993): "Learning, Local Interaction, and Coordination," *Econometrica*, 61, 1047-1071.

Foster, Dean, and H. Peyton Young (1990),"Stochastic Evolutionary Game Dynamics," *Theoretical Population Biology*, 38, 219-232.

Freidlin, Mark, and Alexander Wentzell (1984): *Random Perturbations of Dynamical Systems*. Berlin; Springer-Verlag.

Gauthier, David (1986): *Morals by Agreement*. Clarendon Press: Oxford, 1986.

Kandori, Michihiro, and Rafael Rob (1995): "Evolution of Equilibria in the Long Run: A General Theory and Applications," *Journal of Economic Theory*, 65.

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

North, Douglass C. (1981): *Structure and Change in Economic History*. New York: Norton.

Rawls, John (1971): *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Schelling, Thomas (1971): "Dynamic Models of Segregation," *Journal of Mathematical Sociology*, 1, 143-186.

Schelling, Thomas (1978): *Micromotives and Macrobehavior*. New York: Norton.

Sugden, Robert (1986): *The Evolution of Rights, Cooperation, and Welfare.* NewYork: Basil Blackwell.

Young, H. Peyton. (1993a): "The Evolution of Conventions," *Econometrica* 61, 57-84.

Young, H. Peyton (1993b): "An Evolutionary Model of Bargaining," *Journal of Economic Theory*, 59, 145-168.