ELSEVIER

European Journal of Political Economy
Vol. 11 (1995) 663–681

European Journal of
POLITICAL
ECONOMY

Evolutionary selection of 'chivalrous' conventions in coordination games without common expectations

Pier Luigi Sacco [a,*], Marco Sandri [b]

[a] *Department of Economics, University of Florence, Florence, Italy*
[b] *Institute of Economics, University of Verona, Verona, Italy*

Abstract

We study an evolutionary game-theoretic model where players have to choose within a predetermined set of mixed strategies in a coordination game. Players are of two different kinds, male and female. No common expectations assumption is made; players tend therefore to adopt the strategy that yields larger than average expected payoffs for their kind. In this framework, every stable stationary point of the population dynamics can be interpreted as the emergence of a particular convention. A classification of the possible conventions is provided; conditions for their emergence are determined.

JEL classification: C79

Keywords: Chivalry; Coordination; Social convention; Social learning

1. Introduction

The past few years have witnessed a remarkable upsurge of interest toward the so-called evolutionary approach to game theory; this interest has been partly motivated by the recognition that traditional game theory does not seem to provide a satisfactory explanation for the existence of complex networks of social conven-

* Corresponding author. Department of Economics, University of Florence, Via Montebello, 7, 50123 Firenze, Italy.

SSDI 0176-2680(95)00031-3

tions which are so common and important in everyday life. In particular, it is felt that a satisfactory characterization of strategic rationality cannot do without a careful modelling of the social and cultural facets of the environment in which players are embedded (see Granovetter, 1985).

The acknowledgement of the importance of social and cultural factors as driving forces behind the strategic decisions of players, however, brings about a crucial question: to what extent are players' choices conditioned by the preexisting social environment rather than by explicit optimizing calculations? This tradeoff may be solved in radically different ways. As Elster (1989, p. 97) puts it: "One of the most persisting cleavages in the social sciences is the opposition between the two lines of thought conveniently associated with Adam Smith and Emile Durkheim, between *homo economicus* and *homo sociologicus*. Of these, the former is supposed to be guided by instrumental rationality, while the behaviour of the latter is dictated by social norms. The former is 'pulled' by the prospect of future rewards, whereas the latter is 'pushed' from behind by quasi-inertial forces. The former adapts to changing circumstances, always on the lookout for improvements. The latter is insensitive to circumstances, sticking to the prescribed behavior even if new and apparently better options become available". Granovetter (1992) calls these two positions *under-* and *oversocialized* conceptions of human action, respectively, and argues that this contraposition is in many respects restrictive and even misleading.

It is clear that, to a certain (substantial) degree, individual choices are conditioned by the preexisting social environment, as convincingly argued by the sociopsychological and microsociological literatures (see e.g. Argyle and Henderson, 1985; Goffman, 1974), and that these social conditionings are not completely traceable back to individual optimizing behaviors (see e.g. Elster, 1989). On the other hand, it is very implausible to postulate that individuals are never able to recognize that some courses of action are relatively more rewarding than others and to opt for the more rewarding ones. In other words, the tradeoff between social conditionings and individual optimizing motivations must not be shaped unilaterally as far as the modelling of the determinants of human action is concerned. A careful reflection about the interplay of the two forces is required. As noted by Granovetter (1992, p. 6), "the oversocialized approach has in common with the undersocialized a conception of action uninfluenced by peoples' existing social relations. In the undersocialized account this atomization results from the narrow pursuit of self-interest; in the oversocialized one – which originated as a corrective to the undersocialized one – atomization results nevertheless because behavioral patterns are treated as having been internalized and thus unaffected by ongoing social relations".

The first step toward a correct understanding of such interplay thus calls for a theoretical framework that be able to: (i) explain how the two forces (viz., social and cultural factors vs. optimization) interact and how this interaction feeds back on the 'ongoing social relations'; (ii) explain how the structure of this interaction

tends itself to be modified by 'ongoing social relations'. One possible framework is the following. We postulate that, in a 'short-run' perspective, individual choices tend to be constrained by the social environment in the sense that the individual *choice set* is fixed, i.e. the repertoire of possible individual behaviors is somewhat predetermined by social and cultural factors. Individuals are nevertheless able to choose within this repertoire the course of action that they find more rewarding. This is our preferred interpretation of point (i) above. In a 'long-run' perspective, however, the choice set might itself change for a variety of reasons. For example, the behavioral pattern induced by the existing social environment at the aggregate level might be unsatisfactory or undesirable in some respect, and this fact could be widely recognized by individuals, thus bringing about a 'constitutional change', i.e., an 'official' deliberation to modify the existing social institutions. Clearly, such deliberation should be backed by a strong enough social consensus and should be aimed to induce a more satisfactory/desirable behavioral pattern at the aggregate level in a well specified, and agreed upon, sense. This is our preferred interpretation of point (ii) above.

The aim of this paper is that of building an explicit model of the 'short-run' dynamics of social conventions [to be meant as customary, expected, self-enforcing states of things in the sense of Lewis (1969)] (i.e., point (i) above) in the specific context of a coordination game. The discussion of the 'long-run' dynamics (point (ii)) is outside the scope of the present paper; a tentative analysis of a specific example, although in a different analytical context, is carried out in Sacco (1993a) and Sacco (1993b).

In order to avoid the pitfall of building a model which, in Granovetter's terms, provides a characterization of human action that is 'unaffected by ongoing social relations', it is necessary to explain how the optimizing individual choices within the predetermined choice set are influenced by the interaction, *direct and indirect*, with other members of the society. In other words, it is necessary to explain how the fact that 'socially feasible behaviors' are more or less widespread within the society conditions individual calculations concerning the relative profitability of such behaviors. The evolutionary game theoretic approach provides a natural analytical environment for these phenomena. In this environment, one can construct dynamical models that describe the evolution of behaviors caused by social interaction processes and explain how a specific subset of the original choice set is eventually 'selected' in a self-enforcing way by the social dynamics. This process may be rationalized as the emergence of a 'social convention' in the above specified sense. [1] It is important to notice that the range of possible social conventions is determined by the socially predetermined choice set, but the convention that actually emerges depends entirely on the dynamical interaction of

[1] See also Bicchieri (1990). A static, evolutionary rationale for social conventions has been previously proposed by Sugden (1989).

(i.e. on the ongoing *social* relations between) *individual* optimizing choices. This is in our opinion a reasonable way of shaping the interplay of the sociologically and economically oriented components of human action.

The rest of the paper is organized as follows. Section 2 introduces the model. Section 3 presents the basic results, namely the conditions under which the various possible social conventions are selected by the evolutionary dynamics, and discusses them. Section 4 contains the proofs and a more detailed technical characterization of the dynamics. Section 5 discusses the relationship between individual optimizing choices and the selected social convention. Section 6 discusses the relationships between our results and the existing literature.

2. The model

In this section we translate the discussion of the previous section into a specific game theoretic model. Consider the following coordination game:

$$\begin{array}{cc} & \begin{array}{cc} S & \quad B \end{array} \\ \begin{array}{c} S \\ B \end{array} & \begin{array}{cc} (\gamma_M, \gamma_F) & (0,0) \\ (-\eta_M, -\eta_F) & (\delta_M, \delta_F) \end{array} \end{array} \tag{1}$$

All parameters are nonnegative and moreover $\gamma_M > \delta_M$, $\gamma_F < \delta_F$. This game is a generalization of the well-known battle of the sexes (see e.g. Rasmusen, 1989, p. 34): There are two kinds of players, a male (the row player) and a female one (the column player). Players must choose whether to go to the Stadium or to the Ballroom. The male player definitely prefers the Stadium, whereas the female prefers the Ballroom. On the other hand, both players give priority to the fact of meeting at the same place rather than to going to their respectively favorite place. If, however, players fail to coordinate, they are both worse off when going to the less favorite place rather than to the favorite one.

It is easy to prove that (1) admits a unique Nash equilibrium in mixed strategies, as well as two pure strategy Nash equilibria ((S,S) and (B,B)); nevertheless, the sensibility of mixed strategy Nash equilibrium (and, a fortiori, of the pure strategy Nash equilibria) as a solution concept for this game rests on the so called common expectations assumption, i.e., players share the same joint probability distribution on players' choices (see e.g. Bernheim, 1986; Tan and Werlang, 1988; Binmore, 1990; Hammond, 1992). This assumption is not particularly credible unless one gives an explicit argument for it. The same sort of critique applies to other solution concepts like correlated equilibrium (Brandenburger and Dekel, 1987; Hammond, 1992).

Assume that there is a large population of both male and female players who

are randomly matched to play the coordination game. We assume that the players' choice set is a certain subset of the set of mixed strategies (including of course pure strategies); the actual distribution of strategies within the population at each given moment is known to each player, but, when required to play, neither player knows the opponent's strategy. If not required to play, players are 'spectators' of the game which is going on; more specifically, they are able to observe the actual mixed strategies played in each given game. The distribution of strategies within the population changes in a simple way: Strategies that yield a higher (ex ante) payoff are adopted by an increasing proportion of individuals, at the expenses of less rewarding ones. Under this assumption, it is possible that all players eventually adopt the same mixed strategy, which has therefore been 'selected' on the basis of the 'fitness' criterion of expected payoff. [2]

As explained in Section 1, the existence of social and cultural factors may constrain the choice set faced by players, and there is in principle no compelling reason for assuming that the actual subset of mixed strategies that constitutes the 'socially feasible' choice set must include those strategies that are judged 'equilibrium' strategies in any specific sense by a 'rational outside observer'. The actual content of the socially feasible choice set may be for example the result of accidental historical circumstances that have frozen into an established 'tradition', as explained e.g. by Berger and Berger (1975). For example, for such reasons players could be socially conditioned to consider only a certain set of random mechanisms which have acquired with time a strong ritual meaning, e.g. a given set of urns containing shells of different colors; shells are drawn blindly and the outcome of the drawing is interpreted according to a predetermined code. [3]

To make this point formally, we assume that players' choice sets contain just two alternative socially feasible random mechanisms, i.e. just two mixed strategies, σ_1 and σ_2; at σ_1, S is played with probability α and B with probability $1-\alpha$; at σ_2, S is played with probability β and B with probability $1-\beta$. It is assumed that $\alpha > \beta$. No special assumptions on players' beliefs are made, so σ_1

[2] The assumption that players look at expected payoffs in evolutionary games with a random matching interaction structure is standard in the literature; see e.g. Hofbauer and Sigmund (1988). One might, at the cost of additional technical complications, also consider alternative specifications in which players look at realized payoffs.

[3] Of course, one might argue that, in a 'long-run perspective' in the sense of Section 1 above, the socially predetermined choice set could evolve through a sequence of 'constitutional' changes into a 'fully rational' choice set made only of Nash equilibrium strategies for the coordination game (1). Although certainly plausible, in order to become theoretically compelling this kind of long run dynamics need however an explicit and careful characterization of the adjustment mechanisms that are at work; in particular, the strength of established traditions and customs as barriers to institutional change should not be downplayed in this respect, as argued e.g. by Berger and Luckmann (1966). We therefore leave this difficult point as an open issue for future research.

and σ_2 need not be a Nash equilibrium profile. We could consider richer choice sets but this would complicate the computations without providing further insight. [4]

To be specific, we characterize the two alternative strategies considered by players in terms of their 'chivalry'; more precisely, from the point of view of the male player, strategy σ_1 assigns a relatively larger probability to the preferred place, namely S, whereas strategy σ_2 assigns a relatively larger probability to the place preferred by the female, namely B. So, for the male player σ_2 must be regarded as a 'chivalrous' strategy whereas σ_1 as a 'non-chivalrous' one; exactly the opposite holds for the female player. The proportions of male players and of female players who adopt strategy σ_1 are equal, respectively, to μ and ν. If a player of a given kind plays strategy σ_1, we will say (s)he is a 'type 1' player; an analogous stipulation holds for strategy σ_2.

We ask under what conditions various possible kinds of 'chivalry' may emerge as social conventions, starting from a given initial distribution of types. Notice that three different sorts of 'chivalrous' conventions may emerge: First, 'chivalry' is observed both for male and female players (i.e., all male players are of type 2 whereas all female players are of type 1); we speak in this case of Two-Sided Chivalry. Alternatively, 'chivalry' is observed for males or females only (whereas the other kind plays its 'non-chivalrous' strategy). We then speak of Male and Female Chivalry, respectively.

At this point we need to introduce specific assumptions concerning the population dynamics. In line with the informal remarks of Section 1, we choose to model it as a replicator dynamics; this amounts to assuming that the strategy that yields (ex ante) payoffs above the average increases its proportion within the population at the expense of the other (see e.g. Hofbauer and Sigmund, 1988). [5]

$$\dot{\mu} = \mu(1-\mu)\left[\pi^{\mathrm{M}}(\sigma_1) - \pi^{\mathrm{M}}(\sigma_2)\right], \tag{2}$$

$$\dot{\nu} = \nu(1-\nu)\left[\pi^{\mathrm{F}}(\sigma_1) - \pi^{\mathrm{F}}(\sigma_2)\right], \tag{3}$$

where $\pi^j(\sigma)$, $j = \mathrm{M},\mathrm{F}$, is the (ex ante) payoff to the male and female player, respectively, accruing to strategy σ. Eqs. (2)–(3) may be interpreted as a model of an asynchronous decision making process in which only a tiny (i.e., measure zero) number of individuals choose whether to change their type at each given instant; those individuals who have to choose change their type if and only if the other strategy is more rewarding at that time. The fact that only a negligible number of individuals may change their mind at each given time explains the smoothness of the dynamics; i.e., even if one strategy strictly dominates the other for every possible distribution of types one observes a smooth, rather than a one-shot, population shift toward the more rewarding strategy.

It is apparent that the Cartesian product of $\tilde{\mu} = \{0,1\}$ and $\tilde{\nu} = \{0,1\}$ is a subset of the set of stationary points of the replicator system; i.e., any pattern of 'chivalrous' and 'non-chivalrous' conventions for the two kinds of players is in principle a state of the population that, if reached, is never abandoned unless a perturbation occurs. An interesting question is whether a given stationary point is robust w.r.t. perturbations, i.e., whether it is stable under the replicator dynamics. Notice that stable stationary points correspond to social conventions in the above specified sense, in that they are customary and expected (being stationary points of the ex ante payoff dynamics) as well as self-enforcing (no small subset of players find deviations from the equilibrium distribution of strategies rewarding). One wonders moreover whether there are mixed stationary points in which both types are observed with positive frequencies among players of a given kind, as well as whether such points are stable.

In order to answer these questions, let us compute explicitly the payoffs that are associated to each combination of strategies. For $j = \mathrm{M},\mathrm{F}$, one has

$$\pi^j(\sigma_1|\sigma_1) = \omega_0^j\alpha^2 - \omega_2^j\alpha + \delta_j, \tag{4}$$

$$\pi^j(\sigma_1|\sigma_2) = \omega_0^j\alpha\beta - \omega_1^j\beta + \delta_j(1-\alpha), \tag{5}$$

where $\omega_0^j \equiv \gamma_j + \delta_j + \eta_j$, $\omega_1^j \equiv \delta_j + \eta_j$, $\omega_2^j \equiv 2\delta_j + \eta_j$. After a few tedious computations, it follows that

$$\pi^j(\sigma_1) = \left(\omega_0^j\alpha - \omega_1^j\right)\left[\alpha k + \beta(1-k)\right] + \delta_j(1-\alpha) \tag{6}$$

where $k = \mu$ when $j = \mathrm{F}$, $k = \nu$ when $j = \mathrm{M}$.

Analogously, one has

$$\pi^j(\sigma_2|\sigma_1) = \omega_0^j\alpha\beta - \omega_1^j\alpha + \delta_j(1-\beta), \tag{7}$$

$$\pi^j(\sigma_2|\sigma_2) = \omega_0^j\beta^2 - \omega_2^j\beta + \delta_j, \tag{8}$$

which yields

$$\pi^j(\sigma_2) = \left(\omega_0^j\beta - \omega_1^j\right)\left[\beta(1-k) + \alpha k\right] + \delta_j(1-\beta). \tag{9}$$

It is easy to check that (2)–(3) are now transformed into

$$\dot{\mu} = \mu(1-\mu)(\alpha-\beta)\left[\omega_0^{\mathrm{M}}(\alpha-\beta)\nu + \beta\omega_0^{\mathrm{M}} - \delta_{\mathrm{M}}\right], \tag{10}$$

$$\dot{\nu} = \nu(1-\nu)(\alpha-\beta)\left[\omega_0^{\mathrm{F}}(\alpha-\beta)\mu + \beta\omega_0^{\mathrm{F}} - \delta_{\mathrm{F}}\right], \tag{11}$$

[4] In Section 5 we show how this game can be interpreted as a pure strategy game once available strategies are suitably redefined; we moreover show that, although mixed strategies for the original game need not be Nash equilibria, the equilibrium strategies for the equivalent pure strategy game are always Nash equilibrium strategies. This property is important in that it substantiates the idea that players optimize within the socially predetermined choice set.

[5] We could have chosen different specifications within the larger classes of monotonic selection dynamics as defined by Friedman (1991) [in fact, Friedman calls them order compatible dynamics] or of aggregate monotonic selection dynamics as defined by Samuelson and Zhang (1992). In our specific context, however, this greater generality would have only caused an additional technical complication without any gain in terms of insight.

from which it follows that the interior stationary point, if existing, is unique and has coordinates

$$\hat{\nu} = \frac{\delta_M - \beta\omega_0^M}{\omega_0^M(\alpha - \beta)}, \tag{12}$$

$$\hat{\mu} = \frac{\delta_F - \beta\omega_0^F}{\omega_0^F(\alpha - \beta)}. \tag{13}$$

3. Results

Under our assumptions on parameters, it is apparent that $\omega_0^j(\alpha - \beta) > 0, j = M, F$. In order to have $\hat{\nu} > 0$, $\hat{\mu} > 0$, it is required, respectively, that $\beta < \delta_M/\omega_0^M$, $\beta < \delta_F/\omega_0^F$. On the other hand, it is easy to check that $\hat{\nu} < 1$, $\hat{\mu} < 1$ require, respectively, that $\alpha > \delta_M/\omega_0^M$, $\alpha > \delta_F/\omega_0^F$.

In order to determine the dynamical behavior of the model, it is necessary to distinguish two cases, namely $\delta_M/\omega_0^M > \delta_F/\omega_0^F$ (case I) and $\delta_M/\omega_0^M < \delta_F/\omega_0^F$ (case II). This latter case is the relevant one when the coordination game is close to being symmetric, i.e., when $\gamma_M \approx \delta_F$, $\gamma_F \approx \delta_M$, $\eta_M \approx \eta_F$; the 'classical' coordination game with symmetric payoffs falls therefore within case II.

As to case I, [6] we divide the parameter space into six regions as shown in Fig. 1.

The various regions are characterized as follows:

$$R_1 = \{(\alpha, \beta): \alpha > \beta, \alpha < \delta_F/\omega_0^F, \beta < \delta_F/\omega_0^F\}, \tag{14}$$

$$R_2 = \{(\alpha, \beta): \alpha > \beta, \delta_F/\omega_0^F < \alpha < \delta_M/\omega_0^M, \beta < \delta_F/\omega_0^F\}, \tag{15}$$

$$R_3 = \{(\alpha, \beta): \alpha > \beta, \delta_F/\omega_0^F < \alpha < \delta_M/\omega_0^M, \delta_F/\omega_0^F < \beta < \delta_M/\omega_0^M\}, \tag{16}$$

$$R_4 = \{(\alpha, \beta): \alpha > \beta, \alpha > \delta_M/\omega_0^M, \beta < \delta_F/\omega_0^F\}, \tag{17}$$

$$R_5 = \{(\alpha, \beta): \alpha > \beta, \alpha > \delta_M/\omega_0^M, \delta_F/\omega_0^F < \beta < \delta_M/\omega_0^M\}, \tag{18}$$

$$R_6 = \{(\alpha, \beta): \alpha > \beta, \alpha > \delta_M/\omega_0^M, \beta > \delta_M/\omega_0^M\}. \tag{19}$$

An analogous operation may be carried out for case II, with the obvious caveat that δ_M/ω_0^M and δ_F/ω_0^F must be systematically exchanged in the conditions that define the R_j, $j = 1, \ldots, 6$ (see Fig. 2).

In both cases, an interior stationary point only exists in region R_4. In all other

[6] Notice that assuming $\delta_M/\omega_0^M > \delta_F/\omega_0^F$ amounts to require $\delta_M\omega_1^F > \delta_F\omega_1^M$.

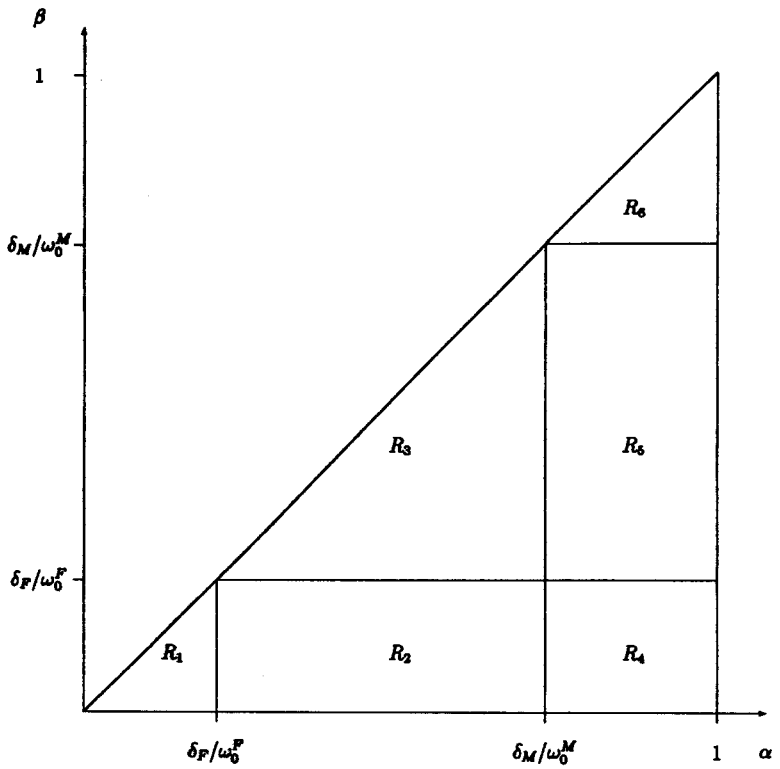

Fig. 1. Case I: $\delta_M/\omega_0^M > \delta_F/\omega_0^F$.

regions, the only stationary points are on the boundary, i.e., one of the available types is always wiped out for each kind of player.

The dynamical properties of our evolutionary process may be completely characterized for both cases I and II. The main features of our results may be summarized in the two propositions below:

Proposition 1. *When* $\delta_F/\omega_0^F < \delta_M/\omega_0^M$ *(case I),* $(\mu = 0, \nu = 0)$ *is globally stable in regions* R_1, R_2 *(i.e., the replicator dynamics select Male Chivalry); in region* R_3, $(\mu = 0, \nu = 1)$ *is globally stable (i.e., Two Sided Chivalry is selected); in regions* R_5, R_6, $(\mu = 1, \nu = 1)$ *is globally stable (i.e., Female Chivalry is selected). Finally, in region* R_4, *one has bistable behavior.*

Proposition 2. *When* $\delta_F/w_0^F > \delta_M/w_0^M$, $(\mu = 0, \nu = 0)$ *is globally stable in regions* R_1, R_2 *(i.e., the replicator dynamics select Male Chivalry); in region* R_3, $(\mu = 1, \nu = 0)$ *is globally stable (i.e., No Chivalry is selected); in regions* R_5, R_6,

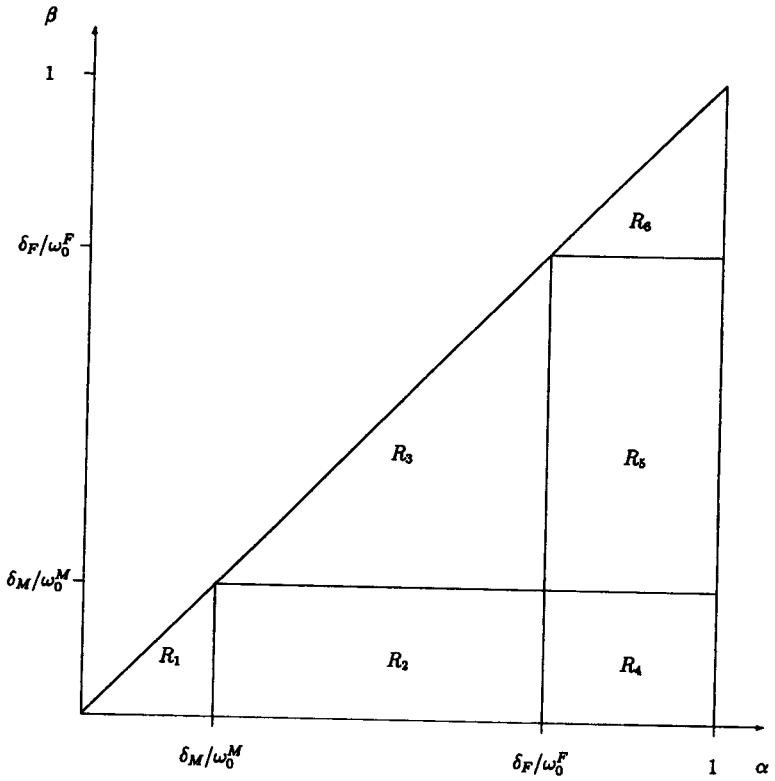

Fig. 2. Case II: $\delta_F/\omega_0^F > \delta_M/\omega_0^M$.

($\mu = 1$, $\nu = 1$) is globally stable (i.e., Female Chivalry is selected). Finally, in region R_4, one has bistable behavior.

By bistable behavior we mean that both ($\mu = 0$, $\nu = 0$) and ($\mu = 1$, $\nu = 1$) are (asymptotically) stable stationary points and that, depending on the initial distribution of strategies (μ^0, ν^0), one of them is eventually reached by the population dynamics. To be precise, this is true for a generic choice of the initial conditions; there is also a 'small' set of initial conditions for which the population dynamics converge to the interior equilibrium (see Section 4 for details). Roughly speaking, then, under bistable behavior either Male or Female Chivalry prevails according to whether at the beginning of the process a relatively large number of both male and female players stick to the former or to the latter convention.

A more detailed characterization of results, together with an illustration of technical details and proofs, will be given in Section 4. In the remainder of this section, we will instead discuss the results just presented.

A first conclusion that emerges from the comparison of Propositions 1 and 2 is

the fact that the only real difference between cases I and II in terms of selection of conventions has to do with region R_3, i.e. with the case in which both α and β lie in an 'intermediate' range of values. In case I, in region R_3 Two Sided Chivalry prevails, whereas in case II it is No Chivalry that prevails. Consequently, it turns out that a necessary condition for the emergence of the Two Sided Chivalry convention is the existence of some degree of asymmetry in the payoff structure of the coordination game; in the classical, symmetric version of the game Two Sided Chivalry cannot be observed as an equilibrium convention.

Figs. 1 and 2 illustrate the dependence of the social convention that is selected by the evolutionary dynamics on the actual structure of the socially feasible choice set faced by players. Notice in particular that whenever the choice set contains a pure Nash equilibrium strategy for the coordination game (1), the corresponding social convention is always selected (i.e., pure Nash equilibrium strategies are asymptotically stable points for the dynamics). [7] It is important to notice that, even when a 'chivalrous' convention is selected, players in this model are not assumed to behave altruistically, in the sense of being concerned with the payoff collected by some other player. When choosing to behave in a chivalrous way, our players are simply choosing a convention that warrants a (relatively) higher probability of meeting the other player, even if at the cost of an unlikely choice of the preferred meeting place. [8]

Let us now see this point in more detail. To fix ideas, consider e.g. the point of view of the male player. [9] If, say, ν is close to one, the vast majority of female players is choosing strategy σ_1, i.e. the strategy that assigns more probability weight (α) to place S. If α is relatively large, i.e. if the fact that most female players choose σ_1 is a relatively clear indication of S as a likely meeting place, then male players will certainly be willing to choose σ_1 in turn and then Female Chivalry will come about. In fact, from Figs. 1 and 2, one sees that α is high in parameter regions R_4–R_6; these are precisely the regions in which Female Chivalry is (asymptotically) stable. One also sees that, the higher β (i.e., the higher the probability of S as a successful meeting place when the other player chooses *the other* strategy, namely σ_2), the more robust the stability of the Female

[7] In particular, when the choice set is made up of the two pure Nash equilibrium strategies (i.e. point (1,0) in the (α, β) space), bistable behavior occurs: the actual equilibrium that is selected depends on the initial distribution of strategies across players.

[8] This is of course does not mean that altruistic behavior cannot be analyzed within an evolutionary framework (see e.g. Hirshleifer, 1982), or even that altruistic motivations do not play a part in the establishment of 'chivalrous' habits in some real-life contexts. The definition of what is precisely meant by 'altruistic' behavior is however always somewhat context-dependent and requires a good deal of care; see e.g. Sacco and Zamagni (1993). For this reason, we do not explore further this point in the present paper.

[9] The discussion that follows may also be rationalized in terms of Eqs. (20)–(23) below.

Chivalry convention. [10] In particular, when β is very low, the dynamics will not converge to Female Chivalry if few male players (and/or not enough female players) choose σ_1. As β (and hence α) increases, the condition that a high enough number of players (male and/or female) initially choose σ_1 becomes less crucial.

An analogous reasoning holds from the point of view of the female player when μ is close to zero: if β is small enough, Male Chivalry will certainly come about. [11]

To sum up, if the characteristics of the available random mechanisms are such that the probability of coordinating on their preferred outcome are small for players of a given type, these players prefer to choose the 'chivalrous' option in order to increase the probability of coordination, on the basis of the (correct) belief that players of the other kind will not choose their 'chivalrous' option.

The interpretation of dynamical behavior for parameters belonging to region R_3 is somewhat more complex. Here neither α nor β are definitely 'small' or 'large'; therefore, there is no clear indication as to the relatively more likely meeting place. A more stringent criterion is then needed. The discriminating condition now becomes whether case I or case II applies; in order to understand whether α (resp., β) is large enough (resp., small enough) to warrant adoption of a non-chivalrous strategy for the male (female) player (or vice versa) one needs to compare its value with the yardstick ratio δ_j/ω_0^j, $j = \mathrm{M}, \mathrm{F}$. One has that, for the sake of likely coordination, α (resp., β) can be considered large enough (resp., small enough), when it is larger than $\delta_{\mathrm{M}}/\omega_0^{\mathrm{M}}$ (resp. smaller than $\delta_{\mathrm{F}}/\omega_0^{\mathrm{F}}$); for a technical derivation, see Section 4 below.

In case I one has $\alpha < \delta_{\mathrm{M}}/\omega_0^{\mathrm{M}}$ and $\beta > \delta_{\mathrm{F}}/\omega_0^{\mathrm{F}}$, i.e., we have that at the same time α is too small to make meeting at place S likely enough for the male player and β is too large to make meeting at place B likely enough for the female player to warrant adoption of the respective non-chivalrous strategies. As a consequence, the Two Sided Chivalry convention is selected. Exactly the opposite happens in case II: both α and $1-\beta$ are large enough to persuade both sorts of players to choose the strategies that assign the larger probability weight to their respectively preferred places, viz., No Chivalry is selected.

Finally notice that, for the special case $\delta_{\mathrm{F}}/\omega_0^{\mathrm{F}} = \delta_{\mathrm{M}}/\omega_0^{\mathrm{M}} \equiv \chi$, regions R_2, R_3 and R_5 collapse; therefore, only instances of one-sided chivalry may emerge in equilibrium, depending on the choice of parameter values. When $\alpha > \chi$, $\beta < \chi$, the result also depends on the initial distribution of strategies for the two kinds.

[10] In terms of Eq. (20) below one sees that α must be regarded as 'large' if it is larger than $\delta_{\mathrm{M}}/\omega_0^{\mathrm{M}}$ and 'too small' otherwise.

[11] In this respect, Eq. (20) below dictates that β must be regarded as 'small' (from the point of view of the male player) if it is smaller than $\delta_{\mathrm{M}}/\omega_0^{\mathrm{M}}$ and 'too large' otherwise. By the same token, on the basis of Eq. (23) below, α and β are 'small' from the point of view of the female player if lower than $\delta_{\mathrm{F}}/\omega_0^{\mathrm{F}}$, and 'large' otherwise.

4. Proofs and technical characterization

This section contains the proofs of the results presented in the previous section as well as a more detailed technical characterization of the dynamics. It is meant for the technically motivated reader. Readers with non-technical interests may skip it without loss of continuity.

In order to determine the stability properties of the replicator system (10)–(11), we compute its Jacobian. In spite of the fact that stationary points typically lie on the boundary, their stability character may be determined by means of the Jacobian because the unrestricted flow associated to (12)–(13) is smooth all over $\Re^2$ (see e.g. Hirsch and Smale, 1974). It turns out that

$$\frac{\partial \dot{\mu}}{\partial \mu} = (1-2\mu)(\alpha-\beta)\left[\omega_0^{\mathrm{M}}(\alpha-\beta)\nu + \omega_0^{\mathrm{M}}\beta - \delta_{\mathrm{M}}\right], \tag{20}$$

$$\frac{\partial \dot{\mu}}{\partial \nu} = \mu(1-\mu)(\alpha-\beta)^2\omega_0^{\mathrm{M}}, \tag{21}$$

$$\frac{\partial \dot{\nu}}{\partial \mu} = \nu(1-\nu)(\alpha-\beta)^2\omega_0^{\mathrm{F}}, \tag{22}$$

$$\frac{\partial \dot{\nu}}{\partial \nu} = (1-2\nu)(\alpha-\beta)\left[\omega_0^{\mathrm{F}}(\alpha-\beta)\mu + \omega_0^{\mathrm{F}}\beta - \delta_{\mathrm{F}}\right]. \tag{23}$$

On the basis of the above information, it is possible to associate to each of the six regions a typical dynamic regime. To this we turn our attention now. It is easily checked that off-diagonal entries are both zero for μ, $\nu = 0, 1$. The Jacobian is therefore always diagonal for every stationary point on the boundary. This implies that eigenvectors may always be chosen as the standard orthonormal basis for $\Re^2$, i.e., as the couple of unit vectors $e_1' = (1,0)$, $e_2' = (0,1)$.

To fix ideas, consider case I and start from region R_1. The main diagonal entries at $(0,0)$ are equal, respectively, to $(\alpha-\beta)(\omega_0^{\mathrm{M}}\beta - \delta_{\mathrm{M}})$, $(\alpha-\beta)(\omega_0^{\mathrm{F}}\beta - \delta_{\mathrm{F}})$. For $(\alpha,\beta) \in R_1$, $(0,0)$ is therefore a stable stationary point under the replicator dynamics. At $(1,1)$, the main diagonal entries are equal to $-(\alpha-\beta)(\omega_0^{\mathrm{M}}\alpha - \delta_{\mathrm{M}})$, $-(\alpha-\beta)(\omega_0^{\mathrm{F}}\alpha - \delta_{\mathrm{F}})$, respectively. For $(\alpha,\beta) \in R_1$, they are both positive, i.e., $(1,1)$ is unstable under the replicator dynamics. Analogously, one finds that for $(0,1)$ and for $(1,0)$ only one main diagonal entry is negative, whereas the other is positive. These points therefore display saddle instability under the replicator dynamics. Since each stationary point on the boundary has eigenvectors which form a standard orthonormal basis, the flow along the boundaries is completely characterized by the sign of the eigenvectors.

We come now to region R_2. Since for $(\alpha,\beta) \in R_2$ it is still true that $\beta < \delta_{\mathrm{M}}/\omega_0^{\mathrm{M}}$, $\beta < \delta_{\mathrm{F}}/\omega_0^{\mathrm{F}}$, one has again that $(0,0)$ is globally stable. It is easy to check, however, that now $(1,1)$ is a saddle point whereas $(1,0)$ has now become globally unstable.

We have therefore shown that

Lemma 1. For $(\alpha, \beta) \in R_1$, $(\mu = 0,\ \nu = 0)$ is a globally stable stationary point. That is, both male and female players are all of type 2 in equilibrium. This means that in this region the replicator dynamics select Male Chivalry. Moreover, $(\mu = 1,\ \nu = 1)$ is a globally unstable point. The other stationary points on the boundary display saddle instability. The same kind of regime is observed in region R_2, except for the fact that $(\mu = 1,\ \nu = 0)$ is now globally unstable whereas $(\mu = 1,\ \nu = 1)$ displays saddle instability.

Lemma 1 says that the Male Chivalry convention is selected for almost all choices of initial conditions. Only in the case where players of some kind are all initially adopting the same strategy different outcomes are possible. For example, in region R_1, when ν is initially equal to one, Two Sided Chivalry is eventually observed, i.e., each kind plays the strategy that assigns a higher probability weight to the place preferred by the other kind.

Consider now region R_3. As to $(0,0)$, although it is still true that $\beta < \delta_M/\omega_0^M$, one has that $\beta > \delta_F/\omega_0^F$. $(0,0)$ is therefore a saddle point. The same can be said for $(1,1)$, since $\alpha < \delta_M/\omega_0^M$ but $\alpha > \delta_M/\omega_0^F$. On the other hand, it is easy to check along the same lines that $(1,0)$ is globally unstable whereas $(0,1)$ is globally stable. We have therefore shown that

Lemma 2. For $(\alpha, \beta) \in R_3$, $(\omega = 0,\ \nu = 1)$ is a globally stable stationary point. That is, male players are all of type 2 in equilibrium whereas female players are all of type 1. This means that in this region the replicator dynamics select Two-Sided Chivalry. Moreover, $(\mu = 1,\ \nu = 0)$ is a globally unstable point. The other stationary points on the boundary display saddle instability.

The following lemma may be proved along the same lines:

Lemma 3. For $(\alpha, \beta) \in R_5$, $(\mu = 1,\ \nu = 1)$ is a globally stable stationary point. That is, both male and female players are all of type 1 in equilibrium. This means that in this region the replicator dynamics select Female Chivalry. Moreover, $(\mu = 1,\ \nu = 1)$ is a globally unstable point. The other stationary points on the boundary display saddle instability. The same kind of regime is observed in region R_6, except for the fact that $(\mu = 0,\ \nu = 0)$ is now globally unstable whereas $(\mu = 1,\ \nu = 0)$ displays saddle instability.

Region R_4 requires a more complex analysis, since an interior stationary point now exists. $\alpha > \delta_M/\omega_0^M$, $\beta < \delta_F/\omega_0^F$ imply that both $(0,0)$ and $(1,1)$ are stable. On the other hand, at $(\hat{\mu}, \hat{\nu})$ the Jacobian is no longer diagonal. More specifically, it is easy to check that the only nonzero entries are now those which are not on the main diagonal. This implies that eigenvalues always have opposite sign, i.e., $(\hat{\mu}, \hat{\nu})$ is a saddle point.

Being Eqs. (10)–(11) separable in μ and ν, it is possible to check that the separatrix between the basins of attraction of $(0,0)$ and $(1,1)$ is given by the implicit function

$$e^{(\alpha-\beta)c}(1-\mu)^{\delta_F-\alpha\omega_0^F}\mu^{\beta\omega_0^F-\delta_F} = (1-\nu)^{\delta_M-\alpha\omega_0^M}\nu^{\beta\omega_0^M-\delta_M} \tag{24}$$

where

$$c = \frac{(\delta_M + \alpha\omega_0^M)\log(1-\hat{\nu}) + (\beta\omega_0^M - \delta_M)\log\hat{\nu}}{\alpha-\beta} - \frac{(\delta_F - \alpha\omega_0^F)\log(1-\hat{\mu}) + (\beta\omega_0^F - \delta_F)\log\hat{\mu}}{\alpha-\beta}. \tag{25}$$

In the special case $\beta = 2\delta_M/\omega_0^M - \alpha$, the separatrix may be given in explicit form as

$$h(\mu) = \begin{cases} 1 + \dfrac{\sqrt{1-4g(\mu)}}{2} & \text{for } \mu \le \hat{\mu}, \\ 1 - \dfrac{\sqrt{1-4g(\mu)}}{2} & \text{for } \mu > \hat{\mu}, \end{cases} \tag{26}$$

where

$$g(\mu) = \left[e^{(2\alpha-2\delta_M/\omega_0^M)c}(1-\mu)^{\delta_F-\alpha\omega_0^F}\mu^{-(\delta_F+\alpha\omega_0^F)+2(\delta_M\omega_0^F/\omega_0^M)}\right]^{1/(\delta_M-\alpha\omega_0^M)}. \tag{27}$$

The above discussion is summarized into

Lemma 4. For $(\alpha, \beta) \in R_4$, $(\mu = 0,\ \nu = 0)$ and $(\mu = 1,\ \nu = 1)$ are both stable stationary points. That is, both male and female players are all of type 1 or 2 in equilibrium depending on whether there is a large enough number of players, either male or female, initially choosing that strategy. This means that in this region the replicator dynamics may select for Male or Female Chivalry depending on initial conditions. The other stationary points on the boundary are globally unstable.

Analogous results for case II, as summarized in Proposition 2 above, may be proved along exactly the same lines.

5. The social coordination game as a pure strategy game

In this section we reformulate our coordination game as a pure strategy game. As it will be readily seen, the outcomes of the mixed strategy evolutionary game

which are asymptotically stable under replicator dynamics all correspond to Nash equilibrium outcomes of the new pure strategy game. [12]

To this purpose, we let C stand for the pure strategy 'play chivalrous' (i.e., choose the stochastic mechanism which assigns higher probability to the less preferred outcome) and NC stand for 'don't play chivalrous' (i.e., choose the stochastic mechanism which assigns higher probability to the preferred outcome). The payoff matrix for this game can be easily built; for example, one has

$$\pi^{M}(NC, C|\alpha, \beta) = \alpha^2\gamma_M + \alpha(1-\alpha)(-\eta_M) + (1-\alpha)^2\delta_M. \quad (28)$$

This is the payoff accruing to the male player when he chooses his non-chivalrous strategy and the female player chooses her chivalrous strategy, parametrized by α and β, i.e., the probabilities that define the characteristics of the two available stochastic mechanisms σ_1 and σ_2. The other entries of the payoff matrix may be built in exactly the same way. For example, one has

$$\pi^{M}(C, C|\alpha, \beta) = \alpha\beta\gamma_M + \alpha(1-\beta)(-\eta_M) + (1-\alpha)(1-\beta)\delta_M. \quad (29)$$

We leave the task of computing the missing entries to the interested reader.

At this point, it is easily checked that the Nash equilibrium strategy pair, as parametrized by (α, β), corresponds to the strategy pair that is selected by the evolutionary dynamics in the mixed strategy game. For example, let us determine under what conditions the strategy pair (C, C) (i.e., Two Sided Chivalry) is a (strict) Nash equilibrium. Using (28) and (29), the (strict) Nash equilibrium condition for the male player becomes

$$(\beta-\alpha)[\alpha\gamma_M + \alpha\eta_M - (1-\alpha)\delta_M] > 0 \quad (30)$$

which is easily seen to be equivalent to

$$\alpha < \frac{\delta_M}{\omega_0^M} \quad (31)$$

The Nash equilibrium condition for the female player may be analogously found to be $\beta > \delta_F/\omega_0^F$; the two Nash equilibrium conditions can be simultaneously met only if $\delta_M/\omega_0^M > \delta_F/\omega_0^F$. Keeping in mind the results of Section 3, we therefore conclude that (C, C) is a Nash equilibrium if and only if Two Sided Chivalry is an asymptotically stable state of the replicator dynamics.

Analogous conclusions can be drawn for all other strategy pairs, as parametrized by (α, β).

6. Relationships with the previous literature

The dynamical behavior of multi-population evolutionary games has been extensively studied in the recent literature. Several important results concerning

[12] More on this in Section 6 below.

the game-theoretic properties of stable stationary points for these dynamics have been obtained. In particular, it has been shown that, for the class of aggregate monotonic selection dynamics [13] (to which the replicator dynamics belong among others) Nash equilibria form a subset of the set of stationary population profiles. Moreover, if a certain population profile is reached starting from an initial distribution of types in which all types are represented, then such profile must be a Nash equilibrium. This implies that all asymptotically stable points of these dynamics must be Nash equilibria, and that the same must be true even for saddle points. In fact, one can say more, namely that only 'almost' strict equilibria [as defined by Samuelson and Zhang (1992)] are asymptotically stable. For the replicator dynamics, this condition becomes more stringent: only strict equilibria are asymptotically stable (see Ritzberger and Vogelsberger, 1990; Friedman, 1991; Samuelson and Zhang, 1992). As an immediate corollary, therefore, interior stationary points cannot be stable in the replicator dynamics.

The results just surveyed are important parts of a larger and difficult puzzle: given a certain interesting class of evolutionary processes, what is their *global* behavior? Convergence to a certain stationary equilibrium is a nice property, but convergence may take time. Therefore, we are not only interested in knowing whether our population of players will converge to a stationary distribution, but also *how* the population is going to arrive at the stationary equilibrium. In addition, in principle the evolutionary dynamics need not converge at all to a stationary equilibrium (in view of the previous discussion, think e.g. of the replicator dynamics for a game which possesses no strict equilibria). In this case, understanding how the dynamics behave far from stationary states becomes even more important. [14] We will refer to the above described problem as to the *direct* problem.

The present paper does not add to the understanding of the direct problem just described. Rather, it addresses a different issue that we term the *inverse* problem, and that can be phrased as follows: given an interesting game-theoretic situation, what are the parameter ranges which correspond to the (partially) known, different regimes of the evolutionary process under consideration, derived as a (partial) solution to the direct problem? Are there aspects of the dynamical behavior of the

[13] Selection dynamics are said to be aggregate monotonic if the vector field is a monotonic (Lipschitz) continuous function of the payoffs associated to each (mixed) population profile of strategies and if each boundary face of the strategy simplex is invariant under the dynamics; see Samuelson and Zhang (1992). In other words, it is required that the dynamics move towards relatively more rewarding (mixed) strategies and away from less rewarding ones and that 'absent' strategies are never adopted.

[14] It is important to remark that Samuelson and Zhang (1992) have shown that in two-population aggregate monotonic dynamics only rationalizable strategies are eventually played. If however rationalizable strategies are the only ones played, no clear inference can be drawn, apart from the results on asymptotic stability listed earlier.

specific process under consideration that cannot be inferred from the known general results? If so, what are they?

The solution of inverse problems is often not easy and can be very important in applied work in evolutionary game theory. It is not enough to explain what are the possible behaviors of a given evolutionary process generated by a certain game-theoretic situation; one needs also to know *when* they occur. Ours is an example of a complete solution of an inverse problem associated to what we believe to be an important class of game-theoretic situations, namely coordination games of the battle of the sexes kind. The basic structure of our results is of course in agreement with the insights that can be drawn from our knowledge of the direct problem: the only asymptotically stable profiles are 'boundary' ones and they correspond to strict Nash equilibria once the game is properly reformulated in pure strategy terms (see Section 5 above). Some other characteristics, however, were not deductible a priori: e.g. the fact that all trajectories eventually converge to a stationary point for every parameter configuration, the characterization of the set of initial conditions under which a certain stable stationary point is reached when there are more than one (and thus the relative 'likelihood' of the competing conventions), the 'thickness' of the parameter ranges corresponding to each regime. Another source of interest is the fact that such results admit a simple and clear formulation that covers a relatively large-dimensional parametric family of games.

Of course, if this sort of analysis has to make practical sense, it is important to choose game-theoretic situations which are of interest per se, i.e. as models, no matter how simple and stylized, of meaningful instances of social interactions. The solution of an inverse problem associated to a game, or family of games, which admits no interesting interpretation is pointless, whereas clearly the same cannot be said as far as the direct problem is concerned. Such considerations are of course perfectly in line with the common wisdom underlying sensible applied analysis. Applied evolutionary analysis does not seem to require a methodological revolution w.r.t. the existing body of applied work; in spite of this, it can add substantially to the understanding of the dynamics of social processes. Much work is certainly still to be done; we need more and more careful and detailed theoretical descriptions of the structure of social interaction as well as a better understanding of the causal links that go from the social context to individual choices and vice versa, possibly along the lines suggested in the introductory section. To this end, an accurate analysis of carefully chosen inverse problems is not less important than further progress toward the understanding of the abstract properties of evolutionary processes.

Acknowledgements

We thank Flavio Delbono, Leonardo Felli, Alfredo Medio, Piero Tedeschi and Gerd Weinrich for useful comments and suggestions. The paper has been moreover substantially improved by the comments of anonymous referees of this journal. The usual disclaimer applies.

References

Argyle, M. and M. Henderson, 1985, The anatomy of relationships (Penguin, London).
Berger, P.L. and B. Berger, 1975, Sociology: A biographical approach, Second edition (Basic Books, New York).
Berger, P.L. and T. Luckmann, 1966, The social construction of reality (Doubleday, New York).
Bernheim, B.D., 1986, Axiomatic characterizations of rational choice in strategic environments, Scandinavian Journal of Economics 88, 473–488.
Bicchieri, C., 1990, Norms of cooperation, Ethics 100, 838–861.
Binmore, K., 1990, Nash equilibrium, in: K. Binmore, Essays on the foundations of game theory (Basil Blackwell, Oxford) 43–77.
Brandenburger, A. and E. Dekel, 1987, Rationalizability and correlated equilibria, Econometrica 55, 1391–1402.
Elster, J., 1989, The cement of society: A study of social order (Cambridge University Press, Cambridge).
Friedman, D., 1991, Evolutionary games in economics, Econometrica 59, 637–666.
Goffman, E., 1974, Frame analysis (Northeastern University Press, Boston, MA).
Granovetter, M., 1985, Economic action and social structure: The problem of embeddedness, American Journal of Sociology 91, 481–510.
Granovetter, M., 1992, Economic institutions as social constructions: A framework for analysis, Acta Sociologica 35, 3–11.
Hammond, P.J., 1992, Aspects of rationalizable behavior, Preprint (Stanford University, Stanford, CA).
Hirsch, M.W. and S. Smale, 1974, Differential equations, dynamical systems and linear algebra (Academic Press, New York).
Hirshleifer, J., 1982, Evolutionary models in economics and the law: Cooperation versus conflict strategies, Research in Law and Economics 4, 1–60.
Hofbauer, J. and K. Sigmund, 1988, the theory of evolution and dynamical systems (Cambridge University Press, London).
Lewis, D., 1969, Convention: A philosophical study (Harvard University Press, Cambridge, MA).
Rasmusen, E., 1989, Games and information: An introduction to game theory (Basil Blackwell, Oxford).
Ritzberger, K. and K. Vogelsberger, 1990, The Nash field, IAS research report no. 263 (Institute for Advanced Studies, Vienna).
Sacco, P.L., 1993a, On the dynamics of social norms, CITG papers on game theory no. 12 (InterUniversity Centre of Game Theory, Florence).
Sacco, P.L., 1993b, Convolution of social norms in a noisy environment, Chapter 3, Ph.D. Dissertation (European University Institute, Florence).
Sacco, P.L. and Zamagni, S., 1993, An evolutionary dynamic approach to altruism, Discussion paper no. 165 (Department of Economics, University of Bologna, Bologna).
Samuelson, L. and J. Zhang, 1992, Evolutionary stability in asymmetric games, Journal of Economic Theory 57, 363–391.
Sugden, R., 1989, Spontaneous order, Journal of Economic Perspectives 3, 85–97.
Tan, T.C. and S.R.D.C. Werlang, 1988, The Bayesian foundations of solution concepts of games, Journal of Economic Theory 45, 370–391.