EVOLUTIONARY STABILITY OF DISCRIMINATING SOCIAL NORMS

Katarzyna Abramczuk[1]
Uniwersytet Warszawski

Abstract: *The paper presents an evolutionary model illustrating the dynamics that give rise to discriminatory social norms i.e. such rules of behaviour that fulfil two conditions: (1) they treat differently actors having the same abilities and technical options, but differing in some arbitrary sense (2) they are supported by socially enforced sanctions. In the presented model both discrimination and social norms are necessary to solve a coordination problem that arises when the situation requires different actors to perform different tasks. The properties of behavioural rules relying on discrimination and leading to various degrees of inequality are analysed. It is demonstrated that in general norms ensuring equal payoffs are easier to stabilize, but unfair norms can also be stable.*

Key words: *game theory, third-party sanctions, social norms, social control, discrimination, privilege assignment, social inequalities.*

EWOLUCYJNA STABILNOŚĆ DYSKRYMINUJĄCYCH NORM SPOŁECZNYCH

Streszczenie: *Praca przedstawia model ewolucyjny ilustrujący mechanizmy przyczyniające się do stabilności dyskryminujących norm społecznych, tj. zasad zachowania spełniających dwa warunki: (1) odmienne traktowanie aktorów o tych samych umiejętnościach i możliwościach, ale różniących się arbitralnymi cechami, (2) wsparcie na sankcjach społecznych. W proponowanym modelu zarówno dyskryminacja, jak i normy społeczne są konieczne, by rozwiązać problem koordynacji powstający, gdy sytuacja wymaga, aby poszczególni aktorzy podjęli różne zadania. Analizowane są własności zasad zachowania opartych na dyskryminacji, skutkujących różnymi poziomami nierówności. Pokazano, że w ogólności zasady zapewniające równe wypłaty są łatwiejsze do ustabilizowania, ale niesprawiedliwe normy także mogą być stabilne.*

[1] Katarzyna Abramczuk, Instytut Socjologii Uniwersytetu Warszawskiego, ul. Karowa 18, 00-927 Warszawa, e-mail: k.abramczuk@uw.edu.pl

 DOI: 10.7206/DEC.1733-0092.77

Słowa kluczowe: *teoria gier, sankcje społeczne, normy społeczne, kontrola społeczna, dyskryminacja, powstawanie przywilejów, nierówności społeczne.*

1. Introduction

Discrimination and resulting inequalities are common to all of the known societies. On many occasions it is of no particular surprise when people belonging to distinct social categories are treated differently. For example in an interaction between a professor and a student the preponderance is determined by the hierarchy of dependence and competence. However, it also happens that some social differences are being sustained even though individuals at hand are very similar in terms of their a priori abilities. In modern times it is most apparent in the case of gender and race discrimination, but can also be applied to assignment of rights related to the birth order or the social group of origin. Oftentimes such discrimination is being normatively justified and enforced by a system of social control that teaches each individual about their right place in the great machinery of a society as a whole. Those who do not adjust, are being criticised and ostracised. This in spite of the fact that such discrimination often causes suboptimal allocation of labour and resources and leads to marginalization of some social groups. Sociologists describe this phenomenon in terms of ascribed status i.e. status that is given rather than achieved and can hardly be shaped by an individual effort.

In what follows the evolutionary game theory is used to build a simple model of this situation. I start with a purposefully unrealistic situation of perfect equality and show how, when certain conditions are met, a discriminatory system can be constructed from scratch. Two factors turn out to be crucial here: existence of publicly observable labels or characteristics of the actors, and the mechanisms of social control. I start with a positive example in which the two combined enable actors to solve off-diagonal coordination dilemmas. I move on to analyse the stability of this solution and its consequences for equality.

The paper contributes to the literature in two ways. First, and this is its primary purpose, it is an abstract exercise. It provides an insight into how social expectations can endow certain categories with meaning. In particular, it shows that no a priori justifiable differences are needed for unequal treatment to become a social norm that will be self-enforcing. Second, it broadens the applicability of the solution proposed in [13] beyond the class of games of enforceable cooperation. I proceed to explain the nature of the two contributions.

 DOI: 10.7206/DEC.1733-0092.77

The inequalities in the following model arise on the basis of some labels or characteristics possessed by actors. It is important to underline that these characteristics are assumed to fulfil two conditions that are rarely applied in the related literature. First, they are not chosen and cannot be changed. They can be thought of as representations of features such as gender, skin colour, or family of origin rather than status symbols, consumption styles, or education level. Second, they are originally completely irrelevant for the game structure. They do not influence in any direct way actors' abilities, options, or payoffs. Their influence results solely from the social mechanisms that endow them with meaning. This approach is close to theories postulating that social categories are constructed and reified via mechanisms of social control. The process is often described in terms of symbolic or ideological power. It stabilizes social structure and makes it appear to be something more than a social construct [26]. The presented model shows in a formal way how this can happen.

Throughout the paper a population of actors is assumed to play an iterated two-person symmetrical game. A special interest is given to games in which the maximal possible total payoff can only be achieved, if the partners choose different rather than the same actions. Such games are sometimes presented as asymmetrical complementarity dilemmas [18, 20], games of specialization [29] or games of division-of-labour [13]. They include many well known dilemmas such as the chicken game, hawk-dove game or the symmetrical battle of the sexes. They cover a wide spectrum of real life situations in which partners are expected to perform different tasks, but the tasks vary in their desirability. For example one is associated with some public or easily transferable good like food production or readiness to concessions, while the other is related to production of goods of a more private character such as education, high status, or greater mobility.

It is important to note that games, for which the most beneficial outcome from the utilitarian point of view lies off the diagonal, are not the only games in which the proposed solution can be applied. The focus on them comes from the fact that they often do not fall in the category of games of enforceable cooperation [13]. For games of enforceable cooperation there is a known stable solution that is also a starting point for the following analyses. This solution rests on the idea that diagonal cooperation can be stable, if there exists an effective punishment that is consistently applied by the whole community and disciplines the norm breakers. In what follows I reiterate the result presented in [13] stating that no cooperation is possible without a normative meta-structure. I start the analysis with showing how this meta-structure can be used to solve other types of dilemmas. The main point is that it may result in assigning meaning to originally meaningless labels and legitimization of inequalities.

 DOI: 10.7206/DEC.1733-0092.77

There is plenty of literature devoted to relations between cultural and normative system and social inequalities. For example Ralf Dahrendorf [21] tried to derive social inequalities from the existence of social norms, while sometime later Edna Ullmann-Margalit [49] argued that it is social inequality which is a simpler and more familiar notion and should serve as an explanans in derivation of social norms. In this paper a different perspective is taken. Neither norms nor inequalities are considered part of the natural order. Instead they are seen as causally intertwined elements of an evolutionary process. Norms play an active role in establishing stability and inequalities are a by-product of this mechanism.

The remainder of the paper is structured as follows. First, I present a short literature review focused on existing formal models of social discrimination. Second, I present the basic assumptions and definitions for the current work. I proceed to discuss the problem of instability of off-diagonal coordination and move on to introduce the existing solution proposed by Bendor and Swistak [13] for diagonal cooperation. Next, I show how an analogous mechanism coupled with some observable actors' characteristics can be used to facilitate off-diagonal coordination. I discuss the properties of this solution and its consequences for equality. The paper closes with a summary and a discussion.

2. Related work

There are several strains of literature that are relevant for the current work. In particular, there is a number of papers on the evolution of cooperation and reputation. The discussion starts with the famous book by Robert Axelrod [7] and continues throughout diverse models of reputation dynamics and various schemas enabling cooperation (e.g. [14, 15, 30, 37, 38, 40, 47]). I will not report on the details of this discussion. An interested reader might consult e.g. [39] for a short review.

An important sub-stream of this literature is constituted by papers concentrated on the role of third party punishment and third party information in stabilizing social systems. Such a widened notion of reputation is present e.g. in [44], where perfect third party information facilitates efficiency in the equilibrium, or in [33], where the effects of transmitting third party information in networks are analysed. A paper from this line of literature that is most important for the current investigation and will be discussed in detail later on is the work of Jonathan Bendor and Piotr Swistak [13]. Their ideas will be used extensively especially in sections 4 and 5. The general problem these authors aimed at solving is the problem of the evolution of social norms i.e. rules of behaviour that are guarded by third party sanctions. In [13] it is

 DOI: 10.7206/DEC.1733-0092.77

shown that this type of rules is, as a matter of fact, indispensable for stability of any rule of behaviour in a wide class of games called games of enforceable cooperation.

Furthermore, various formal models in which discrimination is being discussed are relevant. The definition of discrimination in this paper differs slightly from the vernacular usage. I say that a decision is discriminatory, if it is based on certain characteristics or labels of actors or of the situation. It does not have to necessarily be followed by an unequal distribution of benefits. Yet, typically the main purpose of attempts to model discrimination is to explain how such unequal distributions arise. The primary assumption that is common to all these approaches is that for the inequalities to appear no master plan has to be involved. Paraphrasing Hayek [27, p. 159], we are freed from "inability to conceive of discrimination in human activities without deliberate organization by a commanding intelligence". Emergence is all it takes. Emergence, however, can have it easier or harder.

Many models aiming at explaining inequalities resulting from discrimination assume that the inequalities follow from differences in priorities, abilities, resources, and/or opportunities of different members of the society. Examples of this type of thinking can be found in the rich literature on gender discrimination (e.g. [3]) and in papers based on such theories as human capital [10], resource bargaining [17, 35], economic dependency [2, 17], compensating differentials [32], statistical discrimination [4, 41], or dual labour markets [9]. The current paper starts with an assumption of perfect initial equality. More specifically two basic assumptions are made. First, payoffs do not depend directly on the characteristics/labels of the actors. Second, the acquirement of characteristics is independent of payoffs, and the characteristics are assumed to be fixed. I will discuss these assumptions shortly.

In many models characteristics openly constitute the basis for the calculation of benefits. This approach is common in biological models, in which it is highly fitting. For example, in the generalized hawk-dove game payoff matrices can depend on a relation between sizes of the actors [19, 20]. This is because the size is assumed to be correlated with the chance of winning the fight. A similar approach has also been used in a social context. In [52] an unequal division of housework is studied. The authors assume that the symmetrical (in relation to types of players) configurations of actions yield asymmetrical configurations of payoffs. The assumption is justified by existence of a tie between social embeddedness and a potential cost of breaking up a contract. Yet, the possible sources of such a tie are not specified. These very sources are modelled explicitly in the present work. The tie originates from the social control. The characteristics themselves do not change the symmetrical nature of the game under consideration. Still, identical actors can be treated differently. A similar perspective is assumed for example in [31], where individual actors build models of the society from their experiences hence influencing the social structure.

In some approaches actors' characteristics do not influence payoffs directly but their acquirement becomes an object of the game. This is the case in the games of status [43] and in other games in which payoff is a direct indicator of one's position [42]. In other instances characteristics change randomly. For example, an asymmetrical banknote game in [47, p. 14] differs from the corresponding symmetrical game only in that, that the actors' positions have different names. However, a "cross-cutting asymmetry" is assumed. The position a particular player takes is random. All the actors have a non-zero probability of finding themselves in each of the positions at some point of the game. It does not have to be equal for all the participants and for all the positions, but it cannot be zero for anybody. An example described most elaborately are conventions considering right of way on the crossroads[2]. A very interesting intermediate case is described in [36], where actors' markers (labels) do not have any value themselves but nonetheless are inherited in a similar way as rules of behaviour. The considered markers are largely a matter of choice and involve such traits as style of dress or speech. A new generation follows both the norms and the traits of the most successful individuals from the previous generation.

By comparison, in the current work the characteristics are given to actors when the population is forming and are essentially unchangeable. Also the distribution of characteristics is stable. These two features are typical for such real life traits as gender, race, country of origin. In a limited time frame they are also in effect for birth order, status of the family of origin etc. The meaningfulness of these traits is derived from the social context. It is worth noting though, that the sole fact that they cannot be altered might contribute to their essentialization and reification. For example, in [46] it is argued that social categories insusceptible to modifications are perceived as natural kinds by human cognition and are hence seen as less arbitrary than the categories that can be altered.

3. The model

This paper is based on the model proposed by Bendor and Swistak [13] which utilizes simple social categorizations to stabilize cooperation in what authors call games of enforceable cooperation. A game of enforceable cooperation is a symmetric two-person game with a feasible punishment action. For the purpose of this paper I will define a wider class of games that will be called games of enforceable coordination.

[2] One of the formal advantages of this approach is, that it does not involve interpersonal comparisons of utilities, but rather comparisons of utilities, of the same actor in different positions. Such comparisons, however, are natural in a model that aims at modelling an ideal abstract equality.

 DOI: 10.7206/DEC.1733-0092.77

3.1. Population

I start with an unstructured group of n actors ($n \in \mathbb{N}$), where everybody interacts with everybody else[3]. The set of actors is denoted by $\mathcal{N} = \{1, 2... n\}$. Each actor possesses one of two characteristics (types) from a set $\mathcal{C} = \{c_1; c_2\}$[4]. The profile of characteristics is a vector of types assigned to all the actors. It is denoted by $g \in \mathcal{C}^n$. Two basic assumptions hold. First, the characteristic of a given actor is his/her fixed feature that cannot be changed or imitated by an actor with a different characteristic (g is constant). Second, the characteristics are publicly and infallibly observed. Everybody always perceives the characteristic of a given actor in the same way. Still different actors may assign different meanings to it.

3.2. Game

In a one-shot game each actor can choose an action from a set $\mathcal{A} = \{a_1,... a_k\}$. The payoffs in a one-shot game are described by a payoff function $\mathcal{V} : \mathcal{A}^2 \rightarrow \mathbb{R}$. An actor using action a_k against action a_l ($1 \leq k$; $1 \leq k$) earns payoff $\mathcal{V}(k; l)$. Without loss of generality I assume that the actions are numbered so that $\mathcal{V}(1; 1) \geq \mathcal{V}(2; 2) \geq ... \geq \mathcal{V}(k, k)$. I put $\mathcal{V}(1, 1) = R$. I define a coordination point as a pair of action (a_k*; a_l*) such that:

$$\mathcal{V}(k^*, l^*) + \mathcal{V}(l^*, k^*) = \max_{k,l} \big(\mathcal{V}(k, l) + \mathcal{V}(l, k)\big) = 2T \tag{1}$$

Thus the coordination point is the most desired result from the utilitarian point of view, where the sum of payoffs for both actors is maximized. The coordination point can lie both on and off the diagonal of the game matrix and can reward the two actors with the same or very different payoffs[5]. The game will be classified as a game of enforceable coordination if there exists a punishment action a_p such that:

$$\max_{1 \leq k \leq \kappa} \mathcal{V}(k, p) < \frac{T + R}{2} \tag{2}$$

In one game there might be more than one such punishment action. The most efficient punishment ensures that the punished actor earns no more than his/her minimax payoff. This payoff will be denoted by P. Hence, the condition (2) given above can also be written as $P < \frac{T+R}{2}$. Additionally it is assumed that the game is

[3] For convenience it is also assumed that actors play with themselves.

[4] In principle the number of characteristics could be larger than two or even infinite. This paper, however, is limited to the binary case with varied proportions of the two types. Two characteristics suffice for the current purpose. It should be noted, however, that this scenario limits the class of games in which the proposed solution can be applied. Allowing for a larger number of types would permit revising the definition of games of enforceable coordination that follows.

[5] Note, that a game can have a few coordination points. Some of them may be located on and some of them may be located off the game matrix diagonal. Because the game is symmetric, in the latter case there necessarily exist at least two coordination points on the two sides of the diagonal.

 DOI: 10.7206/DEC.1733-0092.77

not trivial i.e. there is no action that yields the game's maximal payoff regardless of action chosen by his/her partner.

The class of games of enforceable cooperation as characterized by Bendor and Swistak [13] is a subset of the class defined above. A game is a game of enforceable cooperation iff the minimax payoff is smaller than R. One example of such a game is the famous Prisoner's Dilemma. In the case of Prisoner's Dilemma the coordination point (mutual cooperation) lies on the diagonal of the game matrix. However, in some games of enforceable cooperation the coordination point may lay off the diagonal, as some pair of different actions can ensure larger sum of payoffs than (a_1, a_1). To avoid confusion throughout the following discussion I will be using an example of a game that is a game of enforceable coordination, but is not a game of enforceable cooperation - the Complementary Tasks Game (CTG). Its matrix is presented in Table 1.

Table 1
An exemplary payoff matrix of a game of enforceable coordination – CTG

	B	**C**	**N**
B	2 2	4 6	2 2
C	6 4	0 0	0 0
N	2 2	0 0	0 0

In the CTG actors can choose between performing two different tasks and doing nothing. Action *N* stands for the latter – not performing any work. *B* is some basic job that is necessary for any benefits to arise. Each player can consume 2 units from it, no matter who performed the task. *C* is a complementary task that can only bring benefits, if the partner chooses *B*. It gives 4 units of benefit to the person who performs it and 2 units to the partner. The *B* task can be thought of as leading to production of some public or easily transferable good. The *C* task has a more private character or has some additional value in itself e.g. boosts prestige of the performing actor, increases his/her skills etc. There are two coordination points in this game: (*C*, *B*) and (*B*, *C*). The maximal sum of payoffs equals 10. Action *N* is the punishment action. The best one can do against it is to choose *B* and earn 2, which is way less than half of the maximal total payoff. At the same time the game is not a game of enforceable cooperation. The maximal payoff on the diagonal is 2 and there is no punishment action that would ensure that an actor gets payoff smaller than this.

CTG will be used in the following sections as an illustration. In general, however, the presented results are valid for all games of enforceable coordination that serve as stage game in a repeated play. I assume that a stage game is iterated. A standard simplifying assumption is that the payoff matrix stays the same across all iterations.

 DOI: 10.7206/DEC.1733-0092.77

Each interaction is continued independently of all the other interactions with probability δ ($0 < \delta < 1$) that is generally assumed to be *sufficiently high* i.e. sufficiently close to one (see section 3.4). As a consequence the game ends a.s. after a finished number of rounds. Yet, after each iteration it is generally expected to continue [5].

3.3. Strategies

Similarly as in [13] I consider only pure strategies. I also make two non-standard assumptions concerning the strategy space. First, I assume that strategies can be social. Second, I assume they can be discriminatory. Precise definition of what is meant by this requires some additional notation. Let h^t_{ij} denote the action chosen by actor i in his/her interaction with j in round t, where $t \geq 0$. An $n \times n$ matrix of all the actions chosen in round t will be denoted by $H_t = (h^t_{ij})_{i,j \in \mathcal{N}}$. It will be called a t history. For the sake of completeness assume that $H_0 = \{0\}$. In general a strategy of an actor i is a sequence of mappings that in each round t, for each possible combination of a vector of histories up to this round and a vector of actors' characteristics g, determines a vector of actions chosen by i in round t in interactions with all actors in $\mathcal{N}$. It can be written as:

$$S_i((H_\tau)_{\tau=0}^{t-1}, g) = (h^t_{ij})_{j \in \mathcal{N}} \tag{3}$$

The part of a strategy governing the choice of action in interaction with a specific partner j will be denoted by:

$$S_{ij}((H_\tau)_{\tau=0}^{t-1}, g) = h^t_{ij} \tag{4}$$

Please note the differences between this definition of a strategy and a more traditional approach. Traditionally a strategy in a repeated interaction is thought of as a complete plan of a game given the heretofore course of the interaction. The most famous example of a strategy in this meaning is Tit For Tat [7] meant to solve the problem of cooperation in the repeated Prisoner's Dilemma. TFT uses only one piece of information – the information about the most recent action of the partner. It simply states that an actor entering a new interaction should cooperate, and next they should repeat the last action of their partner. In general, the strategies of this type can use information about any of the previous doings of the partner, as long as these are executed within the given pair. In other words, an actor can punish or reward his/her partner for actions taken towards him/her, but cannot interfere with what the partner chooses to do, when interacting with other actors in the group. Strategies with this property will be called dyadic. Formally a strategy of actor i is called dyadic, if for all j and for all t it satisfies the following condition:

$$S_{ij}((H_\tau)_{\tau=0}^{t-1}, g) = S_{ij}((h^\tau_{ij}, h^\tau_{ji})_{\tau=0}^{t-1}, g) \tag{5}$$

 DOI: 10.7206/DEC.1733-0092.77

I follow Bendor and Swistak in revoking the above-mentioned constraint and allow the actors to base their decisions on the whole history of all the interactions in the group. Hence the actors can punish and reward their partners for what they do both to them and to the others. Strategies with this property, i.e. strategies that do not satisfy (5), will be called social. It is important to note that using social strategies does not necessarily imply that actors store away all the events that take place for a future reference. As a matter of fact, it suffices if they use the information to run and update a classification of all the members of the group. The precise nature of this mechanism will be described in the next section.

Further difference lies in allowing the actors to use information about characteristics profile. This means that a strategy may prescribe different choices for actors of different types or against actors of different types. It may also evaluate differently the same action depending on the characteristic of the actor who performed it or the characteristic of his/her partner. Strategies that use this kind of conditioning will be called discriminatory. By contrast a nondiscriminatory strategy does not differentiate between actors of different types. Formally a strategy of actor *i* is called non-disciminatory, if for all possible *g* and for all *t* it satisfies the condition:

$$S_i((H_\tau)_{\tau=0}^{t-1}, g) = S_i((H_\tau)_{\tau=0}^{t-1}) \tag{6}$$

It is important to make a distinction between discriminatory and non-anonymous strategies. The latter allow the actors to vary their actions depending on the identity of their partner. I preclude this type of strategies. The only strategies that will be considered in this paper are anonymous. As long as two actors have the same characteristic and make exactly the same choices, they are treated in the same way. Formally, if we apply a permutation to the numbers of agents in $\mathcal{N}$ and an analogous permutation to characteristics in g, the output vector will contain actions with the same permutation.

3.4. Payoffs

The total payoff of an actor *i* from interaction with an actor *j* equals an expected normalized sum of his/her payoffs throughout the whole such interaction. Any reference to *sufficiently important future* should be interpreted in terms o $\delta > \delta_0$, where $\delta_0 < 1$ is some threshold value that exists. This is mathematically equivalent to computing payoffs given in the following form:

$$\mathcal{V}_i = \lim_{\delta \to 1}(1-\delta)\sum_{t=1}^{\infty}\delta^t \mathcal{V}(h_{ij}^t, h_{ji}^t) = \lim_{\delta \to 1}(1-\delta)\sum_{t=1}^{\infty}\delta^t v_{ij}^t, \tag{7}$$

 DOI: 10.7206/DEC.1733-0092.77

where v^t_{ij} stands for the payoff gained by actor i in interaction with actor j in round t ($t \geq 0$). The expression $(1 - \delta)$ is a normalizaton factor ensuring that the limit is finite.

Because of the symmetry of a game, the symmetric structure of the population, and the anonymity assumption, it is possible to talk about an ecology of strategies rather than a population of actors. I will be replacing the actors' indices with information about their strategies and types. Hence, $V_{S|c}$ will denote a total payoff of strategy S played by type c i.e a total payoff to an actor of type c that uses strategy S. Please note that this payoff depends on the whole composition of the population in terms of both strategies and characteristics. In particular:

$$\mathcal{V}_{S|c} = \sum_{Z,w} p_{Z|w} \times v_{S|c,Z|w}, \tag{8}$$

where Z is a strategy in a strategy set, w is a characteristic in C, $p_{Z|w}$ denotes the frequency of actors that use strategy Z and posses a characteristic w, and $v_{S|c,\ Z|w}$ is a total payoff gained by an actor of type c and using strategy S in an interaction with an actor of type w using strategy Z.

3.5. Stability

Following [13] I will be searching for strategies that are weakly uniformly stable. Here, however, the distribution of characteristics has to be taken into account. Hence, I am interested in strategies, for which there exists at least one binary distribution of characteristics such that, if the strategy is sufficiently common in a population, it does not decrease in frequency regardless of the particular form of the (monotonous[6]) evolutionary dynamics and regardless the strategies used by the remaining actors. For this condition to hold a specific type of unbeatability is needed. It will be defined using some additional notation.

Let $P_S = (p_{S|c1}, p_{S|c2})$ be a vector containing information about frequency of strategy S for both actor types. Let $P = (P_1, P_2, ..., P_L)$, where L is the number of strategies in the ecology, be a vector gathering information about all the used strategies. The sum $p_S = \Sigma_c p_{S|c}$ is a total frequency of strategy S in the population. The sum $p_c = \Sigma_S p_{S|c}$ is a total frequency of type c in the population. The strategy S will be said to be uniformly stable under given distribution of characteristics iff there exists a stabilizing vector P^*_S such that:

1. $0 < p^*_S < 1$,
2. $0 \leq p^*_{S|c} < p_c$, for all c,

[6] In the monotonous evolutionary dynamics actors value more these strategies that lead to better outcomes and are more common in their group.

3. $p^*_{S|c} > 0 \Rightarrow \mathcal{V}_{S|c} \geq \mathcal{V}_{Z|c}$ for any possible strategy Z, for any $c \in \mathcal{C}$, and in any possible P, in which the frequencies of S for various types are equal to or larger than the corresponding $p^*_{S|c}$.

This definition bears some resemblance to a definition of an evolutionarily stable strategy with a uniform invasion barrier [51]. It implies that there is some minimal frequency p^*_S that makes the strategy S robust against any invasion which frequency does not exceed $(1 - p^*_S)$. There are, however, some intricacies that make this notion different.

First, the distribution of characteristics is taken into account. The required frequency of S actors of any type cannot exceed the frequency of this type in the whole population. The strict inequality ensures that a strategy will not be considered stable, if for its stability it is necessary that all actors with certain characteristic use it. If this was the case, a single mutation in this group could lead to losing stability. Second, the possible invasions include invasions consisting of multiple different mutant strategies. I consider it unrealistic to assume only homogeneous mutations. Hence, I allow for several different new strategies to enter the population at the same time in some (originally) small frequency. Third the considered stability is a weak stability. Whenever $p^*_{S|c}$ is larger than zero, S has to be the best possible choice for the c type actors. There cannot be any other strategy that would ensure them higher payoff. Please note that some other strategy may assure equally high payoff to some actors.

In most cases there is more than one vector P^*_S that fulfills the conditions listed above. Two specific cases will be of particular interest. First, I will look for a P^*_S for which the total frequency of S is the lowest. The sum of this vector will be called the total minimal stabilizing frequency of S. Second, I will search for the P^*_S that minimizes the expectations towards any one type i.e. one where $max_c p^*_{S|c}$ is lowest. The sum of this vector will be called the total minimaxing stabilizing frequency[7].

4. The problem of coordination

In this part I use an example to introduce the two building blocks of discriminating social norms i.e discrimination and third-party punishment. The considered example is the CTG i.e. a game that is a game of enforceable coordination but is not a game of enforceable cooperation. I show why achieving an optimal solution in this type of games is problematic. I put forward the intuitive solution in which actors' characteristics are used as coordination devices and show it to be highly unstable.

[7] More precisely, the frequencies at hand give the boundary conditions for stability.

 DOI: 10.7206/DEC.1733-0092.77

Next I present a solution to the problem of stability from [12] that is limited to games of enforceable cooperation.

As already explained in section 3.2, the CTG is a dilemma similar to the battle of the sexes. The optimal combination of actions requires actors to make different choices. There are several possible approaches to solving this problem. One of them is considering mixed strategies that randomize their choices until coordination is achieved and alternate their choices afterwards. This solution is not pursued here, mostly because it is irrelevant for the main goal of the current paper i.e. showing that it is possible to construct discriminative social norms from scratch. It is, however, worth noting that there are two additional reasons to take this approach with care. First, the idea of social norms that expect actors to randomize their choices out of their own accord is sociologically questionable. Rich literature indicates that the concept of randomization is foreign to human cognition. It was shown for example that people do not understand probabilities [8, 48], are basically unable to produce random sequences [45, 50], and that, when faced with random sequences, they keep trying to identify the underlining deterministic patterns [24, 25]. Hence, it is reasonable to assume that sanctioned rules of behaviour will be clear and will preclude decisions governed by chance. Second, considering mixed strategies may lead to a number of nontrivial problems, when uniform stability and thirdparty sanctions are considered. As shown below, third-party sanctions can be used to overcome the (uniform) instability of dyadic strategies[8]. These sanctions, however, require a mechanism allowing for determining, whether a given actor broke the social norm or not. This can prove delicate, when the norm is not deterministic and there is a potential for profitable cheating, which would be the case for unfair mixing rules.

An alternative solution to the problem of coordination is to introduce some publicly observable random signal that would allow the actors to coordinate on the optimal solution. The idea is described in the literature as correlated equilibrium [6]. The signal could indicate, which actor should choose which action or, more subtly, specify who should be the first actor to make a decision. In the latter case one would expect a spontaneous coordination, if the second decision maker uses the stage-game best reply to the choice made by the first decision maker. It is possible to design (socially enforced) strategies based on public signals, that ensure equal payoffs to both partners. Contrary to the first proposed solution, here the randomization is not performed by the actors themselves. The actors simply use a clue provided by the environment. The problem is the origin and the nature of the signal. If we want to understand, how the signal gains its meaning, we need to assume that originally it had none. A nice example of this approach can be found in the aforementioned

[8] The details for the pure strategies case are given in discussion of Theorem 1. The case of mixed strategies has been discussed e.g. in [23, 34].

analysis of the rights of way on the crossroads [47]. In this work I analyse the omitted, but highly relevant from the sociological point of view, case in which the signal used for coordination cannot be changed. Before I show that such signal still can be a coordination device and that it can have a substantial impact on the distribution of payoffs, I give an example of an operating discriminative social norm in CTG.

The distribution of characteristics is assumed to be uniform. Let us assume that actors use the following discriminative strategy called Egalitarian Solution (*ES*):

1. When playing with an actor with the same characteristic play *B*.
2. When playing with an actor with a different characteristic start with *C* (*B* if you are a c_2 actor) and alternate *C* and *B* later on.
3. When your partner breaks the rules given above punish them by playing *N* indefinitely.

It is easy to see that *ES* is a best reply to itself. If you know that your partner is playing according to this strategy, you should do the same. Furthermore, *ES* enables perfect coordination whenever actors of different types meet i.e. half of the time. In the remaining cases the strategy asks the actors to play so as to achieve the maximal diagonal payoff[9]. Finally, the strategy is egalitarian i.e. it does not favour any actor type. The expected payoffs of all its users are the same, regardless their characteristics. In spite of all these advantages, *ES* is not uniformly stable. To see this, imagine that a population, in which virtually everyone uses this strategy, is invaded by two mutants. *RES* is a restarting *ES*. It behaves the same way *ES* does, but it does not punish defectors immediately. Instead, after the norm is broken for the first time, it restarts its interaction with them. It makes the next choice, as though it was the first one. Only after the partner fails to comply for the second time, it punishes them. *MES* is the same as *ES*, but it has an opposite rule for the first choices of the two types, when they meet each other. It expects c_1 actors to start with *B* and c_2 actors to start with *C*, when they meet a partner of a different type. Table 2 shows patterns of interactions for the three strategies, when different types meet[10].

We will be interested in the differences in the course of interactions of *ES* and *RES*. These appear, when they play with *MES*. *ES* punishes *MES* for mixing up the convention and ends up earning nothing in these encounters, as both actors play *N*. *RES*, after the initial miscoordination, retries its original choice and this time it

[9] It is worth noting, that the resulting inefficiency is an artifact resulting from the assumption, that there are only two types. Allowing for a greater number of characteristics, e.g. organized in hierarchies, would allow for eliminating most of it. Furthermore, it can be easily proved, that efficiency is not necessary for a strategy to be uniformly stable [13]. Even very efficient strategies will be unable to invade a population with well sanctioned and sufficiently popular suboptimal norms.

[10] The patterns of interactions for the same types are left out because they are always the same – both partners choose *B* indefinitely.

 DOI: 10.7206/DEC.1733-0092.77

coordinates with *MES*. Their collaboration lasts ever after. As a result the total payoff is higher for *RES* than for *ES*, regardless the actors' characteristics. This suffices to prove, that *ES* is not uniformly stable.

Table 2
Patterns of interactions between actors of different types for every possible pair of strategies in a population consisting of ES, RES and MES

	c_1, c_2	c_2, c_1
ES, ES	CB, BC, CB, BC ...	BC, CB, BC, CB ...
ES, RES	CB, BC, CB, BC ...	BC, CB, BC, CB ...
ES, MES	CC, NN, NN, NN ...	BB, NN, NN, NN ...
RES, RES	CB, BC, CB, BC ...	BC, CB, BC, CB ...
RES, MES	CC, CB, BC, CB ...	BB, BC, CB, BC ...
MES, MES	BC, CB, BC, CB ...	CB, BC, CB, BC ...

The result is not surprising. Dyadic strategies are well known for their instability. The example given above is a reworked version of an example mapped out by Boyd and Lorberbaum [14] to show that contrary to Axelrod's claims [7] TFT is not stable. Their reasoning has been then extended to all mixed strategies using finite number of pure strategies [23], and later to all nondeterministic strategies [34]. Bendor and Swistak [13, p. 1512] have proved the following Theorem[11]:

Theorem 1. *In a symmetric nontrivial iterated two-person game with an important future no dyadic pure strategy can be uniformly stable.*

The original Theorem was formulated with non-discriminative strategies in mind, but it applies to the discriminative strategies as well. It holds, because for every conceivable dyadic strategy it is possible to design a pair of treacherous mutants. One is a dyadically neutral mutant of the original strategy i.e. it behaves the same as the native strategy in its interactions with it. The other is a supporting mutant, that boosts the payoff of the first mutant and lowers the payoff of the original strategy. In the example given above *RES* is the dyadically neutral mutant, while *MES* is the supporting mutant.

The problem is, of course, the consequence of the famous second-order cooperation problem. It can be solved by inflicting punishments iteratively. Not only the norm breakers have to be punished, but also those who do not punish the norm-breakers, those who do not punish those who do not punish the norm-breakers etc. The whole meta-construct was presented neatly in [13] and the final result was wrapped up in the following Theorem [13, p. 1517][12]:

[11] See [11] for a full proof.
[12] See the reference for a full proof.

 DOI: 10.7206/DEC.1733-0092.77

Theorem 2. *Uniformly stable pure (social but non-discriminative)*[13] *strategies exist in a non-trivial symmetric two-person game with sufficiently important future iff the game is a game of enforceable cooperation.*

An exemplary uniformly stable social strategy is *CNF* (Conformity). It consists of the following rules[14]:

1. Categorize all actors as either friends or foes.
2. In the beginning consider everybody friends.
3. Cooperate (play a_1) with all friends and punish (play a_p such that $max_{1 \leq k \leq k} \mathcal{V}(k, p) < \mathcal{V}(1, 1)$) all the foes.
4. After each interaction add to the set of foes all actors to whom at least one of the following applies:

   (a) They did not cooperate with some friend.

   (b) They cooperated with some foe.

CNF is an application of a fairly intuitive logic entailing three simple rules: (1) the friend of my friend is my friend (2) the foe of my friend is my foe (3) the friend of my foe is my foe. It is a social strategy i.e. it interferes with interactions of other actors, even when it is not directly involved in them. Hence, it can be thought of as a social norm. It has two basic features of all norms: it says what actors should be doing, and it is guarded by social sanctions [28].

To see that no mutant can ever fare better than *CNF* in a population dominated by it, note that there are only two possible scenarios. First, a mutant can be completely neutral. Actors using it will always make exactly the same choices as actors using *CNF*. Consequently, they will be treated in the same way and will earn the same payoff. Second, a mutant can at some point make a different choice than *CNF*. It can cooperate with some foe or defect towards some friend. This will result in actors using the mutant strategy to be categorized as foes from this moment on. In most interactions they will receive at most $max_{1 \leq k \leq k} \mathcal{V}(k, p)$, where A_p is the punishment action. This, given the future is suffieciently important, is by definition of a game of enforceable cooperation less than payoff to cooperation achieved by *CNF*.

Unfortunately CTG is not a game of enforceable cooperation and, as stated in Theorem 2, the social norms themselves cannot stabilize any strategy in this game. From Theorem 1 we also know that discrimination is not sufficient to achieve this goal either.

[13] Clarification added.

[14] Similar "moral" strategies were considered earlier e.g. by Boyd and Richerson [16] – p. 182.

 DOI: 10.7206/DEC.1733-0092.77

5. Discriminating social norms

The first main result of the current paper is that, when we couple discrimination and social norms, we can achieve stability in games of enforceable coordination, regardless whether they are or are not games of enforceable cooperation:

Theorem 3. *Pure strategies that are uniformly stable under some binary distribution of characteristics exist in a non-trivial symmetric two-person game with sufficiently important future if and only if the game is a game of enforceable coordination.*

Proof. To prove sufficiency I will give an example of a discriminative social strategy that is stable in games of enforceable coordination under a binary uniform distribution, when delta is sufficiently large. To prove necessity I will show that, when the game is not a game of enforceable coordination, there is no such strategy, regardless the particular form of the binary distribution and the value of delta.

An exemplary uniformly stable strategy will be called *SES* which stands for Social Egalitarian Solution. It applies the following rules:

1. Categorize all actors as either friends or foes.
2. In the beginning consider everybody friends.
3. When playing with friends condition your choice on your characteristic and the characteristic of your partner[15].
   (a) When playing with an actor with the same characteristic as yours choose a_1.
   (b) When playing with an actor with a different characteristic alternate a_{k^*} and a_{l^*}. If you are a c_1 type start with a_{k^*}. Otherwise start with a_{l^*}.
4. When playing with foes always choose the most efficient *ap*.
5. After each interaction add to the set of foes all actors that did not follow the rules listed above.

It is easy to see that *SES* is uniformly stable under a uniform binary distribution of characteristics if the future is sufficiently important. Any other strategy *Z* can either make always the same choices as *SES* or not. In the first case it will earn the same payoff as *SES*, for any δ, any *g*, and any *P*. In the second case, as soon as *Z* makes a choice different from the choice prescribed by *SES*, it will be classified as a foe. Let us assume that $p^*_{SES|c1} = p^*_{SES|c2}$. In that case we know for sure that for sufficiently large δ for all *c*

$$v_{SES|c,SES|w} > v_{Z|c,SES|w} \tag{9}$$

[15] Please note, that, when the coordination point lies on the diagonal, this reduces to the analogous point in the definition of *CNF* in section 4.

 DOI: 10.7206/DEC.1733-0092.77

This implies that, if p_{SES} is sufficiently large, there exists δ such that $\mathcal{V}_{SES|c} > \mathcal{V}_{Z|c}$ for any c and for any non neutral mutant Z. Summing up, under uniform binary distribution of characteristics and for sufficiently large δ there exists P^*_{SES} satisfying the conditions for uniform stability, if a stage game is a game of enforceable coordination.

Let us now assume that in some game, which is not a game of enforceable coordination, there exists some strategy S that is uniformly stable under some binary distribution of characteristics for sufficiently high δ. From Theorems 1 and 2 we know, that it has to be both social and discriminative. Two mutant strategies S_1 and S_2 will be introduced to the population. Let $M = max_{k,\,l}\, \mathcal{V}(k, l)$ be a maximal payoff in the game. Let us assume that S_1 and S_2 interact with each other so that S_1 always receives M. Assume, furthermore, that S_2 always chooses the most efficient a_p, when interacting with S. Hence S_2 is a supporting mutant for S_1. Consequently, for large δ for all $c, w \in \mathcal{C}$ we have:

$$v_{S|c,S2|w} < v_{S1|c,S2|w} \tag{10}$$

If p_{S1} is sufficiently small, this implies that the uniformly stable S has to be a strictly better reply to itself than S_1 for all c[16].

Let S_1 be constructed in such a way, that its minimal payoff in a single interaction with S is not smaller than P. This is always possible, because all the considered strategies are deterministic. The maximal payoff of S in interactions with other S actors equals R, when interacting with the same type, and $2T$, when interacting with another type. However, it is never the case that both types playing S can get the payoff of $2T$. They have to share it. Let q denote the share of $2T$ earned by type c_1, where $q \in \left\langle \frac{min(\mathcal{V}(k*,l*),\mathcal{V}(l*,k*))}{2T}, \frac{max(\mathcal{V}(k*,l*),\mathcal{V}(l*,k*))}{2T} \right\rangle$. For S to be a strictly better reply to itself than S_1 for sufficiently large δ, we need:

$$p_{S|c1} \times R + p_{S|c2} \times q2T > p_S \times P \tag{11}$$

$$p_{S|c1} \times (1-q)2T + p_{S|c2} \times R > p_S \times P \tag{12}$$

This implies:

$$(2T - 2R)\left(\frac{p_{S|c1}}{p_S}\right)^2 - (2T - 2R)\left(\frac{p_{S|c1}}{p_S}\right) + P - R < 0 \tag{13}$$

From the definitions, and because the game is not a game of enforceable coordination, we know that $T \geq P \geq R$. Hence, if anywhere, this condition is fulfilled, if the frequencies of S are equal among both types. In this case however the

[16] Please note that the game is not a game of enforceable cooperation, so no strategy can maintain stability for one type only.

 DOI: 10.7206/DEC.1733-0092.77

inequality above reduces to $P - \frac{R+T}{2} < 0$ which only holds for games of enforceable coordination. Hence, we have a contradiction. □

Having defined a uniformly stable strategy in the general class of games of enforceable coordination, we can now specify it for CTG and investigate shortly its stabilizing frequencies. In CTG (see Table 1) *SES* should choose the basic task, when interacting with the same type actor, and alternate the basic and the complementary tasks, when interacting with another type. It should also use action *N* to punish defectors, as this action ensures that the partner earns no more than 2. To compute the minimal stabilizing frequency of *SES* I will conservatively assume that the minimum *SES* can gain in interactions with other strategies equals the minimal payoff to the punishing actor i.e. 0, while the maximum that a mutant can gain in interactions with other mutants equals the maximal payoff in the game i.e. 6. This gives the following conditions on *SES*'s stability for the two types:

$$p_{SES|c1} \times 2 + p_{SES|c2} \times 5 + (1 - p_{SES}) \times 0 > p_{SES} \times 2 + (1 - p_{SES}) \times 6 \quad (14)$$

$$p_{SES|c1} \times 5 + p_{SES|c2} \times 2 + (1 - p_{SES}) \times 0 > p_{SES} \times 2 + (1 - p_{SES}) \times 6 \quad (15)$$

Figure 1 depicts the minimal frequencies of *SES* playing actors with characteristics c_1 and c_2, that are necessary for the above conditions to hold. Because both conditions have to hold simultaneously, and the maximal sum of the frequencies cannot exceed one, the only combinations that assure *SES*'s stability are located inside of the shaded area (boarders excluded).

When we analyse the figure a few observations become apparent. First, in general the smaller the frequency of *SES* actors of one type, the larger the necessary frequency of *SES* actors of the other type. If egalitarian solution is not very popular with actors with characteristics c_1 or there is very few actors with this characteristic in the first place, it is difficult for the actors with c_2 characteristic to harvest the benefits of coordination. Hence, a larger share of c_2 type actors playing *SES* is needed for *SES* to be profitable enough (especially for the c_2 actors). Second, the total minimal stabilizing frequency and the total minimaxing stabilizing frequency of *SES* in this game coincide and they both equal 0.8. The corresponding vector P^*_{SES} is (0.4, 0.4), where actors playing *SES* earn on average 3.5 with their kins irrespective of their type (a minimum of 2.8 in total, when the interactions with other strategies are taken into account). Third, the vector given above is not always attainable. For instance, when the share of c_1 type equals 0.3 and the share of c_2 type equals 0.7, both total stabilizing frequencies of *SES* equal 0.85 and the corresponding P^*_{SES} vector equals (0.3, 0.55). More generally, when the distribution of characteristics is fixed, it puts further constraints on the feasible stabilizing vectors that have to be taken into account. Fourth, whenever the shares of both types among actors playing *SES* are

 DOI: 10.7206/DEC.1733-0092.77

not equal, the two types earn different payoffs in spite of *SES* being an egalitarian solution. For the exemplary frequencies (0.3, 0.55) the payoff to actors c_1 playing *SES* with other *SES* actors is 3.94 (a minimum of 3.35 in total) while it equals 3.05 (a minimum of 2.6 in total) for the c_2 type. The less common type becomes privileged.

The last observation turns our attention to the question of equality. Please note that in the considered example inequality results in an increase in the minimal stabilizing frequencies. It is not a coincidence. The egalitarian solutions are inherently easier to stabilize. I investigate this problem further in the next section.

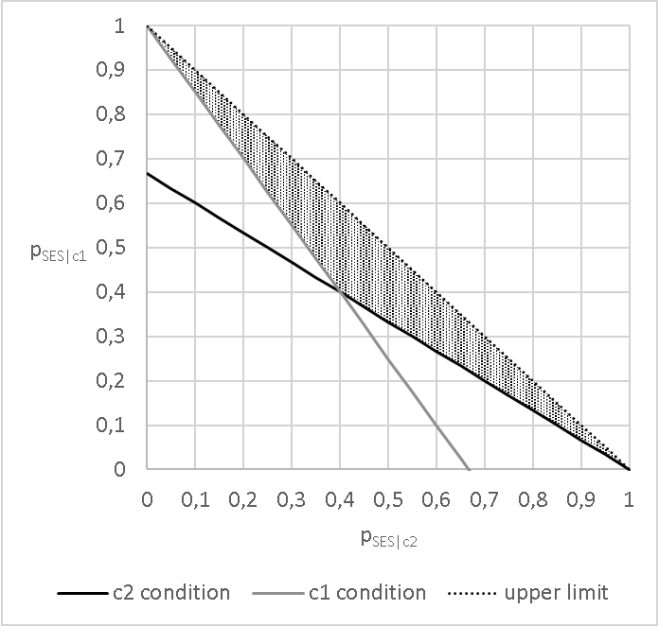

Figure 1. **Stabilizing frequencies of *SES* in CTG.**

6. Stability of unfair norms

"Hunt in the morning, fish in the afternoon, rear cattle in the evening, criticize after dinner, just as I have a mind, without ever becoming a hunter, fisherman, shepherd or critic." This is how Marx and Engels described their ideal non-antagonistic world which was supposed to replace capitalism. In spite of some attempts, we did not enjoy it in practice. However, in the current model it does

 DOI: 10.7206/DEC.1733-0092.77

seem possible. In particular the uniformly stable *SES* presented in the previous section allows all the actors to alternate between different tasks. The possession of certain characteristic is not followed by ascription to some single action. Moreover, in the considered example *SES* has the lowest minimal stabilizing frequency, when the payoffs to both types are identical i.e. both types are equally common among the actors using it. As a result, we get perfect equality in spite of the fact that actors rely on discriminative premises. This may seem somewhat unexpected given the social reality we can observe. In this section I show that equality is only a single point on a long continuum of stable systems of social organization using discriminating social norms. Yet, it is a special point.

Discriminative social norms can result in inequalities for two different reasons: difference in the frequency of the two types (as seen in the previous section) and the nature of the discriminative norm itself. In this section I will introduce an example of a strategy that is uniformly stable in CTG and assigns tasks of different profitability to the partners of different types. I will call this type of strategies unfair and I will analyse their properties. The exemplary strategy is called *SPA* i.e. Social Privilege Assignment. It is a social norm that constructs and reifies inequalities between actors differing only in labels. It states that the complementary task can only be performed by actors who possess the characteristic c_1, while the actors possessing characteristic c_2 should restrict themselves to the basic task. It also punishes all the offenders not willing to follow these rules and everybody who fails to enforce these rules. Hence, it makes the construct all encompassing and impossible to avoid in spite of its complete arbitrariness.

Formally *SPA* is identical to *SES* except for point 3b. In the case of *SPA* it states: When playing with an actor with a different characteristic condition your choice on your type. If you are a c_1 type play *C*. Otherwise play *B*. Under this regime the c_1 type is privileged, as it always gets to choose the more profitable action, when interacting with an actor of a different type. Similarly as in equations (14) and (15) we can write conditions for *SPA*'s stability within the two types:

$$p_{SPA|c1} \times 2 + p_{SPA|c2} \times 6 + (1 - p_{SPA}) \times 0 > p_{SPA} \times 2 + (1 - p_{SPA}) \times 6 \quad (16)$$

$$p_{SPA|c1} \times 4 + p_{SPA|c2} \times 2 + (1 - p_{SPA}) \times 0 > p_{SPA} \times 2 + (1 - p_{SPA}) \times 6 \quad (17)$$

Figure 2 depicts the minimal frequencies of *SPA* playing actors with characteristics c_1 and c_2, that are necessary for the above conditions to hold. Because both conditions have to hold simultaneously and the maximal sum of the frequencies cannot exceed one, the only combinations that assure *SPA*'s stability are located inside of the shaded area (boarders excluded).

 DOI: 10.7206/DEC.1733-0092.77

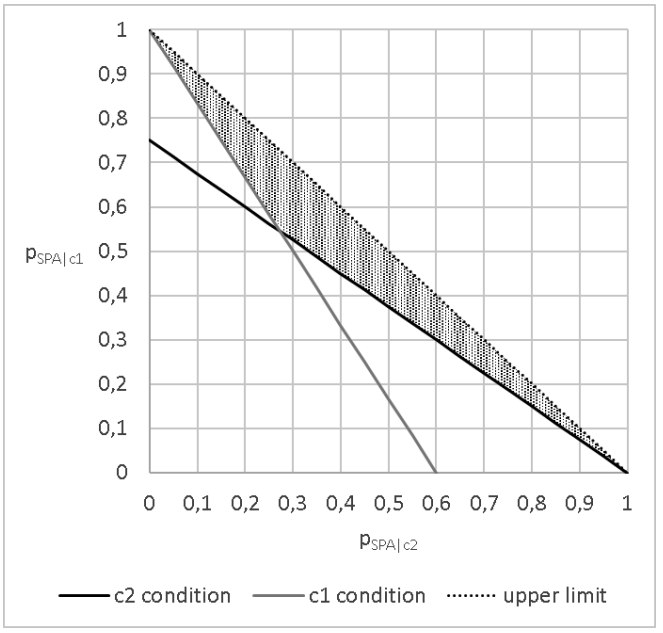

Figure 2. **Stabilizing frequencies of *SPA* in CTG**

In comparison to the egalitarian solution, the c_1 condition (grey line) for *SPA* has moved to the left, while the c_2 condition (black line) has moved up. The privileged c_1 type gets high payoff in a single encounter with the other type and does not need that many encounters to earn enough to repel mutants. The c_2 type on the other hand needs more encounters with the privileged type to achieve the same goal. In spite of the relaxed expectations of the privileged type, the overall effect on stability is negative as it is driven by the more demanding restriction. Hence the total minimal stabilizing frequency is larger than before and equals about 0.82. The corresponding vector is (0.55, 0.27). In this configuration the unfair strategy needs twice as many privileged followers as disadvantaged followers. It is also apparent that this ratio constitutes a condition, under which the expected payoffs for both types are equal. Actors playing *SPA* in their encounters with others also playing *SPA* earn on average 3.3 (a minimum of 2.73 in total). The disadvantaged group can make up for its losses by coordinating more frequently. Yet the total payoffs from interactions with other *SPA* actors for both types are lower than in the case of *SES*. The total minimaxing stabilizing frequency for *SPA* is even larger than the total minimal stabilizing frequency and equals almost 0.86. Its corresponding vector is (0.43, 0.43). Obviously in this case the average total payoff to the privileged group is higher. It equals 4 for interactions with other *SPA* actors (a minimum of 3.43 in

 DOI: 10.7206/DEC.1733-0092.77

total). The average total payoff to the disadvantaged group, on the other hand, is visibly lower. It equals only 3 for interactions with other *SPA* actors (a minimum of 2.57 in total). Had *SES* the same frequencies within the two types, it would ensure the minimal total payoff of 3 to actors with both characteristics.

To investigate the problem of fairness a little more systematically, I will restrict attention to games of enforceable coordination in which the coordination point is off the diagonal and the corresponding payoffs $\mathcal{V}(k^*, l^*)$ and $\mathcal{V}(l^*, k^*)$ are different. This is the only case in which the maximal total payoff can potentially be distributed unequally within a pair playing the game. I will call this type of games – games of enforceable unequal coordination. Furthermore, I will concentrate on strategies similar to *SES* and *SPA* i.e. using the updated friends-foes categorization and punishing the foes indefinitely. I will call them the *SES* family. These strategies differ in their behaviour towards kins i.e. actors using the same strategy. In particular in the *SES* family there is a number of strategies, that distribute the more and less profitable actions constituting the coordination point differently between the partners. For CNF *SES* and *SPA* are limiting cases between which there is for example a strategy that prescribes that actors with characteristic c_1 in their interactions with c_2 actors play a sequence (*C*, *C*, *C*, *B*, *C*, *C*, *C*, *B* ...) while the c_2 actors play the complementary sequence (*B*, *B*, *B*, *C*, *B*, *B*, *B*, *C* ...). As a result the c_1 actors gets on average 0.55 of the sum of payoffs in the coordination point. The maximum share they can get is obviously determined by the game matrix and equals $\frac{max(\mathcal{V}(k*,l*),\breve{\mathcal{V}}(l*,k*))}{2T}$ (0.6 in the CTG). The *SES* family contains also some inefficient strategies that do not coordinate whenever possible or do not choose the optimal A_1, when playing with friends of the same type. The question to be answered is: how should a pair of actors distribute the maximal total payoff between the partners of different types, and what should be the shares of actors of different types among actors using given strategy, to ensure that this strategy has the lowest possible total stabilizing frequency. Bendor and Swistak showed that more efficient social strategies generally have lower stabilizing frequencies [13, p. 1527]. The next Theorem refers to the consequences of this fact for relationship between stability and inequalities.

Theorem 4. *The lowest total minimal and minimaxing stabilizing frequencies among strategies of the SES family in games of enforceable unequal coordination under binary distribution of characteristics is achieved by SES when both types are equally represented among the actors using it.*

Proof. Let *M* denote the highest payoff in a game of enforceable unequal coordination and *U* denote the lowest possible payoff to a punishing actor. Let *S* be a strategy from the *SES* family. Let X_1 and X_2 denote an average payoff that *S* actors with characteristic c_1 and c_2 respectively get when interacting with other *S* actors.

I will construct an environment in which *S* is most susceptible to invasion. I will investigate what should be the shares of actors of different types using *S* and how should the profit from coordination be divided between them to ensure stability for the lowest possible p_S and the lowest possible $max(p_{S|c1}, p_{S|c2})$.

The environment in which *S* is most susceptible to invasion is the one from the proof of Theorem 3. Please recall the construction of mutants S_1 and S_2 where S_2 is a supporting mutant of S_1. I will assume that the frequency of S_1 is negligibly small. Hence, payoffs of *S* and S_1 are determined by what they get with *S* and S_2. The inequality (10) holds. Hence, we get the following two conditions under which the total payoff to *S* is larger than payoff to S_1 in each of the groups:

$$X_1 p_S + p_{S2} U > p_S P + p_{S2} M \tag{18}$$

$$X_2 p_S + p_{S2} U > p_S P + p_{S2} M \tag{19}$$

The left hand sides of these inequalities are growing in *X*'s. As long as the maximal possible *X*'s are larger than *U*, they are also growing in p_S while the right hand sides are decreasing in p_S. When we solve (18) and (19) for the maximal X_1 and X_2 we will find the minimal value of p_S for which the two conditions are met.

The maximal payoff of *S* in interactions with its kin equals *R*, when interacting with the same type, and 2*T* to share with the partner, when interacting with the other type. The latter is only achievable, if *S* is stable for both types i.e. for the smaller of the two *X*'s. If the strategy is stable only within one type, its maximal *X* equals *R*. Hence, as long as the smaller of the two *X*'s is larger than *R*, strategy allowing for coordination between the types will have the smallest total minimal stabilizing frequency. Let us consider strategy that allows for such coordination. Let *q* denote the share of 2*T* earned by type c_1, where $q \in \left\langle \frac{min(\mathcal{V}(k*,l*),\mathcal{V}(l*,k*))}{2T}, \frac{max(\mathcal{V}(k*,l*),\mathcal{V}(l*,k*))}{2T} \right\rangle$. In that case:

$$X_1 = \left(\frac{p_{S|c1}}{p_{S|c1} + p_{S|c2}} R + \frac{p_{S|c2}}{p_{S|c1} + p_{S|c2}} q2T \right) \tag{20}$$

$$X_2 = \left(\frac{p_{S|c1}}{p_{S|c1} + p_{S|c2}} (1-q)2T + \frac{p_{S|c2}}{p_{S|c1} + p_{S|c2}} R \right) \tag{21}$$

Both these expressions are certainly larger than *U*. To make sure they are also larger than *R*, we need $\frac{R}{2T} < q < 1 - \frac{R}{2T}$. In other words privileges cannot be excessive.

Please note that X_1 is growing in *q* and X_2 is decreasing in *q*. To make sure the smaller of the two is possibly large we need them equal. The equality implies that the

 DOI: 10.7206/DEC.1733-0092.77

share of $2T$ assigned to any single type (i.e. q) is determined by the shares of the two types among actors playing S:

$$\frac{p_{S|c1}}{p_{S|c1} + p_{S|c2}} = \frac{2Tq - R}{2(T - R)} \tag{22}$$

If we substitute this in the formula for X_1 and search for maximum in q, we get $q = 0.5$. It implies that the payoff to maximal coordination should be divided equally between the two types. If we substitute this in (22), we learn that the shares of the two types among actors playing S should be the same. Hence, the minimal stabilizing frequency is achieved by *SES* when both types are equally represented among actors using it.

Finally, please note that in order to compute the total minimaxing stabilizing frequencies under any binary distribution of characteristics, we need to assume that $\frac{p_{S|c1}}{p_{S|c1}+p_{S|c2}} = \frac{p_{S|c2}}{p_{S|c1}+p_{S|c2}} = 0.5$ and search for maximal X. This will necessarily lead to the same result. □

Theorem 4 says nothing about accessibility of the optimal solution under some specific distribution of characteristics. When this distribution is unfavourable i.e. the frequencies of the two types are excessively different, the optimal solution might be impossible to achieve. In particular the total minimal and minimaxing stabilizing frequencies of *SES* are given by:

$$p^*_{SES} = \frac{M - U}{M - U + \frac{R+T}{2} - P} \tag{23}$$

If the fraction of any type exceeds half of this expression, *SES*'s minimal and minimaxing stabilizing frequencies will grow. Such a situation took place in one of the examples given at the end of section 5, where the frequency of c_1 type was equal to 0.3. In this case *SES* was easiest to stabilize, when virtually all of the actors of the rarer type used it.

It is worth noting that in such cases some unfair strategy from the *SES* family will have lower minimal stabilizing frequency than *SES* itself. More precisely, it will be the strategy that ensures equal payoffs to both types given the unequal distribution of characteristics. Figure 3 presents the set of minimal stabilizing frequencies for strategies from the *SES* family in CTG and the share of maximal total payoff that should be assigned to actors with characteristic c_1 to achieve this goal. The condition $q = 0.5$ describes the fair *SES* strategy. It is apparent in the figure that its stabilizing frequency is lowest of all and it is achieved, when frequencies of actors using *SES* within both types equal 0.4. When the frequency of c_1 type is equal to 0.3, however, the minimal stabilizing frequency (equal about 0.81) is achieved by a strategy

that gives this type only about 0.42 of the total maximal payoff. There are further problems though. The black part of the line corresponding to the minimal stabilizing frequencies is the only one that is actually accessible in CTG. Please recall that $q \in \left\langle \frac{min(\mathcal{V}(k*,l*),\mathcal{V}(l*,k*))}{2T}, \frac{max(\mathcal{V}(k*,l*),\mathcal{V}(l*,k*))}{2T} \right\rangle$ or in our case $q \in (0.4, 0.6)$. The CTG matrix does not allow for arbitrary distribution of the total maximal payoff. As a result, the total minimal stabilizing frequencies for extremely uneven distributions of characteristics will be even higher than the one presented in Figure 3. These will also be the cases in which actors with different characteristics have different payoff when playing the strategy from the *SES* family with the lowest total minimal stabilizing frequency. Further discussion of this topic is postponed to later work.

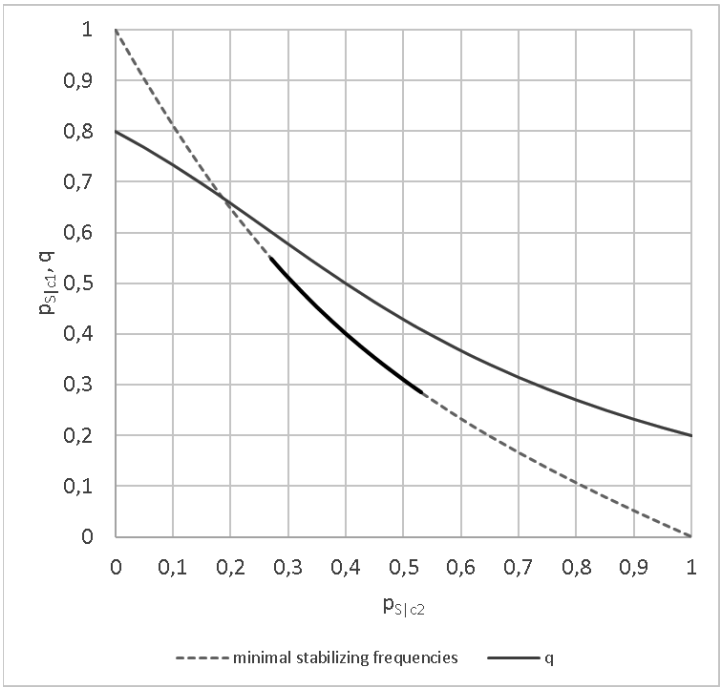

Figure 3. **Minimal stabilizing frequencies of *SES* family in CTG**

One last remark is important, when we consider the workings of unfair social norms for which $q \neq 0.5$. Recall that the definition of SPA required that all the actors using it, regardless whether they were of the privileged or the disadvantaged type, punished all the norm breakers. This implies in particular that the disadvantaged

 DOI: 10.7206/DEC.1733-0092.77

type plays an important role in solidifying and preservation of the social system that cripples it. A good real life illustration of this phenomena would be ”one should know their place” philosophy that is entertained by members of many underprivileged groups. In sociological and psychological literature it is often observed that discriminated people socialize each other to underdog's position (see e.g. [22]) and dispraise those who reach higher. This provokes a question about the importance of this type of attitudes for the stability of the whole discriminative system. In particular, it would be interesting to see how lessening the support for the system on the side of the disadvantaged group can influence the system's stability. This issue is only one of a number of problems that can and should be considered in a follow-up to this paper. Some of the other interesting issues are listed in the last section.

7. Summary and Discussion

In the analyses above I showed that it is possible to construe a discriminative social system from scratch, using only actors' labels as a starting point and enforcing the system via norms guarded by third party sanctions. As a result, originally meaningless and rationally irrelevant labels gain a new socially ascribed meaning that influences actors' choices.

An upside to this is that one can solve the problem of coordination. The solution enables actors to perform complementary tasks ensuring the most effective solution from the utilitarian point of view. A downside is that this mechanism can stabilize a number of payoff structures including also the highly unfair ones. It was shown, however, that solutions ensuring equal payoffs to actors with different labels are generally easier to stabilize. In particular a completely fair strategy that shares the maximal payoff from coordination equally between partners of different types has the lowest minimal stabilizing frequency. Yet, when the distribution of characteristics is unfavourable, solutions ensuring equality can become impossible to realize.

The first question that comes to mind is why isn't the egalitarian solution more common empirically? One possible answer is that some of the observed tasks assignments were originally fair, but the structure of the whole dilemma evolved over time. The changes were not sufficient to deprive the norm of its stability, but they were enough to make it unfair in the sense of leading to different payoffs to different types of actors. More generally, however, it is worth remembering that the set of initial assumptions in the current model, in particular the assumption about perfect original equality, are rather difficult to fulfil in practice. Hence, the model is useful primarily because it clearly presents that no initial differences in abilities

or opportunities are necessary for a discriminative system to come to existence and become socially reified. Consequently it does not reflect social reality in its whole complexity.

One can, of course, think of a number of improvements and additional analyses that could prove interesting in the context of the current model. First of all, the current analysis was limited to binary distributions of characteristics. This restriction is not necessary. Once we remove it, we will be able to consider more complex social structures, including various hierarchies. We will also be able to widen the class of games in which stable solutions exist. In another vain, one can restructure the whole game by assuming for instance that it is only played by actors of different types. At some point it might also prove useful to examine the case of mixed strategies. Furthermore, the assumption of perfect information is not very plausible. Especially in large groups embeddedness is probably far from perfect and it should be taken into account (see [1] for an attempt in this direction). One could also aim at more precise grasping of the learning processes. In reality learning is more tightly connected with the group of origin of a given actor. On one hand, people imitate those who have similar status more commonly than others. On the other hand, they are more prone to follow those who belong to groups of higher status than those who are lower on the social ladder. Finally, one cannot not note that, once gained, privileges make available some additional options and, therefore, the situation of everybody stops being identical. Instead, some kind of market with actors with different purchasing power arises. Not everyone can afford everything, and ascription to some category can become self-enforcing.

The model presented here has some important practical implications. It indicates e.g. that demounting discriminative status quo should be necessarily accompanied by building new normative standards that would allow for more equality without leading to coordinative disaster. It is not enough to tell women they can make a career. One needs also to convince men that doing more housework is not defaming for them. In general, the refinements proposed above can help us understand and shape the crucial social processes at work. No matter how much we want it, we do not live in a world free of discrimination. Our brains were not meant for building individual relationships with everyone we meet. In the majority of our daily interactions others are just labels we assign to them: bus driver, neighbour, woman, old person, son of an actor, guy from Alabama. These labels are not just ornaments. They play vital role in organization of social life. But, as usually is the case, they have also harmful consequences. Hence understanding their mechanics is crucial.

 DOI: 10.7206/DEC.1733-0092.77

References

[1] Abramczuk, K., Obłój, J. (in press). Third party sanctions in games with communication. *Studies in Logic, Grammar and Rhetoric*.

[2] Acker, J. (1988). Class, gender, and the relations of distribution. *Signs*, *13*(3), 473-497.

[3] Anker, R. (1997). Theories of occupational segregation by sex: An overview. *International Labour Review*, *136*, 315.

[4] Arrow, K.J.(1973). The theory of discrimination. In: Ashenfelter, O., Rees, A. (eds.) *Discrimination in Labour Markets* (pp. 3-33). Princeton University Press, Princeton NJ.

[5] Aumann, R.J. (1959). Acceptable points in general cooperative n-person games. *Contributions to the Theory of Games (AM-40)*, *4*, 287.

[6] Aumann, R.J. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, *1*, 67-96.

[7] Axelrod, R.M. (1984). *The Evolution of Cooperation*. Basic Books, New York.

[8] Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211-233.

[9] Barron, R.D., Norris, G.M. (1991). Sexual divisions and the dual labour market. In: Leonard, D., Allen, S. (eds.) *Sexual Divisions Revisited* (pp. 153-177). Springer.

[10] Becker, G.S.(2009). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press, pp. 24.

[11] Bendor, J., Swistak, P. (1998). Evolutionary equilibria: Characterization theorems and their implications. *Theory and Decision*, *45*(2), 99-159.

[12] Bendor, J., Swistak, P. (2000). The impossibility of pure homo economicus. In: *Annual Meeting of the American Political Science Association*. Marriott Wardman Park.

[13] Bendor, J., Swistak, P. (2001). The evolution of norms. *American Journal of Sociology*, *106*(6), 1493-1545.

[14] Boyd, R., Lorberbaum, J.P. (1987). No pure strategy is evolutionarily stable in the repeated prisoner's dilemma game. *Nature*, *327*, 58-59.

[15] Boyd, R., Richerson, P.J. (1989). The evolution of indirect reciprocity. *Social Network*, *11*, 213-36.

[16] Boyd, R., Richerson, P.J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *13*(3), 171-195.

[17] Brines, J. (1993). The exchange value of housework. *Rationality and Society*, *5*(3), 302-340.

[18] Crowley, P.H. (2000). Hawks, doves, and mixed-symmetry games. *Journal of Theoretical Biology*, *204*(4), 543-563.

[19] Crowley, P.H. (2001). Dangerous games and the emergence of social structure: evolving memory-based strategies for the generalized hawk-dove game. *Behavioral Ecology*, *12*(6), 753-760.

[20] Crowley, P.H., Cottrell, T., Garcia, T., Hatch, M., Sargent, R.C., Stokes, B.J., White, J.M. (1998). Solving the complementarity dilemma: evolving strategies for simultaneous hermaphroditism. *Journal of Theoretical Biology*, *195*(1), 13-26.

[21] Dahrendorf, R. (1962). On the origin of social inequality. In: Laslett, P. (ed.) *Philosophy, Politics and Society*. Basil Blackwell, Oxford.

[22] Della Fave, L.R. (1980). The meek shall not inherit the earth: Self-evaluation and the legitimacy of stratication. *American Sociological Review, 45*(6), 955-971.
[23] Farrell, J., Ware, R. (1989). Evolutionary stability in the repeated prisoner's dilemma. *Theoretical Population Biology 36*(2), 161-166.
[24] Gaissmaier, W., Schooler, L.J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416-422.
[25] Gal, I. (1996). Understanding repeated simple choices. *Thinking & Reasoning 2*(1), 81-98.
[26] Haugaard, M. (2003). Reections on seven ways of creating power. *European Journal of Social Theory*, *6*(1), 87-113.
[27] Hayek, F.A. (1960). *The Constitution of Liberty*. The University of Chicago, Chicago, pp. 25.
[28] Homans, G.C. (1950). *The Human Group*. *Brace and Company*. Harcourt.
[29] Ishida, J. (2003). The role of intrahousehold bargaining in gender discrimination. *Rationality and Society*, *15*(3), 361-380.
[30] Kandori, M. (2002). Introduction to repeated games with private monitoring. *Journal of Economic Theory*, *102*(1), 1-15.
[31] Kaneko, M., Matsui, A. (1999). Inductive game theory: discrimination and prejudices. *Journal of Public Economic Theory*, *1*(1), 101-137.
[32] Kilbourne, B.S., England, P., Farkas, G., Beron, K., Weir, D. (1994). Returns to skill, compensating differentials, and gender bias: Effects of occupational characteristics on the wages of white women and men. *American Journalof Sociology*, *100*(3), 689-719.
[33] Lippert, S., Spagnolo, G. (2011). Networks of relations and word-of-mouth communication. *Games and Economic Behavior, 72*(1), 202-217.
[34] Lorberbaum, J. (1994). No strategy is evolutionarily stable in the repeated prisoner's dilemma. *Journal of Theoretical Biology*, *168*(2), 117-130.
[35] McDonald, G.W. (1980). Family power: The assessment of a decade of theory and research, 1970-1979. *Journal of Marriage and the Family*, 841-854.
[36] McElreath, R., Boyd, R., Richerson, P.J. (2003). Shared norms and the evolution of ethnic markers. *Current Anthropology*, *44*(1), 122-130.
[37] Nowak, M., Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, *194*(4), 561-74.
[38] Nowak, M., Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature, 393*, 573-577.
[39] Nowak, M.A. (2006). Five rules for the evolution of cooperation. *Science, 314*(5805), 1560-1563.
[40] Okuno-Fujiwara, M., Postlewaite, A. (1995). Social norms and random matching games. *Games and Economic Behavior 9*(1), 79-109.
[41] Phelps, E.S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, *62*(4), 659-661.
[42] Ponti, G., Seymour, R.M. (1999). Evolutionary stability of inequality structures. *Rationality and Society*, *11*(1), 47-77.
[43] Quint, T., Shubik, M. (2001). Games of status. *Journal of Public Economic Theory*, *3*(4), 349-372.

 DOI: 10.7206/DEC.1733-0092.77

[44] Raub, W., Weesie, J. (1990). Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology*, *96*, 626-654.
[45] Reichenbach, H. (1971). *The Theory of Probability*. University of California Press, pp. 26.
[46] Rothbart, M., Taylor, M. (1992). Category labels and social reality: Do we viewsocial categories as natural kinds? In: Fiedler, K. (ed.) *Language, Interaction and Social Cognition* (pp. 11-36). Sage, Thousand Oaks, CA.
[47] Sudgen, R. (1986). *The Economics of Rights*. Basil Blackwell, Oxford.
[48] Tversky, A., Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review, 90*(4), 293.
[49] Ullmann-Margalit, E. (2015). *The Emergence of Norms*. Oxford University Press, USA.
[50] Wagenaar, W.A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, *77*(1), 65.
[51] Weibull, J.W. (1997). *Evolutionary Game Theory*. MIT press.
[52] Youm, Y., Laumann, E.O. (2003). The effect of structural embeddedness on the division of household labor: A game-theoretic model using a network approach. *Rationality and Society*, *15*(2), 243-280.

 DOI: 10.7206/DEC.1733-0092.77