THE ASSIGNMENT GAME WITH NEGATIVE EXTERNALITIES AND BOUNDED RATIONALITY*

KIMMO ERIKSSON, FREDRIK JANSSON* and THOMAS VETANDER

Centre for the Study of Cultural Evolution
Stockholm University
SE-106 91 Stockholm, Sweden, and
Division of Applied Mathematics
School of Education, Culture and Communication
Mälardalen University
Box 883, SE-721 23 Västerås, Sweden
**fredrik.jansson@intercult.su.se*

We introduce negative externalities in the form of ill will among the players of the classic two-sided assignment game of Shapley and Shubik, by letting each player's utility be negatively correlated with the payoff of all the players in his group. The new game is very complex, but under a certain assumption of bounded rationality we derive a straightforward notion of stable outcomes as certain conjectural equilibria. We prove that several well-known properties of the set of stable outcomes in the assignment game carry over to this new game.

Keywords: Two-sided matching; stable outcome; bounded rationality; assignment game; social preferences; negative externalities.

Subject Classification: C71, C78, D62.

1. Introduction

The standard assumption in game theory that each player cares just about maximizing her own payoff is an oversimplification in many social contexts. Experiments seem to reveal that players hold social preferences as well (cf. Camerer [2003]). A natural approach for game theorists is to incorporate these preferences into the utility functions, a famous example being the work of Fehr and Schmidt [1999] on incorporating fairness into several symmetrical games through a utility function under which a player dislikes situations where other players receive higher payoffs than he does. In this paper we will study another version of such an exotic utility function, applied to a two-sided matching game instead of a symmetrical game.

In two-sided matching there are two disjoint sets of players (called sides), usually thought of as men and women or workers and employers. The players gain utility from forming pairs consisting of one player from each side. Following the

*Electronic version of an article published as International Game Theory Review, Vol. 13, No. 4 (2011) 443-459, DOI: 10.1142/S0219198911003088
http://www.worldscinet.com/igtr

standard reference (Roth and Sotomayor [1990]), there are two main kinds of two-sided matching games. In the *marriage game*, players have a preference ordering over possible partners, and there is no way, such as using side payments, to change these preferences. An outcome in the marriage game is just a matching of the players. In the *assignment game* of Shapley and Shubik [1972], there is money involved, so an outcome is a matching together with a payoff vector.

Sasaki and Toda [1996] seem to have been the first to bring social preferences into the realm of two-sided matching. They study matching games with "externalities" in the sense that the value created (in the assignment game), or the preference lists (in the marriage game), may depend on how other people match.

In their model, stable matchings are guaranteed to exist only if the players consider all matchings to be possible. However, Hafalir [2008] has taken this model one step further for the marriage game by assuming that deviating players include the preference lists of other agents when considering possible new matchings. While Hafalir introduces a notion of expectations that guarantees a nonempty set of stable matchings, both models assume pessimistic players that deviate only if they are better off in the worst possible outcome among the possible new matchings. Maskin [2004] instead studies the formation of coalitions in general restricted to binding agreements and finds that with nonpositive externalities, the solution is unique and the grand coalition forms.

Our aim in this paper is to study what happens in the assignment game when a player experiences externalities from the *payoff* (rather than the matching) of the other players from the same side, and all these externalities are negative. We can interpret this as players feeling ill will toward rivals. A typical case might be a firm that does not want to see a former key employee matching up with a competitor, or a market of movie stars or soccer players, say, where ill will can be interpreted as the level of competition between companies or actors, related to how much they act on the same market. This possibility drastically increases the complexity of the assignment game, so it seems necessary to make an assumption of bounded rationality of the players. In Section 3 we define the new utility functions and the bounded rationality assumption, and derive a notion of stable outcomes for this game that amounts to a strengthening of the stability conditions in the classic assignment game.

Given the new definition of stability, in Section 4 we explore the properties of the set of stable outcomes. We find that many results for the classic assignment game have analogs in this game: All stable payoff vectors are compatible with all stable matchings, and the set of stable outcomes is non-empty with a non-trivial lattice property. In particular, there is a unique stable outcome that is optimal for all players of one side, and we present an algorithm that constructs this outcome.

Finally, in Section 5 we briefly discuss future experimental testing of our assumptions.

2. Preliminaries about the Assignment Game

Let **AG** be short for the assignment game of Shapley and Shubik [1972]. In **AG**, there are two disjoint sets of players, P and Q. Players can be matched in pairs of one P-player and one Q-player, or stay single. Single players are thought of as matched to themselves. In other words, we can define the set of all possible *Pairs* as

$$P \times Q \cup \{(p,p) : p \in P \cup Q\}.$$

The potential *value* of each possible Pair (p,q) is a nonnegative number $\alpha_{pq} \geq 0$, where the possibility of staying alone has a value of zero, that is, $\alpha_{pp} = 0$ for all $p \in P \cup Q$. An instance of **AG** is completely described by the triplet (P, Q, α).

An *outcome* of the game is described by the triplet $\big(\mu, (u,v)\big)$, where μ is a *matching*, and u and v are vectors describing the *payoff* to P-players and Q-players, respectively. A matching is a subset μ of Pairs, such that every player belongs to exactly one Pair in μ. If (p,q) is a matched Pair according to μ, then we write $p = \mu(q)$ and $q = \mu(p)$.

Given a game, the *total value* T_μ of a matching μ is defined as

$$T_\mu = \sum_{(p,q)\in\mu} \alpha_{pq}. \tag{1}$$

A matching μ is *optimal* if $T_\mu \geq T_{\mu'}$ for all matchings μ'. The pair of payoff vectors (u,v) is called a *feasible payoff* for the game if there exists a matching μ satisfying

$$\sum_{p\in P} u_p + \sum_{q\in Q} v_q = T_\mu. \tag{2}$$

In this case we say that the payoff (u,v) and the matching μ are *compatible*, and that $\big(\mu, (u,v)\big)$ is a *feasible outcome*. A feasible outcome $\big(\mu, (u,v)\big)$ is *stable* if, for all $p \in P$ and $q \in Q$,

(S1) $$u_p + v_q \geq \alpha_{pq} \quad \text{for all Pairs } (p,q),$$

that is, when there is no potential value of a pair that is higher than the sum of payoffs of the players in their current matching, so that no player can get a higher payoff by changing partners. Note that the individual rationality conditions $u_p \geq 0$ and $v_q \geq 0$ are obtained from (S1) as special cases, for the Pairs (p,p) and (q,q), respectively.

Define a partial order $\geq_P$ on payoff vectors by $(u',v') \geq_P (u,v)$ if $u'_p \geq u_p$ for all $p \in P$. The following properties of **AG** are well-known (cf. Roth and Sotomayor [1990]):

1. The set of stable outcomes is non-empty.
2. Every stable payoff is compatible with every optimal matching, and no other matching is compatible with a stable payoff.

3. For every pair (p, q) matched in a stable matching, we have $u_p + v_q = \alpha_{pq}$, and every unmatched agent receives a payoff of zero. In other words, in a stable outcome of **AG** there are never any side payments.
4. The set of stable payoffs forms a complete lattice under the partial order $\geq_P$. If (u, v) and (u', v') are stable payoffs, then their meet and join are given by $(u, v) \wedge_P (u', v') = (\min(u, u'), \max(v, v'))$ and $(u, v) \vee_P (u', v') = (\max(u, u'), \min(v, v'))$. Here max and min denote the componentwise maximum and minimum, respectively.
5. In particular, there exists a stable payoff $(\overline{u}, \underline{v})$ that is *P-optimal* in the sense that $(\overline{u}, \underline{v}) \geq_P (u, v)$ for every stable payoff (u, v). Of course, there also exists a Q-optimal stable payoff $(\underline{u}, \overline{v})$ with symmetrical properties.

3. Introducing Negative Externalities and Bounded Rationality

In this section we introduce negative externalities to the assignment game with players of ill will. We then discuss the complex nature of this game and make an explicit assumption regarding the bounded rationality of the players. Given this assumption, we derive a notion of stability.

3.1. *Definition of the assignment game with ill will*

We want to capture the phenomenon that a player p may feel *ill will* toward another player p', by which we mean that the utility of p is negatively correlated with the payoff of p'. (See Remark 3 of Section 3.4 for a brief discussion on the effects of good will.)

We shall here limit the range of ill will to players on the same side of the market. In a real two-sided market, like the labor market, it seems reasonable to assume that players care only about their peers and not about the alien types of the other side.

Finally, for reasons of simplicity we will assume that the utility of any player is a *linear* function of the payoffs of players on the same side, with the player's own payoff having the largest weight. Let $\mathrm{ill}_p(p')$ denote the marginal ill will that p feels toward p'. No player feels any ill will toward himself, that is, $\mathrm{ill}_p(p) = 0$.

Definition 1. The *assignment game with ill will*, **AGI** for short, is an extension of **AG** described by the quadruple $(P, Q, \alpha, \mathrm{ill})$. In **AGI** the utility of a P-player p from an outcome $\big(\mu, (u, v)\big)$ is

$$U_p = u_p - \sum_{p' \in P} \mathrm{ill}_p(p') \cdot u_{p'}, \tag{3}$$

and similarly the utility of a Q-player q is

$$V_q = v_q - \sum_{q' \in Q} \mathrm{ill}_q(q') \cdot v_{q'}. \tag{4}$$

All the coefficients of ill will are assumed to be in the half-open interval $[0, 1)$.

3.2. *Stability and structured bargaining*

It is not obvious how to define stability in **AGI**. For the classic assignment game, the set of stable outcomes coincides with the *core* of the game, that is, the set of feasible outcomes for which there exists no coalition that could have done better for all its members by acting on its own. However, in **AGI**, it is indeed a complex task for a coalition of players to foresee how well they will do on their own, for their utility will depend also on what outcome the other players achieve. It is difficult to define the game as a co-operative game, and thus the core is very difficult to derive.

Dijkstra [2009] has approached this problem by introducing a new notion of the core dealing only with dyads, resulting in a superset to the original core for positive externalities. Negative externalities, however, may change the feasible exchange pattern in this new notion of the core. We will not delve further into redefining the core. Instead, we will define a bounded rationality of the players for which the core will not be well defined, but where we can find the set of stable outcomes and several properties of the set of stable outcomes in **AG** carry over.

In order to derive a reasonable notion of stability, let us imagine a player who is considering to deviate from a certain outcome. There are three possible ways for his utility to increase: by an increase in his own payoff or the payoff of someone toward whom he feels good will, or by a decrease in the payoff of someone toward whom he feels ill will. We want to define stable outcomes as those where no player can find a deviation that, as far as he can see, will increase his utility. In order to make this definition precise, we need to describe exactly what deviations a player can initiate and how far he can foresee the consequences.

Starting with such deviations, we restrict the way the players can act as follows.

Definition 2. A *two-sided bidding game without side payments* is an assignment game with structured bargaining in the sense that each player can negotiate only with players from the other side of the market, and these negotiations are restricted to bids of the kind "Want to be my partner? Your share will be \$XXX." A player may offer more than the value of the pair, thus keeping a negative share. In a *two-sided bidding game with side payments* there is also another kind of bid: "If you find a new partner and stay with that partner, then I'll give you a side payment of \$XXX." Thus, (to possibly increase his utility by making a player on the same side worse off) a player may choose to reduce his own payoff and increase that of a player of the other side.

In such bidding games, a successful bid from p to q given the current outcome $(\mu, (u, v))$ will make q's previous partner $p' = \mu(q)$ partnerless. We can expect p' either to make a counter-bid to q or to make an offer to a new partner, possibly making yet another player partnerless — who in turn will make a new bid, etc. These chains can be very long, increasing the computational powers p needs to foresee the consequences.

Definition 3. If p makes a bid to q, and q is currently matched to p's rival p', then

we say that p' is the *target* of the bid. Define a player p of a two-sided bidding game to be *naïve* if

1. p accepts any bid from the other side that increases his payoff;
2. p believes that among P-players only he and his target $p' = \mu(q)$ will have their payoffs affected by any bid p makes to any Q-player q;
3. p expects that, if forced to move, his target p' will accept any bid from the other side that increases his payoff (rather than utility, thus not taking his target's possible ill will into account).

Finally we can define what we mean by stability in such games.

Definition 4. An outcome of a two-sided bidding game is said to be *naïvely stable* if no naïve player can make a bid that he expects would increase his utility.

Remark 1. A naïvely stable outcome amounts to a so called *conjectural equilibrium*, that is, an outcome in which no agent has an incentive to deviate under a given conjecture about the reactive behavior of the others, and which is itself consistent with the conjecture. In contrast to our "naïve" players, Sasaki and Toda [1996] assume "pessimistic" players in their work on externalities in two-sided matching. A pessimistic player needs to be able to calculate the worst possible conceivable outcome of a move, and hence needs much more computational power than a naïve player.

3.3. *Criteria for naïve stability*

In a world without ill will, naïve stability coincides with classic stability.

Proposition 1. *The set of stable outcomes in* ***AG*** *coincides with the set of naïvely stable outcomes in the corresponding bidding game (with or without side payments).*

Proof. In **AG** there is no ill will, so a naïve player will care only about his own payoff and will never want to make side payments. Let us study possible bids that a player p can make in an outcome $\big(\mu, (u, v)\big)$. A bid of v'_q to q will be accepted if and only if $v'_q > v_q$. It will be rational to make this bid for p if and only if his new payoff is greater than his old payoff, that is, $\alpha_{pq} - v'_q > u_p$. It is possible to satisfy both inequalities if and only if $\alpha_{pq} > u_p + v_q$. Hence, an outcome will be stable if and only if $u_p + v_q \geq \alpha_{pq}$ for every Pair (p, q). □

We shall now derive criteria for naïve stability in **AGIB**, the bidding version of **AGI**. If a naïve player p is considering a bid to some player q, then p must also compute the second best option of the target, that is, the current partner $p' = \mu(q)$ of q.

Definition 5. Given an outcome $\big(\mu, (u, v)\big)$, define the *naïve loss of player* p, denoted by $\mathrm{loss}^{\mu,(u,v)}(p)$, as the loss in payoff that a naïve player expects p to suffer if forced to move from his current partner $\mu(p)$.

When the outcome $\big(\mu,(u,v)\big)$ is evident from the context, we omit it and write just $\mathrm{loss}(p)$.

Lemma 1. *Given an outcome* $\big(\mu,(u,v)\big)$*, the naïve loss of player* p *is given by*

$$\mathrm{loss}(p) = \min_{q\neq\mu(p)} \big\{u_p + v_q - \alpha_{pq}\big\}. \tag{5}$$

Proof. A naïve player expects p to be able to match up with any q as soon as p offers q a payoff v'_q satisfying $v'_q > v_q$. This would give p a payoff of $\alpha_{pq} - v'_q$, so his loss would be $u_p - (\alpha_{pq} - v'_q)$. The infimum of this loss over all Q-players q (except the current partner of p), all offers $v'_q > v_q$, and the option to remain single (in which $v_q = \alpha_{pq} = 0$), yields the naïve loss in the lemma. □

Proposition 2. *A feasible outcome* $\big(\mu,(u,v)\big)$ *in* ***AGIB*** *without side payments is naïvely stable if and only if it satisfies the following two conditions:*

(S2') $u_p + v_q - \alpha_{pq} - \mathrm{ill}_p\big(\mu(q)\big)\cdot\mathrm{loss}\big(\mu(q)\big) \geq 0$ *for all Pairs* (p,q)*,*
(S3') $u_p + v_q - \alpha_{pq} - \mathrm{ill}_q\big(\mu(p)\big)\cdot\mathrm{loss}\big(\mu(p)\big) \geq 0$ *for all Pairs* (p,q)*.*

Proof. If p = $\mu(q)$, then $\mathrm{ill}_p(\mu(q)) = 0$, so (S2') follows from the condition that $(\mu,(u,v))$ is a feasible outcome. Assume $p \neq \mu(q)$. If p were to make an offer to q of $v'_q > v_q$ and his target $p' = \mu(q)$ decided to try his luck somewhere else, then p would expect an increase in utility of size $\alpha_{pq} - v'_q - u_p + \mathrm{ill}_p(p')\cdot\mathrm{loss}_p(p')$. The player p wants to make such an offer if and only if this increase is greater than zero for some $v'_q > v_q$, from which (S2') follows. Condition (S3') follows analogously. □

In the bidding game with side payments, played by naïve players, an offer of a side payment Δ_q to q for leaving his partner will be accepted by q if and only if the offer will compensate for his loss, that is, if and only if $\Delta_q > \mathrm{loss}(q)$.

Theorem 1. *A feasible outcome* $\big(\mu,(u,v)\big)$ *in* ***AGIB*** *with side payments is naïvely stable if and only if it satisfies the following two conditions:*

(S2) $\mathrm{loss}\big(q\big) - \mathrm{ill}_p\big(\mu(q)\big)\cdot\mathrm{loss}\big(\mu(q)\big) \geq 0$ *for all Pairs* (p,q)*,*
(S3) $\mathrm{loss}\big(p\big) - \mathrm{ill}_q\big(\mu(p)\big)\cdot\mathrm{loss}\big(\mu(p)\big) \geq 0$ *for all Pairs* (p,q)*.*

Proof. The naïve loss is nonnegative, so (S2) follows immediately for $p = \mu(q)$. Thus assume $p \neq \mu(q)$. If p were to make an offer to q of a side payment $\Delta_q > \mathrm{loss}\big(q\big)$ for leaving his partner $p' = \mu(q)$ for someone else, then p would expect his utility to increase by $-\Delta_q + \mathrm{ill}_p(p')\cdot\mathrm{loss}(p')$. The player p wants to make such an offer if and only if this increase is greater than zero for some $\Delta_q > \mathrm{loss}\big(q\big)$, from which (S2) follows.

Naïve stability in **AGIB** with side payments implies naïve stability in **AGIB** without side payments, for by Lemma 1, $\mathrm{loss}(q) \leq u_p + v_q - \alpha_{pq}$, so (S2) implies (S2'). □

Corollary 1. *Naïvely stable outcomes of* ***AGIB*** *are always stable outcomes of the corresponding* ***AG****, that is, they satisfy*

(S1) $u_p + v_q - \alpha_{pq} \geq 0$ *for all Pairs* (p, q).

Proof. Choose p and q so that $\text{loss}(p) = \text{loss}(q) = u_p + v_q - \alpha_{pq}$ is minimal. We must show that this expression is nonnegative. This follows from (S2) and the assumption that all coefficients of ill will lie in $[0, 1)$:

$$\begin{aligned} 0 &\leq \text{loss}\big(q\big) - \text{ill}_p\big(\mu(q)\big) \cdot \text{loss}\big(\mu(q)\big) \\ &\leq \text{loss}\big(q\big) - \text{ill}_p\big(\mu(q)\big) \cdot \text{loss}\big(q\big) \\ &= \big[1 - \text{ill}_p\big(\mu(q)\big)\big] \cdot \text{loss}\big(q\big). \end{aligned} \tag{6}$$

□

Hence, if the stability condition in **AG** is satisfied, then the stability condition in **AGIB** is also satisfied. Thus any stable outcome in **AGIB** must also be a stable outcome in **AG**. Hence, the result from **AG** that no side payments occur carries over directly. However, as we have seen, the possibility of such transfers plays an important role in the dynamics of the game.

Remark 2. Another interpretation of the naïve loss of p is the amount that can be transferred from p to $\mu(p)$ without (S1) being violated. We will use this interpretation later on.

3.4. *Examples and remarks*

Let us now look at an example of an assignment bidding game with side payments and check an outcome for naïve stability. Let $P = \{p_1,\ p_2,\ p_3\}$ and $Q = \{q_1,\ q_2,\ q_3\}$ and assume that player p_1 is the only player feeling any ill will, described by the coefficients $\text{ill}_{p_1}(p_2) = 0.4$ and $\text{ill}_{p_1}(p_3) = 0.6$. The value matrix is given below.

$$\alpha = \begin{pmatrix} \alpha_{p_1q_1} & \alpha_{p_1q_2} & \alpha_{p_1q_3} \\ \alpha_{p_2q_1} & \alpha_{p_2q_2} & \alpha_{p_2q_3} \\ \alpha_{p_3q_1} & \alpha_{p_3q_2} & \alpha_{p_3q_3} \end{pmatrix} = \begin{pmatrix} 50 & 120 & 130 \\ 0 & 170 & 150 \\ 0 & 120 & 200 \end{pmatrix}$$

A stable outcome must have $\mu(p_1) = q_1$, $\mu(p_2) = q_2$ and $\mu(p_3) = q_3$, since this is the only optimal matching and thus the only matching for which we can hope to find payoffs that satisfy (S1). To check condition (S2) we need to calculate the naïve loss for p_2, p_3, q_2 and q_3. Condition (S3) will not be restrictive since no Q-player feels any ill will.

Let us check the payoff vectors $u = (u_{p_1},\ u_{p_2},\ u_{p_3}) = (50,\ 100,\ 120)$ and $v = (v_{q_1},\ v_{q_2},\ v_{q_3}) = (0,\ 70,\ 80)$ for stability. These payoff vectors are compatible with the optimal matching, and (S1) is easily seen to be satisfied. (In fact, this is the P-optimal payoff of the corresponding **AG**.) Let us now calculate the naïve loss for the relevant players. The second best option for p_2 is q_3, yielding $\text{loss}(p_2) =$

$u_{p_2} + v_{q_3} - \alpha_{p_2 q_3} = 100 + 80 - 150 = 30$. The second best option of p_3 is q_2, yielding $\text{loss}(p_3) = u_{p_3} + v_{q_2} - \alpha_{p_3 q_2} = 120 + 70 - 120 = 70$. For both q_2 and q_3 their second best options are as good as their best options, so $\text{loss}(q_2) = \text{loss}(q_3) = 0$. With these naïve losses, we see that (S2) is violated by (p_2, q_3) as well as (p_3, q_2), for $0 - 0.4 \cdot 30 \not\geq 0$, and $0 - 0.6 \cdot 70 \not\geq 0$. Hence, this outcome is not naïvely stable. Instead p_1 will make an offer to either q_2 or q_3, expecting their current partners p_2 and p_3 to raise their offers (and thereby lose payoff, increasing the utility of p_1).

In Section 4.6 we will return to this example and see how we can modify this unstable outcome to find payoff vectors that satisfy (S2) as well as (S3).

Remark 3. Let us motivate why the case when the utility of a player p is positively correlated with the payoff of a player p' is less interesting. Assume a positive coefficient $\text{goodwill}_p(p')$, where $U_p = u_p + \text{goodwill}_p(p')u_{p'}$. Assume further that the players are currently matched so that (S1) is satisfied. Now p would like to help p' to obtain a higher payoff, but can he do that? If the naïve loss of $\mu(p')$ is greater than zero, then p could pay another Q-player to attempt to match with p' and hope that $\mu(p')$ would raise her bid to p'. However, what if $\mu(p')$ refuses to do this and decides to match with another P-player? Then p would no longer want to fulfill his offer. Therefore, the deviation is not credible and is really just a bluff that $\mu(p')$ can choose to call. There is no way for a player p to credibly help another P-player obtain a higher payoff.

Remark 4. A naïvely stable outcome of **AGIB** need not be stable with less naïve players, as the following example shows. We have three players on each side, and the only matching that satisfies (S1) is $\mu(p_1) = q_1$, $\mu(p_2) = q_2$ and $\mu(p_3) = q_3$. All the ill will coefficients are zero except $\text{ill}_{p_1}(p_2)$, which is positive. The value matrix is given below.

$$\alpha = \begin{pmatrix} 100 & 90 & 0 \\ 0 & 100 & 90 \\ 0 & 0 & 80 \end{pmatrix}$$

Now consider the payoff vectors $u = (u_{p_1},\ u_{p_2},\ u_{p_3}) = (50,\ 50,\ 30)$ and $v = (v_{q_1},\ v_{q_2},\ v_{q_3}) = (50,\ 50,\ 50)$. This payoff is naïvely stable. However, a less naïve player p_1 may see a possibility to credibly force down the payoff of p_2 and thus drive up his own utility. This is because p_3 would have nowhere to go if he lost q_3, so if p_2 tried to go to his second best option q_3, then p_3 would have to fight for him.

4. Analysis

From now on in this paper, we will study properties of naïvely stable outcomes of **AGIB** with side payments. However, all these results also hold for **AGIB** without side payments, with analogous proofs. For simplicity, we will drop "naïve" and just talk about "stability."

First we will investigate the compatibility between stable matchings and stable payoff vectors. Then we go on to prove that stable outcomes always exist and that

we can always find solutions that are optimal for both sides. After considering an example, we finally prove that the set of stable payoff vectors forms a lattice.

4.1. *Compatibility*

An important feature in **AG** is that every stable payoff is compatible with any optimal matching. We will now prove a similar result for **AGIB**.

Theorem 2. *If* $\big(\mu,(u,v)\big)$ *and* $\big(\mu',(u',v')\big)$ *are stable outcomes of* ***AGIB****, then* $\big(\mu,(u',v')\big)$ *is also a stable outcome.*

Proof. Let us show that $\big(\mu',(u,v)\big)$ is a stable payoff, that is, that (S2) and (S3) are satisfied. It suffices to show that the naïve loss for all players is the same in $\big(\mu',(u,v)\big)$ as in $\big(\mu,(u,v)\big)$.

Look at the naïve loss of an arbitrary P-player p. (The case of Q-players is similar.) We must check that $\text{loss}^{\mu',(u,v)}(p) = \text{loss}^{\mu,(u,v)}(p)$, which is equivalent to

$$\min_{q\neq\mu'(p)}\big\{u_p+v_q-\alpha_{pq}\big\} = \min_{q\neq\mu(p)}\big\{u_p+v_q-\alpha_{pq}\big\}. \tag{7}$$

There are two cases to consider: either $\mu'(p)=\mu(p)$ or $\mu'(p)\neq\mu(p)$. In the first case, the desired equality holds trivially. In the second case, we know that $\big(\mu',(u,v)\big)$ satisfies (S1), since $\big(\mu',(u,v)\big)$ is stable in **AG**. Therefore, we have $u_p+v_{\mu'(p)} = \alpha_{p\mu'(p)}$. From the fact that $\big(\mu,(u,v)\big)$ is stable we have $u_p+v_{\mu(p)}=\alpha_{p\mu(p)}$. It follows that both $\text{loss}^{\mu',(u,v)}(p)$ and $\text{loss}^{\mu,(u,v)}(p)$ are equal to zero. □

4.2. *Existence of stable outcomes*

We shall show that every stable matching μ in **AG** is also a stable matching in **AGIB**, and that the set of stable payoff vectors in **AGIB** is a nonempty subset of the stable payoff vectors in **AG**. This means that ill will among peers will not change the way players match, although it may affect the way that money is distributed among the players. It will turn out that ill will among P-players lowers the maximum amount they can receive in a stable outcome (and similarly for the Q-players).

Theorem 3. *The set of stable matchings is the same for* ***AG*** *and* ***AGIB****. The set of stable payoff vectors associated with a matching* μ *in* ***AGIB*** *is a nonempty subset of the stable payoff vectors associated with* μ *in* ***AG****.*

It is trivially true that any stable matching in **AGIB** is also a stable matching in **AG**, and likewise that any stable pair of payoff vectors associated with a stable matching in **AGIB** also works for the same matching in **AG**. What we need to prove is that from an arbitrary stable matching in **AG**, we can find payoff vectors that are stable for **AGIB**. We will achieve this end by describing an algorithm that constructs such payoff vectors. The proof of the theorem will be completed in Section 4.4.

Before moving on, we shall prove a lemma that will be useful when we examine the algorithm. To begin with, if we define

$$\mathrm{Ill}\big(\mu(q)\big) = \max_p \big\{\mathrm{ill}_p\big(\mu(q)\big)\big\}, \tag{8}$$

the maximum degree of ill will felt by any player p to the partner of q, then we can state the stability conditions in an equivalent but more convenient form:

(S2) $\mathrm{loss}\big(q\big) - \mathrm{Ill}(p) \cdot \mathrm{loss}(p) \geq 0$ for all *matched* Pairs (p, q),
(S3) $\mathrm{loss}\big(p\big) - \mathrm{Ill}\big(q\big) \cdot \mathrm{loss}\big(q\big) \geq 0$ for all *matched* Pairs (p, q).

The essence of our lemma is that if some player makes a bid to q, then no player will make a bid to his partner p.

Lemma 2. *In an outcome in which (S1) holds, let (p, q) be a matched Pair for which (S2) is violated or holds with equality. Then (S3) holds for this Pair.*

Proof. In order to obtain a contradiction, assume that

$$\mathrm{Ill}(p) \cdot \mathrm{loss}(p) \geq \mathrm{loss}(q) \text{ and } \mathrm{Ill}(q) \cdot \mathrm{loss}(q) > \mathrm{loss}(p). \tag{9}$$

Adding the inequalities above gives the inequality

$$0 > \big(1 - \mathrm{Ill}\big(p\big)\big) \cdot \mathrm{loss}(p) + (1 - \mathrm{Ill}(q)) \cdot \mathrm{loss}(q), \tag{10}$$

which is a contradiction since the losses are nonnegative according to (S1), and the ill will coefficients are less than 1. □

4.3. *The algorithm*

We shall now describe an algorithm for constructing a stable outcome in **AGIB**. In the proof, we will use the terms P- and Q-optimality. By a P-optimal outcome we mean a stable outcome such that there exists no other stable outcome in which any P-player receives a higher payoff. (This does not mean that the *utility* of all P-players is maximized.)

Input: A feasible outcome $\big(\mu, (u, v)^0\big)$ of **AGIB** satisfying (S1) and (S3).

Output: An outcome $\big(\mu, (u, v)^*\big)$ that is stable, that is, satisfies (S2) and (S3), and in which (S2) holds with equality for all matched Pairs for which (S2) was not satisfied in $\big(\mu, (u, v)^0\big)$.

Algorithm: In each step we take an outcome $\big(\mu, (u, v)^i\big)$ and construct a new outcome $\big(\mu, (u, v)^{i+1}\big)$ by a transferral of the payoff from p to q in each matched pair of the amount

$$\max\left\{\frac{\mathrm{Ill}(p) \cdot \mathrm{loss}^{\mu,(u,v)^i}(p) - \mathrm{loss}^{\mu,(u,v)^i}(q)}{1 + \mathrm{Ill}(p)}, 0\right\}.$$

Repetition of this procedure yields a sequence $\big\{(u, v)^i\big\}_{i=0}^{\infty}$ of payoff vectors. Define $(u, v)^*$ as the payoff to which the sequence converges. ■

If a transfer of δ is made within a matched pair from p to q, then the naïve losses will change accordingly, so that $\text{loss}^{\text{new}}(p) = \text{loss}^{\text{old}}(p) - \delta$ and $\text{loss}^{\text{new}}(q) = \text{loss}^{\text{old}}(q) + \delta$. Hence, a transfer of

$$\frac{\text{Ill}(p) \cdot \text{loss}^{\mu,(u,v)^i}(p) - \text{loss}^{\mu,(u,v)^i}(q)}{\underbrace{1 + \text{Ill}(p)}_{\geq 1}},$$

leads to the previously unsatisfied (S2) being satisfied with equality. We are now ready to prove the correctness of the algorithm.

Proof. We claim that in each step the conditions (S1) and (S3) still hold. First, for each player p the transfer is strictly less than the naïve loss, so (S1) will not be violated by the transfer. Second, taken by itself any nonzero transfer is designed to achieve equality in (S2) for that matched pair (p, q), and by Lemma 2 we know that (S3) will still hold if (S2) holds with equality. Thus (S3) will also hold in the context of all simultaneous transfers, for they will decrease the payoffs to P-players and increase the payoffs to Q-players, hence not decrease the naïve loss of p and not increase the naïve loss of q. However, condition (S2) might become violated for (p, q) by other transfers.

The sequence of payoff vectors is monotone increasing in v. Since (S1) holds at every step, the sequence of payoff vectors will be bounded by the Q-optimal solution in **AG**. Hence the algorithm converges to a well-defined outcome $(u, v)^*$.

Since (S1) and (S3) are weak inequalities that hold during the entire algorithm, they also hold at the convergence point $(u, v)^*$. It remains for us to show that no pair violates (S2) at $(u, v)^*$. Assuming the opposite, that is, that for some matched pair (p, q) we have

$$\text{Ill}(p) \cdot \text{loss}^{\mu,(u,v)^*}(p) - \text{loss}^{\mu,(u,v)^*}\big(q\big) = \epsilon, \tag{11}$$

for some $\epsilon > 0$, we shall derive a contradiction. Since the sequence converges to $(u, v)^*$, we know that for any $\delta > 0$ we can find a K so that for all $k > K$ we have $u_p^* \leq u_p^k < u_p^* + \delta$ for all players $p \in P$ and $v_q^* - \delta < v_q^k \leq v_q^*$ for all players $q \in Q$. Now look at the matched pair (p, q) for some $k > K$. We know that $\text{loss}^{\mu,(u,v)^k}(p) \geq \text{loss}^{\mu,(u,v)^*}(p) - \delta$ and $\text{loss}^{\mu,(u,v)^k}(q) \leq \text{loss}^{\mu,(u,v)^*}(q) + \delta$. The following amount will be transferred from p to q:

$$\begin{aligned}
&\frac{\text{Ill}(p) \cdot \text{loss}^{\mu,(u,v)^k}\big(p\big) - \text{loss}^{\mu,(u,v)^k}(q)}{1 + \text{Ill}(p)} \\
&\geq \frac{\text{Ill}(p) \cdot \text{loss}^{\mu,(u,v)^*}\big(p\big) - \text{loss}^{\mu,(u,v)^*}(q) - 2\delta}{1 + \text{Ill}(p)} \\
&= \frac{\epsilon - 2\delta}{1 + \text{Ill}(p)}.
\end{aligned} \tag{12}$$

For any $\epsilon > 0$ and δ small enough, the transfer will be greater than δ, which is a contradiction. Thus no pair can violate (S2) when the algorithm has converged.

Finally, it is obvious that no step in the algorithm can make a pair go from violating (S2) to satisfying (S2) with strict inequality. □

Remark 5. The algorithm makes a minimum transfer in the sense that if we start with P-optimality in **AG**, then none of the pairs can make a smaller transfer without violating any of the stability conditions. If we start with the P-optimal payoff $(u, v)^0$ in **AG**, then all transfers of payoff must be from p to q. No pair in $(u, v)^0$ satisfies (S2) with strict inequality, because if it did, then, without violating (S2), we could transfer the payoff

$$\frac{\text{loss}^{\mu,(u,v)^i}(q) - \text{Ill}(p) \cdot \text{loss}^{\mu,(u,v)^i}(p)}{1 + \text{Ill}(p)} < \text{loss}^{\mu,(u,v)^i}(q) = \min_{p \neq \mu(q)} \{u_p + v_q - \alpha_{pq}\}, \quad (13)$$

from q to p without violating (S1), which contradicts the assumption that $(u, v)^0$ is P-optimal.

Since no transfer from p to q within a matched pair can make any other pair stop violating (S2), and all transfers must be made in that direction, there is a unique way to make all the pairs satisfy (S2) with equality, which is also the minimum transfer.

4.4. *Proof of Theorem 3*

Let $(u, v)^0$ be the P-optimal payoff in **AG**. For these payoffs, (S3) cannot be violated by any pair since the naïve loss of all Q-players must be equal to zero. To see why this is true, imagine that $\text{loss}^{\mu,(u,v)^0}(q) > 0$ for some player q. Then we could transfer an amount $\text{loss}^{\mu,(u,v)^0}(q)$ from q to $\mu(q)$ without violating (S1), and thus $(u, v)^0$ cannot have been P-optimal to begin with. Thus both (S1) and (S3) are satisfied at $(u, v)^0$, so the algorithm will produce a stable outcome from this input.

4.5. *Optimal outcomes*

Existence of a P-optimal outcome for **AGIB** will now follow as a consequence of the algorithm.

Theorem 4. *In **AGIB** there exists a P-optimal stable payoff $(\overline{u}, \underline{v})$. (By symmetry, there also exists a Q-optimal outcome.)*

Proof. From condition (S1), all stable outcomes in **AGIB** will be stable in **AG**. Hence, they can be obtained from $(u, v)^0$, the P-optimal outcome of **AG**, by some transferral of payoff from the P-player to the Q-player within matched pairs. No such transfer can ever make another matched pair stop violating (S2), and since the algorithm never transfers more than necessary to satisfy (S2) for the pair in question, this will result in a P-optimal outcome for **AGIB**. □

Remark 6. Jonas Sjöstrand (personal communication) has found that the P-optimal solution $(\overline{u}, \underline{v})$ can also be characterized in the following way. Let $M = \{(u, v) \text{ satisfying (S1) and (S2)}\}$. Then $\overline{u}_p = \max_M\{u_p\}$ and $\underline{v}_q = \min_M\{v_q\}$.

4.6. *The example continued*

Let us apply the algorithm to the example in Section 3.4. We saw that the P-optimal outcome for **AG** was not stable in **AGIB**, since two pairs violated (S2). If we use the notation from the algorithm, then we have $u^0 = (50,\ 100,\ 120)$ and $v^0 = (0,\ 70,\ 80)$. The first step of the algorithm transfers

$$\frac{\mathrm{Ill}(p_2)\cdot \mathrm{loss}^{\mu(u,v)^0}(p_2) - \mathrm{loss}^{\mu,(u,v)^0}(q_2)}{1+\mathrm{Ill}(p_2)} = \frac{0.4\cdot 30 - 0}{1+0.4} = 8 + \frac{4}{7}$$

from p_2 to q_2, and

$$\frac{\mathrm{Ill}(p_3)\cdot \mathrm{loss}^{\mu(u,v)^0}(p_3) - \mathrm{loss}^{\mu,(u,v)^0}(q_3)}{1+\mathrm{Ill}(p_3)} = \frac{0.6\cdot 70 - 0}{1+0.6} = 26 + \frac{1}{4}$$

from p_3 to q_3. This gives us $u^1 = (50,\ 91+\frac{3}{7},\ 93+\frac{3}{4})$ and $v^1 = (0,\ 78+\frac{4}{7},\ 106+\frac{1}{4})$. But both pairs still violate (S2), since the simultaneous transfers increase the naïve losses of the other P-players and decrease them for the Q-players. Therefore we have to keep iterating, calculating new naïve losses and transfer more money to the Q-players. The algorithm will converge to the P-optimal payoffs

$$u^* = (50,\ 82,\ 87),$$
$$v^* = (0,\ 88,\ 113).$$

Note that the ill will of player p_1 has considerably lowered the maximum amount that players p_2 and p_3 can hope to end up with.

4.7. *A lattice property*

In **AG** the set of stable outcomes is a complete lattice under the partial order $\geq_P$. We shall prove that the same thing holds for the stable outcomes of **AGIB**.

The meet and join of two stable payoffs (u, v) and (u', v') in **AG** are given by $(\min(u, u'), \max(v, v'))$ and $(\max(u, u'), \min(v, v'))$, respectively. In **AGIB**, these payoffs are not necessarily stable, but they will work as inputs for the algorithm.

Theorem 5. *The set of stable payoffs in **AGIB** endowed with the partial order $\geq_P$ forms a complete lattice.*

Proof. We will show how to find the meet $(u, v) \wedge_P (u', v')$ of stable payoff vectors (u, v) and (u', v'). Start by defining $(\underline{u}^0, \overline{v}^0) = (\min(u, u'), \max(v, v'))$. This is the meet in **AG**, so in particular this outcome satisfies (S1). To use the algorithm, we also need to show that (S3) is satisfied for these starting payoffs, that is, we must verify the inequality

$$\min_{q''\neq q'} \{\underline{u}^0_{\mu(q')} + \overline{v}^0_{q''} - \alpha_{\mu(q')q''}\} \geq \mathrm{ill}_q(q') \cdot \min_{p\neq\mu(q')} \{\underline{u}^0_p + \overline{v}^0_{q'} - \alpha_{pq'}\}. \tag{14}$$

First note that $\underline{u}^0_{\mu(q')}$ and $\overline{v}^0_{q'}$ are from the same original pair (u, v) or (u', v'). If they are from (u, v) then the inequality above follows from (S3) for (u, v) since $\underline{u}^0_p = u_p$

and $\overline{v}_q^0 \geq v_q$ on the left-hand side, and $\underline{u}_p^0 \leq u_p$ and $\overline{v}_q^0 = v_q$ on the right-hand side. Similarly if they are from (u', v').

Now we have a situation just like when we searched for a stable outcome. Among all stable outcomes that are P-worse than the input outcome $(\underline{u}^0, \overline{v}^0)$, the algorithm will find the unique P-optimal one, that is, a unique meet of (u, v) and (u', v'). The join can be found in the obvious analogous way. Thus the payoffs form a lattice, which will be complete since the set of stable payoffs is compact. □

Let us now look at an example to illustrate this procedure. Let there be four players (p_1, p_2, p_3, and p_4) in P, and similarly in Q. Let the only nonzero ill will be $\mathrm{ill}_{p_1}(p_2) = 0.5$, and let the pair values be given by

$$\alpha = \begin{pmatrix} 60 & 0 & 0 & 0 \\ 0 & 100 & 30 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 40 & 0 & 5 \end{pmatrix}.$$

The only stable matching μ matches $\mu(p_j) = q_j$ for $j = 1, 2, 3, 4$. Consider the following two payoff vectors, (u, v) and (u', v'):

$$u = (30, 50, 10, 0),\ u' = (30, 50, 0, 5),$$
$$v = (30, 50, 0, 5),\quad v' = (30, 50, 10, 0).$$

It is easy to check that both these payoffs are stable. Now define $(\underline{u}^0, \overline{v}^0) = (\min\{u, u'\}, \max\{v, v'\})$, that is, $\underline{u}^0 = (30, 50, 0, 0)$ and $\overline{v}^0 = (30, 50, 10, 5)$. Here (S2) is not satisfied, since $\mathrm{loss}(p_2)^{\mu,(\underline{u}^0,\overline{v}^0)} = 30$, and $\mathrm{loss}(q_2)^{\mu,(\underline{u}^0,\overline{v}^0)} = 10$, and thus

$$\mathrm{ill}_{p_1}(p_2) \cdot \mathrm{loss}^{\mu,(\underline{u}^0,\overline{v}^0)}(p_2) = 0.5 \cdot 30 \not\leq 10 = \mathrm{loss}^{\mu,(\underline{u}^0,\overline{v}^0)}(q_2).$$

We will now use our algorithm to find the meet of (u, v) and (u', v'). This time we only have to iterate one time before the algorithm converges. The transfer of payoff from p_2 to q_2 will be

$$\frac{\mathrm{Ill}(p_2) \cdot \mathrm{loss}^{\mu(\underline{u},\overline{v})^0}(p_2) - \mathrm{loss}^{\mu,(\underline{u},\overline{v})^0}(q_2)}{1 + \mathrm{ill}_{p_2}} = \frac{0.5 \cdot 30 - 10}{1 + 0.5} = 3 + \frac{1}{3},$$

resulting in the payoff vectors $\underline{u} = (30, 46 + \frac{2}{3}, 0, 0)$, $\overline{v} = (30, 53 + \frac{1}{3}, 10, 5)$. Here both (S2) and (S3) are satisfied.

5. Discussion

In this paper, we have introduced negative externalities to the assignment game by including social preferences, in the form of ill will between players on the same side of the game. Under the assumption of "naïve" players (a particular form of bounded rationality), we found that the stability properties of the assignment game do not change very much after the introduction of ill will. The same matchings are stable, but the set of stable payoffs shrinks, eliminating extreme payoffs of individuals.

It is worth observing that with at least three players on each side in **AGIB**, all of them teeming with ill will toward all their rivals, we obtain a social dilemma. If no matches are made, then all the players receive a payoff of zero, and hence a utility of zero. In the case where all the players match up, their payoffs will increase, but each player's utility will have decreased because of all that ill will toward rivals. In prisoners' dilemma terms, all the players would do well if they cooperated (did not find a match), but each individual has an incentive to defect (find a match).

Naïve players have certain expectations about how other players will react to their bids. We have seen that naïve players can always reach a conjectural equilibrium in **AGIB**. It would be interesting to investigate the evolution of such expectations, that is, to study the dynamics of the game when players make bids according to expectations that they then find must be revised.

Our future research plans include taking the assignment game with ill will to the laboratory. There are previous experimental studies on how externalities affect partner choice made by Dijkstra and van Assen [2008]. Their results show that people in a resource dilemma are willing to accept a cost in their choice of partner to avoid being hurt by someone in the future. There are at least two aspects that we wish to investigate in an experimental setting of the assignment game. First, trying to provoke ill will through framing, what size of effects can be attained? Second, with artificial ill will (tying subjects' real payoffs to artificial ill will coefficients), how well will the model of naïve players approximate their behavior? Our working hypothesis is that players are more likely to be naïve than, for example, pessimistic (in the sense of Sasaki and Toda [1996]). A first test could be whether a case similar to the one in remark 4 is stable.

We conclude with a final note on what information each player needs to have access to in the bidding games. To be able to calculate the naïve loss of their rivals, the players need to have access to the payoffs of all the other players, as well as the whole value matrix. However, the coefficients of ill will could be private information, since naïve players will not care about other players' ill will anyway.

Acknowledgements

We thank two anonymous referees for useful comments. This research was supported by the CULTAPTATION project (European Commission contract FP6-2004-NEST-PATH-043434) and the Swedish Research Council.

References

Camerer, C.F. [2003] Behavioral game theory, Princeton University Press, 43–59

Dijkstra, J. [2009] Externalities in Exchange Networks: An Adaptation of Existing Theories of Exchange Networks, *Rationality and Society* **21(4)**, 395–427

Dijkstra, J. and van Assen, M.A.L.M. [2008] Effects of Externalities on Patterns of Exchange, *Sociological Theory and Methods* **23(1)**, 91–110

Dijkstra, J. and van Assen, M.A.L.M. [2008] The Comparison of Four Types of Every-

day Interdependencies: Externalities in Exchange Networks, *Rationality and Society* **20(1)**, 115–143

Fehr E and Schmidt KM [1999] A theory of fairness, competition and cooperation, *Quarterly Journal of Economics* **114**, 817–868

Hafalir, I.E. [2008] Stability of Marriage with Externalities, *International Journal of Game Theory* **37**, 353–369

Maskin, E. [2004] Bargaining, Coalitions and Externalities, unpublished paper at The 13th WZB Conference on Markets and Political Economy, Berlin, Germany

Roth, A.E., and Sotomayor M.A.O. [1990] Two-sided matching, Cambridge University Press

Sasaki, H. and Toda, M. [1996] Two-sided matching problems with externalities, *Journal of Economic Theory* **70**, 93–108

Shapley, L.S. and Shubik, M. [1972] The assignment game I: the core, *International Journal of Game Theory* **1**, 111–130