Strong implementation with partially honest individuals

Foivos Savva†

Abstract

In this paper we provide sufficient conditions for a social choice rule to be implementable in strong Nash equilibrium in the presence of partially honest agents, that is, agents who break ties in favor of a truthful message when they face indifference between outcomes. In this way, we achieve a relaxation in the condition of Korpela (2013), namely the Axiom of Sufficient Reason. Our new condition, Weak Pareto Dominance is shown to be sufficient along with Weak Pareto Optimality and Holocaust Alternative. We finally provide an application of our result in a pure matching environment.

1 Introduction

Implementation theory studies the relationship between social goals and institutions. Specifically, it aims to examine the effect of institutional design to the attainment of socially desirable outcomes. For example, suppose that a group of people have agreed on the desirable social outcomes as a function of their preferences. How can they make sure that they can indeed obtain those outcomes, when some or all of them may potentially benefit by misrepresenting their preferences? They thus have to rely on designing an institution (formally, mechanism or game form) through which they will interact, that will ensure the optimality of the outcomes reached through this interaction. More formally, for any collective choice rule that assigns some socially optimal outcomes as a function of individual preferences, implementation is achieved when, for any profile of preferences, the set of socially optimal outcomes coincides with the set of outcomes attained in the equilibrium of the game induced by the mechanism.

*Newer version accepted for publication at the Journal of Mathematical Economics. I want to thank my supervisors, Michele Lombardi, Takashi Hayashi and Anna Bogomolnaia for their invaluable comments and support.
†Adam Smith Business School, University of Glasgow, email: f.savva.2@research.gla.ac.uk
While most of the classic literature on the subject relies on the assumption that agents have a purely consequentialist nature, that is, they only care about the final outcomes, the strand of behavioural implementation theory typically assumes that agents may also have procedural concerns. One recent subfield in particular, takes into account the fact that agents may have an intrinsic preference for honesty. In the field of mechanism design this weak preference for honesty is usually modelled in the following manner: Suppose that an agent is indifferent between two outcomes. Then she will strictly prefer to send a truthful message rather than an untruthful one. This type of rationale is typically referred to as partial honesty or minimal honesty and can be supported by the experimental findings of Hurkens and Kartik (2009) for example, who show that subjects either are always honest, or tend to lie only when they gain by doing so. Despite being rather weak, partial honesty is shown to bear a significant positive effect for the set of implementable rules and limitations imposed by Maskin monotonicity\(^1\) in particular. In their seminal paper, Dutta and Sen (2012) show that in the presence of just one partially honest agent in the society, Maskin monotonicity is no longer necessary and No Veto Power alone becomes sufficient when society consists of more than three players.

Overall, the results on Nash implementation with partial honesty have been positive. An important question that remains though is whether these possibilities can be extended to other, possibly stronger, equilibrium concepts. For example, in many situations, the social planner cannot exclude the possibility of pre play communication between the agents and thus the mechanism may be vulnerable to group deviations. In such settings the natural solution concept to use is strong Nash equilibrium\(^2\) (Aumann (1959)), that is robust to deviations by any possible coalition of agents.

The current paper identifies sufficient conditions for strong implementation when all agents are partially honest. Instead of a full characterization, we chose to follow the work of Korpela (2013) in providing simple sufficient conditions that have a more intuitive appeal and are generally easier to check in applications. First, we identify sufficient conditions for strong implementation when all agents are partially honest and prove their sufficiency. Specifically, we show that if a social choice rule satisfies Weak Pareto Optimality (WPO), Holocaust Alternative (HA) and Weak Pareto Dominance (WPD), then it can be implemented in strong equilibrium. In this way we achieve a relaxation in the condition of Korpela.

\(^1\)Maskin (1999) in his seminal paper identified a condition now known as Maskin monotonicity as necessary and almost sufficient for Nash implementation. It roughly says that if an optimal outcome at some state does not fall in anyone’s ranking when switching to another state, then it should still be selected as optimal. A formal definition will be given later.

\(^2\)From now on we will use the terms strong equilibrium and strong Nash equilibrium interchangeably. The same applies for the respective implementation concepts.
(2013), namely the Axiom of Sufficient Reason (ASR). Our new condition, WPD roughly requires that if an outcome is optimal at some state, and if there is another outcome that is weakly preferable (weakly Pareto dominates it) for all agents in the same state, then the latter should be optimal as well. WPD is implied by ASR, therefore our condition is weaker. Next, we provide an application of our result in pure matching environments. More specifically, we show that the man (or woman) optimal solution in a pure matching environment is strongly implementable, when agents are partially honest.

The remainder of the paper is organized as follows: In section 2, we review the relevant literature. In section 3, we present the basic implementation setting and formal definitions. In section 4, we provide the definitions of our conditions, our main theorem and results. Section 5 is our applications section. Finally, in section 6 we conclude discussing our results and providing some points for further research. All main proofs are in the appendix, which is in section 7.

2 Related Literature

The problem of strong implementation has primarily been studied by Maskin (1978), providing necessary conditions, as well as some impossibility results. Moulin and Peleg (1982) also study the same issue with the use of effectivity functions. A complete characterization of strongly implementable social choice rules is due to Dutta and Sen (1991). Suh (1996) generalizes the latter result by allowing the planner to possibly exclude some coalition formation \textit{ex ante}, so in this more general setting not all coalitions are feasible. If the planner though cannot obtain such information, the relevant implementation concept is double implementation in Nash and strong equilibrium. Suh (1997) provides general results in this case as well. While complete characterizations are of high theoretical significance, they can be hard to apply to more specific settings. This motivates the more recent work by Korpela (2013) to identify simple sufficient conditions for strong implementation.

On the issue of partial honesty in implementation, the pioneering work of Dutta and Sen (2012) shows that \textit{No Veto Power} (NVP) alone becomes sufficient for Nash implementation in the presence of at least one partially honest agent\(^3\). Their results are generalized by Lombardi and Yoshihara (2014), who provide a full characterization of Nash implementable rules in the presence of partial honesty. Kartik et al. (2014) focus on environments with economic interest and identify sufficient conditions for implementation in two rounds of iterative deletion of strictly dominated strategies by “simple” mechanisms, that is, without utilizing the usual

\(^3\)In contrast with the case of no partial honesty, where NVP along with Maskin monotonicity are sufficient. The famous result is due to Maskin (1999).
canonical mechanisms\textsuperscript{4}. In other solution concepts with complete information, Saporiti (2014) shows that with partial honesty strategy-proofness is necessary and sufficient for secure implementation, which essentially requires implementation in dominant strategies and Nash equilibrium. Hagiwara (2016) also shows that NVP is sufficient with at least one, and unanimity is sufficient with at least two partially honest agents for double implementation in Nash and undominated Nash equilibria. Finally, Lombardi and Yoshihara (2016) explore the possibility of strategy space reduction in partially honest Nash implementation.

Partial honesty can yield positive results in incomplete information environments as well. For example, Matsushima (2008) shows that incentive compatibility is sufficient for implementation in strong iterative dominance and Korpela (2014) proves that incentive compatibility and NVP are sufficient for implementation in Bayes Nash equilibrium. Studies with alternative solution concepts include Ortner (2015), who provides more positive results with partial honesty in \textit{fault-tolerant Nash equilibrium}\textsuperscript{5} and \textit{stochastically stable equilibrium}.

The issue of implementation with partial honesty nevertheless can be put in the broader context of implementation with motives, where it is typically assumed that agents may also give significance to motives as procedural concerns, apart from the final outcomes. Along this line of research, it is worth mentioning a concept related to partial honesty, namely that of “social responsibility”. In Lombardi and Yoshihara (2017), the effect of social responsibility is explored with regards to natural implementation. In a different environment, Dogan (2017) shows that the unique socially optimal allocation of objects to agents can be Nash implemented, when at least three agents have a social responsibility motive. Some general results on motives as tie-breaking rules with regards to Nash implementation are in Kimya (2017). Other significant contributions to the literature of motives in implementation include Glazer and Rubinstein (1998) and Bierbrauer and Netzer (2016).

3 Preliminaries

Our society consists of a finite set of individuals $N = \{1, \ldots, n\}$ with $|N| = n \geq 3$. By $C \subseteq N$ we will denote a coalition of agents. The set of all possible social outcomes is denoted by $A$ and we typically assume that $|A| \geq 2$. Each agent $i$\textsuperscript{4} Jackson (1992) criticizes the use of canonical mechanisms in implementation theory as unrealistic and too permissive due to their unbounded strategy spaces. Instead, he derives a necessary condition for implementation with bounded mechanisms in undominated strategies. In a similar context, Mukherjee and Muto (2016) fully characterize the set of implementable rules in undominated strategies with bounded mechanisms when all agents are partially honest.

\textsuperscript{5}Fault-tolerant Nash equilibrium was first introduced by Eliaz (2002) as an equilibrium concept which is robust to the bounded rationality of a number of agents.
is endowed with a preference ordering (complete, reflexive and transitive binary relation) over \( A \) that is denoted by \( R_i \). We denote the set of all such possible orderings for \( i \) by \( \mathcal{R}_i \) and, as usual, by \( P_i \) and \( I_i \) we denote the asymmetric and symmetric part of \( R_i \) respectively. Define \( \mathcal{R} \equiv \times_{i \in N i} R_i \) with a typical element \( R = (R_1, \ldots, R_n) \) which we call a preference profile or simply, state. For each \( i \in N \) let \( L_i(a, R) = \{ b \in A | a R_i b \} \) be agent \( i \)'s lower contour set of outcome \( a \) in state \( R \). A Social Choice Rule (SCR) \( f \) is a correspondence \( f : \mathcal{R} \rightarrow A \) such that for all \( R \in \mathcal{R} \), \( \emptyset \neq f(R) \subseteq A \). A Social Choice Function (SCF) is a single-valued SCR. For any \( R \in \mathcal{R} \), we call \( f(R) \) the set of \( f \)-optimal outcomes in state \( R \).

A mechanism \( G \) is a pair \( (S, g) \), which consists of a strategy space \( S = \times_{i \in N S_i} \), with \( S_i \) being the set of available strategies for each \( i \in N \), and an outcome function \( g : S \rightarrow A \), that maps each strategy profile \( s = (s_1, \ldots, s_n) \in S \) to an outcome in \( A \). For convenience, let \( (s_i', s_{-i}) \) be the strategy profile where agent \( i \) plays the strategy \( s_i' \) while all other players \( j \neq i \) play the strategy \( s_j \). In a similar manner, let \( (s_i', s_N \setminus \{c_i\}) \) be the strategy profile where all \( i \in C \) play \( s_i' \), and all \( j \in N \setminus C \) play \( s_j \). Any mechanism \( (S, g) \) with a preference profile \( R \) define a normal form game \((S, g, R)\). We focus on the case of complete information, that is, the state \( R \) is common knowledge among the agents, while not to the planner.

In our setting, we assume that agents do not only care about the social outcomes, but also give some importance (although small) to the procedure that leads to those outcomes. More specifically, we assume that agents are partially honest in the following sense: If an agent is indifferent between two outcomes and she can attain those outcomes with two different strategies with one being “honest” and the other being “dis-honest”, then she strongly prefers to follow the honest strategy. More formally, in order for honesty to be meaningful in our setting, we define the strategy set of each \( i \in N \) to be \( S_i = \mathcal{R} \times M_i \). That is, each agent is required to announce a preference profile \( R \in \mathcal{R} \) and an arbitrary message \( m_i \in M_i \). Then, given mechanism \( G \), for any \( i \in N \) we define \( i \)'s truthful correspondence as \( T_i^G : \mathcal{R} \Rightarrow S_i \) such that for each agent \( i \), state \( R \) and message \( m_i \), \( T_i^G(R) = \{ R \} \times M_i \). The truthful correspondence represents the truthful strategies for each agent \( i \) in state \( R \), which essentially consist of announcing the “true” state. We now define agent \( i \)'s extended preferences on the strategy space \( S \) as follows. Given a mechanism \( G \), \( \forall i \in N, \forall R \in \mathcal{R} \), define \( \succeq_i^R \) as a complete, transitive and reflexive binary relation on \( S \). An extended preference profile in state \( R \) is defined as \( \succeq^R = (\succeq_1^R, \ldots, \succeq_n^R) \). We are now ready to proceed to the formal definition of partial honesty.

Given mechanism \( G \), an agent \( i \) is partially honest if \( \forall s_i, s'_i \in S_i, \forall s_{-i} \in S_{-i} \):

- \( [s_i \in T_i^G(R), s'_i \notin T_i^G(R) \text{ and } g(s_i, s_{-i}) R_i g(s'_i, s_{-i})] \Rightarrow (s_i, s_{-i}) \succeq_i^R (s'_i, s_{-i}) \).
- In all other cases, \( g(s_i, s_{-i}) R_i g(s'_i, s_{-i}) \iff (s_i, s_{-i}) \succeq_i^R (s'_i, s_{-i}) \)
An agent $i$ is not partially honest if $\forall s_i, s_i' \in S_i, \forall s_{-i} \in S_{-i}$:

- $g(s_i, s_{-i}) R_i g(s_i', s_{-i}) \iff (s_i, s_{-i}) \succeq_i (s_i', s_{-i})$

In other words, an agent cares about honesty in a lexicographic manner: First she “consults” her ordering over outcomes, and if she is indifferent between some, she consults her ordering over strategies, strongly preferring the honest strategies if they exist. That is, her partial honesty serves the purpose of a tie-breaking rule when she faces indifference. On the other hand, an agent that is not partially honest cares only about the outcomes and does not give significance to her strategies.

A mechanism $(S, g)$ with an extended preference profile $\succeq R$ in state $R$ define a game in normal form $(S, g, \succeq R)$. We typically assume that in our society there can be partially honest and not partially honest agents and we denote the set of partially honest agents by $H$. For the planner however, we only assume that he knows the class of all conceivable sets of partially honest agents, $H \subseteq 2^N$, without knowing which set is the actual one.

Regarding the solution concept, since we assume that players are allowed to collude, the equilibrium notion that we use is strong equilibrium. Formally, $s \in S$ is a strong equilibrium in the game $(S, g, R)$, if $\forall C \subseteq N, \forall s_C' \in S_C$, there exists an agent $i \in C$ such that $g(s) \succeq_i R_i g(s_C', s_{N\setminus C})$. In other words, a strategy profile is a strong equilibrium if there is no coalition that can deviate from it and make all of its members strictly better off. Let the set of strong equilibria of $(S, g, \succeq R)$ be $SE(S, g, R) = \{ s \in S | s$ is a strong equilibrium in $(S, g, R) \}$. We say that the mechanism $G = (S, g)$ implements the SCR $f$ in strong equilibrium, if in any state $R \in \mathcal{R}$, $SE(S, g, R) \subseteq f(R)$, and $SE(S, g, R) \supseteq f(R)$. The SCR $f$ is strongly implementable if there exists a mechanism that implements it in strong equilibrium.

The previous formal setting can be interpreted as follows. First of all, the SCR (or SCF) represents the collective choice rule that our society utilizes in order to make collective decisions. It can also be interpreted as the constitution of the society designed in an ex ante stage. A mechanism on the other hand represents the institution through which the agents in the society interact with each other, that is, it determines the rules and the outcomes of the interaction. A hypothetical benevolent social planner wishes to implement the SCR, however, he does not know the true state, hence, he relies on the agents in order to obtain this information. On the other hand, truthful revelation of the the state may not be in the best interests of some agents. Therefore, the goal of the social planner is to construct a mechanism that will lead to the optimal according to the SCR outcome, for any realization of the agents’ preferences, that is, for any preference profile. We thus require any optimal outcome to be attainable by some strong equilibrium and any strong equilibrium to lead to an optimal outcome.
4 Results

In this section, we present our main results. Before proceeding though, it would be helpful first to review the result of Korpela (2013). This will enable us to outline the weakening of the sufficient conditions for strong implementation when we adopt the partial honesty assumption. The conditions are the following:

**Holocaust Alternative (HA):** \( \exists a_H \in A, \text{ with } a_H \notin f(R), \forall R \in \mathcal{R}, \forall a \in A \setminus \{a_H\}. \)

**Weak Pareto Optimality (WPO):** \( \forall R \in \mathcal{R}, f(R) \subseteq wPO(A, R), \) where \( wPO(A, R) = \{a \in A | \exists b \in A \text{ such that } bP_i a, \forall i \in N\}. \)

**Axiom of Sufficient Reason (ASR):** \( \forall R, R' \in \mathcal{R}, \forall a \in f(R), \forall c \in A:\)

\[
\forall i \in N, L_i(a, R) \subseteq L_i(c, R') \Rightarrow c \in f(R')
\]

Intuitively, HA can be thought of as a worse alternative for all agents, that cannot ever be selected as an optimal outcome. It is a significant restriction on the preference domain, however, as will be shown later, it can be meaningful in several applications. It essentially allows us to overcome more complicated assumptions regarding the intersection of the lower contour sets, as in the condition of Dutta and Sen (1991). WPO restricts the range of the SCR to weakly Pareto optimal outcomes. Note that weak Pareto optimality is also a necessary condition for strong implementation. ASR can be interpreted as follows: Let an outcome \( a \) be selected as \( f \)-optimal for some preference profile \( R \). Now imagine an outcome \( b \) and profile \( R' \) such that \( b \) is weakly more preferable to \( a \) in \( R' \) by all agents. Then, \( b \) should be \( f \)-optimal in \( R' \). In other words, if every reason for \( a \) to be \( f \)-optimal in \( R \) is also a reason for \( b \) to be \( f \)-optimal in \( R' \), and \( a \) is indeed selected as an optimal outcome in \( R \), then \( b \) should be selected as an optimal outcome in \( R' \) as well. It is useful to note that ASR is stronger than Maskin monotonicity (MON) and unanimity (U) and it implies both. We review the formal definitions below:

**Maskin Monotonicity (MON):** \( \forall R, R' \in \mathcal{R}, \forall i \in N, \forall a \in f(R): \)

\[
\forall i \in N, L_i(a, R) \subseteq L_i(a, R') \Rightarrow a \in f(R')
\]

**Unanimity (U):** \( \forall R \in \mathcal{R}, \forall a \in A: \)

\[
\forall i \in N, A \subseteq L_i(a, R) \Rightarrow a \in f(R)
\]

For example, note that we obtain MON if in the definition of ASR we set \( c = a \). To see that it implies U, suppose that ASR holds, and for some state \( R \)
and outcome \( a \) we have that for all \( i \), \( A \subseteq L_i(a, R) \). Then, for any state \( R' \) and any outcome \( c \in f(R') \) it trivially holds that for all \( i \), \( L_i(c, R') \subseteq A \subseteq L_i(a, R) \), and from \( \text{ASR} \), \( a \in f(R) \) obtains. We are now ready to present Korpela’s theorem:

**Theorem 1** (Korpela (2013)). If a SCR \( f \) satisfies \( \text{HA} \), \( \text{WPO} \) and \( \text{ASR} \) then it is strongly implementable.

Theorem 1 makes no assumptions with regards to the partial honesty motive. Its significance lies on the simplicity and intuitive appeal of the conditions. Now proceeding to our results, we will utilize the following assumption, which summarize the knowledge of the social planner regarding the number of partially honest agents in the society.

**Assumption 1:** All agents in \( N \) are partially honest and the planner knows that.

As in the case of the Dutta and Sen (2012) in Nash implementation, our goal is to examine the effect of the presence of partially honest agents in the implementation problem. Moreover, we aim to determine whether partial honesty bears analogous significant impact in the case of strong implementation as in Nash implementation, given that the sufficient conditions for the former are much stronger than in the case of the latter. For our first result, we identify sufficient conditions for strong implementation when all agents are partially honest. A key condition is the following:

**Weak Pareto Dominance (WPD):** \( \forall R \in \mathcal{R}, \forall a \in f(R), \forall c \in A: \)

\[
L_i(a, R) \subseteq L_i(c, R), \forall i \in N \Rightarrow c \in f(R)
\]

Our condition dictates that any outcome that weakly Pareto dominates an \( f \)-optimal outcome must be selected by the SCR as \( f \)-optimal as well. Notice that our condition is implied by \( \text{ASR} \) and implies \( \text{U}^6 \). Along with \( \text{HA} \) and \( \text{WPO} \) our condition becomes sufficient for strong implementation, when all agents are partially honest, which is stated in our main theorem:

**Theorem 2.** Suppose that Assumption 1 holds. If a SCR \( f \) satisfies \( \text{WPO} \), \( \text{HA} \), \( \text{WPD} \), then it is strongly implementable.

*Proof.* See appendix.

Two comments are worth noting in this particular theorem. First, there is a significant weakening of the \( \text{ASR} \) which reduces to a Pareto related condition. The

\( ^6 \)To see that it is implied by \( \text{ASR} \) simply set \( R = R' \) in the definition of \( \text{ASR} \). To show that it implies \( \text{U} \), a similar argument as in the \( \text{ASR} \Rightarrow \text{U} \) implication applies.
second point to note is that if we only allow for linear orderings\footnote{Formally, let $\mathcal{L}_i$ be the set of all linear, that is, complete, transitive and antisymmetric, orders on $A$ for each agent $i$ and let $\mathcal{L} \equiv \times_{i \in N} \mathcal{L}_i$. Let the space of admissible preferences be $\mathcal{R}^A$. So, in this case we set $\mathcal{R}^A = \mathcal{L}$.}, then, WPD is implied by WPO. Therefore, WPD becomes redundant as a sufficient condition. Below we provide a formal proof and in Corollary 1 we state the sufficiency theorem for the case of linear preferences.

**Proposition 1.** If $\mathcal{R}^A = \mathcal{L}$, then WPO implies WPD.

*Proof.* Suppose that $\mathcal{R}^A = \mathcal{L}$ and assume that WPO holds. Now consider $R \in \mathcal{R}$, $c \in A$ and $a \in f(R)$ such that $\forall i \in N, L_i(a, R) \subseteq L_i(c, R)$. We distinguish two cases:

- $a = c$: Then, $c \in f(R)$ and WPD holds.
- $a \neq c$: Then, since we have linear preferences, $\forall i \in N, L_i(a, R) \subseteq L_i(c, R)$ implies that $\forall i \in N, cP_i a$. But, because $f$ satisfies WPO, this contradicts that $a \in f(R)$, so it cannot be the case.

\[\Box\]

**Corollary 1.** Let $\mathcal{R}^A = \mathcal{L}$ and assumption 2 hold. If a SCR $f$ satisfies WPO and HA, then it is strongly implementable.

*Proof.* Immediate implication of Theorem 2 and Proposition 1. \[\Box\]

5 Application in Pure Matching Environments

In this section we provide an application of our Theorem 1, in pure matching environments, that is, one-to-one matching environments where staying unmatched is not a feasible alternative, or it is the worst alternative for every agent. For example, a manager in a firm might want to match people from two groups in pairs, to undertake projects. In this case it might be reasonable to assume that staying unmatched is not feasible (as it might lead to redundancies). We show that when all agents are partially honest, the man-optimal (or woman-optimal) stable solution is strongly implementable. This is to be compared with the results of Tadenuma and Toda (1998), who show that with more than three agents in each group, while the whole stable solution in pure matching problems is Nash implementable, no single-valued subsolution of it is. Lombardi and Yoshihara (2014) show that partial honesty can resolve this issue for Nash implementation, as the man (or woman) optimal solution become Nash implementable in the presence of partial
honesty. With regards to strong implementation in matching environments, the
strong implementability of the stable rule in pure matching environments is shown
in Korpela (2013). What is more, Shin and Suh (1996) construct a mechanism for
strong implementation of the stable rule in one-to-one matching problems.

We start by defining the formal pure matching environment. Let $M, W$ be two
fixed and disjoint finite sets, such that $|M| = |W| \geq 2$. For all $i \in M$, $P_i$ is a
linear order on $W \cup \{i\}$, and for all $i \in W$, $P_i$ is a linear order on $M \cup \{i\}$. A
matching is a function $\mu : M \cup W \rightarrow M \cup W$ such that for any $i \in M \cup W$ the
following hold:

- $i \in M \& \mu(i) \neq i \Rightarrow \mu(i) \in W$,
- $i \in W \& \mu(i) \neq i \Rightarrow \mu(i) \in M$, and
- $\mu(\mu(i)) = i$.

Let $\mathcal{M}$ be the set of all matchings. We now extend the relation $P_i$ to the set of
matchings by defining a new relation $R_i$ as follows:

$\forall i \in M \cup W, \forall \mu, \mu' \in \mathcal{M}, \mu R_i \mu' \iff \mu(i) P_i \mu'(i)$ or $\mu(i) = \mu'(i)$

Let the set of all preferences over $\mathcal{M}$ of each agent $i$ be $R_i$. We then define
$R \equiv \times_{i \in M \cup W} R_i$. As usual, $R \in R$. Now we make the following assumption,
which makes our environment one of pure matching:

**Assumption 3:** We assume that $\forall m \in M, \forall w \in W, \forall \mu \in \mathcal{M}, wP_m m \& mP_w w$.

A solution (or SCR) is a correspondence $\varphi : R \Rightarrow \mathcal{M}$ such that for all
$R \in R, \varphi(R) \subseteq \mathcal{M}$. A pair $(m, w) \in M \times W$ blocks $\mu \in \mathcal{M}$ in $R \in R$ if
$wP_m \mu(m)$ and $mP_w \mu(w)$. A matching $\mu \in \mathcal{M}$ is stable in $R \in \mathcal{R}$, if there is
no pair $(m, w) \in M \times W$ such that $(m, w)$ blocks $\mu$ in $R$. Let $S(R)$ be the set of
all stable matchings in $R \in \mathcal{R}$. The stable matching rule is a rule $f^S : R \Rightarrow \mathcal{M}$
such that for every $R \in \mathcal{R}$, $f^S(R) = S(R)$. We say that $\mu^M \in \mathcal{M}$ is the man-
optimal stable matching in state $R \in \mathcal{R}$ if $\mu^M \in S(R)$ and for every $\mu' \in S(R)$ and
$m \in M$, we have that $\mu(m) P^M m(\mu')$. The man-optimal stable rule $f^M$ is a
function $f^M : R \rightarrow \mathcal{M}$ such that for every $R \in \mathcal{R}, f(R) = \mu^M$. In a similar
manner, we can define the woman-optimal stable matching and rule. We now proceed
by stating our possibility result for the pure matching environment.

**Proposition 2.** Let assumptions 1 and 3 hold. The man-optimal stable rule $f^M$
is strongly implementable.

**Proof.** It suffices to show that $f^M$ satisfies WPO, WPD, HA.

**Claim 1:** $f^M$ satisfies HA.
Proof. By the construction of the pure matching environment, we have assumed that staying single is the worst alternative for every $i \in M \cup W$. So, we can set $a_H = \mu_H$, where for all $i \in M \cup W$, $\mu_H(i) = i$. So, our environment satisfies HA.

Claim 2: $f^M$ satisfies WPO.

Proof. Suppose not. Then, consider $R \in \mathcal{R}$ and $\mu, \mu' \in \mathcal{M}$ with $\mu \neq \mu'$, such that $f^M(R) = \mu$ and for all $i \in M \cup W$, $\mu'(i) P_i \mu(i)$. Let $m \in M$ and $w \in W$ such that $\mu'(m) = w$. Consequently, it must be the case that $m R_m m$ and $w R_w \mu(w)$. However, $(m, w)$ would block $\mu$ in the first place which contradicts the stability of $\mu$. Therefore, $f^M$ satisfies WPO.

Claim 3: $f^M$ satisfies WPD.

Proof. Consider $R \in \mathcal{R}$ and let $f^M(R) = \mu^M$. Since the man-optimal stable rule $f^M$ is a function, it suffices to show that for any $\mu \in \mathcal{M}$, $\forall i \in M \cup W, L_i(\mu^M, R) \subseteq L_i(\mu, R) \Rightarrow \mu = \mu^M$. Suppose not. That is, suppose that there exists $\mu \neq \mu^M$ such that for all $i \in M \cup W$, $L_i(\mu^M, R) \subseteq L_i(\mu, R)$. Then, because $\mu \neq \mu^M$, there must exist $(m, w) \in M \cup W$, such that $\mu(m) P_m \mu^M(m)$ and $\mu(w) P_w \mu^M(w)$. This however contradicts the stability of $\mu^M$. So $f^M$ satisfies WPD.

We have showed that the man-optimal stable rule satisfies the sufficient conditions for strong implementation when all agents are partially honest and is therefore strongly implementable. This completes the proof.

6 Concluding remarks

We have provided a sufficiency theorem for strong implementation when all agents are partially honest and one for $k$-strong implementation. Our goal was to extend the positive results that have been obtained in partially honest Nash implementation to the solution concept of strong equilibrium. Our sufficient conditions are much stronger than in the case of Nash implementation and this is due to the much more demanding solution concept, as well as due to the attempt to provide simple sufficient conditions rather than a complete characterization.

As applications of our main theorem, we showed that the man (or woman) optimal stable solution in a pure matching environment is strongly implementable when all agents are partially honest. However, as noted before, it is not strongly implementable when there are no partially honest agents, therefore our results show the expansion of strongly implementable rules when the motive of minimal honesty is assumed.
In our view, the application of our theorem provide an insight into the possibilities that arise in implementation theory when non-consequentialist motives are taken into account, and emphasize the importance of procedural concerns in mechanism design and social choice theory. An interpretation of our results could also point to the significance of honesty as a human motive for the stability of a society or group.

Further research on the issue could potentially include a full characterization of strong implementation in the presence of partial honesty. In that way, the domain restriction of HA could be avoided and more clear-cut results on the strongly implementable rules can be obtained.

7 Appendix

7.1 Mechanism

For the proofs of the theorems 2 and 3 we will utilize the following mechanism $G = (S, g)$:

For all $i \in N$, $S_i = A \times \mathcal{R} \times \{NF, F\} \times \mathbb{N}_+$. The outcome function $g$ is defined as follows:

1) If $\forall i \in N, s_i = (a, R, NF, \cdot)$ and $a \in f(R)$, then $g(s) = a$.

2) If $\exists C \subset N, \forall i \in N \setminus C, s_i = (a, R, NF, \cdot)$ with $a \in f(R)$, and $\forall j \in C, s_j = (a^j, R^j, F, n^j)$, then:

   • If $k = \min\{\arg\max_{j \in C} n^j\}$ and $a^k \in \cup_{j \in C} L_j(a, R)$, then $g(s) = a^k$

   • Otherwise, $g(s) = a$

3) If $\forall i \in N, s_i = (a^i, R^i, F, n^i)$, then $k = \min\{\arg\max_{j \in N} n^j\}$ and set $g(s) = a^k$.

4) If none of the above apply, set $g(s) = a_H$.

7.2 Proof of theorem 2

We will show that a SCR $f$ that satisfies our premises, namely HA, WPD and WPO can be implemented by mechanism $G$ and we break the proof into two parts:

**Part 1:** $\forall R \in \mathcal{R}, f(R) \subseteq SE(R)$

Let the true state be $R^*$. Consider the strategy profile where $\forall i \in N, s_i = (a, R^*, NF, \cdot)$. If $j \in N$ deviates she will obtain any $b \in L_j(a, R^*)$. So, $g(S_j, s_{N \setminus \{j\}}) = L_j(a, R^*)$. If any $C \subset N$ deviates, the obtained outcome will be in $L_j(a, R^*)$ for at least one $j \in C$. Finally, if $N$ deviate, there cannot be an improvement for all
\( i \in N \) since \( f \) satisfies WPO. Therefore, \( s \) is a strong equilibrium in \( R^* \).

**Part 2:** \( \forall R \in \mathcal{R}, SE(R) \subseteq f(R) \)

Let the true state be \( R^* \). We proceed by first proving three useful claims:

**Claim 1:** There is no strong equilibrium under rule 1 where \( \forall i \in N, R_i \neq R^* \).

*Proof.* Suppose there exists a strong equilibrium under rule 1, where \( \forall i \in N, s_i = (a, R, NF, \cdot) \) with \( a \in f(R) \) and \( R \neq R^* \). By rule 1 the outcome is \( a \). Then, \( \forall i \in N, s_i \notin T_i^G(R^*) \), so, any \( i \in N \) can deviate to \( s'_i = (a, R^*, F, n^i) \in T_i^G(R^*) \) inducing rule 2 while announcing the true state and not changing the outcome. Therefore, \( s \) cannot be a strong equilibrium. \( \square \)

**Claim 2:** There is no strong equilibrium under rule 2 where \( \exists i \in N \setminus C \) such that \( R^* \neq R^* \).

*Proof.* Suppose there exists a strong equilibrium under rule 2 where \( \exists i \in N \setminus C, s_i = (a, R, NF, \cdot) \) with \( a \in f(R) \), \( R \neq R^* \), and \( \forall j \in C, s_j = (a_j, R_j, F, n^j) \) and let \( g(s) = b \). Then, we have that \( s_i \notin T_i^G(R^*) \). We break the proof into two cases:

**Case 1:** \( |N \setminus C| \geq 2 \)

- If \( b = a \): Then, since by definition \( a \in L_i(a, R) \) holds, \( i \) can play \( s'_i = (a, R^*, F, n^i) \in T_i^G(R^*) \) with a sufficiently high integer without changing the outcome and become strictly better off by Rule 2.

- If \( b \neq a \): Then, again, since \( b \in \bigcup_{j \in C} L_j(a, R) \) it must hold that \( b \in \bigcup_{j \in C \cup \{i\}} L_j(a, R) \), so agent \( i \) can play \( s'_i = (b, R^*, F, n^i) \in T_i^G(R^*) \) with a sufficiently high integer without changing the outcome and become strictly better off by Rule 2.

**Case 2:** \( N \setminus C = \{i\} \)

In this case \( i \) can play \( s'_i = (b, R^*, F, n^i) \in T_i^G(R^*) \) with a sufficiently high integer without changing the outcome and become strictly better off by Rule 3.

Therefore, there is no strong equilibrium under rule 2, where for some \( i \in N \setminus C, R^i \neq R^* \). \( \square \)

**Claim 3:** There is no strong equilibrium under rule 2 where \( \exists i \in C, \) with \( R^i \neq R^* \).

*Proof.* Suppose this is not the case, that is, there exists a strong equilibrium under rule 2 such that \( \exists i \in C, \) with \( R^i \neq R^* \). Also, by Claim 2, we have established that in any strong equilibrium that falls in Rule 2, \( \forall j \in N \setminus C, R^j = R^* \). So, we consider a case where \( \forall j \in N \setminus C, s_j = (a, R^*, NF, \cdot) \) with \( a \in f(R^*) \) and
\( \forall k \in C, s_k = (a^k, R^k, F, n^k) \) such that \( R^k \neq R^* \) for some \( i \in C \), that is, \( \exists i \in C \) such that \( s_i \notin T_i^G(R^*) \). Moreover, let \( g(s) = b \). Now we take two mutually exclusive cases:

**Case 1:** \(|C| \geq 2\)

- If \( b = a \), then, since we have that \( a \in L_i(a, R^*) \) by definition, agent \( i \) can play \( s'_i = (a, R^*, F, n^i) \in T_i^G(R^*) \) with a sufficiently high \( n^i \) inducing Rule 2 without changing the outcome and becoming strictly better off.

- If \( b \neq a \), where \( l = \min\{\arg\max_{j \in C} n^j\} \), we distinguish two cases:
  - \( l \neq i \): In this case, since \( a \in \cup_{j \in C} L_j(a, R^*) \), agent \( i \) can deviate to \( s'_i = (b, R^*, F, n^i) \in T_i^G(R^*) \), win the integer game for a sufficiently high integer without affecting the outcome, and thus become better off by Rule 2.
  - \( l = i \): Again, \( a \in \cup_{j \in C} L_j(a, R^*) \), so \( i \) can play \( s'_i = (b, R^*, F, n^i) \in T_i^G(R^*) \) and again become better off by Rule 2.

**Case 2:** \( C = \{i\} \).

- If \( b = a \), then \( i \) can deviate to \( s'_i = (a, R^*, NF, \cdot) \in T_i^G(R^*) \) inducing Rule 1 and become better off by announcing the truth.

- If \( b \neq a \), then it must be that \( b = a^i \). So, since \( b \in L_i(a, R^*) \), \( i \) can revert to truth-telling by playing \( s'_i = (b, R^*, F, n^i) \in T_i^G(R^*) \) and become better off by Rule 2.

Therefore, there is no strong equilibrium under rule 2 where \( \exists i \in C \) such that \( R^i \neq R^* \).

**Claim 4:** There is no strong equilibrium under rule 3 where \( \exists i \in N \), with \( R^i \neq R^* \).

**Proof.** Suppose there exists a strong equilibrium under rule 3 where \( \forall j \in N, s_j = (a^j, R^j, F, n^j) \), \( g(s) = b \) and let \( R^i \neq R^* \) for some \( i \in N \), that is, \( \exists i \in N \) such that \( s_i \notin T_i^G(R^*) \). Then, \( h \), can deviate to \( s'_h = (b, R^*, F, n^h) \in T_i^G(R^*) \) and obtain \( b \) while announcing the true state \( R^* \), for a sufficiently high integer \( n^h \). Therefore, \( s \) cannot be a strong equilibrium.

**Corollary 2.** In any strong equilibrium \( s \) of the mechanism \( G \), it holds that \( \forall i \in N, R^i = R^* \).

**Proof.** Immediate implication of claims 1-4 as well as of the fact that there cannot exist any strong equilibria under rule 4.
By the above arguments, we can restrict attention to strong equilibria where \( \forall i \in N, R_i = R^* \).

Consider a strong equilibrium under rule:

1. That is, \( \forall i \in N, s_i = (a, R^*, NF, \cdot) \). Then \( g(s) = a \in f(R^*) \).

2. That is, \( \forall i \in N \setminus C, s_i = (a, R^*, F, \cdot) \) with \( a \in f(R^*) \), and \( \forall j \in C, s_j = (a_j, R^*, F, n_j) \). Let \( g(s) = b \). We distinguish two cases:

   \[ |N \setminus C| \geq 2: \] Then, it must be that \( \forall i \in N \setminus C, g(S_i, s_{N \setminus \{i\}}) = \bigcup_{j \in C} L_j(a, R^*) \) and \( \forall j \in C, g(S_j, s_{N \setminus \{j\}}) = \bigcup_{j \in C} L_j(a, R^*) \), from Rule 2. For \( s \) to be a strong equilibrium, it must hold that \( \forall i \in N \setminus C, L_i(a, R^*) \subseteq \bigcup_{j \in C} L_j(a, R^*) \) and, \( \forall j \in C, L_j(a, R^*) \subseteq \bigcup_{j \in C} L_j(a, R^*) \subseteq L_j(b, R^*) \). So, for any \( i \in N \) we have that \( L_i(a, R^*) \subseteq L_i(b, R^*) \). From \textbf{WPD} it follows that \( b \in f(R^*) \).

   \[ N \setminus C = \{i\}: \] Then, for \( i \) it must hold that \( g(S_i, s_{N \setminus \{i\}}) = A \) from rule 3, and \( \forall j \in C \) it must hold that \( g(S_j, s_{N \setminus \{j\}}) = \bigcup_{j \in C} L_j(a, R^*) \) by rule 2. For \( s \) to be a strong equilibrium, it must hold that \( \forall i \in N \setminus C, L_i(a, R^*) \subseteq A \subseteq L_i(b, R^*) \) and \( \forall j \in C, L_j(a, R^*) \subseteq \bigcup_{j \in C} L_j(a, R^*) \subseteq L_j(b, R^*) \). So for all \( i \in N \) it holds that \( L_i(a, R^*) \subseteq L_i(a, R^*) \). Again, from \textbf{WPD} we must have that \( b \in f(R^*) \).

3. That is, \( s_i = (a^i, R^*, F, n^i) \), \( \forall i \in N \) and let \( g(s) = b \). Then, \( \forall i \in N \), it must hold that \( g(S_i, s_{N \setminus \{i\}}) = A \). Since \( f(R^*) \neq \emptyset \), there exists \( a \in f(R^*) \) such that \( \forall i \in N, L_i(a, R^*) \subseteq A \). Moreover, for \( s \) to be a strong equilibrium it must be that \( \forall i \in N, L_i(a, R^*) \subseteq A \subseteq L_i(b, R^*) \). Then, from \textbf{WPD} it must hold that \( b \in f(R^*) \).

This completes the proof.
References

Aumann, R. (1959), ‘Acceptable points in general cooperative n-person games, contributions to the theory of games iv, annals of mathematics studies (vol. 40)’.

Bierbrauer, F. and Netzer, N. (2016), ‘Mechanism design and intentions’, Journal of Economic Theory 163, 557–603.

Dogan, B. (2017), ‘Eliciting the socially optimal allocation from responsible agents’.

Dutta, B. and Sen, A. (1991), ‘Implementation under strong equilibrium: A complete characterization’, Journal of Mathematical Economics 20(1), 49–67.

Dutta, B. and Sen, A. (2012), ‘Nash implementation with partially honest individuals’, Games and Economic Behavior 74(1), 154–169.

Eliaz, K. (2002), ‘Fault tolerant implementation’, The Review of Economic Studies 69(3), 589–610.

Glazer, J. and Rubinstein, A. (1998), ‘Motives and implementation: On the design of mechanisms to elicit opinions’, journal of economic theory 79(2), 157–173.

Hagiwara, M. (2016), ‘Double implementation with partially honest agents’.

Hurkens, S. and Kartik, N. (2009), ‘Would i lie to you? on social preferences and lying aversion’, Experimental Economics 12(2), 180–192.

Jackson, M. O. (1992), ‘Implementation in undominated strategies: A look at bounded mechanisms’, The Review of Economic Studies 59(4), 757–775.

Kartik, N., Tercieux, O. and Holden, R. (2014), ‘Simple mechanisms and preferences for honesty’, Games and Economic Behavior 83, 284–290.

Kimya, M. (2017), ‘Nash implementation and tie-breaking rules’, Games and Economic Behavior 102, 138–146.

Korpela, V. (2013), ‘A simple sufficient condition for strong implementation’, Journal of Economic Theory 148(5), 2183–2193.

Korpela, V. (2014), ‘Bayesian implementation with partially honest individuals’, Social Choice and Welfare 43(3), 647–658.

Lombardi, M. and Yoshihara, N. (2014), ‘Partially honest nash implementation: A full characterization’. 16
Lombardi, M. and Yoshihara, N. (2016), ‘Treading a fine line:(im) possibilities for nash implementation with partially-honest individuals’.

Lombardi, M. and Yoshihara, N. (2017), ‘Natural implementation with semi-responsible agents in pure exchange economies’, *International Journal of Game Theory* pp. 1–22.

Maskin, E. (1978), ‘Implementation and strong nash equilibrium’.

Maskin, E. (1999), ‘Nash equilibrium and welfare optimality’, *The Review of Economic Studies* 66(1), 23–38.

Matsushima, H. (2008), ‘Role of honesty in full implementation’, *Journal of Economic Theory* 139(1), 353–359.

Moulin, H. and Peleg, B. (1982), ‘Cores of effectivity functions and implementation theory’, *Journal of Mathematical Economics* 10(1), 115–145.

Mukherjee, S. and Muto, N. (2016), Implementation in undominated strategies with partially honest agents, Technical report, mimeo.

Ortner, J. (2015), ‘Direct implementation with minimally honest individuals’, *Games and Economic Behavior* 90, 1–16.

Saporiti, A. (2014), ‘Securely implementable social choice rules with partially honest agents’, *Journal of Economic Theory* 154, 216–228.

Shin, S. and Suh, S.-C. (1996), ‘A mechanism implementing the stable rule in marriage problems’, *Economics Letters* 51(2), 185–189.

Suh, S.-C. (1996), ‘Implementation with coalition formation: A complete characterization’, *Journal of Mathematical Economics* 26(4), 409–428.

Suh, S.-C. (1997), ‘Double implementation in nash and strong nash equilibria’, *Social Choice and Welfare* 14(3), 439–447.

Tadenuma, K. and Toda, M. (1998), ‘Implementable stable solutions to pure matching problems’, *Mathematical Social Sciences* 35(2), 121–132.

Vartiainen, H. (2007), ‘Nash implementation and the bargaining problem’, *Social Choice and Welfare* 29(2), 333–351.