FAIRNESS IN MECHANISM DESIGN

YARON AZRIELI AND RITESH JAIN

February 13, 2015

Abstract. In a standard Bayesian environment with independent private values and two possible alternatives, it is shown that a social choice function which maximizes utilitarian welfare in the class of incentive compatible and anonymous social choice functions only if it is a qualified weighted majority rule. We also introduce a notion of fairness at the interim stage. It is shown that only interim fair SCF's are constant functions. Finally, we show that we can implement any incentive social choice function with a symmetric mechanism, perhaps indirect.

Keywords: Qualified Weighted majority rules; Pareto efficiency; Incentive compatibility.

JEL Classification: C72, D01, D02, D72, D82.

Department of Economics, The Ohio State University, 1945 North High street, Columbus, OH 43210.
e-mails: azrieli.2@osu.edu, jain.224@osu.edu

1. Introduction

What is the tradeoff between efficiency and fariness, when agents hold relevant private information? This is one of the most basic and important questions analyzed by the extensive mechanism design literature of the last several decades. In this paper we address this question in what is probably the simplest non-trivial environment: Society needs to choose between two alternatives, say *reform* or *status-quo*. Each agent is privately informed about his utility for each of the alternatives. The uncertainty of agents about the types of their opponents is captured by a common prior distribution, which we assume to be independent across agents. Monetary transfers are prohibited.

While the environment we analyze is simple, we emphasize that no symmetry between the agents or the alternatives is assumed. Even at the ex-ante stage, different agents may have different utility distributions, and these distributions may be biased in favor of one of the alternatives. We believe that this is an important feature of our analysis, since many real-life environments are inherently asymmetric. Examples include representative democracies with heterogenous district sizes, publicly held firms with institutional and private shareholders, and faculty hiring decisions in which the job candidate has closer research interests to some faculty members than others.

The simplicity of the environment enables us to get a clear and precise answer to the above basic question, by characterizing the class of efficient and fair mechanisms subject to incentive compatibility. In our analysis an important family of anonymous voting rules play a major role. These are qualified majority rules.

Let us be more explicit regarding what we call a qualified weighted majority rule. A Social Choice Function (SCF) f is a mapping from type profiles to lotteries over {reform, status-quo}. We say that f is a qualified weighted majority rule if we can find a positive quota q, such that under f the reform is implemented if the number of agents that prefer the reform exceeds q, and the status-quo prevails if this sum is less than q. Ties are allowed to be broken in an arbitrary way.

The most natural way to think about fariness is considering anonymous mechanisms. Anonymity means that the chosen alternative depends only on the reported types of agents and not on their names. In other words, anonymity requires the symmetry of the social choice function f

Our first result (Theorem 1) characterizes SCFs that maximize (ex-ante, utilitarian) social welfare subject to incentive compatibility and anonymity.[1] We show that a SCF is a solution to this optimization problem if and only if it is a qualified weighted majority rule. In particular, incentive compatibility constraints prevent effective use of any information about *realized* intensity of preferences; only the ordinal ranking of the two alternatives as reported by the agents matters for the outcome.

We then move on to study fairness at the interim stage. At the interim stage an agent knows his realized type but does not know the type of others. We call the interim fariness as equal treatment of equals at the interim stage. Equal treatment of equals demand that for any two agents with same type realization should have similar interim allocations. This property can be seen as a weaker version of anonymity at the interim stage.

Our second result (Theorem 2) characterizes SCFs that maximize (ex-ante, utilitarian) social welfare subject to incentive compatibility and satisfy interim fairness. We show that from the point of utilitarianism the set of incentive compatible and interim fair SCF's are constant functions.

Finally we consider the use of indirect mechanisms. Restrictring attention to direct mechanisms is with loss of generality in this environment. The reason is that, with anonymity, the revelation principle is no longer valid. Instead of talking about a fair SCF, we can think about a fair mechanism, perhaps indirect. What are the class of social choice functions which can be implemented by an anonymous mechanism? We give a very clear answer to this question, Theorem 3, by showing that the set of of incentive compatible social choice functions is precicely the set of SCF's that can be implemented by a symmetric mechanism. This presents a sharp contrast between the possibilites of using a Direct verses indirect mechanims.

The environment we study has very unique features (two alternatives, independent private values, no transfers)However, from a practical point of view, the two alternatives case is perhaps the most interesting one to study, since binary decision problems are frequent. The no transfers assumption is also realistic, since in many cases they are infeasible or excluded for ethical reasons. Type independence is more restrictive in our context, but given the pervasiveness of this assumption in the literature we think that it is an interesting benchmark to study.

[1] By the revelation principle, and assuming that agents play Bayes-Nash equilibrium, incentive compatibility constraints exactly characterize the class of feasible SCFs.

2. Environment

We consider a standard Bayesian environment à la Harsanyi. The set of agents is $N = \{1, 2, \ldots, n\}$ with $n \geq 1$. For each $i \in N$, T_i is a finite set of possible types of agent i, and t_i denotes a typical element of T_i. The type of agent i is a random variable $\hat{t}_i$ with values in T_i. The distribution of $\hat{t}_i$ is[2] $\mu_i \in \Delta(T_i)$, which we assume has full support. Let $T = T_1 \times \cdots \times T_n$ be the set of type profiles. We assume that types are independent across agents, so the distribution of $\hat{t} = (\hat{t}_1, \ldots, \hat{t}_n)$ is the product distribution $\mu = \mu_1 \times \cdots \times \mu_n \in \Delta(T)$. As usual, a subscript $-i$ means that the ith coordinate of a vector is excluded.

Let $A = \{\text{reform}, \text{status-quo}\} = \{r, s\}$ be the set of alternatives. The utility of each agent depends on the chosen alternative and on his own type only (private values). Specifically, the utility of agent i is given by the function $u_i : T_i \times A \to \mathbb{R}$. For ease of notation we write $u_i^r(t_i) = u_i(t_i, r)$, $u_i^s(t_i) = u_i(t_i, s)$ and $u_i(t_i) = (u_i^r(t_i), u_i^s(t_i))$. For expositional purposes, we assume that no agent is ever indifferent between the two alternatives, that is $u_i^r(t_i) \neq u_i^s(t_i)$ for every $t_i \in T_i$ and every $i \in N$. In addition, we assume that for every i there are $t_i, t_i' \in T_i$ such that $u_i^r(t_i) > u_i^s(t_i)$ and $u_i^r(t_i') < u_i^s(t_i')$. We normailize the utility of the agents by assuming $u_i^s(t_i') = 0$.

Since randomization over alternatives will be considered, we need to extend each $u_i(t_i, \cdot)$ to $\Delta(A)$. We identify $\Delta(A)$ with the interval $\{(p, 1-p) \ : \ 0 \leq p \leq 1\} \subseteq \mathbb{R}^2$, where the first coordinate corresponds to the probability of r and the second coordinate to the probability of s. With abuse of notation we write $u_i(t_i, (p, 1-p)) = p u_i^r(t_i) + (1-p) u_i^s(t_i)$ for $0 \leq p \leq 1$.

A Social Choice Function (SCF) is a mapping $f : T \to \Delta(A)$. The set of all SCFs is denoted F. It will be useful to think about F as a (convex, compact) subset of the linear space $\mathbb{R}^{2|T|}$. Thus, if $f, g \in F$ and $\alpha \in [0, 1]$ then $\alpha f + (1-\alpha) g \in F$ is defined by $(\alpha f + (1-\alpha) g)(t) = \alpha f(t) + (1-\alpha) g(t) \in \Delta(A)$.

For every agent i, type $t_i \in T_i$ and SCF f we denote by $U_i(f|t_i)$ the interim expected utility of agent i under f conditional on him being of type t_i:

$$U_i(f|t_i) = \mathbb{E}\left(u_i\left(\hat{t}_i, f\left(\hat{t}\right)\right) \ | \ \hat{t}_i = t_i\right) = u_i(t_i) \cdot \mathbb{E}\left(f(t_i, \hat{t}_{-i})\right),$$

[2] For every finite set X, $\Delta(X)$ denotes the set of probability measures on X.

where $x \cdot y$ denotes the inner product of the vectors x and y. The ex-ante utility of agent i under f is

$$U_i(f) = \mathbb{E}\left(u_i\left(\hat{t}_i, f\left(\hat{t}\right)\right)\right) = \sum_{t_i \in T_i} \mu_i(t_i) U_i(f|t_i).$$

Definition 1. *A SCF f is* Incentive Compatible (IC) *if truth-telling is a Bayesian equilibrium of the direct revelation mechanism associated with f. Namely, if for all $i \in N$ and all $t_i, t_i' \in T_i$, we have*

$$u_i(t_i) \cdot \left(\mathbb{E}\left(f(t_i, \hat{t}_{-i})\right) - \mathbb{E}\left(f(t_i', \hat{t}_{-i})\right)\right) \geq 0. \tag{1}$$

The set of all IC SCFs is denoted F^{IC}.

Voting rules that discriminate between voters, in the sense of giving more power to one group of voters over another, are often excluded as violating the basic fairness criterion of “one person, one vote”. We emphasize that we do *not* assume that the environment is symmetric between the agents – different agents may have different utility distributions.

We restrict attention to SCFs which satisfies the following property. Let all agents have the same set of type labels, we call a SCF anonymous if a permutation of agents' reports does not change the outcome. In other words the outcome depends only on the reported types and not the name of the agents.

Definition 2. *A SCF f is* anonymous *if for every permutation π of agents $f(t_1, ..., t_n) = f(t_{\pi(1)}, ..., t_{\pi(1)})$ for every $t \in T$. The class of anonymous SCFs is denoted F^{ANO}.*

2.1. **Ordinal rules and weighted majority rules.** We now define a class of SCFs that will have an important part in the analysis below. For each agent i, let P_i be the partition of T_i into the two (non-empty) sets

$$\begin{aligned} T_i^r &= \{t_i \in T_i : u_i^r(t_i) > 0\}, \\ T_i^s &= \{t_i \in T_i : u_i^r(t_i) < 0\}. \end{aligned}$$

Recall that agents are never indifferent, so every type t_i is in exactly one of these sets. The partition P_i reflects the ordinal preferences of agent i over the alternatives. Let P be the partition of T which is the product of all the P_i's: t and t' are in the same element of P if and only if t_i and t_i' are in the same element of P_i for every $i \in N$. As usual, let $P(t)$ be the element of the partition P that contains the type profile t.

Definition 3. *A SCF f is* ordinal *if it is P-measurable, i.e., if $f(t) = f(t')$ whenever $P(t) = P(t')$. The set of all ordinal SCFs is denoted F^{ORD}.*

Thus, an ordinal SCF depends only on the ordinal information in the reported type profile, and is not affected by changes in the expressed intensity of preference.

We discuss an imporatnt class of ordinal social choice function induced by f, this is the class of conditional expectation function of f w.r.t. partition P. We define a P measurable function $g : T \to \Delta(A)$

$$g(t) = \mathbb{E}[f|P](t) \tag{2}$$

Let the set of all SCF's g be G. The relation between these two functions will be explored in the following sections.

3. The utilitarian rule

We start the analysis by considering the problem of maximizing social welfare, i.e., the sum of ex-ante expected utilities of all the agents. It will be convenient to denote $v^r(t) = \sum_{i \in N} u_i^r(t_i)$, $v^s(t) = \sum_{i \in N} u_i^s(t_i)$ and $v(t) = (v^r(t), v^s(t))$ for every $t = (t_1, \ldots, t_n) \in T$. These are the welfare totals for each of the two alternatives when t is the realized type profile.

Definition 4. *The* (ex-ante) social welfare *of a SCF f is* $V(f) = \sum_{i \in N} U_i(f) = \mathbb{E}\left(v\left(\hat{t}\right) \cdot f\left(\hat{t}\right)\right)$.

Without requiring incentive compatibility and anonymity, a maximizer of social welfare simply chooses r whenever $v^r(t) > v^s(t)$ and chooses s if the other inequality holds (anything, including randomization, can be chosen when $v^r(t) = v^s(t)$). However, such SCFs will typically not be IC and will be "unfair", since agents will have an incentive to exaggerate the intensity of their preference. Moreover such SCF's include dictatorial rules. The following theorem characterizes maximizers of social welfare subject to incentive compatibility and anonymity.

Before we state the theorem, we would give the following remarks

Remark 1. *The one to one mapping between F and G is preserved if we add incentive compatibility. In other words, for every incentive compatible social choice function there is a ordinal incentive compatible social choice function generating the same welfare. This is Lemma 2 in* [Azrieli and Kim, 2014]. *This lemma makes the optimization problem simpler.*

Remark 2. *This mapping is not preserved if we add anonymity. An incentive compatible, anonymous SCF may induce a non-anonymous SCF.*

Example 1. $T_1 = T_2 = \{2, 1, -2, -1\}$
$\mu_1 = \{1/6, 1/6, 1/6, 1/2\}; \mu_2 = \{1/8, 1/8, 1/4, 1/2\}$

T	*2*	*1*	*-1*	*-2*
2	*1*	*1*	*1*	*0*
1	*1*	*1*	*1*	*0*
-1	*1*	*1*	*1*	*0*
-2	*0*	*0*	*0*	*1*

T	*2*	*1*	*-1*	*-2*
2	*1*	*1*	*1/3*	*1/3*
1	*1*	*1*	*1/3*	*1/3*
-1	*1/4*	*1/4*	*7/12*	*7/12*
-2	*1/4*	*1/4*	*7/12*	*7/12*

Theorem 1. *For 2 agent economies i.e. N=2. A SCF f is a maximizer of V in the class $F^{IC} \cap F^{ANO}$ only if f is a qualified majority rule.*

The extension of this case to more than 2 agents does not follow directly from the case of two agents. It is work in progress at this point. We have results from some special cases.

4. Interim-fariness

In this section we consider another notion of fairness defined at the interim stage. At the interim stage an agent knows his type but does not know the types of other agents. We formalize interim fairness using the idea of equal treatment of equals. Interim fairness demands that for any two agents with same type should recieve the same interim allocations. This means that if two agents prefer reform by 2 utils then the chance of choosing reform should be similar for both the agents. The following theorem characterizes the class of SCF's which are incentive compatibile and satisfy interim-anonymity.

Lemma 1. *There is a one to one mapping from the space $F^{IC} \cap F^{I-ANO}$ to $G^{IC} \cap G^{I-ANO}$. This is a version of Lemma 2 in* [Azrieli and Kim, 2014]

Theorem 2. *A SCF f is a maximizer of V in the class $F^{IC} \cap F^{I-ANO}$ if and only if f is a constant function*

Interim anonymity tunrs out to be a strong property on SCF's. It is very simple to show that a constant function satisfies interim anonymity. The main message of this theorem is that these are the "only" interim fair SCF's.

5. Symmetric Implementation

Usually fairness considerations, both practical and legal, restrict the mechanism designer from exploiting the heterogeneity of agents in maximizing the utilitarian welfare For e.g. in auction theory, an optimal auction uses bidder heterogeneity to maximize welfare. In our context the optimal voting rule may not be anonymous. It is natural to ask if there exists a symmetric mechanism, perhaps indirect, which implements a SCF. This approach has taken by [Deb and Pai, 2014] in the context of private value auctions. We say that a SCF is symmetrically implemented if there exists a mechanism which implements it. Our next theorem characterizes the class of SCF's which can be symmetrically implemented.

Definition 5. *A mechanism is a tuple (M,g), where* $M = \times_{i=1}^{n} M_i$ *and* $g : M \mapsto \Delta(X)$ Once a state of the world realizes i.e. $t \in T$ this game form induces a game of incomplete information among the agents. We assume that people play Bayes Nash equilibrium of this game of incomplete information.

Definition 6. *Let* $BNE : T \mapsto 2^{\Delta(X)}$*, be a mapping which specifies the set of Bayes Nash equilibria of the game for every* $t \in T$

Definition 7 (Symmetric Implementation)**.** *A mechanismn* (M, g) *is said to implement a SCF* $f : T \mapsto \Delta(X)$ *, if it satisfies the following conditions*

(1) $(\forall t \in T)(BNE(t) \bigcap f(t) \neq \emptyset)$
(2) $g : M \mapsto \Delta(X)$ *is symmetric, i.e. only the message matters and not the identity of the agents*

Example 2. $T_1 = T_2 = \{2, 1, -2, -1\}$
$\mu_1 = \{1/6, 1/6, 1/6, 1/2\}; \mu_2 = \{1/8, 1/8, 1/4, 1/2\}$

T	$+$	$+$	-	-
$+$	1	1	1	1
$+$	1	1	1	1
-	0	0	0	0
-	0	0	0	0

M	W_1^r	W_1^s	W_2^r	W_2^s
W_1^r	p_1	p_1	1	1
W_1^s	p_1	p_1	0	0
W_2^r	1	0	0	0
W_2^s	1	0	0	0

Theorem 3. *A SCF f can be implemented symmetrically if and only if f is incentive compatible*

Proof. We provide a mechanism which implements any incentive compatible social choice function.

$M_i = \{T_i^r, T_i^s\} \times \{1, 2, ..., n\}$

$$g(m) = \begin{cases} f(t) & if \quad \text{everyone announces a different number} \\ \mathrm{s}_i & if \quad \text{The message contains } n-1 \text{ distinct numbers and } i \text{ is missing} \\ c & if \quad O.W \end{cases}$$

Now, the mechanism is symmetric by construction. What we need to show is that any incentive compatible social choice function can be implemented symmetrically. □

References

[1] J. Apesteguia, M.A. Ballester and R. Ferrer (2011), On the justice of decision rules, *Review of Economic Studies* **78**, 1-16.

[2] W.W. Badger (1972), Political individualism, positional preferences, and optimal decision rules, In *Probability Models of Collective Decision Making*, edited by R.G. Niemi and H.F. Weisberg. Merrill, Columbus, Ohio.

[3] C. Beisbart and L. Bovens (2007), Welfarist evaluations of decision rules for boards of representatives, *Social Choice and Welfare* **29**, 581-608.

[4] S. Barberà and M.O. Jackson (2004), Choosing how to choose: Self-stable majority rules and constitutions, *Quarterly Journal of Economics* **119**, 1011-1048.

[5] S. Barberà and M.O. Jackson (2006), On the weights of nations: Assiging voting weights in a heterogeneous union, *The Journal of Political Economy* **114**, 317-339.

[6] T. Börgers and P. Postl (2009), Efficient compromising, *Journal of Economic Theory* **144**, 2057-2076.

[7] R.B. Curtis (1972), Decision rules and collective values in constitutional choice, In *Probability Models of Collective Decision Making*, edited by R.G. Niemi and H.F. Weisberg. Merrill, Columbus, Ohio.

[8] M. Fleurbaey (2008), One stake one vote, working paper.

[9] D. Gale (1960), The theory of linear economic models, McGraw-Hill Book Company.

[10] A. Gershkov, B. Moldovanu and X. Shi (2013), Optimal mechanism design without money, working paper.

[11] B. Holmström and R.B. Myerson (1983), Efficient and durable decision rules with incomplete information, *Econometrica* **51**, 1799-1819.

[12] M.O. Jackson (1991), Bayesian implementation, *Econometrica* **59**, 461-477.

[13] M.O. Jackson and H.F. Sonnenschein (2007), Overcoming incentive constraints by linking decisions, *Econometrica* **75**, 241-257.

[14] S. Kim (2012), Ordinal versus cardinal voting rules: A mechanism design approach, The Ohio State University working paper.

[15] D.M. Kreps and R. Wilson (1982), Sequential equilibria, *Econometrica* **50**, 863-894.

[16] D. Majumdar and A. Sen (2004), Ordinally Bayesian incentive compatible voting rules, *Econometrica* **72**, 523-540.

[17] A. Miralles (2012), Cardinal Bayesian allocation mechanisms without transfers, *Journal of Economic Theory* **147**, 179-206.

[18] K. Nehring (2004), The veil of public ignorance, *Journal of Economic Theory* **119**, 247-270.

[19] J. von-Neumann and O. Morgenstern (1944), Theory of games and economic behavior, Princeton University Press.

[20] T.R. Palfrey and S. Srivastava (1989), Mechanism design with incomplete information: A solution to the implementation problem, *The Journal of Political Economy* **97**, 668-691.

[21] D.W. Rae (1969), Decision-rules and individual values in constitutional choice, *The American Political Science Review* **63**, 40-56.

[22] R.T. Rockafellar (1970), Convex analysis, Princeton University Press, New Jersey.

[23] S.M. Samuels (1965), On the number of successes in independent trials, *Annals of Mathematical Statistics*, **36**, 1272-1278.

[24] P.W. Schmitz and T. Tröger (2006), Garbled elections, CEPR discussion paper No. 195.

[25] P.W. Schmitz and T. Tröger (2011), The (sub-)optimality of the majority rule, *Games and Economic Behavior*, **74**, 651-665.

[26] L.S. Shapley and M. Shubik (1954), A method for evaluating the distribution of power in a committee system, *American Political Science Review* **48**, 787-792.

[27] J.A. Weymark (2005), Measurment theory and the foundations of utilitarianism, *Social Choice and Wefare* **25**, 527-555.

References

[Azrieli and Kim, 2014] Azrieli, Y. and Kim, S. (2014). Pareto efficiency and weighted majority rules. *International Economic Review*, 55(4):1067–1088.

[Deb and Pai, 2014] Deb, R. and Pai, M. (2014). Symmetric auctions. *Working Paper, UPENN.*