ON THE SELF-(IN)STABILITY OF WEIGHTED MAJORITY RULES

YARON AZRIELI* AND SEMIN KIM**

ABSTRACT. A voting rule f is *self-stable* (Barberà and Jackson [4]) if any alternative rule g does not have sufficient support in the society to replace f, where the decision between f and g is based on the rule f itself. While Barberà and Jackson focused on anonymous rules in which all agents have the same voting power, we consider here the larger class of weighted majority rules. Our main result is a characterization of self-stability in this setup, which shows that only few rules of a very particular form satisfy this criterion. This result provides a possible explanation for the tendency of societies to use more conservative rules when it comes to changing the voting rule. We discuss self-stability in this latter case, where a different rule F may be used to decide between f and g.

Keywords: Voting rules, weighted majority rules, self-stability.

JEL Classification: D72.

June 22, 2015

*Dept. of Economics, The Ohio State University. Email: azrieli.2@osu.edu.
**Dept. of Economics, National Taiwan University. Email: seminkim@ntu.edu.tw.

1. Introduction

The voting rule used by a society to choose between different alternatives impacts which alternative is eventually chosen. Therefore, agents' preferences over alternatives naturally translate into preferences over voting rules. Given that members of a society have preferences over voting rules, and that members can propose to replace the current rule by a different one, which rules are likely to survive in the long-run? And how does the answer depend on the characteristics of the society?

Barberà and Jackson [4] (BJ henceforth) develop a theoretical framework to study these questions, a theory that is based on the endogenous preferences of agents over voting rules. The key concept, which characterizes voting rules that can withstand proposed changes to the rule, is called *self-stability*. Roughly speaking, a rule f is self-stable if, given any proposed alternative rule g, the coalition of agents who prefer g to f is not large enough to win the vote on replacing f by g, where the rule used for this latter decision is f itself. Thus, the idea underlying this concept is that the same voting rule (f) governs ordinary decisions as well as decisions about changes to the voting rule itself.[1]

Our contribution in this paper is to extend the analysis of BJ to a larger class of voting rules. Specifically, BJ study the case of anonymous rules in which a reform passes if and only if the number of its supporters exceeds the threshold specified by the rule. While this is the type of rules most often used in the real world, there are also many institutions that use more complicated rules in which different agents have different voting weights. Examples of such institutions include the United Nations Security Council, the Council of the European Union, and the International Monetary Fund, among many others. It is therefore important to understand what type of rules are self-stable when anonymity is not assumed.

In our formal model, a voting rule is any mapping from preference profiles over $\{Reform,\ Status\ quo\} = \{R, S\}$ to lotteries over this set. A weighted majority rule is a voting rule that can be described by assigning weights to the agents and setting a quota, such that R is chosen if the total weight of the agents that support R exceeds the quota, and S is chosen otherwise. Our main result (Theorem 1) is a characterization of the class of self-stable weighted majority rules. In other words, we characterize those weighted majority rules that cannot be defeated by any arbitrary voting rule.[2]

[1] BJ also consider the case of constitutions, where a different rule F can be used for the decision between f and g. We discuss this possibility in Section 5.

[2] Note that there may be weighted majority rules that cannot be defeated by any other weighted majority rule, but are defeated by a rule that is not a weighted majority rule.

Our characterization shows that only few rules of a very particular form are self-stable. Namely, each self-stable rule partitions the society into at most three groups, where the weights of agents within each group are the same. The first group contains 'veto' players, i.e. agents that can single-handedly vote down any reform. The second group contains 'null' players with zero weight whose vote never affect the outcome. The last group contains the rest of the society, and we refer to these agents as 'normal' players. Put differently, according to a self-stable rule a Reform passes if and only if the coalition of agents who support it contains all the veto players and at least a certain number of normal players. There are additional constraints on the numbers of veto players and normal players that need to be satisfied in order for the rule to be self-stable. These constraints vary with the characteristics of the society, but in every society self-stability implies that the rule has the form described above.

Our results imply that self-stability is an extremely restrictive requirement, and that voting rules used in the real world almost never satisfy this criterion.[3] It is therefore not surprising that societies usually use different voting rules for everyday decisions than for decisions involving changes to the rules. Our results strengthen the conclusions of BJ, who argue that [4, page 1011] "...constitutions where the voting rule used to amend the constitution is the same as the voting rule used for ordinary business are dangerously simplistic". We therefore extend our analysis to the case of constitutions in which a rule F (possibly different than f) is used to decide whether the ordinary business rule f will be replaced. While we do not have a complete characterization of self-stability in this case, we do obtain several necessary conditions which suggest that even in this setup self-stability is quite restrictive. In particular, we show that self-stability implies that F is more conservative than f, in the sense that if a coalition T is not sufficiently large to pass a Reform according to f then T is also not sufficiently large to replace f by another rule (according to F).

We now discuss how this paper is related to previous literature. The model we use, in which agents' preferences over voting rules are endogenously determined from their assessments regarding their future preferences over alternatives, was first suggested in early papers by Rae [16], Badger [2], and Curtis [7]. These papers only consider anonymous voting rules with the same weight to all agents.

[3] In the United Nations Security Council the five permanent member states (China, France, Russia, the UK, USA) are veto players and the other ten states are normal players. To pass a resolution, the support of all veto players and at least four normal players is needed (this assumes no abstentions). However, there is no society for which this rule satisfies the constraints on the number of normal players implied by self-stability. See Theorem 1 for details.

The theoretical investigation of weighted majority rules appears already in the seminal book of von-Neumann and Morgenstern [14, Section 5], who are mainly interested in measures of the voting power of agents under the rule. A common scenario leading to heterogeneous voting weights is that of a representative democracy with heterogenous district sizes. An early paper on this topic is Penrose [15]. More recent papers are Barberà and Jackson [5] and Fleurbaey [8], who point out the advantage of weighted majority rules from a utilitarian point of view. In our recent work [1] we show that, in a standard mechanism design setup, weighted majority rules naturally arise from considerations of efficiency and incentive compatibility.

Several papers extend the analysis of BJ's self-stability concept in various directions. The closest one to ours is Sosnowska [17] who considers a model in which the voting weights of agents are fixed but the quota is subject to amendments. She provides examples that demonstrate the differences between the case of heterogenous voting weights and the case of equal voting power as in BJ. Wakayama [18] considers self-stability under the possibility that agents can abstain from voting. Kultti and Miettinen [13] study self-stability in a model of constitutions that contain several layers of voting rules, where the voting rule in each layer is used to decide on changes to the voting rule of the previous one. The same authors consider in [12] a setup with a continuum of agents and analyze stability of voting rules using the 'stable set' concept from the theory of cooperative games. Coelho [6] uses the maximin criterion instead of self-stability for the evaluation of voting rules. Our paper is different than these previous works in that we keep the same self-stability concept as BJ and characterize it in a larger class of voting rules.

Finally, the idea that the same voting rule used to choose between alternatives is also used to choose between voting rules resembles the concept of self-selection for social choice functions introduced by Koray [10]. See also Barberà and Bevià [3] and Koray and Slinko [11].

The rest of paper is organized as follows. In the next section we describe the voting environment and give the definition of self-stability. In Section 3 we formulate our main results on the characterization of self-stable weighted majority rules. The proofs of the main results immediately follow in Section 4. Section 5 contains extensions of the analysis to societies in which agents have heterogenous utility functions, and to the case of constitutions that use a different voting rule for rule changes than for other decisions. We conclude in Section 6. Proofs that do not appear in the text are in the Appendix.

2. The voting environment and self-stability

A society faces a binary decision of whether to implement a Reform (R) or to keep the Status-quo (S). There are $n \geq 2$ agents in the society indexed by $i \in [n] := \{1, \ldots, n\}$. Each agent can either prefer R or S. The ex-ante probability that agent i prefers R is $p_i \in (0,1)$, and with the complement probability $1 - p_i$ he prefers S. We assume throughout the paper that agents' types are independent, and for every subset of agents (coalition) $T \subseteq [n]$ we denote $p(T) = \prod_{i \in T} p_i \prod_{i \notin T}(1 - p_i)$ the probability that the agents in T are exactly those who like R. All agents have the same utility function, parameterized by $r > 0$: If R is implemented, then an agent who prefers R gets a utility of r and an agent who prefers S gets a utility of -1; if the Status quo prevails then both types get a utility of zero.[4] A society can therefore be characterized by the pair $\left(\underline{p}, r\right)$, where $\underline{p} = (p_1, \ldots, p_n)$.

A *voting rule* is used to aggregate the preferences of the agents into a decision. Formally, a voting rule is any mapping $f : 2^{[n]} \to [0,1]$, with the interpretation that, for any coalition T, $f(T)$ is the probability that R is chosen when the members of T are those who prefer R. Given a voting rule f, the (ex-ante) expected utility of agent i is given by

$$U_i(f) = r \sum_{\{T \,:\, i \in T\}} p(T) f(T) - \sum_{\{T \,:\, i \notin T\}} p(T) f(T). \tag{1}$$

We will use the following standard terminology. A voting rule f is *deterministic* if $f(T) \in \{0,1\}$ for all T. Given a deterministic rule f, T is a *winning coalition* if $f(T) = 1$ and is a *minimal winning coalition* if it is winning and no strict sub-coalition of it is winning. Agent i is called a *veto player* if it is included in every winning coalition, and is called a *null player* if it is not included in any minimal winning coalition.

In this paper we focus our attention to weighted majority rules. These are relatively simple voting rules and we refer the reader to our previous work [1] for a discussion of the importance of these rules based on efficiency considerations. The formal definition is as follows.

Definition 1. The voting rule f is a *weighted majority rule* if there are non-negative weights $\underline{w} = (w_1, \ldots, w_n)$ summing up to 1 and a quota $0 < q < 1$ such that

$$f(T) = \begin{cases} 1, & \text{if } \sum_{i \in T} w_i > q \\ 0, & \text{if } \sum_{i \in T} w_i \leq q. \end{cases}$$

[4] In Section 5 we discuss how our results generalize to the case of heterogenous utility functions.

Note first that a weighted majority rule is deterministic. Second, in the above definition we impose a tie-breaking rule that favors the Status-quo. This is done for expositional reasons – our results hold unchanged if the tie-breaking rule favors the Reform. However, we cannot allow a tie-breaking rule that varies with the coalition who supports the Reform. Third, the weights and quota that define a weighted majority rule are typically not unique, that is the same rule f may be represented by different sets of weights and quotas. The concepts we study do not depend on the particular representation used. We write $f = (\underline{w}, q)$ if f can be represented by these weights and quota. To simplify notation, given a weighted majority rule $(\underline{w}, q)$ and a coalition T we write $w(T) := \sum_{i \in T} w_i$ for its total weight. Finally, if $f = (\underline{w}, q)$ then the expected utility of i under f in (1) can be rewritten as

$$U_i(f) = r \sum_{\{T \,:\, w(T)>q,\ i \in T\}} p(T) - \sum_{\{T \,:\, w(T)>q,\ i \notin T\}} p(T). \tag{2}$$

Let f be a weighted majority rule and let g be an arbitrary voting rule. Denote by $T(g, f) = \{i \in [n] \ : \ U_i(g) > U_i(f)\}$ the coalition of agents for which rule g yields a strictly higher expected utility than rule f. We can now define the concept of self-stability.

Definition 2. Given a society $\left(\underline{p}, r\right)$, a weighted majority rule f is *self-stable* if $f(T(g, f)) = 0$ for any voting rule g. Equivalently, if $f = (\underline{w}, q)$, then self-stability means that $w(\{i \ : \ U_i(g) > U_i(f)\}) \leq q$ for any rule g.

In words, self-stability of an incumbent rule f means that no alternative rule g (the reform) would have sufficient support to replace f if the voting rule used to determine the winner is f itself.

3. Characterization of self-stability

Our analysis of self-stability starts with the following observation.

Proposition 1. The set of self-stable weighted majority rules does not depend on the vector of probabilities $\underline{p}$. That is, for any $\underline{p}, \underline{p}', r$, a weighted majority rule f is self-stable in $\left(\underline{p}, r\right)$ if and only if it is self stable in $\left(\underline{p}', r\right)$.

Given Proposition 1, when we consider the self-stability of a certain voting rule we only need to specify the parameter r for the society and not the vector of probabilities $\underline{p}$. In particular, we can restrict attention to societies with ex-ante symmetric agents, which would be very helpful in the proof of our main theorem. The fact that self-stability is

not affected by the probabilities $\underline{p}$ is also interesting in its own right, as it demonstrates the significant difference between self-stability in our setup and in the setup of BJ [4], who are mainly interested in understanding the connection between self-stability and the probabilities $\underline{p}$ that define the society.

The following is the main result of the paper.

Theorem 1. Let f be a weighted majority rule.

(i) For $r \notin \{1, 1/2, 1/3, \ldots\}$, f is self-stable if and only if it has one of the following three forms:

(A) There is a non-empty coalition V with $|V| \leq \frac{1}{r} + 1$ such that

$$f(T) = \begin{cases} 1, & \text{if } V \subseteq T \\ 0, & \text{otherwise.} \end{cases}$$

(B) There are non-empty and disjoint coalitions V, M with $|V| \leq \frac{1}{r}$ and $|M| \geq 2$ such that,

$$f(T) = \begin{cases} 1, & \text{if } V \subseteq T \text{ and } |T \cap M| \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

(C) There are non-empty and disjoint coalitions V, M with $|V \cup M| \leq \frac{1}{r} + 2$ and $|M| \geq 2$ such that,

$$f(T) = \begin{cases} 1, & \text{if } V \subseteq T \text{ and } |T \cap M| \geq |M| - 1 \\ 0, & \text{otherwise.} \end{cases}$$

(ii) For $r \in \{1, 1/2, 1/3, \ldots\}$, f is self-stable if and only if it is either in one of the forms (A), (B) or (C) above, or if there is a coalition M of size $|M| = \frac{1}{r} + 2$ such that

$$f(T) = \begin{cases} 1, & \text{if } |T \cap M| \geq |M| - 1 \\ 0, & \text{otherwise.} \end{cases}$$

In words, self-stability implies that the voting rule must have a very particular form: The first possibility (form (A)) is that there is a non-empty group V of at most $\frac{1}{r} + 1$ agents, such that the Reform passes if and only if all the members of V support it. In other words, V is the unique minimal winning coalition, all agents in V are veto players, and all other agents are null. The second option (form (B)) is that the coalition V cannot pass the reform by itself, but requires one additional agent from the group M. Thus, the agents in V are veto players, the agents in M are neither veto nor null, and the rest of the agents (if there are any) are null. The third option is similar to the second, but requires that in addition to V all the agents in M except at most one support the Reform

in order for it to pass. Note that forms (B) and (C) coincide when $|M| = 2$. Finally, if $r = \frac{1}{k}$ for some integer k, then another type of rules becomes self-stable. These are the rules in which there is a group of $m := k + 2$ agents such that a coalition is winning if and only if it contains at least $m - 1$ of them. In particular, only for such values of r there exists a self-stable rule without any veto players.

Remark. From the description of the voting rules in Theorem 1 it is not immediately clear that these are indeed weighted majority rules. To represent f of form (A) as a weighted majority rule, set $w_i = \frac{1}{|V|}$ for each $i \in V$, $w_i = 0$ for each $i \notin V$, and choose $q > 1 - \frac{1}{|V|}$. For form (B), choose $\epsilon > 0$ sufficiently small so that $1 - \epsilon > \epsilon|V|$, and define $w_i = \frac{1-\epsilon}{|V|}$ for $i \in V$, $w_i = \frac{\epsilon}{|M|}$ for $i \in M$, $w_i = 0$ for $i \notin M \cup V$, and $q = 1 - \epsilon$. For form (C) use the same weights as in the previous sentence, but increase the quota to $q = 1 - \frac{2\epsilon}{|M|}$. The voting rule in part (ii) of the theorem can be represented by letting $w_i = \frac{1}{|M|}$ for each $i \in M$ (and zero weights to all other agents), and setting $q = 1 - \frac{2}{|M|}$.

When r is relatively large self-stability becomes particularly restrictive. In the following definition and two corollaries we describe the class of self-stable rules in 'reform-biased' societies ($r > 1$) and in 'neutral' societies ($r = 1$). Since these corollaries immediately follow from Theorem 1 we omit the proofs.

Definition 3. A weighted majority rule f is
1. *dictatorial* if there is $i \in [n]$ such that $f(T) = 1$ if and only if $i \in T$.
2. *quasi-dictatorial* if there is at least one veto player and all minimal winning coalitions are of size two.
3. *triumvirate* if there are three agents $i, j, k \in [n]$ such that $f(T) = 1$ whenever $|T \cap \{i, j, k\}| \geq 2$ and $f(T) = 0$ otherwise.

Corollary 1. In 'reform-biased' societies ($r > 1$), a weighted majority rule is self-stable if and only if it is dictatorial.

Corollary 2. For 'neutral' societies ($r = 1$), a weighted majority rule is self-stable if and only if it is dictatorial or quasi-dictatorial or triumvirate.

4. Proofs

This section contains the proofs for the results of Section 3. We start with the proof of Proposition 1. We then prove a sequence of lemmas which constitute key steps in the proof of Theorem 1. The proof of Theorem 1 itself ends the section.

4.1. **Proof of Proposition 1.** Fix a society $\left(\underline{p}, r\right)$, an agent i, and two voting rules f, g. Assume that in this society agent i prefers g over f, that is $U_i(g) > U_i(f)$. Consider now an alternative society $\left(\underline{p}', r\right)$. For each coalition T denote $p'(T) = \prod_{i\in T} p'_i \prod_{i\notin T}(1-p'_i)$ and $\alpha_T = \frac{p(T)}{p'(T)}$. Let $\alpha = \max_T \alpha_T$. Define the voting rule g' by $g'(T) = \frac{\alpha_T}{\alpha} g(T) + \left(1 - \frac{\alpha_T}{\alpha}\right) f(T)$. We claim that agent i prefers the rule g' over f in the new society $\left(\underline{p}', r\right)$. Indeed,

$$
\begin{aligned}
0 < U_i(g) - U_i(f) \;=\; & r \sum_{\{T\,:\, i\in T\}} p'(T)\alpha_T[g(T) - f(T)] - \sum_{\{T\,:\, i\notin T\}} p'(T)\alpha_T[g(T) - f(T)] = \\
& r \sum_{\{T\,:\, i\in T\}} p'(T)\alpha \left[\frac{\alpha_T}{\alpha} g(T) + \left(1 - \frac{\alpha_T}{\alpha}\right) f(T) - f(T)\right] - \\
& \sum_{\{T\,:\, i\notin T\}} p'(T)\alpha \left[\frac{\alpha_T}{\alpha} g(T) + \left(1 - \frac{\alpha_T}{\alpha}\right) f(T) - f(T)\right] = \\
& \alpha \left(r \sum_{\{T\,:\, i\in T\}} p'(T)\left[g'(T) - f(T)\right] - \sum_{\{T\,:\, i\notin T\}} p'(T)\left[g'(T) - f(T)\right] \right) = \\
& \alpha \left[U'_i(g') - U'_i(f)\right],
\end{aligned}
$$

so $U'_i(g') > U'_i(f)$. Note that the rule g' does not depend on the particular agent i considered. Thus, if there is a rule g for which $T(g, f)$ is a winning coalition under f in the original society, then there is also a rule g' for which $T(g', f)$ is a winning coalition under f in the new society. Obviously this also implies that if f is self-stable in the original society then it is self-stable in the alternative society. □

4.2. **Preliminary lemmas.** For all of the lemmas in this section we assume that agents are ex-ante symmetric, specifically that $p_i = \frac{1}{2}$ for all $i \in [n]$.

Lemma 1. If $f = (\underline{w}, q)$ and $w_i \geq w_j$ then $U_i(f) \geq U_j(f)$. Conversely, if f is a weighted majority rule and $U_i(f) = U_j(f)$ then f can be represented by $(\underline{w}, q)$ satisfying $w_i = w_j$.

Proof. Let i, j be two agents and fix some weighted majority rule f. By (2) we have

$$\begin{aligned}
U_i(f) - U_j(f) &= r\left(\sum_{\{T\,:\,w(T)>q,\ i\in T\}} p(T) - \sum_{\{T\,:\,w(T)>q,\ j\in T\}} p(T)\right) + \\
&\left(\sum_{\{T\,:\,w(T)>q,\ j\notin T\}} p(T) - \sum_{\{T\,:\,w(T)>q,\ i\notin T\}} p(T)\right) = \\
&r\left(\sum_{\{T\,:\,w(T)>q,\ i\in T,\ j\notin T\}} p(T) - \sum_{\{T\,:\,w(T)>q,\ j\in T,\ i\notin T\}} p(T)\right) + \\
&\left(\sum_{\{T\,:\,w(T)>q,\ j\notin T,\ i\in T\}} p(T) - \sum_{\{T\,:\,w(T)>q,\ i\notin T,\ j\in T\}} p(T)\right) = \\
&(r+1)\left(\sum_{\{T\,:\,w(T)>q,\ i\in T,\ j\notin T\}} p(T) - \sum_{\{T\,:\,w(T)>q,\ j\in T,\ i\notin T\}} p(T)\right) = \\
&(r+1)\left(\sum_{\{T\subseteq[n]\setminus\{i,j\}\,:\,w(T)+w_i>q\}} p(T\cup\{i\}) - \sum_{\{T\subseteq[n]\setminus\{i,j\}\,:\,w(T)+w_j>q\}} p(T\cup\{j\})\right).
\end{aligned}$$

Thus, if $w_i \geq w_j$ then any coalition T that appears in the second sum also appears in the first, and $p(T\cup\{i\}) = p(T\cup\{j\})$ by symmetry, so we get $U_i(f) \geq U_j(f)$.

In the other direction, assume $U_i(f) = U_j(f)$ and that $f = (\underline{w}, q)$. We claim that replacing both w_i and w_j by $\frac{w_i+w_j}{2}$ does not change f. This is clearly the case for coalitions T that contain both i and j and for coalitions T which contain neither. Assume (w.l.o.g.) that $w_i \geq w_j$. Then as before any coalition T in the second sum appears also in the first. It follows then from $U_i(f) = U_j(f)$ that these two collections of coalitions are equal. Thus, $w(T) + w_i > q \iff w(T) + w_j > q$. Since $w_j \leq \frac{w_i+w_j}{2} \leq w_i$ the set of coalitions with weight greater than the quota has not changed by averaging the weights. □

Lemma 2. If f is a self-stable weighted majority rule then all minimal winning coalitions are of the same size.

Proof. Assume not. Then there are T_1, T_2 minimal winning coalitions with $|T_1| > |T_2|$. We claim that there must be an agent $\bar{i} \in T_2 \setminus T_1$ and an agent $\bar{j} \in T_1 \setminus T_2$ such that $U_{\bar{i}}(f) > U_{\bar{j}}(f)$. Otherwise, by Lemma 1 above, f can be represented with weights that

satisfy $w_i \leq w_j$ for every $i \in T_2 \setminus T_1$ and $j \in T_1 \setminus T_2$. But then we get that

$$\begin{aligned} w(T_1) \;\geq\; & w(T_1 \cap T_2) + |T_1 \setminus T_2| \left(\min_{j \in T_1 \setminus T_2} w_j \right) \geq w(T_1 \cap T_2) + (|T_2 \setminus T_1| + 1) \left(\min_{j \in T_1 \setminus T_2} w_j \right) \geq \\ & w(T_1 \cap T_2) + |T_2 \setminus T_1| \left(\max_{i \in T_2 \setminus T_1} w_i \right) + \left(\min_{j \in T_1 \setminus T_2} w_j \right) \geq w(T_2) + \left(\min_{j \in T_1 \setminus T_2} w_j \right) > \\ & q + \left(\min_{j \in T_1 \setminus T_2} w_j \right). \end{aligned}$$

Thus, $w(T_1) - \left(\min_{j \in T_1 \setminus T_2} w_j\right) > q$ which contradicts the minimality of T_1.

Now, let h be the rule that is identical to f except that $h(T_1 \setminus \{\bar{j}\}) = 1$. Then $U_j(h) > U_j(f)$ for every $j \in T_1 \setminus \{\bar{j}\}$. Let h' be the rule that switches the weights of $\bar{i}$ and $\bar{j}$ and keeps everything else unchanged. By the symmetry between the agents, the utilities of all agents except $\bar{i}$ and $\bar{j}$ are unaffected by this switch, and the utilities of agents $\bar{i}$ and $\bar{j}$ are switched. Therefore, $U_{\bar{j}}(h') > U_{\bar{j}}(f)$ and $U_j(h') = U_j(f)$ for every other agent $j \in T_1$. Since utilities are linear in the voting rule, it follows that for sufficiently small $\epsilon > 0$ the rule $g = \epsilon h + (1-\epsilon)h'$ improves the utilities of all agents in T_1. □

Lemma 3. If f is a self-stable weighted majority rule and i, j are two agents which are neither veto players nor null players then $U_i(f) = U_j(f)$.

Proof. Denote by M the set of 'normal' players, i.e. players which are neither veto nor null, and we may assume that $M \neq \emptyset$ (otherwise there is nothing to prove). Let $\bar{x} = \max\{U_i(f) : \; i \in M\}$ and let $\bar{i} \in M$ be an agent that achieves this maximum, i.e. $U_{\bar{i}}(f) = \bar{x}$. We need to show that $U_i(f) = \bar{x}$ for every $i \in M$.

Assume that this is not the case, so there is a normal player with utility less than $\bar{x}$. We first claim that there must be a minimal winning coalition T and a normal player $\underline{i}$ with $U_{\underline{i}}(f) < \bar{x}$, such that $\underline{i} \in T$ and $\bar{i} \notin T$. Indeed, consider a representation of f where agents with the same expected utilities have the same weight (we know this exists by Lemma 1). Since $\bar{i}$ is not a veto player, there is some minimal winning coalition T' that does not contain $\bar{i}$. If there is a normal player in T' with utility less than $\bar{x}$ then we are done. Otherwise, all normal players in T' have weight $w_{\bar{i}}$, which is the highest weight of a normal player. Consider the coalition T which is identical to T' except that it replaces one of the normal players in T by a normal player $\underline{i}$ of lower weight (by Lemma 2 all minimal winning coalitions have the same number of normal players, so in particular T' contains at least one such player). We claim that T must be a winning coalition: By Lemma 2, all minimal winning coalitions are of size $|T'| = |T|$. But from all the coalitions of this size that contain $\underline{i}$, T has the largest possible weight (except $\underline{i}$, it may contain

only veto players and players from M with the highest weight). Since we know that $\underline{i}$ is included in some minimal winning coalition, it therefore must be that T is winning.

We now show that f is not self-stable. Let T and $\underline{i}$ be as in the previous paragraph. Let h be identical to f except that $h(T \setminus \{\underline{i}\}) = 1$. Let h' be the rule that switches the weights of $\underline{i}$ and $\bar{i}$ (and keeping everything else unchanged). Then as in the proof of the previous lemma, for $\epsilon > 0$ sufficiently small, $g = \epsilon h + (1-\epsilon)h'$ increases the utilities of all agents in T relative to f. □

Lemma 4. If f is a self-stable weighted majority rule and T is a minimal winning coalition then $|T| \leq 1 + \frac{1}{r}$.

Proof. Assume by contradiction that f has a minimal winning coalition $\bar{T}$ with $|\bar{T}| > 1 + \frac{1}{r}$. Consider the rule g which is identical to f, except that $g(S) = 1$ for every $S \subseteq \bar{T}$ with $|S| = |\bar{T}| - 1$. Then for any $i \in \bar{T}$,

$$\begin{aligned} U_i(g) - U_i(f) &= r \sum_{\{T \,:\, i \in T\}} p(T)[g(T) - f(T)] - \sum_{\{T \,:\, i \notin T\}} p(T)[g(T) - f(T)] = \\ & r \sum_{\{S \,:\, i \in S \subseteq \bar{T},\ |S| = |\bar{T}|-1\}} p(S) - \sum_{\{S \,:\, i \notin S \subseteq \bar{T},\ |S| = |\bar{T}|-1\}} p(S) = \\ & r \left(\frac{1}{2}\right)^n (|\bar{T}| - 1) - \left(\frac{1}{2}\right)^n = \left(\frac{1}{2}\right)^n \left[r(|\bar{T}| - 1) - 1\right] > 0. \end{aligned}$$

Thus, every agent in $\bar{T}$ prefers g over f which contradicts the self-stability of f. □

Lemma 5. Assume that f is a self-stable weighted majority rule, and let $\bar{i}$ be an agent with the highest weight in some representation of f. If T is a coalition such that $\bar{i} \notin T$ and $|T| < \frac{1}{r} + 1$ then $f(T) = 0$.

Proof. Assume by contradiction that $\bar{i} \notin T$ and $|T| < \frac{1}{r} + 1$ but $f(T) = 1$. Then, since $\bar{i}$ has (weakly) higher weight than any other agent, it follows that for every $i \in T$ the coalition $S_i := (T \cup \{\bar{i}\}) \setminus \{i\}$ must be winning as well. Consider the rule g that is identical to f except that $g(S_i) = 0$ for each of those coalitions S_i. Then for every $j \in T$,

$$\begin{aligned} U_j(g) - U_j(f) &= -r \sum_{\{S_i \,:\, j \in S_i\}} p(S_i) + \sum_{\{S_i \,:\, j \notin S_i\}} p(S_i) = \\ & -r \left(\frac{1}{2}\right)^n (|T| - 1) + \left(\frac{1}{2}\right)^n = \left(\frac{1}{2}\right)^n [1 - r(|T| - 1)] > 0, \end{aligned}$$

so all agents in T prefer g over f. □

Lemma 6. If $r \notin \{1, 1/2, 1/3, \ldots\}$ and f is a self-stable weighted majority rule then f has a veto player.

Proof. Let $\bar{i}$ be an agent with the highest weight in some representation of f, and let T be an arbitrary minimal winning coalition. We claim that $\bar{i} \in T$. Indeed, if $\bar{i} \notin T$ then by Lemma 5 we have $|T| \geq \frac{1}{r} + 1$. By Lemma 4 we have $|T| \leq \frac{1}{r} + 1$. Thus, it must be that $|T| = \frac{1}{r} + 1$, but since $r \notin \{1, 1/2, 1/3, \ldots\}$ this is not an integer. It follows that $\bar{i}$ belongs to every minimal winning coalition, so he is a veto player. □

Lemma 7. Assume that f is a self-stable weighted majority and let M be the set of agents that are neither veto nor null players. If T is a minimal winning coalition then $|M \cap T| \in \{0, 1, |M| - 1\}$.

Proof. First, note that it follows from Lemma 2 that $|M \cap T|$ is the same for all minimal winning coalitions T, and that it follows from Lemmas 1 and 3 that all agents in M have the same weight. Now, fix some minimal coalition T. It can't be that $|M \cap T| = |M|$ since this would imply that T is the unique minimal winning coalition and therefore that the agents in M are veto players. We now show that it can't be the case that $|M \cap T| \in \{2, 3, \ldots, |M| - 2\}$.

Assume by contradiction that $|M \cap T| \in \{2, 3, \ldots, |M| - 2\}$, and rename the agents in $M \cap T$ if needed so that $M \cap T = \{1, 2, \ldots, |M \cap T|\}$. By assumption there are (at least) two agents $j, k \in M \setminus T$. Let the rule g be identical to f except that (1) $g(T \setminus \{i\}) = 1$ for every $i = 1, \ldots, |M \cap T|$; and (2) $g(\{j, k\} \cup (T \setminus \{i, i + 1\})) = 0$ for every $i = 1, \ldots, |M \cap T| - 1$.

We claim that g improves the utilities of all agents in T. Consider first an agent $l \in T \setminus M$ (if there is such agent): Whenever l prefers S the rules f and g are identical. When l prefers R he gains $r\left(\frac{1}{2}\right)^n |M \cap T|$ due to the change (1) and loses $r\left(\frac{1}{2}\right)^n (|M \cap T| - 1)$ due to the change (2). Thus,

$$U_l(g) - U_l(f) = r\left(\frac{1}{2}\right)^n [|M \cap T| - (|M \cap T| - 1)] = r\left(\frac{1}{2}\right)^n > 0.$$

Consider now an agent $l \in M \cap T$: Due to change (1) agent l gains $r\left(\frac{1}{2}\right)^n (|M \cap T| - 1)$, and loses $\left(\frac{1}{2}\right)^n$. Due to change (2) l gains either $\left(\frac{1}{2}\right)^n$ (if $l = 1,\ |M \cap T|$) or $2\left(\frac{1}{2}\right)^n$ (if $l = 2, \ldots, |M \cap T| - 1$), and loses either $r\left(\frac{1}{2}\right)^n (|M \cap T| - 2)$ (if $l = 1,\ |M \cap T|$) or $r\left(\frac{1}{2}\right)^n (|M \cap T| - 3)$ (if $l = 2, \ldots, |M \cap T| - 1$). In total,

$$U_l(g) - U_l(f) \geq r\left(\frac{1}{2}\right)^n (|M \cap T| - 1) - \left(\frac{1}{2}\right)^n + \left(\frac{1}{2}\right)^n - r\left(\frac{1}{2}\right)^n (|M \cap T| - 2) = r\left(\frac{1}{2}\right)^n > 0.$$

This contradicts the self-stability of f. □

4.3. **Proof of 'Only If' direction in Theorem 1.** Consider first part **(i)** of the theorem where $r \notin \{1, 1/2, 1/3, \ldots\}$. Assume that f is a self-stable weighted majority rule. By Proposition 1 we may assume that $p_i = \frac{1}{2}$ for all i. Denote by V the set of veto players of f and by M the set of 'normal' players, i.e. players which are neither veto nor null. From Lemma 6 we know that $V \neq \emptyset$. From Lemma 3 it follows that there is a number b such that T is a winning coalition if and only if $V \subseteq T$ and $|M \cap T| \geq b$. From Lemma 7 we know that $b \in \{0, 1, |M| - 1\}$.

If $b = 0$ then $M = \emptyset$ and V is the unique minimal winning coalition. By Lemma 4 $|V| \leq \frac{1}{r} + 1$. Thus, f is of the form (A). If $b = 1$ then we must have $|M| \geq 2$, since otherwise there is only one minimal winning coalition. By Lemma 4 $|V| + 1 \leq \frac{1}{r} + 1$, or $|V| \leq \frac{1}{r}$. Thus, f is of the form (B). Finally, if $b = |M| - 1$ then again it must be $|M| \geq 2$ and by Lemma 4 $|V \cup M| - 1 \leq \frac{1}{r} + 1$, or $|V \cup M| \leq \frac{1}{r} + 2$. This means that f is of the form (C). This concludes the proof of part **(i)**.

Consider now part **(ii)**. The only place where we used the assumption that $r \notin \{1, 1/2, 1/3, \ldots\}$ was to conclude that there must be a veto player. We now show that if f is self-stable and has no veto player then it must be in the from described in **(ii)**.

As in part **(i)** above, it follows from Lemma 3 that there is a number b such that T is a winning coalition if and only if $|M \cap T| \geq b$, and from Lemma 7 that $b \in \{0, 1, |M| - 1\}$. We claim that $b = \frac{1}{r} + 1$. Indeed, by Lemma 4 $b \leq \frac{1}{r} + 1$ and from Lemma 5 it can't be that $b < \frac{1}{r} + 1$ since there is a normal player (a player with the highest weight) that is not included in some minimal winning coalition. Therefore, $b = 0$ and $b = 1$ are impossible, and we are left with $b = |M| - 1$. We get $|M| - 1 = \frac{1}{r} + 1$, or $|M| = \frac{1}{r} + 2$ as needed. □

4.4. **Proof of 'If' direction in Theorem 1.** To prove this direction we will show that if f has one of the forms in the theorem then for every minimal winning coalition T there are non-negative numbers $\{\lambda_i\}_{i \in T}$ (not all zero) such that f is a maximizer of $\sum_{i \in T} \lambda_i U_i(g)$ among all voting rules g. This would imply that no rule g can strictly improve the utilities of all members of some minimal winning coalition, and hence that f is self-stable.

Consider first f of the form (A). Then V is the unique minimal winning coalition. Let $\lambda_i = 1$ for every $i \in V$, that is we are looking for a maximizer of the sum of utilities of the veto players. But note that f of form (A) chooses the alternative that maximizes the sum of utilities of the agents in V at every type profile. Indeed, if all agents in V prefer R then R is chosen, and if at least one of them prefers S then, since $|V| \leq \frac{1}{r} + 1$, S is the alternative that maximizes the sum of utilities and it is chosen by f. Thus, f is self-stable.

For form (B), every minimal winning coalition contains the players in V and one additional player from M. Given such a coalition, define $\lambda_i = 1$ for each $i \in V$ and $\lambda_i = r|V|$ for the single player from M. Then again f maximizes this weighted sum of utilities at every type profile: If all agents in this coalition prefer R then R is chosen. If all agents in V prefer R and the agent from M prefers S then both alternatives yield the same weighted sum of utilities, since choosing R gives a utility of r for each agent in V and a utility of -1 for the agent from M, so the weighted total is $|V|r - |V|r = 0$. Any probability of choosing R is therefore optimal in such type profiles. Finally, if at least one of the veto players prefers S then the weighted sum of utilities from choosing R is at most $(|V| - 1)r + |V|r^2 - 1 \leq \left(\frac{1}{r} - 1\right) r + r - 1 = 0$, where the inequality is due to $|V| \leq \frac{1}{r}$. For any such type profile f chooses S.

The proof for form (C) is similar, where the weight of each player in V is $\lambda_i = 1$ and the weight of each player from M is $\lambda_i = \frac{|V|r}{1-r(|M|-2)}$. It is straightforward to check that f maximizes the weighted sum of utilities at each type profile.

For part **(ii)**, the minimal winning coalitions are those of size $|M| - 1$ of players from M. Choose one such coalition and assign a weight of $\lambda_i = 1$ for each agent i. Then if exactly $|M| - 2$ out of the agents support R then the sum of utilities from choosing R is $(|M| - 2)r - 1 = 1 - 1 = 0$, where the first equality follows from $|M| = \frac{1}{r} + 2$. The rule chooses S in this case, which is optimal. For any smaller number of supporters of R the rule (optimally) chooses S, and if all agents support R then R is chosen. It follows that f is self-stable in this case as well. □

5. Extensions

5.1. **Self-stable constitutions.** In reality, it is often the case that the voting rule used for everyday decisions, say f, is different than the rule used to make procedural amendments such as replacing f by another rule. Following BJ, we call a pair of weighted majority rules (f, F) a *constitution*. The interpretation is that F is the rule used to determine whether f will be replaced by another rule, and f is used for all other decisions. For the following definition recall that $T(g, f) = \{i \in [n] \ : \ U_i(g) > U_i(f)\}$.

Definition 4. A constitution (f, F) is self-stable if $F(T(g, f)) = 0$ for every voting rule g.

Thus, a constitution (f, F) is self-stable if no alternative rule g can get sufficient support to replace f, where the decision between f and g is based on F. Note that this definition generalizes the definition of a self-stable voting rule (Definition 2) in the sense that f is self-stable if and only if the constitution (f, f) is self-stable.

While we do not have a complete characterization of self-stable constitutions, we collect in this section several results on the properties of such constitutions. Notice first that, as in the case of self-stable weighted majority rules discussed so far, self-stability of a constitution does not depend on the vector of probabilities $\underline{p}$ that characterizes the society. The proof of this is identical to the proof of Proposition 1 and is therefore omitted.

Second, observe that a necessary condition for self-stability is that F is at least as conservative as f, a property that to our knowledge is satisfied by almost every real-world constitution.[5]

Proposition 2. If (f, F) is self-stable and $f(T) = 0$ then $F(T) = 0$.

Proof. Assume that there is T such that $f(T) = 0$ but $F(T) = 1$. Let g be the rule that is identical to f except that $g(T) = 1$. Then clearly g improves the utilities of all members of T relative to f. Since $F(T) = 1$ we get that (f, F) is not self-stable. □

The following two propositions provide additional necessary conditions for self-stability of constitutions. Proofs can be found in Appendix A.

Proposition 3. If (f, F) is self-stable and f has a representation in which all agents have the same weight, then $F(T) = 0$ for every T satisfying $|T| \leq n - 2$.

Proposition 4. If (f, F) is self-stable and there is $\bar{i} \in [n]$ such that $w_{\bar{i}} > w_i$ for every other agent i in any representation of f, then $\bar{i}$ is a veto player of F.[6]

According to Proposition 3, if all agents have the same weight under f, as is typically the case in reality, then for self-stability to hold it must be the case that nearly unanimous agreement is required to change f: At least $n - 1$ out of the n agents must support the change in order for it to pass. In the opposite case where f favors one agent over all others, Proposition 4 implies that for self-stability we need that this favored agent can veto any proposed change to f; for an alternative rule g to replace f it must further increase the utility of this agent.

It is not hard to formulate additional simple necessary conditions for self-stability of a constitution. For example, if f is not weakly Pareto efficient,[7] then for any F the constitution (f, F) is not self-stable. It is also easy to construct examples of self-stable

[5] Barberà and Jackson [4, page 1012] give an example from California where f is more conservative than F. We are not aware of other such examples.

[6] The proof of this proposition shows that we can slightly strengthen the result, namely, we can allow for one additional agent to have the same weight as agent $\bar{i}$.

[7] That is, there is a rule g such that $U_i(g) > U_i(f)$ for all $i \in [n]$.

constitutions, for instance by taking an arbitrary self-stable rule f and considering any constitution (f, F) such that F is more conservative than f. However, we believe that it would be difficult to formally describe the set of self-stable constitutions, and we do not pursue this direction further.

5.2. **Asymmetric utilities.** In this section we show how our main result generalizes to the case where agents have heterogenous utility functions. Specifically, we now assume that if agent i prefers the Reform and the Reform is chosen then his utility is $r_i > 0$. Thus, agents' biases towards the alternatives may be different. All other ingredients of the model are the same as before. The utility of agent i under voting rule f is the same as in (1) except that r is replaced by r_i.

Looking at the proof of Proposition 1, it is easy to see that the same result holds in this more general setup. To generalize Theorem 1 we need to introduce the following notation. For every coalition T we let $r(T)$ be the unique number satisfying $\frac{1}{1+r(T)} = \frac{1}{|T|}\sum_{i\in T}\frac{1}{1+r_i}$. Note that $r(T) > 0$, that $r(\{i\}) = r_i$, and that if $r_i = r$ for all $i \in T$ then $r(T) = r$.

Theorem 2. Let f be a weighted majority rule.

(i) Assume that there is no coalition M for which $|M| = \frac{1}{r_i} + 2$ holds for all $i \in M$. Then f is self-stable if and only if it has one of the following three forms:
(A) There is a non-empty coalition V satisfying $|V| \leq \frac{1}{r(V)} + 1$ such that

$$f(T) = \begin{cases} 1, & \text{if } V \subseteq T \\ 0, & \text{otherwise.} \end{cases}$$

(B) There are non-empty and disjoint coalitions V, M with $|M| \geq 2$ and $|V| \leq \frac{1}{r(V\cup\{i\})}$ for each $i \in M$ such that,

$$f(T) = \begin{cases} 1, & \text{if } V \subseteq T \text{ and } |T \cap M| \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

(C) There are non-empty and disjoint coalitions V, M with $|M| \geq 2$ and $|V \cup M| \leq \frac{1}{r(V\cup(M\setminus\{i\}))} + 2$ for each $i \in M$ such that,

$$f(T) = \begin{cases} 1, & \text{if } V \subseteq T \text{ and } |T \cap M| \geq |M| - 1 \\ 0, & \text{otherwise.} \end{cases}$$

(ii) If there is a coalition M for which $|M| = \frac{1}{r_i} + 2$ holds for all $i \in M$, then f is self-stable if and only if it is either in one of the forms (A), (B) or (C) above, or if for

some such M,

$$f(T) = \begin{cases} 1, & \text{if } |T \cap M| \geq |M| - 1 \\ 0, & \text{otherwise.} \end{cases}$$

A sketch of the the proof of Theorem 2 can be found in Appendix B. It is similar to the proof for the homogenous case (Theorem 1), with some modifications needed to account for the heterogeneity. It is not hard to see that this theorem boils down to Theorem 1 when all agents have the same utility function.

6. Final comments

We have characterized the set of self-stable weighted majority rules in a model where agents can propose arbitrary voting rules as an alternative to the incumbent rule. Self-stability is extremely restrictive in this model, and in particular it typically requires the existence of veto players. It would be interesting to consider the case where only weighted majority rules can be proposed as alternatives to the existing rule. Our proofs heavily rely on randomized and other non-weighted majority rules to obtain necessary conditions for self-stability, so it is likely that such modification will significantly increase the set of self-stable rules. From a theoretical perspective it is also interesting to ask which other voting rules, beside weighted majority rules, are self-stable in the model we have considered.

Another direction which we find intriguing is based on the idea that agents sometimes already know their own preferences when a change to the voting rule is proposed, but they can only conjecture the preferences of other agents. Defining stability of voting rules in such a setup with private information is a non-trivial task, since when an agent considers whether he should vote for rule f or rule g he should take into account that other agents' votes depend on their private information, which in turn is going to affect their voting behavior in the subsequent vote between the alternatives. In our previous work [1, Section 7] we briefly discussed this issue, building on the concept of *durability* due to Holmström and Myerson [9]. We leave this issue for future research.

References

[1] Y. Azrieli and S. Kim (2014), Pareto efficiency and weighted majority rules, *International Economic Review* **55**, 1065-1086.

[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] S. Barberà and C. Bevià (2002), Self-selection consistent functions, *Journal of Economic Theory* **105**, 263-277.

[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] D. Coelho (2005), Maximin choice of voting rules for committees, *Economics of Governance* **6**, 159-175.

[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), Weighted majority and democratic theory, working paper.

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

[10] S. Koray (2000), Self-selective social choice functions verify Arrow and Gibbard-Satterthwaite theorems, *Econometrica* **68**, 981-995.

[11] S. Koray and A. Slinko (2008), Self-selective social choice functions, *Social Choice and Welfare* **31**, 129-149.

[12] K. Kultti and P. Miettinen (2007), Stable set and voting rules, *Mathematical Social Sciences* **53**, 164-171.

[13] K. Kultti and P. Miettinen (2009), Stability of constitutions, *Journal of Public Economic Theory* **11**, 891-896.

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

[15] L.S. Penrose (1946), The elementary statistics of majority voting, *Journal of the Royal Statistical Society* **109**, 53-57.

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

[17] H. Sosnowska (2002), Self-stable majority rules for weighted voting games, working paper.

[18] T. Wakayama (2002), Endogenous choice of voting rules with abstention, working paper.

Appendix A. Proof of Propositions 3 and 4

Since the set of self-stable constitutions does not depend on $\underline{p}$, we may assume that $p_i = p_j$ for all $i, j \in [n]$. The two propositions are now immediate consequences of the following lemma.

Lemma 8. Assume $p_i = p_j$ for all $i, j \in [n]$ and that $n \geq 3$. Let f be a weighted majority rule, and order the agents so that $U_1(f) \geq U_2(f) \geq \ldots \geq U_n(f)$. Then there is a voting rule g such that $U_i(g) > U_i(f)$ for every $i = 3, 4, \ldots, n$. Furthermore, if $U_1(f) > U_2(f)$

or $U_1(f) = U_2(f) > U_3(f)$ then there is a voting rule g such that $U_i(g) > U_i(f)$ for every $i = 2, 3, \ldots, n$.

Proof. **Case (1)**: $U_1(f) > U_2(f)$. For each agent $i \geq 2$ consider the rule f_i defined by switching the weights of agent 1 with the weight of agent i, and keeping everything else unchanged. Then the utilities of agent i and agent 1 are switched under f_i, while the utility of any other agent remains unaffected. Therefore, for the voting rule $g = \frac{1}{n-1}\sum_{i=2}^{n} f_i$ we have that $U_i(g) > U_i(f)$ for every $i \neq 1$.

Case (2): $U_1(f) = U_2(f) > U_3(f)$. For each $i \geq 3$ we can use the rule f_i as above. Let h be the dictatorial rule with agent 2 being the dictator. Since $U_1(f) = U_2(f)$ it can't be that f is dictatorial, so it must be that $U_2(f) < U_2(h)$. This is true since under h agent 2 gets his preferred alternative at any realization of types but not so under f. Define the rule g by $g = \epsilon h + (1-\epsilon)\frac{1}{n-2}\sum_{i=3}^{n} f_i$, where $\epsilon > 0$ is sufficiently small. Then every agent except 1 prefers g over f.

Case (3): $U_1(f) = U_2(f) = U_3(f)$. Let $k^*(f) = \max\{1 \leq k \leq n \ : \ U_k(f) = U_1(f)\}$ be the number of agents with maximal expected utility under f and let $K^*(f) = \{3, \ldots, k^*(f)\}$. By our assumption $K^*(f) \neq \emptyset$. Let h be the weighted majority rule where the weight of each $i \in K^*(f)$ is $w_i = \frac{1}{k^*(f)-2}$, the weights of all other agents are zero, and $q = \frac{1}{1+r}$.

We claim that $U_i(f) < U_i(h)$ for any $i \in K^*(f)$. Indeed, consider the problem of maximizing the sum of expected utilities of the agents in this set. A solution to this optimization problem is the rule h, since a simple calculation shows that h chooses R if and only if

$$r|\{i \in K^*(f) \ : \ i \textit{ prefers } R\}| - |\{i \in K^*(f) \ : \ i \textit{ prefers } S\}| > 0, \tag{3}$$

hence h chooses the alternative that gives higher sum of utilities in $K^*(f)$ at any type profile. Now, the sum of expected utilities in $K^*(f)$ under f must be strictly lower than under h, since there is at least one realization of types at which (3) holds but f selects S instead of R, or where the strict reverse inequality of (3) holds but f selects R instead of S. To see this, let t be the minimal number of agents in $K^*(f)$ supporting R needed for (3) to hold, and consider a type profile in which exactly t agents in $K^*(f)$ prefer R and everyone else (also outside of $K^*(f)$) prefer S. If f selects S at this profile we are done. Otherwise (if f selects R), choose two out of the t agents in $K^*(f)$ who prefer R (if $t = 1$ take only that one agent), and consider the profile in which these two agents prefer S instead of R but agents 1 and 2 prefer R instead of S (if $t = 1$ just

change agent's 1 preference); types of all other agents remain unchanged. By Lemma 1 f can be represented with weights satisfying $w_1 = w_2 = \ldots = w_{k^*(f)}$, so the outcome of f in this modified profile is still R. But by the definition of t it now must be the case that the strict reverse inequality of (3) holds.[8] To conclude, we have shown that $\sum_{i\in K^*(f)} U_i(f) < \sum_{i\in K^*(f)} U_i(h)$, but since under both rules all agents in $K^*(f)$ have the same utility it follows that $U_i(f) < U_i(h)$ for every $i \in K^*(f)$.

Define the voting rule g by $g = \epsilon h + (1-\epsilon)\frac{1}{n-k^*(f)} \sum_{i=k^*(f)+1}^{n} f_i$ for $\epsilon > 0$ small. Then $U_i(g) > U_i(f)$ for every $i \geq 3$. □

Appendix B. Proof of Theorem 2

In this section we sketch the main steps in of the proof of Theorem 2. The proof follows the footsteps of the proof of Theorem 1. Complete details of the proof are available from the authors upon request.

By Proposition 1 we may (and we will) assume that $p_i = \frac{1}{2}$ for all i.

Lemma 9. Assume $f = (\underline{w}, q)$ with $w_i \leq w_j$. Let g be the rule obtained from f by switching the weights of i and j. Then $U_i(f) \leq U_i(g)$ and if $U_i(f) = U_i(g)$ then f can be represented by weights satisfying $w_i = w_j$.

Proof. Similar to Lemma 1. □

Lemma 10. If f is a self-stable weighted majority rule then all minimal winning coalitions are of the same size.

Proof. Similar to Lemma 2, but using Lemma 9 instead of Lemma 1. □

Lemma 11. If f is a self-stable weighted majority rule and i, j are two agents which are neither veto players nor null players then f can be represented by weights satisfying $w_i = w_j$.

Proof. Similar to Lemma 3, but using Lemmas 9 and 10 instead of Lemmas 1 and 2. □

Lemma 12. If f is a self-stable weighted majority rule and T is a minimal winning coalition then $|T| \leq 1 + \frac{1}{r(T)}$.

Proof. Similar to Lemma 4, but the function g used to upset f is defined by $g\left(\bar{T} \setminus \{i\}\right) = \frac{r_i}{1+r_i}$ for each $i \in \bar{T}$. □

[8]For this argument to work in the case $t \geq 2$ we must switch the preferences of two agents in $K^*(f)$, since if we switch only one we may end up with a profile in which the left-hand side of (3) is exactly zero. This is the case for example if $r = 1$ and $K^*(f)$ has four agents, since then $t = 3$ and the resulting profile after the switch has two agents supporting each alternative. When $t = 1$ there cannot be a tie.

Lemma 13. Assume that f is a self-stable weighted majority rule, and let $\bar{i}$ be an agent with the highest weight in some representation of f. If T is a coalition such that $\bar{i} \notin T$ and $|T| < \frac{1}{r(T)} + 1$ then $f(T) = 0$.

Proof. Similar to Lemma 5, but the function g used to upset f is defined by $g(S_i) = 1 - \frac{r_i}{1+r_i}$ for each $i \in T$. □

Lemma 14. If f is a self-stable weighted majority rule and there is no coalition M for which $|M| = \frac{1}{r_i} + 2$ holds for all $i \in M$ then f has a veto player.

Proof. From Lemmas 12 and 13 above we know that if a minimal winning coalition T does not contain an agent with the highest weight then $|T| = \frac{1}{r(T)} + 1$. If there is no veto player then (by Lemma 11) all 'normal' players in M have the highest weight, and each minimal winning coalition does not contain one of them, so $|T| = \frac{1}{r(T)} + 1$ holds for all minimal winning coalitions T. This implies together with Lemma 7 (which holds unchanged) that the minimal winning coalitions are exactly those who contain $|M| - 1$ agents from M. Thus, $|M| - 1 = \frac{1}{r(T)} + 1$ for every $T \subseteq M$ satisfying $|T| = |M| - 1$. But since $r(T)$ is the average of $\frac{1}{1+r_i}$ over $i \in T$ we get that r_i must be the same for all $i \in M$. Thus, $|M| = \frac{1}{r_i} + 2$ holds for all $i \in M$. □

Proof of 'Only If' direction in Theorem 2:
The proof is essentially the same as that of theorem 1. If f has veto players then, by the above lemmas, minimal winning coalitions are those containing V and either 0, 1, or $|M| - 1$ of the 'normal' players from M. This translates into forms (A), (B), and (C), respectively. The bounds on the sizes of V and M come from Lemma 12. If f does not have veto players then it follows from Lemmas 14 and 7 that f must look as in (**ii**). □

Proof of 'If' direction in Theorem 2:
As in the proof of Theorem 1, given f in one of the forms of the theorem and given a minimal winning coalition T we find weights $\{\lambda_i\}_{i \in T}$ (not all zero) such that f is a maximizer of $\sum_{i \in T} \lambda_i U_i(g)$ among all voting rules g.

For form (A) set $\lambda_i = \frac{1}{1+r_i}$ for each $i \in V$. For form (B) let $\lambda_i = \frac{1}{1+r_i}$ for each $i \in V$, and $\lambda_i = \sum_{j \in V} \frac{r_j}{1+r_j}$ for the single player from M in T. For form (C) set $\lambda_i = 1$ for all $i \in V$, and $\lambda_i = \frac{\sum_{j \in V} \frac{r_j}{1+r_j}}{1 - \sum_{j \in M \cap T} \frac{r_j}{1+r_j}}$ for each $i \in M \cap T$. For f in the form of part **(ii)** simply let $\lambda_i = 1$ for all $i \in T$. □