Potentials in Social Environments

Thomas Demuynck* P. Jean-Jacques Herings† Christian Seel‡

February 8, 2024

Abstract

We develop and extend notions of potentials for normal-form games (Monderer and Shapley, 1996) to present a unified approach for the general class of social environments. The different potentials and corresponding social environments can be ordered in terms of their permissiveness. We classify different methods to construct potentials and we characterize potentials for specific examples such as matching problems, vote trading, multilateral trade, TU games, and various pillage games.

KEYWORDS: *Potential games, social environments.*

JEL-CLASSIFICATION: *C70, C71.*

*Ecares, Université Libre de Bruxelles, Belgium. E-Mail: Thomas.Demuynck@ulb.ac.be

†Department of Econometrics and Operations Research, Tilburg University, Tilburg, The Netherlands. E-Mail: P.J.J.Herings@tilburguniversity.edu

‡Department of Economics, Maastricht University, Maastricht, The Netherlands. E-Mail: C.Seel@maastrichtuniversity.nl

1 Introduction

In a potential game, the incentives of all players can be represented by one single function, called the potential. Since the introduction of potential games by Monderer and Shapley (1996), the amount of literature on the topic has been enormous, both within economics but also in related fields such as computer science (e.g., Yamamoto, 2015) and evolutionary biology (e.g., Szabó and Fáth, 2007). Over time, different notions of the original potential have appeared (Voorneveld, 2000; Dubey, Haimanko, and Zapechelnyuk, 2006) and convergence properties of different dynamics have been studied for sub-classes of potential games such as aggregative games (Selten, 1970) and congestion games (Rosenthal, 1973); see, e.g., Jensen (2010) and Chien and Sinclair (2011).

In this paper, we develop various notions of potential for the general class of social environments (Chwe, 1994). A social environment consists of four components: a set of individuals, a set of states, preferences for each individual over the set of states, and an effectivity correspondence that specifies which coalitions of individuals can switch from one state to another.

Social environments provide a unified framework for many popular models in economics. For instance, matching models, models of network formation, coalitional games, and normal-form games can all be represented as a social environment (Demuynck et al., 2019). When framing normal-form games as a particular kind of social environment, the notion of a Nash equilibrium generalizes naturally to the notion of a core element. As we will show, under mild conditions, potential social environments have a non-empty core as any state that maximizes the potential function is a core element. Similarly, in terms of learning dynamics, potential social environments tend to have attractive convergence properties towards the core of the social environment.

Our paper serves several different additional purposes. First, we connect different strands of literature to provide a common and general structure in which potentials are useful. This structure facilitates the knowledge transfer across environments. Second, we provide a classification which allows us to compare different applications in terms of their potentials and thereby their convergence/stability levels. Third, we discuss two methods which can help to find potentials in unexplored social environments. Finally, we explore novel social environments by studying some variants of pillage games (Jordan, 2006).

We classify potentials along two dimensions: better reply versus coalitional best reply versus individual best reply potentials and strong versus weak potentials. Better reply potentials are the ones that increase when coalitions take better replies while best reply

potentials only need to increase when taking best replies. We differentiate between strong and weak potentials based on whether the potential is required to increase for *every* better or best reply (strong) or only for *some* better or best replies (weak).

In normal-form games, best replies are naturally defined on the individual level. For social environments, defining best replies is less straightforward, as coalitions consisting of multiple individuals can deviate. To define best replies in such settings, we distinguish between two types: coalitional and individual best replies. Coalitional best replies require that a coalition makes a profitable deviation that is Pareto optimal among all deviations feasible for this coalition. Here Pareto efficiency is restricted to hold only for the members of the deviating coalition. For individual best replies, we require that the deviation is optimal for some individual among all profitable deviations involving all coalitions that contain this individual. Coalitional best replies could be thought of as the result of a discussion in a fixed group, while individual best replies take the view of an individual who initiates a group and proposes a move in her best interest.

Based on these dimensions, i.e., strong versus weak and better reply versus coalitional best reply versus individual best reply, we obtain six types of potentials for social environments. These different potentials can be ordered in terms of their permissiveness. The strong better reply potential is most demanding, followed by the strong coalitional best reply, the strong individual best reply, the weak individual best reply, the weak coalitional best reply and finally, our most permissive concept, the weak better reply potential.

The nested structure allows to interpret social environments which admit a more demanding potential as having nicer stability and convergence properties than social environments which only admit a more permissive potential. Indeed, if a finite social environment admits a strong better reply potential, every path of better replies will converge in finite time to a core element, while a weak better reply potential only implies the existence of one such better reply path. The latter property relates to the weak finite improvement property in games as defined by Friedman and Mezzetti (2001), while the former relates to the finite improvement property in games (Monderer and Shapley, 1996).

When we consider normal-form games as a special type of social environment, the concept of a generalized ordinal potential game (Monderer and Shapley, 1996) coincides with our notion of a strong better reply potential when applied to normal-form games. The notion of a best reply potential in games (Voorneveld, 2000) is more demanding than our notion of a strong best reply potential. Under some weak technical conditions, our strong best reply potential is in turn more demanding than the pseudo-potential (Dubey, Haimanko, and Zapelchelnyuk, 2006), which is more demanding than our weak best reply potential.

We apply our notions to different examples of social environments. In a pure exchange economy with multilateral trade, the additive social welfare function is a strong better reply potential as any Pareto-improving trade increases that function. In a model of group formation model with transferable utilities, the potential of the cooperative game (Hart and Mas-Colell, 1989) is a strong better reply potential of the related social environment.

Even in the framework of normal-form games, there is no easy recipe how to find a potential. For social environments, we provide possible approaches to finding potentials by distilling and adapting two methods from the literatures on matching and vote trading. The first method of a sequential potential splits up the state space and focuses on local weak potential properties which can then, under some conditions, be combined into a global potential. The second method finds several potentials and then prioritizes between them based on a lexicographic order. We illustrate the usefulness of these two methods by providing tight characterizations of potentials for the matching model by Gale and Shapley (1962) and the model of vote trading by Casella and Palfrey (2019).

Finally, we analyze the potentials admitted by different types of pillage games. This analysis covers both the original pillage game discussed in Jordan (2006) as well as two new versions of the game. The variant with the strongest individual attachments (pillaging by gangs) admits a more demanding potential than the other two versions.

We proceed as follows. In Section 2, we introduce social environments and develop different potentials for that framework. The relations between our concepts and the convergence properties of both better and best reply dynamics are studied in Section 3. Section 4 compares our potentials for social environments to the existing notions for normal-form games. We then discuss two existing applications for which potentials are easy to find in Section 5. Section 6 introduces two different techniques to construct potentials and shows how to use them in applications to matching and vote trading. In Section 7, we study a less explored environment in the form of pillage games, including some new versions of that model. Section 8 provides a short conclusion. Proofs and some of the technical details are relegated to appendices.

2 Potential Social Environments

This section provides an overview of the notions of social environment and core. We also define and compare the various notions of a potential social environment. At the end of the section, we discuss to which extend our notions deviate from existing notions of potential in normal-form games.

Let N be the finite set of all individuals. A coalition is a subset of N. The collection of all non-empty coalitions is denoted by $\mathcal{N}$. A social environment is determined by the set of individuals N, a metric space (X, d) where X is a non-empty space of states and d is a metric on X, an effectivity correspondence E that associates to each pair of states $(x, y) \in X \times X$ a, possibly empty, subset of $\mathcal{N}$, and a tuple of utility functions $u = (u^i)_{i \in N}$ where, for each $i \in N$, $u^i : X \to \mathbb{R}$. The notation $S \in E(x, y)$ means that coalition S is able to move from state x to state y whereas $u^i(x)$ gives the utility of individual i in state x. We denote a social environment by $\Gamma = (N, (X, d), E, u)$.

A social environment $\Gamma = (N, (X, d), E, u)$ is said to be *finite* if the set of states X is finite. We next develop a dominance relation between the states in X.

Definition 2.1. A state $y \in X$ *dominates* a state $x \in X$ via coalition $S \in \mathcal{N}$ if $S \in E(x, y)$, for every $i \in S$, $u^i(y) \geq u^i(x)$, and, for some $j \in S$, $u^j(y) > u^j(x)$.

We say that $y \in X$ dominates $x \in X$ if y dominates x via some coalition. The subset of X consisting of all states that dominate $x \in X$ via coalition $S \in \mathcal{N}$, together with the state x itself, is denoted by $f_S(x)$, i.e., $f_S(x) = \{x\} \cup \{y \in X \mid y \text{ dominates } x \text{ via coalition } S\}$. We denote by $f(x)$ the set of all states that dominate x together with x itself, $f(x) = \bigcup_{S \in \mathcal{N}} f_S(x)$.

Definition 2.2. Let $\Gamma = (N, (X, d), E, u)$ be a social environment. The *core* of Γ is given by all states $x \in X$ such that $f(x) = \{x\}$.

Thus, a state belongs to the core if no coalition can move to another state which gives at least the same utility to all and higher utility to some coalition members. It is well-known that for some social environments the core is empty.[1]

2.1 Best Replies in Social Environments

A key choice in extending the concept of a best reply potential from normal-form games to social environments lies in finding a suitable extension of the notion of a best reply from a single player to a coalition. For this purpose, we define the set of feasible states for coalition $S \in \mathcal{N}$ at state $x \in X$ by $F_S(x) = \{y \in X \mid S \in E(x, y)\}$. The following

[1] Our definition of the core generalizes the standard definition of the core as defined for coalitional games to the setting of social environments. As explained in Osborne and Rubinstein (1994): "The core is a solution concept for coalitional games that requires that no set of players be able to break away and take a joint action that makes all of them better off." Our concept of domination extends this reasoning to social environments in a straightforward way. Our requirement that at least one player gets higher utility rather than all players turns out to be technically more convenient.

definition formalizes that a coalition of players has the option to change the state and chooses to do so in an optimal way for its members. It thereby extends the corresponding best reply notion for normal-form games.

Definition 2.3. Let $\Gamma = (N, (X, d), E, u)$ be a social environment. A state $y \in X$ is a *coalitional best reply* to $x \in X$ by coalition $S \in \mathcal{N}$ if $y \in f_S(x)$ and there is no state $z \in F_S(x)$ such that, for every $i \in S$, $u^i(z) \geq u^i(y)$ and, for some $j \in S$, $u^j(z) > u^j(y)$.

In case y is a coalitional best reply to x, we write $y \in \mathrm{BR}_S(x)$. We also denote the collection of all best replies to x by

$$\mathrm{BR}(x) = \bigcup_{S \in \mathcal{N}} \mathrm{BR}_S(x).$$

Under technical conditions on the correspondence F_S and the utility functions, the set of best replies by a coalition turns out to be non-empty.

Proposition 2.4. *Let $\Gamma = (N, (X, d), E, u)$ be a social environment. Let $x \in X$ and $S \in \mathcal{N}$ be such that $F_S(x)$ is compact and, for every $i \in S$, the utility function u^i is continuous. Then $\mathrm{BR}_S(x)$ is non-empty.*

As is typical for notions of coalitional domination, the set $\mathrm{BR}_S(x)$ may fail to be closed. The proof of Proposition 2.4 takes care of this subtlety by focusing on a relevant compact subset of $\mathrm{BR}_S(x)$.

An alternative, more individualistic, view on extending best replies is that one individual takes the initiative to organize a coalition for a state change that is mutually beneficial, but in her best interest. This leads to the following definition of an individual best reply.

Definition 2.5. Let $\Gamma = (N, (X, d), E, u)$ be a social environment. A state $y^* \in X$ is an *individual best reply* to $x \in X$ for an individual $i \in N$ if

$$y^* \in \operatorname*{argmax}_{y \in \bigcup_{\{S \in \mathcal{N} \mid i \in S\}} \mathrm{BR}_S(x)} u^i(y).$$

In case y^* is an individual best reply to x for individual i, we write $y^* \in \mathrm{IBR}_i(x)$. We also denote the set of all individual best replies to x by $\mathrm{IBR}(x) = \bigcup_{i \in N} \mathrm{IBR}_i(x)$. Note that for all $x \in X$, $IBR(x) \subseteq BR(x) \subseteq f(x)$.

The following result shows that the set of individual best replies is non-empty under an appropriate strengthening of the conditions in Proposition 2.4.

Proposition 2.6. *Let $\Gamma = (N, (X, d), E, u)$ be a social environment. Let $i \in N$ and $x \in X$. If for all $S \in \mathcal{N}$ such that $i \in S$, the set $F_S(x)$ is compact and, for every $j \in N$, the utility function u^j is continuous, then $\mathrm{IBR}_i(x)$ is non-empty.*

2.2 Strong and Weak Potentials

We start with the definition of a strong better reply potential. As we will see later on, this generalizes the notion of a generalized ordinal potential (Monderer and Shapley, 1996) from games to social environments.

Definition 2.7. Let $\Gamma = (N, (X, d), E, u)$ be a social environment. A function $P : X \to \mathbb{R}$ is a *strong better reply potential* for Γ if for all states $x, y \in X$,

$$y \in f(x) \setminus \{x\} \implies P(y) > P(x).$$

The social environment Γ is a *strong better reply potential social environment* if it admits a strong better reply potential.

We can also define natural analogues for strong best reply potentials.

Definition 2.8. Let $\Gamma = (N, (X, d), E, u)$ be a social environment. A function $P : X \to \mathbb{R}$ is a *strong coalitional best reply* (*strong individual best reply*) *potential* for Γ if, for every state $x \in X$,

$$y \in \mathrm{BR}(x) \setminus \{x\} \ (y \in \mathrm{IBR}(x) \setminus \{x\}) \implies P(y) > P(x).$$

The social environment Γ is a *strong coalitional best reply* (*strong individual best reply*) *potential social environment* if it admits a strong coalitional best reply (strong individual best reply) potential.

Note that the potential P might also increase if some members of a deviating coalition increase their payoffs, whereas other coalition members decrease their payoffs. Yet, the latter part of the coalition would not consent and hence the deviation should not be considered a better (or best) reply. As such, using a two-sided implication in Definitions 2.7 and 2.8 would place an undesirable restriction on deviations, thereby motivating our modeling choice of a one-sided implication.

Our final set of definitions of potentials for social environments reduces the requirements on the potential. In particular, we now only require that whenever a state is dominated

by another state, in terms of better reply, coalitional best reply, or individual best reply, then there is at least one domination which increases the potential.

These weaker notions retain many appealing properties of its stronger counterparts. As we shall see, if one of these weak potentials exist, the core is non-empty. Furthermore, optima of a potential function correspond to core elements, which might facilitate its computation. Moreover, the introduction of weak potentials allows us to draw a formal connection between results on better-reply dynamics (e.g. Friedman and Mezzetti, 2001) and the literature on potentials.

Definition 2.9. Let $\Gamma = (N, (X, d), E, u)$ be a social environment. A function $P : X \to \mathbb{R}$ is a *weak better reply* (*weak coalitional best reply*) [*weak individual best reply*] *potential* for Γ if for every $x \in X$ such that $f(x) \setminus \{x\} \neq \emptyset$, there exists $y \in f(x)$ ($y \in \mathrm{BR}(x)$) [$y \in \mathrm{IBR}(x)$] such that $P(y) > P(x)$.

The social environment is a *weak better reply* (*weak coalitional best reply*) [*weak individual best reply*] *potential social environment* if it admits a weak better reply (weak coalitional best reply) [weak individual best reply] potential.

Under mild assumptions, we can order the various potentials.

Proposition 2.10. *Let $\Gamma = (N, (X, d), E, u)$ be a social environment such that, for every $x \in X$, for every $S \in \mathcal{N}$, $\mathrm{BR}(x) \neq \emptyset$, and, for every $i \in N$, $\mathrm{IBR}_i(x) \neq \emptyset$. We have the following implications for a potential P for Γ:*

strong better reply $\Rightarrow$ strong coalitional best reply,
strong coalitional best reply $\Rightarrow$ strong individual best reply,
strong individual best reply $\Rightarrow$ weak individual best reply,
weak individual best reply $\Rightarrow$ weak coalitional best reply,
weak coalitional best reply $\Rightarrow$ weak better reply.

Given the nestedness of the concepts, we say that a social environment *admits a potential* if and only if it has a weak better reply potential.

3 Potentials: Basic Properties and Relations

Under weak conditions, every potential social environment has a non-empty core.

Proposition 3.1. *Let $\Gamma = (N, (X, d), E, u)$ be a social environment. If Γ admits a potential that reaches a maximum on X, then its core is non-empty. In particular, sufficient conditions are that X is compact and that the potential is continuous.*

Thus, all finite potential social environments have a non-empty core. Yet, they differ in the way different dynamic processes converge to the core. We now present one possible dynamic foundation which will facilitate the proofs. Our approach is similar in spirit to the dynamics obtained from analyzing acyclic best-reply graphs in games (Young, 1993, p.64) but we also look at better reply dynamics and extend the framework to social environments. A different, non-determistic foundation via Markov chains is discussed in Appendix B.

For our characterization in terms of paths, we need some additional notation. Let $\mathbb{N} = \{1, 2, \ldots\}$ denote the set of positive integers and let $\mathcal{K} = \{\{1\}, \{1, 2\}, 1, 2, 3, \ldots\} \cup \{\mathbb{N}\}$ be the collection of index sets. For $K \in \mathcal{K}$, if $K = \mathbb{N}$, then we define $K^- = K$, and if there is $m \in \mathbb{N}$ such that $K = \{1, \ldots, m\}$, then define $K^- = K \setminus \{m\}$ as the set that results from K by leaving out its highest element.

Definition 3.2. Let $\Gamma = (N, (X, d), u)$ be a social environment and $K \in \mathcal{K}$. The path $(x_k)_{k \in K} \in X^K$ is

a *better reply path* if for all $k \in K^-$, $x_{k+1} \in f(x_k) \setminus \{x_k\}$,
a *coalitional best reply path* if for all $k \in K^-$, $x_{k+1} \in \mathrm{BR}(x_k) \setminus \{x_k\}$,
an *individual best reply path* if for all $k \in K^-$, $x_{k+1} \in \mathrm{IBR}(x_k) \setminus \{x_k\}$.

We have the following characterization of the various finite potential social environments.

Proposition 3.3. *Let $\Gamma = (N, (X, d), u)$ be a finite social environment. Then,*

- *Γ has a strong better reply (coalitional best reply) [individual best reply] potential if and only if every better reply (coalitional best reply) [individual best reply] path is finite.*
- *Γ has a weak better reply (coalitional best reply) [individual best reply] potential if and only if from every state $x \in X$ there is a finite better reply (coalitional best reply) [individual best reply] path that starts at x and ends at a core element.*

In addition to being helpful in the characterization of potentials, Proposition 3.3 sheds light on the properties of dynamic processes within a finite social environment. In the presence of a strong better reply potential, better reply dynamics is guaranteed to converge to a core element in a finite number of iterations without visiting any state outside the core more than once. Analogous properties hold for coalitional (individual) best reply dynamics in the presence of a strong coalitional (individual) best reply potential. For social environments with a weak potential, the convergence properties become probabilistic: the various types

Game 1

	A	B	C
A	(4,4)	(0,0)	(0,0)
B	(0,0)	(2,2)	(0,0)
C	(0,0)	(0,0)	(2,2)

Game 2

	A	B	C
A	(4,4)	(4,0)	(4,0)
B	(0,4)	(2,2)	(3,1)
C	(0,4)	(3,1)	(2,2)

Game 3

	A	B	C
A	(4,4)	(0,0)	(0,0)
B	(0,0)	(2,2)	(3,1)
C	(0,0)	(3,1)	(2,2)

Game 4

	A	B	C
A	(4,4)	(0,0)	(0,0)
B	(0,0)	(3,1)	(5,0)
C	(0,0)	(5,2)	(2,3)

Game 5

	A	B	C
A	(3,2)	(0,0)	(0,0)
B	(0,0)	(1,1)	(4,0)
C	(0,0)	(4,0)	(2,4)

Game 6

	A	B	C
A	(2,2)	(3,0)	(0,0)
B	(0,0)	(0,3)	(4,0)
C	(0,0)	(4,0)	(0,3)

Figure 1: Coalitional normal-form games to illustrate Proposition 3.3.

of dynamics converge to a core element in finite time with probability one and it is possible to return to a previously visited state outside the core; see Appendix B for details.

We illustrate Proposition 3.3 by the six social environments that are induced by the two-player coalitional normal-form games in Figure 1. We add the adjective coalitional to emphasize that also non-singleton coalitions are allowed to deviate.[2] More precisely, for every $x, y \in X$, it holds that $\{1\} \in E(x, y)$ if and only if $x_2 = y_2$, $\{2\} \in E(x, y)$ if and only if $x_1 = y_1$, and $\{1, 2\} \in E(x, y)$. The core of the induced social environment corresponds to the set of strong Nash equilibria (Aumann, 1959).

The function $P = u^1 + u^2$ is a strong better reply potential for Game 1. It follows that Game 1 has a strong Nash equilibrium and that better reply dynamics converges to it in a finite number of iterations.

Game 2 has a better reply cycle (B,B), (C,B), (C,C), (B,C), (B,B), so it cannot have a strong better reply potential by Proposition 3.3. Best replies of the singleton coalition $\{1\}$ always lead to states where player 1 chooses A and best replies of the singleton coalition $\{2\}$ always lead to states where player 2 chooses A. From any state different from (A,A), the unique best reply of coalition $\{1, 2\}$ is state (A,A). State (A,A) itself is not dominated. Every coalitional best reply path is therefore finite. By virtue of Proposition 3.3, Game 2 has a strong coalitional best reply potential.

In Game 3, best replies by singleton coalitions lead to a coalitional best reply cycle (B,B), (C,B), (C,C), (B,C), (B,B). By Proposition 3.3, Game 3 does not admit a strong coalitional

[2]If we restrict the effectivity correspondence to singleton deviations, Games 3, 4, and 5 do not admit a potential.

best reply potential. At any state different from (A,A), both players 1 and 2 have a unique individual best reply, which is to establish coalition $\{1,2\}$ and deviate to state (A,A). As state (A,A) is not dominated, every individual best reply path is finite, so Game 3 possesses a strong individual best reply potential.

Game 4 has an individual best reply cycle (B,B), (C,B), (C,C), (B,C), (B,B) with singleton moving coalitions. Thus, this game has no strong individual best reply potential. Nevertheless, it is an individual best reply for player 2 to form a coalition $\{1,2\}$ and deviate from any state different from (B,C), (C,B) and (A,A) to (A,A). As the state (A,A) is in the core, Game 4 has a weak individual best reply potential.

The path (B,B), (C,B), (C,C), (B,C), (B,B) is an individual best reply cycle in Game 5. Unlike in Game 4, (A,A) is no longer an individual best reply of player 2 to (B,B) and (C,C). Thus, the game has no finite individual best reply path that ends at a core element from any of the states (B,B), (C,B), (C,C), and (B,C), thereby ruling out a weak individual best reply potential. For any state outside (A,A), (B,C), (C,B), and (C,C), i.e., in particular for state (B,B), it is a coalitional best reply for $\{1,2\}$ to deviate to (A,A). Since moreover state (A,A) is in the core, Game 5 admits a weak coalitional best reply potential.

Finally, in Game 6, there is no finite coalitional best reply path that ends at a core element (A,A) from any of the states in the cycle (B,B), (C,B), (C,C), (B,C), (B,B). Since (A,B) is a better reply to (B,B) for player 1 and (A,A) is a better reply (A,B) for player 2, it is possible to construct a finite better reply path that ends at a core element from any of the states (B,B), (C,B), (C,C), and (B,C). From the other states (A,B), (A,C), (B,A), and (C,A), the core element can be reached by means of a single better reply. Thus, Game 6 has a weak better reply potential.

If we only change the payoff associated to (A,B) from (0,3) to (0,0) in Game 6, there is no finite better reply path to (A,A) from any of the states (B,B), (C,B), (C,C), and (B,C). Thus, the non-vacuity of the core is not a sufficient condition for the existence of a weak better reply potential.

4 Potential Normal-form Games

This section covers the relation of our concepts to the previous literature in the specific social environment of normal-form games. For that purpose, we first provide a formal introduction of normal-form games and how they fit into the social environment framework.

Let $G = (N, (X^i)_{i\in N}, (u^i)_{i\in N})$ be a normal-form game, where N is the set of players,

X^i is the strategy space of player i, and $u^i : \prod_{j\in N} X^j \to \mathbb{R}$ is the utility function of player i. This game induces the social environment $\Gamma^G = (N, (X, d), E, (u^i)_{i\in N})$ in the following way. The set of players N coincides with the set of individuals in the social environment. The state space $X = \prod_{i\in N} X^i$ is the set of all strategy profiles, and d is a suitable metric on X. We denote by X^{-i} the set of strategy profiles for the set of individuals in $N \setminus \{i\}$ and we write elements of X^{-i} accordingly as x^{-i}. The utility functions in the social environment coincide with those in the normal-form game. The effectivity correspondence E is defined by $S \in E(x, y)$ if and only if there is $i \in S$ such that $S = \{i\}$ and $x^{-i} = y^{-i}$, where x and y are two arbitrary strategy profiles in X. In particular, it holds that $E(x, y) = \{\{i\} : i \in N, x^{-i} = y^{-i}\}$. The restriction to single-player deviations ensures that the core coincides with the set of all pure-strategy Nash equilibria. We distinguish between the game G and its social environment representation by writing the latter as Γ^G.

The starting point from the literature on potential games is the definition of a generalized ordinal potential game by Monderer and Shapley (1996).

Definition 4.1. Let $G = (N, (X^i)_{i\in N}, (u^i)_{i\in N})$ be a normal-form game. A function $P : X \to \mathbb{R}$ is a *generalized ordinal potential* for G if, for every $i \in N$, for every $x^{-i} \in X^{-i}$ and for every $y^i, z^i \in X^i$, we have

$$u^i(y^i, x^{-i}) > u^i(z^i, x^{-i}) \implies P(y^i, x^{-i}) > P(z^i, x^{-i}).$$

The game G is a *generalized ordinal potential game* if it admits a generalized ordinal potential.

The literature on potential games has provided several variations of potential games based on best replies. Based on their prominence in the literature and closeness to our concepts, we will focus on the notions of a best reply potential (Voorneveld, 2000) and a pseudo-potential (Dubey, Haimanko, and Zapelchelnyuk, 2006). We call a function $P : X \to \mathbb{R}$ *regular* if for all $i \in N$ and $x^{-i} \in X^{-i}$, we have $\operatorname{argmax}_{x^i\in X^i} P(x^i, x^{-i}) \neq \emptyset$.

Definition 4.2. Let $G = (N, (X^i)_{i\in N}, (u^i)_{i\in N})$ be a normal-form game.

- A function $P : X \to \mathbb{R}$ is a *best reply potential* for G if it is regular and, for every $i \in N$, for every $x^{-i} \in X^{-i}$, we have

$$\arg\max_{x^i\in X^i} u^i(x^i, x^{-i}) = \arg\max_{x^i\in X^i} P(x^i, x^{-i}).$$

  The game G is a *best reply potential game* if it admits a best reply potential.

- The function $P : X \to \mathbb{R}$ is a *pseudo-potential* for the game G if it is regular and, for every $i \in N$, for every $x^{-i} \in X^{-i}$, we have

$$\arg\max_{x^i \in X^i} u^i(x^i, x^{-i}) \supset \arg\max_{x^i \in X^i} P(x^i, x^{-i}).$$

  The game G is a *pseudo-potential game* if it admits a pseudo-potential.

Normal-form games restrict attention to single-player deviations. Thus, the notions of individual best reply and coalitional best reply coincide and we simply we refer to best reply when analyzing normal-form games through the lens of social environments. The following proposition derives implications between the different concepts.

Proposition 4.3. *Let $G = (N, (X^i)_{i \in N}, (u^i)_{i \in N})$ be a normal-form game.*

- *The game G is a generalized ordinal potential game if and only if Γ^G is a strong better reply potential social environment.*
- *If G is a best reply potential game, then Γ^G is a strong best reply potential social environment.*
- *If Γ^G is a strong best reply potential social environment, it has a regular potential, and, for every $i \in N$, for every $x \in X$, $BR_{\{i\}}(x) \neq \emptyset$, then G is a pseudo-potential game.*
- *If G is a pseudo-potential game, then Γ^G is a weak best reply potential social environment.*

Our least permissive concept of a strong better social environment directly extends the notion of a generalized ordinal potential game. Voorneveld (2000) shows that a generalized ordinal potential game might not be a best reply potential game and vice versa. As such, we cannot rank these two concepts. The remaining three implications show that, under some technical assumptions, the other concepts have a nested structure: a best reply potential game imposes more demanding restrictions than our strong best reply notions for social environments, while the pseudo-potential game is in between our strong and weak best reply notions for social environments in terms of permissiveness.

The following examples illustrate that the second, third and fourth implication of Proposition 4.3 are not equivalences. To show this for the second implication of Proposition 4.3, consider the normal-form game G in Figure 2. The environment Γ^G is a strong best reply potential social environment with the best reply potential P defined by $P(A, A) =$

$P(A,B) = P(B,B) = 0$ and $P(B,A) = 1$. If G admitted a best reply potential, then $P(A,A) = P(A,B)$, $P(A,A) = P(B,A)$, and $P(A,B) = P(B,B)$ by the conditions resulting from the best replies, i.e., the best reply potential would have to be constant over all strategy profiles. However, A is a the unique best reply of the column player if the row player chooses B, which implies $P(B,A) > P(B,B)$, leading to a contradiction.

	A	B
A	(0,0)	(0,0)
B	(0,1)	(0,0)

Figure 2: A game that is not a best reply potential game, but induces a strong best reply potential social environment.

To show that the third implication is not an equivalence, consider Game G in Figure 3. This game has a pseudo-potential P defined by $P(A,A) = 2$, $P(A,B) = 1$, $P(A,C) = 1$, $P(B,A) = 3$, $P(B,B) = 2$, and $P(B,C) = 0$. Suppose there is a strong best reply potential for Γ^G. From the conditions on the individual best replies, we have $P(A,B) > P(A,C) > P(B,C) > P(B,B) > P(A,B)$, a contradiction.

	A	B	C
A	(0,1)	(0,1)	(1, 0)
B	(1,1)	(1,0)	(0, 1)

Figure 3: A pseudo-potential game which does not induce a strong best reply potential social environment.

Finally, to show that the last implication of Proposition 4.3 is not an equivalence, consider the three-player game G in Figure 4.[3] Note that Γ^G is a weak best reply social environment with potential P defined by $P(B,A,A) = 1$, $P(B,B,A) = 2$, $P(A,B,A) = 3$, $P(A,A,A) = 4$, $P(A,A,B) = 5$, $P(A,B,B) = P(B,A,B) = P(B,B,B) = 0$. Any pseudo-potential for G would require that $P(B,A,A) > P(A,A,A) > P(A,B,A) > P(B,B,A) > P(B,A,A)$, i.e., G does not admit a pseudo-potential.

Intuitively, the pseudo-potential requires that every player has a best reply which increases the potential, whereas the weak best reply potential allows flexibility in choosing the player. According to the following proposition, the two notions coincide for finite two-player games.

[3]Player 1 is the row player, player 2 is the column player, and player 3 chooses the payoff matrix.

	A	
	A	B
A	(0,1,0)	(1,0,0)
B	(1,0,0)	(0,1,0)

	B	
	A	B
A	(1,1,1)	(0,0,0)
B	(0,0,0)	(0,0,0)

Figure 4: A game that is not a pseudo-potential game, but induces a weak best reply potential social environment.

Proposition 4.4. *Let $G = (N, (X^i)_{i\in N}, (u^i)_{i\in N})$ be a finite normal-form game with two players. If Γ^G is a weak best reply potential social environment, then G is a pseudo-potential game.*

5 Two Simple Applications

This section characterizes potentials for two prominent social environments in the literature, multilateral trade and transferable utility games.

5.1 Multilateral Trade

Consider an exchange economy $\mathcal{E} = (N, \omega, \widetilde{u})$ consisting of a set of individuals N, aggregate initial endowments $\omega = (\omega_1, \omega_2, \ldots, \omega_L) \geq 0$ of the L goods in the economy, and, for every $i \in N$, a utility function $\widetilde{u}^i : \mathbb{R}^L_+ \to \mathbb{R}$. A state is given by an allocation $x = (x^i)_{i\in N}$, where, for every $iinN$, $x^i \in \mathbb{R}^L_+$, and, for every $\ell = 1, \ldots, L$, $\sum_{i\in N} x^i_\ell = \omega_\ell$. The state space X consists of all such allocations.

The utility function $u^i : X \to \mathbb{R}$ of an individual $i \in N$ in the social environment is defined by $u^i(x) = \widetilde{u}^i(x^i)$, so it depends only on x^i. Individuals repeatedly meet in groups and have the possibility to exchange their current consumption bundles. The effectivity correspondence only allows for redistribution inside the trading coalition, i.e., for every $x, y \in X$ it holds that $S \in E(x, y)$ if and only if $\sum_{i\in S} x^i = \sum_{i\in S} y^i$ and, for every $j \in N \setminus S$, $x^j = y^j$. The resulting social environment is denoted by $\Gamma^{\mathcal{E}}$.

Proposition 5.1. *Let $\mathcal{E} = (N, \omega, u)$ be an exchange economy. The additive social welfare function $P : X \to \mathbb{R}$ defined by $P(x) = \sum_{i\in N} u^i(x^i)$, $x \in X$, is a strong better reply potential for the social environment $\Gamma^{\mathcal{E}}$.*

Any domination increases the sum of utilities of the members of the moving coalition, while keeping the utility of the remaining individuals constant. Thus, the sum of all individual

utilities increases when moving from a state to another state that dominates it. A better reply dynamics leads to an increase in the potential in every iteration. The corresponding non-tâtonnement process therefore exhibits attractive dynamic behavior where utilities of individuals are monotonically increasing until a core element is reached. Note that the resulting core element is not necessarily the maximum of the additive social welfare function. However, as long as the social planner is purely interested in Pareto efficiency and individual rationality, there is no reason to intervene in an exchange economy.

5.2 Shapley Value

In the next setting, we analyze group formation where the surplus of the group is split according to the Shapley value. Each non-member can decide to join and each member can decide to leave the group. More formally, we have a transferable utility game (N, v), where $v(\emptyset) = 0$ and, for every $S \in \mathcal{N}$, $v(S) \in \mathbb{R}$ is the worth of coalition S. For $S \in \mathcal{N}$, we denote the subgame of (N, v) restricted to coalition S by $(S, v_{|S})$ and the subgame payoffs as determined by the Shapley value by $\varphi(S, v_{|S}) \in \mathbb{R}^S$. The symmetric difference between two sets is denoted by Δ, so if $S, T \in \mathcal{N}$, then $S \Delta T = (S \setminus T) \cup (T \setminus S)$.

The social environment $\Gamma^{(N,v)} = (N, (X, d), E, u)$ is obtained by taking the state space equal to the collection of subsets of N, so $X = \mathcal{N} \cup \{\emptyset\}$. Given two states $x, y \in X$, it holds that $S \in E(x, y)$ if and only if $|S| = 1$ and $x \Delta y = S$, so only single players are effective and they can either leave or join an existing coalition. Finally, for every $x \in X$, for every $i \in N$, $u^i(x) = 0$ if $i \notin x$ and $u^i(x) = \varphi^i(x, v_{|x})$ if $i \in x$.

Proposition 5.2. *Let (N, v) be a transferable utility game. The function $P : X \to \mathbb{R}$ defined by*

$$P(x) = \sum_{T \subset x} \frac{(|T| - 1)!(|x| - |T|)!}{|x|!} v(T), \quad x \in X,$$

is a strong better reply potential for the social environment $\Gamma^{(N,v)}$.

The expression for $P(x)$ in Proposition 5.2 is equal to the potential of the cooperative game $(x, v_{|x})$ as defined in Hart and Mas-Colell (1989). Proposition 5.2 thereby illustrates how our definition of potential for social environments naturally incorporates the Hart and Mas-Colell (1989) potential as defined in cooperative game theory. Monderer and Shapley (1996) obtain a similar relation between the Hart and Mas-Colell (1989) potential and the potential of a non-cooperative participation game.

As a consequence of Proposition 5.2, better reply dynamics converges quickly to a stable coalition when the revenues of the team are allocated via the Shapley value.

6 Methods to Find Weak Potentials

For some applications, finding a potential might be difficult. In this section, we provide a toolkit by conceptualizing two methods to find potentials.

6.1 Sequential Potentials

The next definition constructs potential functions locally that lead "step-by-step" towards a core element.

Definition 6.1. Let $\Gamma = (N, (X, d), E, u)$ be a social environment. The profile of sets and functions $\mathcal{AP} = ((A_1, P_1), \ldots, (A_\ell, P_\ell))$, where

$$X = A_1 \supseteq \cdots \supseteq A_{\ell-1} \supseteq A_\ell \supseteq A_{\ell+1} = \emptyset$$

and, for every $k \in \{1, \ldots, \ell\}$, $P_k : X \to \mathbb{R}$, is a *sequential weak better reply (sequential weak coalitional best reply) [sequential weak individual best reply] potential system* for Γ if, for every $k \in \{1, \ldots, \ell\}$, for every $x \in A_k \setminus A_{k+1}$, $f(x) = \{x\}$ or

$$\text{there exists } y \in f(x) \cap A_k \ (y \in \mathrm{BR}(x) \cap A_k) \ [y \in \mathrm{IBR}(x) \cap A_k] \text{ such that } P_k(y) > P_k(x).$$

This definition states that if $x \in A_k \setminus A_{k+1}$ does not belong to the core, then there must always be a dominating state in A_k that increases the potential P_k. In the definition, we employ weak rather than strict set inclusions to account for subcases where, for example due to a particular preference profile, two adjacent sets might coincide.

For finite X, either x belongs to the core, or there is an improving path to a core element in $A_k \setminus A_{k+1}$, or there is an improving path to an element in A_{k+1}. This observation is useful in proving the following result.

Proposition 6.2. *Let Γ be a finite social environment. If $\mathcal{AP}$ is a sequential weak better reply (coalitional best reply) [individual best reply] potential system for Γ, then Γ has a weak better reply (coalitional best reply) [individual best reply] potential.*

The converse of Proposition 6.2 holds as well: if Γ has a weak better reply (coalitional best reply) [individual best reply] potential, then there is a profile of sets and functions $\mathcal{AP}$ which is a sequential weak better reply (sequential coalitional best reply) [sequential individual best reply] potential system for Γ. Indeed, it suffices to take $\ell = 1$ and choose P_1 to be the potential.

Definition 6.1 and Proposition 6.2 can be extended to strong potentials. It suffices to replace there exists $y \in f(x) \cap A_k$ ($y \in \mathrm{BR}(x) \cap A_k$) [$y \in \mathrm{IBR}(x) \cap A_k$] by for every $y \in f(x)$ ($y \in \mathrm{BR}(x)$) [$y \in \mathrm{IBR}(x)$] and to require that $y \in A_k$.

We illustrate the usefulness of sequential potential systems in a matching application.

Gale-Shapley Matching Consider a matching problem $\mathcal{M} = (M, W, u)$ with a set M of men and a disjoint set W of women. The set of individuals equals $N = M \cup W$. The state space X consists of the set of all possible matchings, which are functions $x : N \to N$ such that, for every $m \in M$, $x(m) \in W \cup \{m\}$, for every $w \in W$, $x(w) \in M \cup \{w\}$, and for every $i \in N$, $x(x(i)) = i$. The effectivity correspondence E allows only for pairwise and singleton deviations of the following type: Every $(m, w) \in M \times W$ can deviate from a matching x to a matching $y = x + (m, w)$ where they form a couple and their possible previous partners become single. Other individuals are not affected by such a deviation. Every $m \in M$ and every $w \in W$ who is part of a couple (m, w) in x can unilaterally deviate to a matching $y = x - (m, w)$ where both members of the previous couple become single and other individuals are not affected. The utilities $u^m : X \to \mathbb{R}$ and $u^w : X \to \mathbb{R}$ depend only on the own partner. Without loss of generality, the range of the utility functions is contained in $(0, 1)$. All preferences are strict. The resulting social environment is denoted by $\Gamma^{\mathcal{M}}$.

A matching $x \in X$ is *individually rational* if no matched individual prefers to be single. The pair $(m, w) \in M \times W$ is a *blocking pair* in x if the matching $y = x + (m, w)$ is such that $u^m(y) > u^m(x)$ and $u^w(y) > u^w(x)$. The matching x is *stable* if it is individual rational and has no blocking pair, which is equivalent to $f(x) = \{x\}$. Using Definition 2.2, a matching is stable if and only if it belongs to the core of the social environment.

We borrow the following example from the proof of Theorem 4.1 in Ackermann, Goldberg, Mirrokni, Röglin, and Vöcking (2011) to show that there are matching problems $\mathcal{M}$ for which the associated social environment $\Gamma^{\mathcal{M}}$ has no strong individual best reply potential.

Example 6.3. Let $M = \{m_1, m_2, m_3\}$ and $W = \{w_1, w_2, w_3\}$ and let utility functions be in accordance with the following preferences:

$\underline{m_1}$	$\underline{m_2}$	$\underline{m_3}$		$\underline{w_1}$	$\underline{w_2}$	$\underline{w_3}$
w_1	w_2	w_1		m_2	m_1	m_3
w_3	w_1	w_2		m_3	m_2	m_1
w_2	w_3	w_3		m_1	m_3	m_2
m_1	m_2	m_3		w_1	w_2	w_3.

Let x_1 be the matching where m_1 is single, m_2 is matched to w_2, and m_3 to w_3, see Figure 5 for an illustration. Let $x_2 = x_1 + (m_3, w_1)$, $x_3 = x_2 + (m_1, w_2)$, $x_4 = x_3 + (m_1, w_3)$, $x_5 = x_4 + (m_2, w_1)$, $x_6 = x_5 + (m_2, w_2)$, and $x_1 = x_7 = x_6 + (m_3, w_3)$.

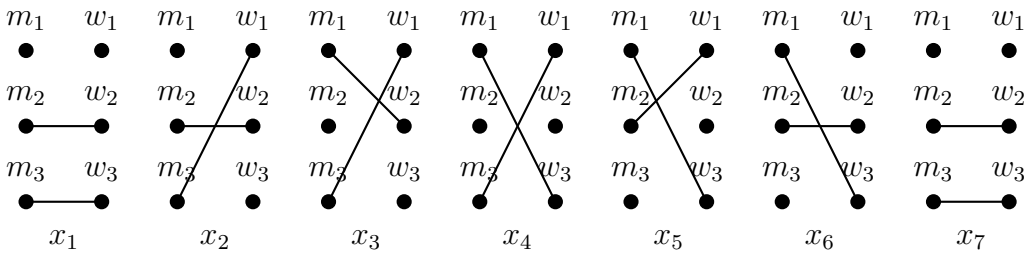

Figure 5: Infinite individual best reply path.

It clearly holds that $x_2 \in \mathrm{IBR}_{m_3}(x_1)$ and $x_3 \in \mathrm{IBR}_{w_2}(x_2)$. As w_1 prefers her match m_3 at x_3 to m_1, it holds that w_3 is the best possible partner in a blocking pair for m_1, so $x_4 \in \mathrm{IBR}_{m_1}(x_3)$. It clearly holds that $x_5 \in \mathrm{IBR}_{w_1}(x_4)$, $x_6 \in \mathrm{IBR}_{m_2}(x_5)$, and $x_7 \in \mathrm{IBR}_{w_3}(x_6)$. This example therefore admits an infinite individual best reply path. We use Proposition 3.3 to conclude that $\Gamma^{\mathcal{M}}$ does not have a strong individual best reply potential. It follows that $\Gamma^{\mathcal{M}}$ does not admit a strong coalitional or a strong better reply potential. Note, however, that this argument does not rule out the existence of a weak individual best reply potential. For the example, let $\tilde{x}_3 = x_2 + (m_1, w_3)$ and note that $\tilde{x}_3 \in \mathrm{IBR}_{w_3}(x_2)$ and that $\tilde{x}_3$ is a stable matching. Thus, the social environment might admit a weak individual best reply potential. $\diamond$

To show that matching problems indeed admit a weak individual best reply potential, we construct a sequential weak individual best reply potential system.

Let $\mathcal{M} = (M, W, u)$ be a matching problem. For a matching $x \in X$, we denote by $S_{\mathrm{M}}(x)$ the set of single men and by $S_{\mathrm{W}}(x)$ the set of single women. We define A_1 as the set of all matchings, A_2 as the set of all matchings where no matched man prefers to be single, A_3 as the set of all individually rational matchings without blocking pairs involving married women, and A_4 as the empty set. Notice that $X = A_1 \supseteq A_2 \supseteq A_3 \supseteq A_4 = \emptyset$.

For every $x \in X$, let

$$\begin{array}{rcl} P_1(x) & = & \sum_{m \in M \setminus S_{\mathrm{M}}(x)} (u^m(x) - 1), \\ P_2(x) & = & \sum_{w \in W \setminus S_{\mathrm{W}}(x)} (u^w(x) - 1), \\ P_3(x) & = & \sum_{m \in M} u^m(x). \end{array} \tag{1}$$

The function P_1 increases in the payoff of the matched men, but also increases if more men become single. The function P_2 increases in the payoffs of the matched women, but also

increases if more women become single. The function P_3 corresponds to the total payoffs of the men.

Our next result verifies that the above sets and potentials form a sequential weak individual best reply potential system. To get an intuition, start with a matching x which is not stable. The matching can be in $A_1 \setminus A_2$, $A_2 \setminus A_3$, or $A_3 \setminus A_4$. Let $k \in \{1, 2, 3\}$ be such that $x \in A_k \setminus A_{k+1}$. We construct an individual best reply which belongs to the set A_k and increases the corresponding potential. For example, if $x \in A_1 \setminus A_2$, some men prefer to be single. A best reply of such a man increases P_1 and remains in A_1. We provide a similar argument for each set and the corresponding potential which implies the following result.

Proposition 6.4. *Let $\mathcal{M} = (M, W, u)$ be a matching problem with associated social environment $\Gamma^{\mathcal{M}}$. Then $\Gamma^{\mathcal{M}}$ admits a weak individual best reply potential.*

The result is an extension of known results the literature on better/best-reply paths in matching. Roth and Vande Vate (1990) show that from every matching there is a better reply path to a stable matching. In our language, every one-to-one matching model has a weak better reply potential. Our result strengthens Theorem 4.2 in Ackerman et al. (2011) who also consider individual best replies, but they restrict preferences such that each individual prefers any partner over being unmatched.

We conclude that there is an individual best reply path to a stable matching, or, using Appendix B, that the individual best reply dynamics is guaranteed to reach a stable matching in finite time with probability one.

6.2 Lexicographic Potentials

A lexicographic potential system consists of a set of potentials with a certain priority. Let $\Gamma = (N, (X, d), E, u)$ be a social environment and let $\mathcal{P} = (P_1, \dots, P_\ell)$ be a finite profile of functions, where, for every $k \in \{1, \dots, \ell\}$, $P_k : X \to \mathbb{R}$. Consider the strict partial order on X derived from the lexicographic relationship between the functions in $\mathcal{P}$,

$$x \succ_{\mathcal{P}} y \iff \exists k' \in \{1, \dots, \ell\},\ \forall k < k',\ P_k(x) = P_k(y) \text{ and } P_{k'}(x) > P_{k'}(y).$$

Definition 6.5 (Lexicographic potential system)**.** The profile of functions $\mathcal{P} = (P_1, \dots, P_\ell)$, where, for every $k \in \{1, \dots, \ell\}$, $P_k : X \to \mathbb{R}$ is a *lexicographic weak better reply (weak coalitional best reply) [weak individual best reply] potential system* for the social environment

$\Gamma = (N, (X, d), E, u)$ if, for every $x \in X$,

$$f(x) = \{x\} \text{ or there exists } y \in f(x) \ (y \in \mathrm{BR}(x)) \ [y \in \mathrm{IBR}(x)] \text{ such that } y \succ_{\mathcal{P}} x.$$

Proposition 6.6. *Let Γ be a finite social environment. If $\mathcal{P}$ is a lexicographic weak better reply (weak coalitional best reply) [weak individual best reply] potential system for Γ, then Γ has a weak better reply (weak coalitional best reply) [weak individual best reply] potential.*

It is straightforward to extend Definition 6.5 and Proposition 6.6 to strong potentials.

Vote Trading To illustrate the usefulness of lexicographic potential systems, we consider the vote trading model of Casella and Palfrey (2019). In their model, N is a finite set of voters, who have to vote on a finite set of binary proposals R. For every $i \in N$, for every $r \in R$, $x_r^{0i} \in \mathbb{N}_0$ is the number of votes that i can initially cast on proposal r.[4] However, it is possible to change x^0 by trading votes. A distribution of votes belongs to the finite set

$$X = \left\{ x \in \mathbb{N}_0^{NR} \middle| \text{for every } r \in R, \ \textstyle\sum_{i \in N} x_r^i = \sum_{i \in N} x_r^{0i} \right\}.$$

For each proposal r, the total number of votes $\sum_{i \in N} x_r^{0i}$ is odd in order to avoid ties.

Let $r \in R$ be a proposal. Voter i has a non-zero intensity $z_r^i \in \mathbb{R}$ for proposal r, where $z_r^i > 0$ if i is in favor of r and $z_r^i < 0$ if i is against proposal r. We define $N_r^+ = \{i \in N \mid z_r^i > 0\}$ and $N_r^- = \{i \in N \mid z_r^i < 0\}$ as the voters that are in favor, respectively against, the implementation of proposal r. Voter $i \in N$ casts all votes x_r^i in favor of r if $z_r^i > 0$ and all votes x_r^i against r if $z_r^i < 0$. Let $x \in X$. We obtain $v_r^+(x) = \sum_{i \in N_r^+} x_r^i$ and $v_r^-(x) = \sum_{i \in N_r^-} x_r^i$ as the number of votes in favor, respectively against, proposal r. Let $A(x) \subset R$ be the set of all proposals that are accepted in state x, which are the proposals that obtain a majority of votes in favor in state x, so $A(x) = \{r \in R \mid v_r^+(x) > v_r^-(x)\}$. We call $A(x)$ the outcome of the vote.

The utility of voter $i \in N$ at state $x \in X$ is given by $u^i(x) = \sum_{r \in A(x)} z_r^i$. Preferences are assumed to be strict, i.e., for every $x, y \in X$ such that $A(x) \neq A(y)$, for every $i \in N$, it holds that $u^i(x) \neq u^i(y)$. The score $s^i(x)$ of voter $i \in N$ at state $x \in X$ is given by $s^i(x) = \sum_{r \in R} x_r^i |z_r^i|$.

The proposals that are decided by a minimal majority are collected in the set $M(x)$ defined by $M(x) = \{r \in R \mid |v_r^+(x) - v_r^-(x)| = 1\}$. If a proposal r belongs to $M(x)$, then moving one vote from the majority to the minority changes the outcome on proposal r.

[4] We use the notation $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$.

A coalition can move from one state to another by redistributing the votes within the coalition. More formally, let $x, y \in X$. It holds that $S \in E(x, y)$ if and only if $y^{-S} = x^{-S}$. Due to the assumptions on utility functions, it makes no difference whether we define domination as in Definition 2.1 or we require improvements to be strict. The social environment for the vote trading model is denoted by Γ^{vt}.

Casella and Palfrey (2019) show that the vote trading model admits a sequence of payoff-improving trades leading to a core allocation, i.e., a finite better reply path in our language. By Proposition 3.3, the social environment Γ^{vt} admits a weak better reply potential. In their Example 4, they construct a cycle for another path of better replies, thereby showing that the game admits no strong better reply potential. In the rest of this section, we tighten their characterization to include our remaining potentials.

Our next example shows that there are social environments Γ^{vt} for which no strong individual best reply potential exists. This example is a small variation on Example 4 in Casella and Palfrey (2019).[5]

Example 6.7. Consider the vote trading situation displayed in Table 1.

Table 1: Each cell shows the preference intensity z_r^i.

Proposal \ Individual	1	2	3	4	5	6	7
A	2.3	-1	-1.1	-1.1	1	1	1
B	-1	2.3	-1.1	-1.1	1	1	1
C	-1.1	-1.1	2.3	-1	1	1	1
D	-1.1	-1.1	-1	2.3	1	1	1

In the initial distribution of votes, denoted by x_1, each voter is assumed to have one vote on each proposal. Thus, each proposal passes with minimal majority, so $A(x_1) = M(x_1) = \{\text{A}, \text{B}, \text{C}, \text{D}\}$. In terms of utilities, we have $u^1(x_1) = u^2(x_1) = u^3(x_1) = u^4(x_1) = -0.9$ and $u^5(x_1) = u^6(x_1) = u^7(x_1) = 4$.

Voters 5, 6, and 7 cannot be part of a deviating coalition as they want all proposals to pass and they obtain the maximum possible utility at x_1. Any domination changes the outcome

[5] As in Casella and Palfrey (2019), the preferences in Example 6.7 are not strict for expositional clarity; the example can be easily modified to one with strict preferences.

of at least one proposal, so any domination requires that at least one of the voters 1, 2, 3, or 4 gives up a vote on his preferred proposal, which is then rejected. The utility of such a voter can only be increased if all proposals are rejected. Any domination therefore results in the outcome where all proposals are rejected. One such domination is achieved as follows: voter 1 trades one vote on A for one vote on B with voter 2 and voter 3 trades his vote on C for one vote on D with voter 4. The resulting distribution of votes is denoted by x_2. Note that $x_2 \in \mathrm{IBR}_1(x_1)$. All proposals are rejected in x_2 and all proposals are decided by minimal majorities, so $A(x_2) = \emptyset$ and $M(x_2) = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$. In terms of utilities, we have that $u^1(x_2) = \cdots = u^7(x_2) = 0$.

Any individual best reply to x_2 by voter 1 has to lead to a majority for proposal A. At the same time, a trade needs to change the majority on one of the other proposals, where B is least painful to 1. Thus, 1 and 2 trading one vote back leads to distribution of votes x_3. It holds that $x_3 \in \mathrm{IBR}_1(x_2)$, the only proposals that pass are A and B, and both do so with minimal majority. Utilities are equal to $u^1(x_3) = u^2(x_3) = 1.3$, $u^3(x_3) = u^4(x_3) = -2.2$, and $u^5(x_3) = u^6(x_3) = u^7(x_3) = 2$.

From x_3, individuals 1 and 2 cannot be involved in any trade which leads to a majority in favor of proposal C or D. Any best reply to x_3 by individual 3 needs to reestablish a majority for C. Note that 5, 6, 7 vote in favor of C anyway, so 3 needs to get at least one vote on C from 4. Thus, given the incentives of voter 4, any individual best reply by voter 3 leads to an acceptance of proposals C and D. Let voters 3 and 4 exchange one vote on proposals C and D. The resulting distribution of votes is given by $x_4 \in \mathrm{IBR}_3(x_3)$. Notice that $x_4 = x_1$. Thus, we have a cycle of individual best replies, i.e., the social environment has no strong individual best reply potential by Proposition 3.3. ◇

The more difficult part of our characterization is to construct a lexicographic weak individual best reply potential system. For that purpose, let the profile of functions $\mathcal{P} = (P_1, \ldots, P_{n+1})$ be defined by

$$\begin{array}{rcll} P_1(x) & = & |M(x)|, & x \in X, \\ P_k(x) & = & s^{k-1}(x), & k \in \{2, \ldots, n+1\},\ x \in X. \end{array} \tag{2}$$

We need to show that whenever there is a domination, at least one such domination is an individual best reply and increases the potential in (2). We construct such a domination by either increasing the number of proposals which are decided by minimal majority to increase P_1 or by shifting votes to increase the score of the deviating coalition member with the lowest index k to increase P_k.

Proposition 6.8. *The profile of functions $\mathcal{P}$ as defined in (2) is a lexicographic weak individual best reply potential system for* Γ^{vt}.

Thus, individual best reply dynamics converges to a core element in finite time with probability one in the vote trading social environment. Yet, particular distributions of votes might show up multiple times before a stable distribution of votes is reached.

7 Pillage Games

To show some novel applications of potentials, this section considers modifications of the wealth-equals-power pillage game. We first discuss the basic setup of Jordan (2006) and the potential properties under the original assumptions. The remainder of the section analyzes potentials in new variants of the game. The comparison between the different results shows that more coalitional attachment when pillaging leads to the admission of more demanding potentials.

7.1 Basic Setup

Consider a game with a finite set of players $N = \{1, \ldots, n\}$, where $n \geq 2$. The set of allocations of wealth is denoted by $\Omega = \{\omega \in \mathbb{R}_+^N \mid \sum_{i \in N} \omega^i = 1\}$. The *power* of a coalition $S \in \mathcal{N}$ at allocation $\omega \in \Omega$ is denoted by $\alpha(\omega, S)$ and serves as an input for the effectivity correspondence, which in turn specifies which pillages, i.e., reallocations of wealth, are possible.

Intuitively, richer and larger coalitions are more powerful than poorer and smaller ones. Jordan (2006) formalizes this idea by imposing the following three properties on the power function. Let $x, y \in \Omega$ and $S, T \in \mathcal{N}$. (1) If $S \subset T$, then $\alpha(x, S) \leq \alpha(x, T)$. (2) If, for every $i \in S$, $x^i \leq y^i$, then $\alpha(x, S) \leq \alpha(y, S)$. (3) If, for every $i \in S$, $x^i < y^i$, then $\alpha(x, S) < \alpha(y, S)$. For the rest of this section, we impose the more specific assumption that the power of a coalition $S \in \mathcal{N}$ at allocation $\omega \in \Omega$ is equal to its wealth $\alpha(\omega, S) = \sum_{i \in S} \omega^i$ as discussed in Section 3 of Jordan (2006).

7.2 Pillaging with no Coalitional Attachment

A pillage game with no coalitional attachment is denoted by $\mathcal{PN} = (N, \Omega, u)$ and follows Jordan (2006), who assumes that endowment changes by a pillage are permanent, but

coalitions are temporary, i.e., the state is determined only by the distribution of wealth. In particular, former partners-in-crime are not bound together and can freely pillage each other in the future. The state space equals $X = \Omega$.

Let $x \in X$. We assume that individuals maximize their wealth, so, for every $i \in N$, we have $u^i(x) = x^i$. We define $Z(x) = \{i \in N | x^i = 0\}$ as the set of players with zero wealth at x. The total wealth of a coalition $S \in \mathcal{N}$ at x is denoted by $x(S) = \sum_{i \in S} x_i$.

Let $x, y \in X$. We set $S \in E(x, y)$ if and only if there exists a coalition $T \in \mathcal{N}$ with $S \cap T = \emptyset$ such that

1. $x(S) > x(T)$,
2. for all $i \notin S \cup T$, $y^i = x^i$.

We call T the pillaged coalition. The first condition requires the power of the pillaging coalition S to be bigger than the power of the pillaged coalition T. The second condition restricts a redistribution of resources to $S \cup T$. Note that we allow coalition S to refrain from pillaging all resources of coalition T.

The social environment induced by a pillage game $\mathcal{PN}$ with no coalitional attachment is denoted by $\Gamma^{\mathcal{PN}}$. We obtain the following result.

Proposition 7.1. *Let $\mathcal{PN} = (N, \Omega, u)$ be a pillage game with no coalitional attachment. The function $P_1 : X \to \mathbb{R}$ defined by*

$$P_1(x) = |Z(x)|, \quad x \in X,$$

is a weak individual best reply potential for the social environment $\Gamma^{\mathcal{PN}}$. For $n = 2$, the function $P_2 : X \to \mathbb{R}$ defined by

$$P_2(x) = |x^1 - x^2|, \quad x \in X,$$

is a strong better reply potential for $\Gamma^{\mathcal{PN}}$. For $n \geq 3$, $\Gamma^{\mathcal{PN}}$ has no strong individual best reply potential.

7.3 Partners-in-Crime

A pillage game with partners-in-crime is denoted by $\mathcal{PP} = (N, \Omega, u)$. Unlike in the case with no coalitional attachment, one-time partners-in-crime form a bond and are, be it

morally or by a group-specific contract, not able to pillage each other henceforth. For example, one might think of settlers who suppress inhabitants or of gang membership. In the former example, which we will analyze in this section, it seems natural to restrict the bond to the relation between pillagers. In the latter example, which we address in the next section, gangs also offer protection to their members when being pillaged.

In this section, a state is described by a profile of wealth levels ω together with a profile of coalitions $C = (C^i)_{i \in N}$. Here $C^i \subset N$ gives for each individual i the players with whom i has pillaged in the past. Clearly, $j \in C^i$ if and only if $i \in C^j$ and we impose the convention $i \in C^i$ for all $i \in N$. We let $\mathcal{C}$ be the set of all such possible profiles.

A state $x \in X = \Omega \times \mathcal{C}$ consists of a profile of wealth levels $\omega(x) \in \Omega$ and a profile of coalitions $C(x) \in \mathcal{C}$. As before, the power of a coalition is equal to the aggregate wealth of the members in the coalition. A coalition $S \in \mathcal{N}$ can move from a state $x \in X$ to a state $y \in X$, i.e., $S \in E(x, y)$, if and only if there exists a coalition $T \in \mathcal{N}$ such that, for every $i \in S$, $C^i(x) \cap T = \emptyset$ and

1. $\omega(x)(S) > \omega(x)(T)$,
2. for all $i \notin S \cup T$, $\omega^i(y) = \omega^i(x)$,
3. for all $i \in S$, $C^i(y) = C^i(x) \cup S$ and for all $i \notin S$, $C^i(y) = C^i(x)$.

There are two crucial modifications compared to the previous model. First, the members of the pillaging coalition may no longer pillage former partners-in-crime. Second, by Condition 3, current partners-in-crime become attached to each other. The social environment induced by a pillage game $\mathcal{PP}$ with partners-in-crime networks is denoted by $\Gamma^{\mathcal{PP}}$.

Proposition 7.2. *Let $\mathcal{PP} = (N, \Omega, u)$ be a pillage game with partners-in-crime. The function $P_1 : X \to \mathbb{R}$ defined by*

$$P_1(x) = |Z(\omega(x))|, \quad x \in X,$$

is a weak individual best reply potential for the social environment $\Gamma^{\mathcal{PP}}$. For $n = 2$, the function $P_2 : X \to \mathbb{R}$ defined by

$$P_2(x) = |\omega^1(x) - \omega^2(x)|, \quad x \in X,$$

is a strong better reply potential for $\Gamma^{\mathcal{PP}}$. For $n = 3$, the function $P_3 : X \to \mathbb{R}$ defined by

$$P_3(x) = |Z(\omega(x))| + \sum_{j \in N} \sum_{i \in C^j(x)} \omega^i(x), \quad x \in X,$$

is a strong coalitional best reply potential for $\Gamma^{\mathcal{PP}}$*, but* $\Gamma^{\mathcal{PP}}$ *has no strong better reply potential. For* $n \geq 4$*,* $\Gamma^{\mathcal{PP}}$ *has no strong individual best reply potential.*

7.4 Pillaging by Gangs

A pillage game with gang formation is denoted by $\mathcal{PG} = (N, \Omega, u)$. In such a game, a pillage with a gang member leads to gang membership. A state consists of the distribution of wealth and a partition of the set of players into gangs. More formally, if we denote by Π the collection of partitions of N, we define a state $x \in X = \Omega \times \Pi$ to consist of a profile of wealth levels $\omega(x) \in \Omega$ and a partition $\pi(x) \in \Pi$. As before, the power of a coalition $S \in \mathcal{N}$ is equal to the aggregate wealth $\omega(x)(S)$ of the members in the coalition.

The pillaging coalition S may contain members of different gangs. If so, then all the gangs involved merge into a new gang. If coalition S attacks coalition T, the gang members $G(x,T) = \bigcup_{\{C \in \pi(x) | C \cap T \neq \emptyset\}} C$ of T offer protection to the individuals in T. A coalition $S \in \mathcal{N}$ can therefore move from a state $x \in X$ to a state $y \in X$, i.e., $S \in E(x,y)$, if and only if there exists a coalition $T \in \mathcal{N}$ with $S \cap G(x,T) = \emptyset$ such that

1. $\omega(x)(S) > \omega(x)(G(x,T))$,
2. $\forall i \notin S \cup T$, $\omega^i(y) = \omega^i(x)$,
3. $G(x,S) \in \pi(y)$ and $\pi(y) \setminus \{G(x,S)\} \subset \pi(x)$.

The requirement $S \cap G(x,T) = \emptyset$ states that the pillaged coalition cannot contain any gang members of the pillagers. By the modified third condition, all gang members of the pillaging coalition become part of the new coalitional structure in y and other gangs stay intact. We denote the social environment induced by a pillage game with gang formation by $\Gamma^{\mathcal{PG}}$.

Proposition 7.3. *Let* $\mathcal{PG} = (N, \Omega, u)$ *be a pillage game with gang formation. For* $n = 2$*, the function* $P_1 : X \to \mathbb{R}$ *defined by*

$$P_1(x) = |\omega^1(x) - \omega^2(x)|, \quad x \in X,$$

is a strong better reply potential for $\Gamma^{\mathcal{PG}}$*. For* $n \geq 3$*, the function* $P_2 : X \to \mathbb{R}$ *defined by*

$$P_2(x) = -|\pi(x)| + \sum_{C \in \pi(x)} (\omega(x)(C))^2, \quad x \in X,$$

is a strong coalitional best reply potential for the social environment $\Gamma^{\mathcal{PG}}$. *For* $n \geq 3$, $\Gamma^{\mathcal{PG}}$ *has no strong better reply potential.*

When we compare the three variants of pillage games, the version with no coalitional attachment admits less demanding types of potential than the variant with partners-in-crime, which in turn admits less demanding types of potential than the one with pillaging by gangs. In line with intuition, more coalitional attachment after joint pillages makes it easier to reach a stable state. For the most general case of $n \geq 3$, the jump from partners-in-crime to gangs is substantial, thereby illustrating the appeal of a stronger coalitional attachment for stability.

8 Conclusion

We have provided several notions of potential which are suitable for the general class of social environments. The general ordinal potential function from the literature on normal-form games (Monderer and Shapley, 1996) has a direct analogue for social environments. Other concepts in games such as the best reply potential (Voorneveld, 2000) and the pseudo-potential (Dubey, Haimanko, and Zapelchelnyuk, 2006) are nested with our notions when we consider the social environment induced by normal-form games.

The notions of better and best reply are straightforward to define for single-player deviations in the context of normal-form games. A key contribution of our paper is to provide notions of potential when we allow for deviations of coalitions consisting of multiple individuals and a more general framework of social environments. The nested structure of the different potentials allows us to compare different environments in terms of their permissiveness to different types of potential. Furthermore, we provide a toolkit in the form of two different methods to construct such potentials.

Our characterizations of potentials for the different social environments such as multilateral trade, group formation, matching, vote trading, and pillage games are tight. On the one hand, these findings have closed some gaps in the individual literatures. On the other hand, we can conclude that the studied environments in multilateral trade and coalition formation have stronger stability properties than the matching and vote trading environments. We can also compare the stability across different pillage games and conclude that the variant with gang formation is particularly conducive to stability due to stronger coalitional attachments.

We see several avenues for future research. For instance, one might construct potentials in other unexplored environments, address the tightness of such potentials in these en-

vironments, and analyze the comparative statics of potentials for different environments under restrictions on the preferences or the effectivity correspondence. A necessary condition for the existence of a potential is the non-emptiness of the core. As such, Knuth matching problems (Knuth, 1976 and Tamura, 1993) and roommate problems do not admit such a potential without restrictions on preferences, but may well do so under appropriate conditions on the primitives.

A Proofs

A.1 Proofs of Section 2

Proof of Proposition 2.4 If $f_S(x) = \{x\}$, then $x \in \mathrm{BR}_S(x)$, so we are done. Consider now the case where there is $z \in f_S(x) \setminus \{x\}$. Then $\sum_{i\in S} u^i(z) > \sum_{i\in S} u^i(x)$, so we can find an $\varepsilon > 0$ such that $\sum_{i\in S} u^i(z) \geq \sum_{i\in S} u^i(x) + \varepsilon$. We define

$$A = \left\{ y \in f_S(x) \,\middle|\, \sum_{i\in S} u^i(y) \geq \sum_{i\in S} u^i(x) + \varepsilon \right\}.$$

Note that $z \in A$, so A is non-empty. Let us show next that A is closed. Let $(y_k)_{k\in\mathbb{N}}$ be a sequence in A and assume that $y_k \to \overline{y}$. Then as $F_S(x) = \{y \in X \mid S \in E(x, y)\}$ is compact, it follows that $S \in E(x, \overline{y})$. Next, for every $k \in \mathbb{N}$, $\sum_{i\in S} u^i(y_k) \geq \sum_{i\in S} u^i(x) + \varepsilon$. By continuity of the function $\sum_{i\in S} u^i$, it follows that $\sum_{i\in S} u^i(\overline{y}) \geq \sum_{i\in S} u^i(x) + \varepsilon$. Finally, for every $i \in S$, for every $k \in \mathbb{N}$, we have $u^i(y_k) \geq u^i(x)$ as $y_k \in f_S(x)$. By continuity of u^i, it follows that $u^i(\overline{y}) \geq u^i(x)$. As such, $\overline{y} \in A$, as we wanted to show.

As A is a closed subset of the compact set $F_S(x)$, the set A is compact.

Let

$$y^* \in \operatorname*{argmax}_{y\in A} \sum_{i\in S} u^i(y).$$

As A is non-empty and compact and the function $\sum_{i\in S} u^i$ is continuous, such a y^* exists. Let us show that $y^* \in \mathrm{BR}_S(x)$. First, we have $y^* \in f_S(x)$. As y^* maximizes the sum of utilities on A, there is no $z \in f_S(x)$ such that, for every $i \in S$, $u^i(z) \geq u^i(y^*)$, and, for some $j \in S$, $u^j(z) > u^j(y)$. This completes the proof.

Proof of Proposition 2.6 From Proposition 2.4, for all $S \in \mathcal{N}$ such that $i \in S$, the set $\mathrm{BR}_S(x)$ is non-empty.

If for all coalitions $S \in \mathcal{N}$ that contain i and for all states $y \in \mathrm{BR}_S(x)$ it holds that $u^i(y) =$

$u^i(x)$, then we immediately have that $\mathrm{IBR}_i(x) = \bigcup_{\{S\in\mathcal{N}|i\in S\}} \mathrm{BR}_S(x)$. Non-emptiness of $\mathrm{IBR}_i(x)$ follows as every set $\mathrm{BR}_S(x)$ is non-empty.

Otherwise, there is at least one coalition $S \in \mathcal{N}$ that contains i and there is at least one state $z \in \mathrm{BR}_S(x)$ such that $u^i(z) > u^i(x)$. In particular, we can find $\varepsilon > 0$ such that $u^i(z) \geq u^i(x) + \varepsilon$.

Consider the set

$$A = \left\{ y \in \bigcup_{\{S\in\mathcal{N}|i\in S\}} f_S(x) \middle| u^i(y) \geq u^i(x) + \varepsilon \right\}.$$

The set A is non-empty as it contains z.

We show next that A is closed. Let $(y_k)_{k\in\mathbb{N}}$ be a sequence in A that converges to $\overline{y} \in X$. As $\mathcal{N}$ is finite, we can assume without loss of generality that the coalition S is fixed along the sequence. Then, for every $j \in S$, for every $k \in \mathbb{N}$, we have $u^j(y_k) \geq u^j(x)$. As u^j is continuous, we obtain $u^j(\overline{y}) \geq u^j(x)$. Also $u^i(y_k) \geq u^i(x) + \varepsilon$, so taking the limit we find $u^i(\overline{y}) \geq u^i(x) + \varepsilon$. The compactness of $F_S(x)$ implies $\overline{y} \in F_S(x)$. It follows that $\overline{y} \in A$.

For every $S \in \mathcal{N}$ such that $i \in S$ it holds that $F_S(x) \cup \{x\}$ is compact, so $\cup_{\{S\in\mathcal{N}|i\in S\}} F_S(x) \cup \{x\}$ is compact as well. Since A is a closed subset of the compact set $\cup_{\{S\in\mathcal{N}|i\in S\}} F_S(x) \cup \{x\}$, it follows that A is compact.

Let

$$y^i \in \operatorname*{argmax}_{y\in A} u^i(y).$$

As A is non-empty and compact and u^i is continuous, such a y^i exists. The state y^i may fail to be in $\mathrm{IBR}_i(x)$ as it may not be in $\bigcup_{\{S\in\mathcal{N}|i\in S\}} \mathrm{BR}_S(x)$. We use y^i to find a state y^* with $u^i(y^*) = u^i(y^i)$ that belongs to the latter set.

Let $T \in \mathcal{N}$ with $i \in T$ be such that $y^i \in f_T(x)$. We now define the set of states B as those in $f_T(x)$ such that individual i attains the utility level corresponding to y^i, so

$$B = \{y \in f_T(x) \mid u^i(y) = u^i(y^i)\}.$$

The set B can be written as the intersection of the compact set $f_T(x)$, the set $(u^i)^{-1}(\{u^i(y^i)\})$, and, for $j \in S \setminus \{i\}$, the sets $(u^j)^{-1}([u^j(x), \infty))$, where the latter sets are closed as the preimage by a continuous function of a closed set. As an intersection of a compact set and closed sets, the set B is compact. It is non-empty as it contains y^i.

Let

$$y^* \in \operatorname*{argmax}_{y\in B} \sum_{j\in T} u^j(y).$$

As B is non-empty and compact and the function $\sum_{j\in T} u^j$ is continuous, such a y^* exists. Let us show that $y^* \in \mathrm{BR}_T(x)$. First, we have $y^* \in f_T(x)$. Second, as y^* maximizes the sum of utilities on B, there is no $z \in f_T(x)$ such that, for every $j \in T$, $u^j(z) \geq u^j(y^*)$, and, for some $j \in T$, $u^j(z) > u^j(y^*)$.

Since y^* maximizes u^i on the set A, there cannot be a coalition $S \in \mathcal{N}$ containing i and a $y \in \mathrm{BR}_S(x)$ such that $u^i(y) > u^i(y^*)$. We have shown that $y^* \in \mathrm{IBR}_i(x)$.

Proof of Proposition 2.10 Let P be a strong individual best reply potential for Γ. Let $x \in X$ be such that $f(x) \setminus \{x\} \neq \emptyset$. Then there is $S \in \mathcal{N}$ such that $f_S(x) \setminus \{x\} \neq \emptyset$. By assumption, $\mathrm{BR}_S(x) \neq \emptyset$. Let $y \in \mathrm{BR}_S(x)$. Since $f_S(x) \setminus \{x\} \neq \emptyset$, there is $j \in S$ such that $u^j(y) > u^j(x)$. By assumption, $\mathrm{IBR}_j(x) \neq \emptyset$. Let $z \in \mathrm{IBR}_j(x)$. Then $u^j(z) \geq u^j(y) > u^j(x)$, so $z \in \mathrm{IBR}(x) \setminus \{x\}$. Definition 2.8 implies $P(z) > P(x)$. It follows that P is a weak individual best reply potential for Γ.

Since every individual best reply is also a coalitional best reply and every coalitional best reply is also a better reply, the other implications in the proposition follow immediately from the definitions.

A.2 Proofs of Section 3

Proof of Proposition 3.1 Let P be a weak better reply potential for Γ that reaches a maximum on X. Let $x^* \in \mathrm{argmax}_{x\in X} P(x)$.

Suppose x^* is not in the core. Then we have $f(x^*) \setminus \{x^*\} \neq \emptyset$. By the definition of a weak better reply potential, there exists $y \in f(x^*)$ such that $P(y) > P(x^*)$, a contradiction to the definition of x^*. Consequently x^* is in the core.

Proof of Proposition 3.3 We give the proof for the strong better reply and the weak better reply potentials. The other cases are similar and those proofs are hence omitted.

Let P be a strong better reply potential for Γ and let $(x_k)_{k\in K}$ with $K \in \mathcal{K}$ be a better reply path. We need to show that K is finite. Towards a contradiction assume that it is not, so $K = \mathbb{N}$. For all $k \in K^-$, we have $x_{k+1} \in f(x_k) \setminus \{x_k\}$, so as P is a strong better reply potential it follows that $P(x_{k+1}) > P(x_k)$. As X is finite, there must be $k^1, k^2 \in \mathbb{N}$ with $k^1 < k^2$ such that $x_{k^1} = x_{k^2}$, but this contradicts $P(x_{k^1}) < P(x_{k^2})$. Consequently, K is finite.

For the reverse, assume that every better reply path is finite. Define the binary relation $\succ$

on X by $y \succ x$ if and only if $y \in f(x) \setminus \{x\}$. We show that $\succ$ is an acyclic relation.

To obtain a contradiction, suppose that there is a sequence $(x_k)_{k=1}^m$ such that

$$x_1 \succ x_m \succ x_{m-1} \succ \cdots \succ x_2 \succ x_1.$$

But then, the infinite path $(x_k)_{k\in\mathbb{N}}$ with $x_k = x_j$ whenever $k \mod m = j$ is an infinite better reply path, which gives the desired contradiction. Consequently, $\succ$ is an acyclic relation.

As $\succ$ is acyclic, we can use Szpilrajn's theorem (Szpilrajn, 1930) to extend it to a total linear order, say $\succ^*$, i.e., $x \succ y$ implies $x \succ^* y$ and $\succ^*$ is transitive, asymmetric, and total. We define the function $P : X \to \mathbb{R}$ by

$$P(x) = |\{y \in X | x \succ^* y\}|, \quad x \in X,$$

so $P(x)$ corresponds to the number of states domined by x under $\succ^*$. Let us show that P is a strong better reply potential. Let $y \in f(x) \setminus \{x\}$. Then $y \succ x$, so $y \succ^* x$ and, for every $z \in X$, if $x \succ^* z$ then by transitivity of $\succ^*$ also $y \succ^* z$. As such, $P(y) > P(x)$ as we wanted to show.

Let us now prove the second part. Let P be a weak better reply potential for Γ and let $x \in X$. We want to construct a finite better reply path that starts at x and ends at a core element. Let $x_1 = x$. If x_1 is in the core, the path (x_1) has the desired properties. If x_1 does not belong to the core, then there is $x_2 \in f(x_1) \setminus \{x_1\}$ such that $P(x_2) > P(x_1)$. If x_2 is in the core, then (x_1, x_2) is the desired path. Else, there is $x_3 \in f(x_2) \setminus \{x_2\}$ such that $P(x_3) > P(x_2)$. Continue this procedure to obtain a path $(x_k)_{k\in K}$ with $K \in \mathcal{K}$ such that for all $k \in K^-$, $x_{k+1} \in f(x_k) \setminus \{x_k\}$ and $P(x_{k+1}) > P(x_k)$. This path cannot be longer than $|X| - 1$, the number of states minus one, which is finite. The last state on the path must therefore be in the core.

For the reverse, assume that from every state $x \in X$ there is a finite better reply path that starts at x and ends at a core element. For every $x \in X$, define the potential of x by the length of the shortest better reply path from x to a core element of Γ times -1.

Let $x \in X$ be such that $f(x) \setminus \{x\} \neq \emptyset$. Let $(x_1, \ldots, x_m)$ with $x_1 = x$ and x_m a core element be such a shortest better reply path. Since x is not a core element, it holds that $m \geq 2$ and that $x_2 \in f(x_1)$. It holds that $P(x_2) = P(x_1) + 1$. We have shown that there is $y \in f(x)$ such that $P(y) > P(x)$. It follows that P is a weak better potential for Γ.

A.3 Proofs of Section 4

Proof of Proposition 4.3 The first equivalence follows directly from the definitions.

For the second implication, let G be a best reply potential game with potential P. Towards a contradiction, assume Γ^G is not a strong best reply potential social environment. This means that there exists $i \in N$, a strategy profile $x = (x^i, x^{-i}) \in X$ and a strategy profile $z = (z^i, x^{-i}) \in \mathrm{BR}_i(x) \setminus \{x\}$ for which $P(z) \leq P(x)$. From the definition of $\mathrm{BR}_i(x)$ it follows that:

$$z^i \in \arg\max_{y^i \in X^i} u^i(y^i, x^{-i}).$$

and $u^i(z^i, x^{-i}) > u^i(x^i, x^{-i})$. As G is a best reply potential game, we have

$$z^i \in \arg\max_{y^i \in X^i} P(y^i, x^{-i}).$$

In particular, it holds that $P(z) \geq P(x)$. We conclude that $P(z) = P(x)$, so

$$x^i \in \arg\max_{y^i \in X^i} P(y^i, x^{-i}).$$

Again, using the fact that G is a best reply potential game, this implies that $x^i \in \mathrm{argmax}_{y^i \in X^i} u^i(y^i, x^{-i})$, contradicting $u^i(z^i, x^{-i}) > u^i(x^i, x^{-i})$.

We now show the third implication. Let Γ^G be a strong best reply social environment with a regular potential P. Towards a contradiction, suppose that P is not a pseudo-potential for G. Then there is an $i \in N$, $x^{-i} \in X^{-i}$, and $y^i \in \mathrm{argmax}_{x^i \in X^i} P(x^i, x^{-i})$ such that $y^i \notin \mathrm{argmax}_{x^i \in X^i} u^i(x^i, x^{-i})$. Next, since by assumption $\mathrm{BR}_i(x) \neq \emptyset$, we can take $z^i \in \mathrm{argmax}_{x^i \in X^i} u^i(x^i, x^{-i})$. It follows that $(z^i, x^{-i}) \neq (y^i, x^{-i})$. In particular, $(z^i, x^{-i}) \in \mathrm{BR}_i(y^i, x^{-i}) \setminus \{(y^i, x^{-i})\}$. As P is a strong best reply potential for Γ^G we can conclude that $P(z^i, x^{-i}) > P(y^i, x^{-i})$. This contradicts $y^i \in \mathrm{argmax}_{x^i \in X^i} P(x^i, x^{-i})$.

For the last implication, let G be a pseudo-potential game and let P be a pseudo-potential for G. We show that P is a weak best reply potential for the social environment Γ^G. Let $x \in X$ be such that $f(x) \setminus \{x\} \neq \emptyset$. Then there is $i \in N$ such that $\mathrm{BR}_i(x) \setminus \{x\} \neq \emptyset$. Let $Z = \mathrm{argmax}_{y^i \in X^i} P(y^i, x^{-i})$, which is a non-empty set by the regularity of P. Then as G is a pseudo-potential game, $Z \subset \mathrm{argmax}_{y^i \in X^i} u(y^i, x^{-i})$. This implies that $x^i \notin Z$. For $z^i \in Z$, we have that $P(z^i, x^{-i}) > P(x)$ and $(z^i, x^{-i}) \in \mathrm{BR}_{\{i\}}(x) \subset \mathrm{BR}(x)$, which concludes the proof.

Proof of Proposition 4.4 Let Γ^G be a weak best reply potential social environment with potential P. Without loss of generality, we can assume that, for every $x \in X$, $P(x) > 0$.

Let Z be the set of all strategies $x \in X$ such that $x \in \mathrm{BR}_{\{1\}}(x)$ or $x \in \mathrm{BR}_{\{2\}}(x)$. Define the function $\widetilde{P} : X \to \mathbb{R}$ by

$$\widetilde{P}(x) = \begin{cases} P(x) & \text{if } x \in Z \\ 0 & \text{if } x \notin Z. \end{cases}$$

Let us show that $\widetilde{P}$ is a pseudo-potential for the game G. Because G is a finite game, $\widetilde{P}$ is regular. It remains to be shown that, for every $i \in \{1, 2\}$, for every $x^{-i} \in X^{-i}$, $\mathrm{argmax}_{y^i \in X^i} \widetilde{P}(y^i, x^{-i}) \subset \mathrm{argmax}_{y^i \in X^i} u^i(y^i, x^{-i})$. We prove this for $i = 1$. The proof for $i = 2$ is analogous.

Let $x^1 \in \mathrm{argmax}_{y^1 \in X^1} \widetilde{P}(y^1, x^2)$. We have that $Z \cap \{(y^1, x^2) | y^1 \in X^1\}$ is non-empty, so $(x^1, x^2) \in Z$. Towards a contradiction, assume that $x^1 \notin \mathrm{argmax}_{y^1 \in X^1} u^1(y^1, x^2)$. As $(x^1, x^2) \in Z$ it must therefore be that $(x^1, x^2) \in \mathrm{BR}_{\{2\}}(x^1, x^2)$, which implies that $x^2 \in \mathrm{argmax}_{y^2 \in X^2} u^2(x^1, y^2)$. As Γ^G is a weak best reply potential social environment and $\mathrm{BR}(x_1, x_2) \setminus \{(x_1, x_2)\} \neq \emptyset$ we must have that there is $(z^1, z^2) \in \mathrm{BR}(x^1, x^2)$ for which $P(z^1, z^2) > P(x_1, x_2)$. Notice that this also implies that $(z^1, z^2) \in Z$, so $\widetilde{P}(z^1, z^2) = P(z^1, z^2) > P(x^1, x^2) = \widetilde{P}(x^1, x^2)$.

Given that $(x^1, x^2) \in \mathrm{BR}_{\{2\}}(x^1, x^2)$ it follows that $(z^1, z^2) \in \mathrm{BR}_{\{1\}}(x^1, x^2)$ so $z^2 = x^2$ and therefore $\widetilde{P}(z^1, x^2) > \widetilde{P}(x^1, x^2)$. This, however, contradicts the assumption that $x^1 \in \mathrm{argmax}_{y^1 \in X^1} \widetilde{P}(y^1, x^2)$.

A.4 Proofs of Section 5

Proof of Proposition 5.1 Let $x, y \in X$ be such that $y \in f(x) \setminus \{x\}$. Then there is $S \in E(x, y)$ such that, for every $i \in S$, $u^i(y) \geq u^i(x)$, and, for some $j \in S$, $u^j(y) > u^j(x)$. In addition, for every $i \in N \setminus S$, we have $x^i = y^i$. It follows that

$$P(x) = \sum_{i \in N} u^i(x^i) = \sum_{i \in S} u^i(x^i) + \sum_{i \in N \setminus S} u^i(y^i) < \sum_{i \in S} u^i(y^i) + \sum_{i \in N \setminus S} u^i(y^i) = P(y).$$

Hence P is a strong better reply potential.

Proof of Proposition 5.2 Let $x, y \in X$ be such that $y \in f(x) \setminus \{x\}$.

First, consider the case where there is $i \in N$ such that $y = x \cup \{i\}$, so y is a larger

coalition. Since $y \in f(x) \setminus \{x\}$, it holds that $u^i(y) > u^i(x) = 0$, where the equality holds since $i \notin x$. It follows that $\varphi^i(y, v_{|y}) = u^i(y) > 0$. By Hart and Mas-Colell (1989), it holds that $\varphi^i(y, v_{|y}) = P(y) - P(x)$, so $P(y) - P(x) > 0$ as was to be shown.

Second, consider the case where there is $i \in N$ such that $y = x \setminus \{i\}$, so y is a smaller coalition. Since $y \in f(x) \setminus \{x\}$, it holds that $0 = u^i(y) > u^i(x)$, where the equality holds since $i \notin y$. It follows that $\varphi^i(x, v_{|x}) = u^i(x) < 0$. By Hart and Mas-Colell (1989), it holds that $\varphi^i(x, v_{|x}) = P(x) - P(y)$, so $P(y) - P(x) > 0$ as was to be shown.

A.5 Proofs of Section 6

Proof of Proposition 6.2 We give the proof for the sequential weak better reply potential system. The proofs for the other cases are similar and hence omitted.

Let $\mathcal{AP} = ((A_1, P_1), \ldots, (A_\ell, P_\ell))$ be a sequential weak better reply potential system. We construct from any initial state a finite better reply path to a core element. The proof then follows from Proposition 3.3.

Let $x \in A_k \neq \emptyset$. We show that we can reach a state inside the core or a state in A_{k+1} in at most finitely many iterations. If $x \in A_{k+1}$, we are done. If $x \in A_k \setminus A_{k+1}$, either $f(x) = \{x\}$ and we are done or $f(x) \setminus \{x\} \neq \emptyset$ and there is $y \in A_k$ such that $P_k(y) > P_k(x)$. After a finite number of iterations, we find a state z such that $f(z) = \{z\}$, in which case z is in the core, or $f(z) \setminus \{z\} \neq \emptyset$ and $z \in A_{k+1}$. Since the number of sets A_k is finite, we reach a state z' such that $f(z') = \{z'\}$ after at most finitely many iterations. Thus, from every initial state there is a finite better reply path to a core element.

Proof of Proposition 6.4 To establish the proposition, we show that $\mathcal{AP}$ as defined in (1) is a sequential weak individual best reply potential system.

Let $x \in A_1 \setminus A_2$. Then there is a man m who is matched in x but prefers to be single. Either m matches with his most preferred women among all blocking pairs that involve m or m becomes single, in case he prefers this. Let y be the new matching. Notice that $y \in \mathrm{IBR}_m(x) \cap A_2$.

If m forms a blocking pair, the potential increases by $u^m(y) - u^m(x) > 0$. If due to the formation of a blocking pair another man m' becomes single, the potential further increases by $1 - u^{m'}(x) > 0$. If m becomes single, the potential increases by $1 - u^m(x) > 0$.

Let $x \in A_2 \setminus A_3$. Thus, no man who is matched in x prefers to be single and there is a married woman w that does not satisfy the individual rationality condition or can form

a blocking pair. Pick such a woman w. Either w matches with her most preferred man among all blocking pairs that involve w or w becomes single, in case she prefers this. Let y be the new matching. Notice that $y \in \text{IBR}_w(x)$.

If w decides to become single and was previously married, the potential increases by $1 - u^w(x) > 0$. Instead, if w forms a blocking pair with m, then the potential increases by $u^w(y) - u^w(x) > 0$. If due to the formation of the blocking pair (m, w) some other women w' becomes single, the potential further increases by $1 - u^{w'}(x) > 0$.

We need to show that $y \in A_2$, so it is individually rational for all men. The only potential new match that was formed was (m, w). As m weakly prefers his situation in x over being single, and as he prefers to be married to w compared to his situation in x, he prefers to be married in y, i.e., the matching y is individually rational for men.

Let $x \in A_3 \setminus A_4 = A_3$ and assume that x is not stable. This means that x satisfies individual rationality and has no blocking pairs that involve a married woman. As x is not stable, there should be a blocking pair (m, w) for which w is single.

Let m' be the most preferred man for w among all blocking pairs that involve w. Let

$$y = x + (m', w).$$

Notice that $y \in \text{IBR}_w(x)$. As w is single in x, the only man whose utility changes is m'. His utility $u^{m'}(y) - u^{m'}(x) > 0$ increases, thereby increasing the potential.

We need to show that $y \in A_3$. First, as the matching x satisfies individual rationality and only one new blocking pair is formed, individual rationality is also satisfied in y. Next, we need to show that there is no blocking pair in y involving a married women. Suppose there is one and recall that $u^m(y) \geq u^m(x)$ for all men m.

If the blocking pair is $(\tilde{m}, \tilde{w})$, where $\tilde{w} \neq w$ is married in x, then $(\tilde{m}, \tilde{w})$ is also a blocking pair in x, contradicting the assumption that $x \in A_3$.

If the blocking pair involves w, say $(\tilde{m}, w)$, man m' cannot have been the most preferred partner for w among all blocking pairs, a contradiction.

Proof of Proposition 6.6 We give the proof for the weak better reply potential. The other cases are similar and hence omitted. Let $\mathcal{P}$ be a lexicographic weak better reply potential system for Γ. We construct, from any initial state, a finite better reply path to a core element. The proof then follows from Proposition 3.3.

Let $x \in X$. If x is in the core, there is nothing to prove. If not, then there is a $y \in f(x) \setminus \{x\}$ such that $y \succ_{\mathcal{P}} x$. If y is in the core, we are done. If not, then there is $z \in f(y) \setminus \{y\}$ such

that $z \succ_{\mathcal{P}} y$. Iterating this argument, as $\succ_{\mathcal{P}}$ is acyclic and X is finite, we arrive at a core element in a finite number of steps.

Proof of Proposition 6.8 We show that $\mathcal{P}$ as defined in (2) is a lexicographic weak individual best reply potential system for Γ^{vt}. The proof consists of three steps. The first step shows that if a state is dominated by another state and there is a proposal which was not decided by a minimal majority for which the outcome changes, then we can also find a domination with the same set of accepted proposals as in the earlier domination and a higher number of proposals which are decided by a minimal majority. The second step demonstrates that if a state is dominated by another state such that the outcome only changes for proposals which are decided by a minimal majority, then for every voter in the deviating coalition there is a domination with the same set of accepted proposals as in the earlier domination, the same set of proposals which are decided by a minimal majority as in the original state, and such that the score of the given voter increases. The third step establishes the statement of the proposition.

Step 1. If there is $S \in \mathcal{N}$ and $x' \in f_S(x)$ such that $(A(x)\Delta A(x')) \setminus M(x) \neq \emptyset$, then there is $y \in f(x)$ such that $A(y) = A(x')$ and $P_1(y) > P_1(x)$.
Let $S \in \mathcal{N}$, $x \in X$, and $x' \in f_S(x)$ be such that $(A(x)\Delta A(x')) \setminus M(x) \neq \emptyset$. Since $(A(x)\Delta A(x')) \setminus M(x) \neq \emptyset$, there is at least one proposal in $A(x)\Delta A(x')$ which involves a non-minimal majority at x. We show that there is $y \in f_S(x)$ such that $A(y) = A(x')$ and $M(x) \subsetneq M(y)$, i.e., coalition S has a deviation that results in the same set of accepted proposals as x' and increases the number of proposals which are decided by a minimal majority.

For every proposal $r \in R \setminus (A(x)\Delta A(x'))$, the outcome does not change, so we define $y_r = x_r$, thereby guaranteeing that if $r \in M(x)$, then $r \in M(y)$, and if $r \in A(x')$, then $r \in A(y)$.

Next consider the case where $A(x) \setminus A(x') \neq \emptyset$ and $r \in A(x) \setminus A(x')$. It holds that $v_r^+(x) > v_r^-(x)$ and $v_r^+(x') < v_r^-(x')$. For every $i \in N \setminus S$, we define $y_r^i = x_r^i = x_r'^i$. For every $i \in S$, we take $y_r^i \in \mathbb{N}_0$ to be such that $\sum_{j \in N_r^+} y_r^j = \sum_{j \in N_r^-} y_r^j - 1$. This can always be achieved by transferring less votes from voters in $S \cap N_r^+$ to voters in $S \cap N_r^-$ if necessary. We have that $r \notin A(y)$ and $r \in M(y)$.

Finally consider the case where $A(x') \setminus A(x) \neq \emptyset$ and $r \in A(x') \setminus A(x)$. It holds that $v_r^+(x) < v_r^-(x)$ and $v_r^+(x') > v_r^-(x')$. For every $i \in N \setminus S$, we define $y_r^i = x_r^i = x_r'^i$. For every $i \in S$, we take $y_r^i \in \mathbb{N}_0$ to be such that $\sum_{j \in N_r^+} y_r^j = \sum_{j \in N_r^-} y_r^j + 1$. This can always be

achieved by transferring less votes from voters in $S \cap N_r^-$ to voters in $S \cap N_r^+$ if necessary. We have that $r \in A(y)$ and $r \in M(y)$.

Our construction ensures that $A(y) = A(x')$.

It holds that $S \in E(x, y)$ and, for every $i \in S$, $u^i(y) = u^i(x') > u^i(x)$, so $y \in f_S(x)$. Moreover, it holds that $M(x) \subset M(y)$ and $\emptyset \neq (A(x)\Delta A(x')) \setminus M(x) \subset M(y)$, so $M(x) \subsetneq M(y)$ and $P_1(y) > P_1(x)$.

Step 2. If there is $S \in \mathcal{N}$ and $x' \in f_S(x) \setminus \{x\}$ such that $(A(x)\Delta A(x')) \subset M(x)$, then, for every $i \in S$, there is $y \in f_S(x)$ such that $A(y) = A(x')$, $M(y) = M(x)$, and $P_{i+1}(y) > P_{i+1}(x)$.
Let $S \in \mathcal{N}$, $x \in X$, and $x' \in f_S(x) \setminus \{x\}$ be such that $(A(x)\Delta A(x')) \subset M(x)$, so the proposals for which the outcome changes are decided by a minimal majority in x. For every $i \in S$, let $R^{\mathrm{b}}(i)$ be the set of proposals where i benefits from changing state x to state x' and $R^{\ell}(i)$ be the set of proposals where i loses from changing state x to state x', so

$$\begin{aligned} R^{\mathrm{b}}(i) &= \{r \in R \mid [z_r^i > 0 \text{ and } r \in A(x') \setminus A(x)] \text{ or } [z_r^i < 0 \text{ and } r \in A(x) \setminus A(x')]\}, \\ R^{\ell}(i) &= \{r \in R \mid [z_r^i > 0 \text{ and } r \in A(x) \setminus A(x')] \text{ or } [z_r^i < 0 \text{ and } r \in A(x') \setminus A(x)]\}. \end{aligned}$$

For every $i \in S$, since $u^i(x') > u^i(x)$, it holds that $R^{\mathrm{b}}(i) \neq \emptyset$. Moreover, for every $r \in A(x)\Delta A(x')$, there is $i \in S$ such that $r \in R^{\mathrm{b}}(i)$ and there is $i \in S$ such that $r \in R^{\ell}(i)$ and $x_r^i > 0$ as otherwise it would not be possible that the outcome on proposal r changes when going from state x to state x'.

Let $i \in S$. We now construct the state y with the desired properties as stated in Step 2. For proposals r outside $A(x)\Delta A(x')$, we define $y_r = x_r$. For voters j outside S, we define $y^j = x^j = x'^j$. For every proposal $r \in R^{\mathrm{b}}(i)$, let $y_r^i = x_r^i + 1$, choose a voter $j \in S$ such that $r \in R^{\ell}(j)$ and $x_r^j > 0$, define $y_r^j = x_r^j - 1$, and keep the votes on proposal r unchanged for other coalition members. For every proposal $r \in R^{\ell}(i)$ such that $x_r^i > 0$, define $y_r^i = x_r^i - 1$, choose a voter $j \in S$ such that $r \in R^{\mathrm{b}}(j)$, define $y_r^j = x_r^j + 1$, and keep the number of votes on proposal r unchanged for the other coalition members. For every proposal $r \in R^{\ell}(i)$ such that $x_r^i = 0$, define $y_r^i = x_r^i$ and transfer one vote from a voter $i^1 \in S$ such that $r \in R^{\ell}(i^1)$ and $x_r^{i^1} > 0$ to a voter $i^2 \in S$ such that $r \in R^{\mathrm{b}}(i^2)$ and keep the number of votes on proposal r unchanged for the other coalition members. Since $(A(x)\Delta A(x')) \subset M(x)$, an appropriate transfer of a single vote is sufficient to guarantee that $A(y) = A(x')$. It holds

that

$$\begin{aligned} s^i(y) - s^i(x) &= \sum_{r \in R^{\flat}(i)} (y^i_r - x^i_r)|z^i| + \sum_{r \in R^{\ell}(i)} (y^i_r - x^i_r)|z^i| \\ &\geq \sum_{r \in R^{\flat}(i)} |z^i| - \sum_{r \in R^{\ell}(i)} |z^i| \\ &= u^i(y) - u^i(x) > 0. \end{aligned}$$

STEP 3. $\mathcal{P}$ is a lexicographic weak individual best reply potential system.
Let $x \in X$. If $f(x) = \{x\}$, then we are done, so consider the case where $f(x) \setminus \{x\} \neq \emptyset$.

Let $i \in N$ be the voter with the lowest index for which there is $S \in \mathcal{N}$ with $i \in S$ such that $f_S(x) \setminus \{x\} \neq \emptyset$. Since the state space is finite, it holds that $\mathrm{IBR}_i(x) \neq \emptyset$. Let $x' \in \mathrm{IBR}_i(x)$ and $S \in \mathcal{N}$ be such that $x' \in f_S(x)$. From the choice of i we have that $x' \neq x$. If $(A(x)\Delta A(x')) \setminus M(x) \neq \emptyset$, then we use the construction in Step 1 to find $y \in f_S(x)$ such that $A(y) = A(x')$ and $P_1(y) > P_1(x)$. This implies $y \in \mathrm{IBR}_i(x)$ and $y \succ_{\mathcal{P}} x$. Otherwise, it holds that $(A(x)\Delta A(x')) \subset M(x)$. Following the construction in Step 2, we find $y \in f_S(x)$ such that $A(y) = A(x')$, $M(y) = M(x)$, and $s^i(y) > s^i(x)$. For every $j \in \{1, \ldots, i-1\}$, it holds that $j \notin S$ and $y^j = x^j$, so $s^j(y) = s^j(x)$. It follows that $y \in \mathrm{IBR}_i(x)$ and $y \succ_{\mathcal{P}} x$.

A.6 Proofs of Section 7

Proof of Proposition 7.1 We first show that P_1 is a weak individual best reply potential. Let $x \in X$ be such that $f(x) \setminus \{x\} \neq \emptyset$. Then there is $S \in \mathcal{N}$ such that $f_S(x) \setminus \{x\} \neq \emptyset$. Let $i \in S$ be such that $x^i > 0$. Such an individual exists, since otherwise coalition S cannot pillage another coalition. It holds that $y \in \mathrm{IBR}_i(x)$ if and only if there is $T^* \subset N \setminus \{i\}$ which attains the maximum value of $x(T)$ among all coalitions $T \subset N \setminus \{i\}$ such that $x(T^*) < x(N \setminus T^*)$, $y^i = x^i + x(T^*)$, for every $j \in T^*$, $y^j = 0$, and for every $j \in N \setminus (T^* \cup \{i\})$, $y^j = x^j$. Notice that the finiteness of the collection of subsets of $N \setminus \{i\}$ ensures the existence of T^* and the non-emptiness of $\mathrm{IBR}_i(x)$. Since $f_S(x) \setminus \{x\} \neq \emptyset$ and $i \in S$, there is $j \in T^*$ such that $x^j > 0$, as otherwise i would not improve from the pillage. It holds that $y^j = 0$, so we conclude that $P_1(y) = |Z(y)| > |Z(x)| = P_1(x)$. We have shown that P_1 is a weak individual best reply potential.

We show next that for $n = 2$, the function P_2 is a strong better reply potential. Let $x \in X$ and $y \in f(x) \setminus \{x\}$, so $x \neq (1/2, 1/2)$. Without loss of generality, we can assume that $x^1 > x^2$. It holds that $y^1 > x^1 > x^2 > y^2$, so

$$P_2(y) = |y^1 - y^2| > |x^1 - x^2| = P_2(x).$$

We have shown that P_2 is a strong better reply potential.

Finally, let $n \geq 3$. We construct an individual best reply path which cycles, at odds with the existence of a strong individual best reply potential by Proposition 3.3. Let $x_1 \in X$ be such that $x_1^1 = 2/3$, $x_1^2 = 1/3$, and, for every $i \in N \setminus \{1,2\}$, $x_1^i = 0$. The unique individual best reply for individual 3 is to form a coalition with individual 1 and pillage individual 2, resulting in the state $x_2 \in X$ where $x_2^1 = 2/3$, $x_2^3 = 1/3$, and, for every $i \in N \setminus \{1,3\}$, $x_2^i = 0$. Next, the unique individual best reply for individual 2 is to form a coalition with individual 1 and pillage individual 3, resulting in the state $x_3 = x_1$.

Proof of Proposition 7.2 We first show that P_1 is a weak individual best reply potential for $\Gamma^{\mathcal{PP}}$. Let $x \in X$ be such that $f(x) \setminus \{x\} \neq \emptyset$. Then there is $S \in \mathcal{N}$ such that $f_S(x) \setminus \{x\} \neq \emptyset$. Let $i \in S$ be such that $x^i > 0$. Such an individual exists, since otherwise coalition S cannot pillage another coalition. The neighborhood of a coalition $T \in \mathcal{N}$ in $C \in \mathcal{C}$ is denoted by $N^T(C) = \cup_{i \in T} C^i$. It holds that $y \in \mathrm{IBR}_i(x)$ if and only if there is $T^* \in \mathcal{N}$ which attains the maximum value of $\omega(x)(T)$ among all coalitions $T \in \mathcal{N}$ such that $i \notin N^T(C(x))$ and $x(T) < x(N \setminus N^T(C(x)))$ and, moreover, $y^i = x^i + x(T^*)$, for every $j \in T^*$, $y^j = 0$, and for every $j \in N \setminus (T^* \cup \{i\})$, $y^j = x^j$. The finiteness of $\mathcal{N}$ together with the fact that $f_S(x) \setminus \{x\} \neq \emptyset$ and $i \in S$ ensures the existence of a non-empty T^* and the non-emptiness of $\mathrm{IBR}_i(x)$. Since $f_S(x) \setminus \{x\} \neq \emptyset$, there is $j \in T^*$ such that $x^j > 0$. We use that $y^i > x^i > 0$ and $y^j = 0 < x^j$ to derive that $Z(\omega(x)) \subsetneq Z(\omega(y))$. We conclude that $P_1(y) = |Z(\omega(y))| > |Z(\omega(x))| = P_1(x)$, so P_1 is a weak individual best reply potential for $\Gamma^{\mathcal{PP}}$.

We show next that for $n = 2$, the function P_2 is a strong better reply potential for $\Gamma^{\mathcal{PP}}$. Let $x \in X$ and $y \in f(x) \setminus \{x\}$, so $\omega(x) \neq (1/2, 1/2)$, $C^1(x) = \{1\}$, and $C^2(x) = \{2\}$. Without loss of generality, we can assume that $\omega^1(x) > \omega^2(x)$. It holds that $\omega^1(y) > \omega^1(x) > \omega^2(x) > \omega^2(y)$, so

$$P_2(y) = |\omega^1(y) - \omega^2(y)| > |\omega^1(x) - \omega^2(x)| = P_2(x),$$

so P_2 is a strong better reply potential for $\Gamma^{\mathcal{PP}}$.

Let $n = 3$. We show that P_3 is a strong coalitional best reply potential for $\Gamma^{\mathcal{PP}}$.

Let $x \in X$ be such that $C^1(x) = \{1\}$, $C^2(x) = \{2\}$, and $C^3(x) = \{3\}$. Consider a coalition $S \in \mathcal{N}$ consisting of a single player. Let $y \in \mathrm{BR}_S(x) \setminus \{x\}$. The individual in S must have positive wealth, whereas the wealth of any pillaged individual becomes zero, so

$|Z(\omega(y))| \geq |Z(\omega(x))| + 1$. Since $C^1(y) = \{1\}$, $C^2(y) = \{2\}$, and $C^3(y) = \{3\}$, we have

$$\sum_{j \in N} \sum_{i \in C^j(y)} \omega^i(y) = \sum_{i \in N} \omega^i(y) = \sum_{i \in N} \omega^i(x) = \sum_{j \in N} \sum_{i \in C^j(x)} \omega^i(x),$$

so $P_3(y) > P_3(x)$. Consider a coalition $S \in \mathcal{N}$ with at least two players. Let $y \in \mathrm{BR}_S(x) \setminus \{x\}$. We use $n = 3$ to conclude that coalition S must have exactly two members. At most one member of S can have zero wealth, as otherwise coalition S is not able to pillage another coalition. The wealth of the pillaged individual goes from positive to zero, so $|Z(\omega(y))| \geq |Z(\omega(x))|$. The total wealth of coalition S increases, so

$$\begin{array}{rcl} \sum_{j \in N} \sum_{i \in C^j(y)} \omega^i(y) & = & 2 \sum_{i \in S} \omega^i(y) + \sum_{i \in N \setminus S} \omega^i(y) = 2 \sum_{i \in S} \omega^i(y) = 2 \sum_{i \in N} \omega^i(x) \\ & > & \sum_{i \in N} \omega^i(x) = \sum_{j \in N} \sum_{i \in C^j(x)} \omega^i(x). \end{array}$$

It follows that $P_3(y) > P_3(x)$.

Let $x \in X$ be such that, for some $i \in N$, $|C^i(x)| \geq 2$. Let $y \in \mathrm{BR}(x) \setminus \{x\}$. It follows that two of the coalitions in $C(x)$ contain two elements and one coalition in $C(x)$ is a singleton. Without loss of generality assume that $C^1(x) = C^2(x) = \{1, 2\}$ and $C^3(x) = \{3\}$. It now holds that, for every $i \in N$, $C^i(y) = C^i(x)$. Let $S \in \mathcal{N}$ be such that $y \in \mathrm{BR}_S(x)$. It holds that S either consists of a single player or $S = \{1, 2\}$. If $S = \{1\}$ or $S = \{2\}$, then individual 3 is pillaged, $|Z(\omega(y))| = |Z(\omega(x))| + 1$ and

$$\begin{array}{rcl} \sum_{j \in N} \sum_{i \in C^j(y)} \omega^i(y) & = & 2(\omega^1(y) + \omega^2(y)) + \omega^3(y) > 2(\omega^1(y) + \omega^2(y)) = 2 \sum_{i \in N} \omega^i(x) \\ & > & 2(\omega^1(x) + \omega^2(x)) + \omega^3(x) = \sum_{j \in N} \sum_{i \in C^j(x)} \omega^i(x), \end{array}$$

so $P_3(y) > P_3(x)$. If $S = \{3\}$, then $|Z(\omega(y))| \geq |Z(\omega(x))| + 1$, whereas $\omega^1(x) + \omega^2(x) < 1$, so

$$\begin{array}{rcl} P_3(y) & = & |Z(\omega(y))| + 2(\omega^1(y) + \omega^2(y)) + \omega^3(y) \\ & \geq & |Z(\omega(x))| + 1 + 2(\omega^1(y) + \omega^2(y)) + \omega^3(y) \\ & \geq & |Z(\omega(x))| + 1 + \sum_{i \in N} \omega^i(x) \\ & > & |Z(\omega(x))| + 2(\omega^1(x) + \omega^2(x)) + \omega^3(x) \\ & = & P_3(x). \end{array}$$

Finally, if $S = \{1, 2\}$, then $\omega^3(y) = 0$, so $|Z(\omega(y))| \geq |Z(\omega(x))|$, and

$$\begin{array}{rcl} \sum_{j \in N} \sum_{i \in C^j(y)} \omega^i(y) & = & 2 \sum_{i \in S} \omega^i(y) + \sum_{i \in N \setminus S} \omega^i(y) = 2 \sum_{i \in S} \omega^i(y) = 2 \sum_{i \in N} \omega^i(x) \\ & > & \sum_{i \in N} \omega^i(x) = \sum_{j \in N} \sum_{i \in C^j(x)} \omega^i(x), \end{array}$$

and therefore $P_3(y) > P_3(x)$.

We now construct a better reply path with cycles, which rules out the existence of a strong better reply potential by Proposition 3.3. Let $x_1 \in X$ be such that $C^1(x_1) = C^2(x_1) = \{1,2\}$, $C^3(x_1) = \{3\}$, and $\omega(x_1) = (4/7, 1/7, 2/7)$. It holds that x_2 defined by $C^1(x_2) = C^2(x_2) = \{1,2\}$, $C^3(x_2) = \{3\}$, and $\omega(x_2) = (4/7, 0, 3/7)$ belongs to $f_{\{3\}}(x_1)$. In fact, x_2 is an individual best reply for individual 3. Since $x_3 = x_1 \in f_{\{1,2\}}(x_2)$, we have found the desired better reply path with cycles.

Let $n \geq 4$. We construct an individual best reply path with cycles, which rules out the existence of a strong individual best reply potential by Proposition 3.3. Let x_1 be such that $C^1(x_1) = C^2(x_1) = \{1,2\}$, $C^3(x_1) = C^4(x_1) = \{3,4\}$, for every $i \in N \setminus \{1,2,3,4\}$, $C^i(x_1) = \{i\}$, $\omega^1(x_1) = 2/5$, $\omega^2(x_1) = 1/5$, and $\omega^3(x_1) = 2/5$. Since $\omega(x) \in \Omega$, the wealth levels of all other individuals are equal to zero. The only possibility for individual 4 is to form a coalition with individual 3 and pillage individual 2. From the perspective of individual 4, the best option is to transfer all the wealth of individual 2 to individual 4, resulting in x_2 such that $C(x_2) = C(x_1)$, $\omega^1(x_2) = 2/5$, $\omega^3(x_2) = 2/5$, and $\omega^4(x_2) = 1/5$. By an analogous argument, we find that $x_3 = x_1 \in \mathrm{IBR}_2(x_2)$ which yields the individual best reply path with cycles.

Proof of Proposition 7.3 For $n = 2$, apart from the notation for gang membership, there is no difference between the social environments $\Gamma^{\mathcal{PP}}$ and $\Gamma^{\mathcal{PG}}$. The result therefore follows from Proposition 7.2.

Let $n \geq 3$. We show that P_2 is a strong coalitional best reply potential. For every $x \in X$, it holds that

$$\tfrac{1}{n} \leq \sum_{C \in \pi(x)} (\omega(x)(C))^2 \leq 1,$$

where the minimum value $1/n$ is attained for the partition of the players into singletons and a uniform wealth distribution and the maximum value 1 is obtained when a single coalition has wealth 1. Let $x, y \in X$ be such that $y \in \mathrm{BR}(x) \setminus \{x\}$. From the definition of the effectivity correspondence, it follows that $|\pi(y)| \leq |\pi(x)|$.

Consider the case where $|\pi(y)| < |\pi(x)|$ which implies $|\pi(y)| \leq |\pi(x)| - 1$. We have that

$$\begin{aligned} P(x) &= -|\pi(x)| + \textstyle\sum_{C \in \pi(x)} (\omega(x)(C))^2 \\ &\leq -|\pi(x)| + 1 \\ &< -|\pi(y)| + \tfrac{1}{n} \\ &\leq -|\pi(y)| + \textstyle\sum_{C \in \pi(y)} (\omega(y)(C))^2 \\ &= P(y). \end{aligned}$$

Next consider the case where $|\pi(y)| = |\pi(x)|$. Let $S \in \mathcal{N}$ be such that $y \in \mathrm{BR}_S(x)$. Let $T \in \mathcal{N}$ be the pillaged coalition. Since $\omega(x)(S) > \omega(x)(G(x,T))$, we can assume that $T = G(x,T)$. Let $C \in \pi(x)$ be such that $S \subset C$. Let $D_1, \ldots, D_\ell \in \pi(x)$ be such that $G(x,T) = \cup_{k=1}^{\ell} D_k$. We define $a = \omega(x)(C) > 0$ and, for $k = 1, \ldots, \ell$, $b_k = \omega(x)(D_k)$. It follows that $\omega(y)(C) = a + \sum_{k=1}^{\ell} b_k$ and, for every $k = 1, \ldots, \ell$, $\omega(y)(D_k) = 0$. Since $y \in \mathrm{BR}(x) \setminus \{x\}$, there is $k \in \{1, \ldots, \ell\}$ such that $b_k > 0$. We have that

$$\begin{array}{rcl} P(y) - P(x) & = & -|\pi(y)| + \sum_{C \in \pi(y)} (\omega(y)(C))^2 + |\pi(x)| - \sum_{C \in \pi(x)} (\omega(x)(C))^2 \\ & = & (a + \sum_{k=1}^{\ell} b_k)^2 - a^2 - \sum_{k=1}^{\ell} b_k^2 \ > 0. \end{array}$$

We finally show that when $n \geq 3$, a strong better reply potential does not exist.

Suppose $P : X \to \mathbb{R}$ is a strong better reply potential. Fix an arbitrary partition $\pi \in \Pi$ for which $\{1\}, \{2\}, \{3\} \in \pi$. We define

$$\overline{X} = \{x \in X \mid \pi(x) = \pi \text{ and, for every } i \in N \setminus \{1,2,3\},\ \omega^i(x) = 0\}.$$

Let $x \in \overline{X}$ be such that $\omega^1(x) > 1/2$ and $\omega^2(x) > \omega^3(x)$. For every $y \in \overline{X}$ such that $\omega^1(y) = \omega^1(x)$ and $\omega^2(y) > \omega^2(x)$, it holds that $y \in f(x) \setminus \{x\}$, so $P(y) > P(x)$, as coalition $\{2\}$ can pillage coalition $\{3\}$. For every $y \in \overline{X}$ such that $\omega^1(y) > \omega^1(x)$, it holds that $y \in f(x) \setminus \{x\}$, so $P(y) > P(x)$, as coalition $\{1\}$ can pillage coalition $\{2,3\}$.

For every $a \in (1/2, 1)$, we define

$$Q(a) = \{P(x) \in \mathbb{R} \mid x \in \overline{X},\ \omega^1(x) = a,\ \omega^2(x) > \omega^3(x)\}.$$

It holds that $Q(a)$ has infinitely many elements, so there is a rational number $q(a) \in \mathbb{Q}$ for which there is $\underline{p} \in Q(a)$ with $\underline{p} \leq q(a)$ and there is $\overline{p} \in Q(a)$ with $\overline{p} \geq q(a)$. If $a_1 < a_2$, then $q(a_1) < q(a_2)$, so the set

$$\widetilde{Q} = \{q(a) \in Q \mid 1/2 < a < 1\}$$

contains uncountably many rational numbers, a contradiction. Consequently, $\Gamma^{\mathcal{PG}}$ has no strong better reply potential.

B Markov Chains

Instead of deterministic paths, we can also use a stochastic framework to enable transfers of results from literatures like evolutionary game theory. Let X be a finite set of states. A Markov chain M on X associates with every pair of states $(x, y) \in X \times X$ a non-negative number $M(x, y)$ such that, for all $x \in X$, we have $\sum_{y \in X} M(x, y) = 1$, where $M(x, y)$ gives the probability of going from state x to state y in one step.

Definition B.1. Let $\Gamma = (N, (X, d), E, u)$ be a social environment with a finite state space. A Markov chain M on X is a *better reply Markov chain* for Γ if for all $x \in X$ the following two statements hold:

- If $f(x) = \{x\}$ then $M(x, x) = 1$.
- If $f(x) \setminus \{x\} \neq \emptyset$ then for all $y \in f(x) \setminus \{x\}$, $M(x, y) > 0$ and $\sum_{y \in f(x) \setminus \{x\}} M(x, y) = 1$.

To define a *coalitional best reply Markov chain* (*individual best reply Markov chain*), we replace $f(x)$ by $\mathrm{BR}(x)$ ($\mathrm{IBR}(x)$) in the definition of a better reply Markov chain.

A better reply Markov chain which starts at a state x induces a better reply path as defined in Definition 3.2. As such, if the social environment has a strong better reply potential, any strong better reply path is finite and it takes at most $|X| - 1$ steps to reach a core element. If x is in the core then it will stay there forever as $M(x, x) = 1$. A similar observation can be made for the best reply potentials.

For a weak potential, there is a positive probability to reach a core element in a finite number of steps from every state. As such, the process gets absorbed at a core element with probability one. We formalize the previous statements in the following proposition.

Proposition B.2. *Let $\Gamma = (N, (X, d), E, u)$ be a social environment with a finite state space.*

- *If M is a better reply (coalitional best reply) [individual best reply] Markov chain for Γ and if Γ has a strong better reply (coalitional best reply) [individual best reply] potential, then any path of the Markov chain will reach a core element in at most $|X| - 1$ steps.*
- *If M is a better reply Markov chain and if Γ admits a potential, then the Markov chain will reach a core element with probability 1.*

Proof. The first statement follows from Proposition 3.3 and the second one follows as there is a finite path from every state to an (absorbing) state in the core. □

References

Ackermann, H., P.W. Goldberg, V.S. Mirrokni, H. Röglin, B. Vöcking (2011), "Uncoordinated Two-Sided Matching Markets" *SIAM Journal on Computing, 40,* 92–106.

Aumann, R. J. (1959): "Acceptable Points in General Cooperative n-Person Games," in A.W. Tucker and R.D. Luce (eds.), *Contributions to the Theory of Games, Volume IV,* Princeton University Press, Princeton, pp. 287–324.

Casella, A., and T. Palfrey (2019), "Trading Votes for Votes: A Dynamic Theory," *Econometrica, 87,* 631–651.

Chien, S., and A. Sinclair (2011), "Convergence to Approximate Nash Equilibria in Congestion Games," *Games and Economic Behavior, 71,* 315-327.

Chwe, M. S.-Y. (1994), "Farsighted Coalitional Stability," *Journal of Economic Theory, 53,* 299–325.

Demuynck, T., Herings, J.-J., Saulle, R. and C. Seel (2019), "The Myopic Stable Set in Social Environments," *Econometrica, 87,* 111–138.

Dubey, P., O. Haimanko, and A. Zapechelnyuk (2006), "Strategic Complements and Substitutes, and Potential Games," *Games and Economic Behavior, 54,* 77–94.

Friedman, J. W., and C. Mezzetti (2001): "Learning in Games by Random Sampling," *Journal of Economic Theory, 98,* 55–84.

Gale, D., and L.S. Shapley (1962), "College Admissions and the Stability of Marriage," *American Mathematical Monthly, 69,* 9–15.

Hart, S., and A. Mas-Colell (1989), "Potential, Value, and Consistency," *Econometrica, 57,* 589–614.

Jensen, M.K. (2010), "Aggregative Games and Best Reply Potentials," *Economic Theory, 43,* 45-66.

Jordan, J.S. (2006), "Pillage and Property," *Journal of Economic Theory, 131,* 26–44.

Knuth, D.E. (1976), *Marriages Stables,* Les Presses de l'Université de Montreal, Montreal.

Monderer, D., and L.S. Shapley (1996), "Potential Games," *Games and Economic Behavior, 14,* 124–143.

Osborne, M.J., and A. Rubinstein (1994), *A Course in Game Theory,* MIT Press, Cambridge, Massachusetts.

Rosenthal, R. W. (1973): "A Class of Games Possessing Pure-Strategy Nash Equilibria," *International Journal of Game Theory, 2,* 65–67.

Roth, A. E., and J. H. Vande Vate (1990): "Random Paths to Stability in Two-sided Matching," *Econometrica, 58,* 1475–1480.

Selten, R. (1970): "Preispolitik der Mehrproduktunternehmen in der statistischen Theorie," Springer-Verlag.

Szabó, G., and G. Fáth (2007): "Evolutionary Games on Graphs," *Physics Reports, 446,* 97–216.

Szpilrajn E. (1930) "Sur l'Extension de l'Ordre Partiel," *Fundamenta Mathematicae, 16,* 386–389.

Tamura A. (1993) "Transformation from Arbitrary Matching to Stable Matchings," *Journal of Combinatorial Theory, Series A,* 310–323.

Voorneveld, M. (2000), "Best-response Potential Games," *Economics Letters, 66,* 289–295.

Yamamoto, K. (2015), "A Comprehensive Survey of Potential Game Approaches to Wireless Networks," *IEICE Transactions on Communications, 98,* 1804-1823.

Young, P. (1993), "The Evolution of Conventions," *Econometrica, 61,* 57–84.