Conventions and Coalitions in Repeated Games[*]

S. Nageeb Ali[†] Ce Liu[‡]

January 31, 2020

Abstract

We develop a theory of repeated interaction for coalitional behavior. We consider stage games where both individuals and coalitions may deviate. However, coalition members cannot commit to long-run behavior, and anticipate that today's actions influence tomorrow's behavior. We evaluate the degree to which history-dependence can deter coalitional deviations. If monitoring is perfect, every feasible and strictly individually rational payoff can be supported by history-dependent conventions. By contrast, if players can make secret side-payments to each other, every coalition achieves a coalitional minmax value, potentially reducing the set of supportable payoffs to the core of the stage game.

[*] We thank (in random order) Daniel Barron, Alex Wolitzky, Debraj Ray, Scott Kominers, Elliot Lipnowski, Joel Watson, Ben Golub, Vijay Krishna, Mike Powell, George Mailath, Federico Echenique, Ron Siegel, Drew Fudenberg, Hideo Konishi, Laura Doval, Matt Elliott, Maciej Kotowski, and various audiences for helpful comments.

[†] Department of Economics, Pennsylvania State University. Email: nageeb@psu.edu.

[‡] Department of Economics, Michigan State University. Email: celiu@msu.edu.

Contents

1 Introduction

The theory of repeated games is central to our study of dynamic incentives. It models strategic players who follow conventions that are history-dependent and self-enforcing. These conventions reflect a shared understanding of how past and current choices influence future behavior, and given that understanding, each individual is deterred from deviating. The main approach for studying repeated interaction is non-cooperative, relying on individual optimization, and without the possibility for joint deviations.

But in a number of settings, the natural units of analysis are not just individuals but also coalitions. For example, matching theory studies matches where no set of players gains from jointly deviating ("stable matches"). Analyses of political economy focus on outcomes that are not overruled by decisive coalitions of voters ("Condorcet winners"). The study of networks focuses on graphs where no pair of individuals wishes to jointly deviate in their selection of neighbors ("stable networks"). In principle, in these settings, one could have modeled coalitional behavior using a non-cooperative extensive-form game in which players can make offers to others, choose to accept or reject those offers, etc. But this requires a complete specification of the set of feasible actions as well as the order, timing, and observability of moves. There is thus a convenience of taking a "cooperative approach" that abstracts from how coalitions form and instead focuses on the payoffs of coalitional moves.

Our objective is to tractably combine this cooperative approach with a repeated-games understanding of dynamic incentives. When cooperative environments are repeated, what is the appropriate notion of stability? To what degree and when does the power of a shared understanding influence the incentives and stability of coalitions? What kinds of carrots and sticks are themselves immune to coalitional deviations? These questions motivate this paper.[1]

We study self-enforcing conventions of behavior when both individuals and coalitions may deviate in the repeated play of an abstract stage game. Special cases of this stage game are strategic-form games (in which players choose actions) and partitional games (in which players partition into groups). Payoffs accrue to players based on outcomes of the stage game, and players share a common discount factor. We use the language of *effectivity correspondences* to describe coalitional moves. We consider both non-transferable and transferable utility environments.

In the spirit of repeated games, we adhere to the principle that individuals and coalitions cannot commit by external means to their long-run behavior, neither on the path of play nor in their deviations. But the stage game is cooperative: coalitions may act together within a single period. Our goal is to study behavior that is self-enforcing through the power of expectations and a shared understanding of the future, just as in the standard theory of repeated games, despite

[1] Answering these questions is germane not only for repeated *cooperative* games but also for studying coalitional deviations in repeated *non-cooperative* games. In practice, players may find ways to communicate, coordinate, and collude so that groups of them jointly deviate, and just as in cooperative game theory, it may be useful to study when such joint deviations are profitable without fully specifying how these joint deviations are coordinated.

the prospect of these coalitional deviations.

Because there is no “off-the-shelf” solution-concept, we develop one that is consistent with our motivation by building on approaches to farsighted stability in cooperative games (surveyed in Ray and Vohra 2015a). We define a *convention* as a mapping from the history of outcomes to a prescription for today. Such conventions reflect the players’ shared understanding of how the future unfolds in response to past and current choices. A convention is *stable* if given this shared understanding, no coalition has a profitable one-shot deviation at any history. We then ask the question: *What can stable conventions implement?*

Result for Perfect Monitoring: We pose this question first in a standard setting in which all behavior by individuals and coalitions is perfectly observed. Our first observation is that history-dependence is a source of stability.[2] When behavior is history-dependent, a farsighted coalition that has a myopic incentive to deviate may not find it in its best interest to do so. We elucidate this logic using simple examples in Section 2 where we show that (1) in a repeated roommates-matching problem, every efficient allocation can be supported by a stable convention even if the one-shot interaction has no stable match, and (2) in a repeated division problem, one can use reversion to the core of a stage-game to build a stable convention, just like Nash-reversion in repeated (non-cooperative) games.

Given these possibilities, we investigate the limits of history-dependence in Section 3. How much can it support? We find few limits to what a convention can credibly implement in both non-transferable utility and transferable utility environments (Theorems 1 and 2).

> **A Folk Theorem For Perfect Monitoring.** *For every payoff vector that is feasible and strictly individually rational, there exists a* $\underline{\delta} < 1$ *such that if* $\delta > \underline{\delta}$*, then there is a stable convention that achieves that payoff.*

The set of supportable payoffs identified in this folk theorem coincides with that of Fudenberg and Maskin (1986), although we allow for coalitional actions and deviations. Thus, we find that coalitional deviations do not refine the set of supportable outcomes beyond individual deviations when players are patient; dynamic incentives effectively ward off coalitional deviations. This result has a simple intuition: to ward off coalitional deviations, it suffices to punish an individual member of each coalition as if she were the sole deviator. This logic applies even when players can transfer utility, effectively bribing others to join their coalition, because the convention can then punish players for paying or receiving bribes.

Secret Transfers: The previous result exploits the observability of transfers. But in many contexts, the power of bribes and side payments comes from their secrecy and the inability to

[2] We are not the first to note that history-dependence can be a source of coalitional stability: Hyndman and Ray (2007), Vartiainen (2011), and Dutta and Vartiainen (2019) offer similar conclusions in different contexts.

punish people for paying or accepting bribes. Our second set of results, exposited in Section 4, finds a contrasting conclusion when coalitions can use secret side-payments.

Specifically, suppose that for any coalition that blocks an outcome, its members can transfer utility to each other secretly. In other words, the convention cannot condition future continuation play on these transfers, although it can condition behavior on the identity and actions of the deviating coalition. In this setting, players can effectively bribe others to join a deviating coalition; although the convention identifies who deviated and how (in terms of actions), it does not identify who made or received the side-payments. We find that this is an important imperfection: secret transfers severely undermine dynamic incentives, potentially limiting behavior to the core of the stage game, regardless of players' patience.

To describe our result, let us define coalition C's *coalitional minmax* to be the lowest total stage-game payoff (adding across its constituents) that coalition C can be pushed down to by others when it can best-respond. This coalitional payoff is analogous to the individual minmax, except that it treats the coalition as a single entity whose payoff is the sum of payoffs of its constituents. In cooperative games without externalities, the coalitional minmax of a coalition equals its value specified by the characteristic function. We prove the following result (Theorem 3).

> **An Anti-Folk Theorem For Secret Transfers.** *For each $\delta < 1$, a stable convention implements only those payoffs that give each coalition at least its coalitional minmax.*

For cooperative games without externalities, the result implies that the set of sustainable payoffs are those within the core of the stage game, regardless of players' patience. Here, dynamic incentives fail to sustain any outcome that could not have been sustained in the one-shot game. When externalities are present, then the coalitional minmax involves others outside the coalition taking actions to minimize the gains of the deviating coalition. In this case, our result relates to a variation of the core to permit externalities: the β-characteristic function suggested by Von Neumann and Morgenstern (1945) derives the value of a coalition C based on that coalition being minmaxed, and the β-core is the core corresponding to that characteristic function. Our result implies that stable conventions can implement payoffs only within the β-core of the game, which generally is a strict subset of the set of feasible and individually rational payoffs.[3]

Why do secret transfers matter? The key idea is that once transfers are secret, a deviating coalition can structure their transfers to ensure that if it collectively gains from deviating, then so does each individual member without changing the continuation play. Thus, the convention can no longer single out a member of that coalition to credibly punish and must instead do its best to punish the entire deviating coalition. More formally, we prove that with secret transfers, a *One-Shot Coalitional Deviation Principle* (Lemma 1) applies: a coalition lacks a profitable one-shot deviation from a convention if and only if it lacks a profitable multi-shot deviation.

[3] The set of of β-core payoffs is non-empty if and only if the $\beta-$characteristic function satisfies the conditions of the Bondareva-Shapley Theorem.

This result is the crux of the Anti-Folk Theorem: any convention that sustains an outcome below a coalitional minmax is susceptible to these multi-shot deviations and therefore by this principle, has a profitable one-shot deviation. Hence, such a convention is unstable. This result illustrates that once coalitions can make secret transfers, long-term commitments are no longer necessary for coalitions to capitalize on long-term gains; such gains can be appropriated using short-term commitments and secret side-payments.

We iterate this logic in a setting where *all* coalitions can make secret transfers and the grand coalition is omnipotent in that it can choose any feasible alternative. Because the grand coalition can also guarantee itself a coalitional minmax, efficient actions must then be chosen after every history. Define the *efficient β-core* to be the set of payoffs that are both (i) efficient, and (ii) give each coalition higher than its coalitional minmax in a reduced game where *only* efficient alternatives may be chosen. We prove in Theorem 4 that for every discount factor, stable conventions support payoffs only within the efficient β-core of the game and that for patient players, every payoff within the relative interior of that set can be supported.

Once coalitions can make secret transfers, the appropriate analysis treats each non-singleton coalition as a fictitious entity, expanding the number of players from n to $2^n - 1$. The efficient β-core emerges as the relevant folk theorem for this set of "players." The β-core is often criticized on the grounds that it is unclear as to why individuals outside of a coalition would wish to minimize the payoffs of those within a blocking coalition (Ray 2007). That criticism is apt in one-shot interactions where those outside a blocking coalition would not hurt themselves to punish deviators. But a repeated game can reward players for punishing others. If transfers can be made secretly within blocking coalitions, the efficient β-core emerges as the set of supportable outcomes.

While we delineate our results on perfecly observable vs. secret transfers as two separate cases, both of these extremes are special cases of a more general result. In Section 5, we study a setting where only a subset of coalitions can make secret transfers. We show then that only these coalitions are guaranteed their coalitional minmax. Moreover, any payoff vector that delivers strictly more than these minmaxes and is strictly individually rational is supportable when players are sufficiently patient (Theorem 5).

An Application: We apply our ideas to pure division problems—*simple games* (Von Neumann and Morgenstern 1945)—in which some players are elites in that they have veto power. We study the degree to which repeated interaction can motivate elite players to share resources with non-elite players. We use self-generation approaches to show that when side-payments are perfectly observable, then even for fixed discount factors, substantial sharing with non-elite citizens can be supported by stable conventions. However, once elites can make secret side-payments to co-opt others, then elites always obtain all of the surplus.

1.1 Related Literature

This paper is part of a growing effort to combines elements from cooperative and non-cooperative game theory; for example, with respect to incomplete information, see Liu, Mailath, Postlewaite and Samuelson (2014) and Liu (2018), or with respect to reasoning, see Ambrus (2006, 2009) and Lipnowski and Sadler (2019).[4] We develop new notions of coalitional stability when those coalitions act under the shadow of the future. Accordingly, we build on important precursors in cooperative and repeated games, and describe some of the most closely related papers below.

Our work builds on the broad study of farsighted stability in coalitional games, surveyed in Ray and Vohra (2015a). A closely related strand, initiated by Konishi and Ray (2003), studies real-time coalition-formation processes, and our solution-concept builds on theirs.[5] They study the dynamics of coalitional structures where payoffs accrue in real time, and coalitions evaluate their moves according to a recursive continuation value, just like our formulation of a stable convention in Definition 3.[6] Behavior in this setting is "Markov," where coalitions condition their behavior only on the current payoff-relevant state and not how it was reached. Hyndman and Ray (2007) introduce history-dependence with long-term binding agreements that can be renegotiated only by all affected parties. Vartiainen (2011) establishes existence of history-dependent absorbing deterministic farsightedly stable processes in a variation of this game without discounting.

We build on this strand with several differences. We study an abstract repeated game—which embeds both coalitional and strategic-form games—where all alliances are temporary and the only intertemporal interlinkage is the publicly observed history. We investigate the power and limits of history-dependence, with and without transfers. Incidentally, the direction in which we proceed is suggested by Ray (2007, pp. 301) as being potentially important for future research:

> *It would be of interest to investigate dynamic noncooperative games with (nonbinding) coalition formation...one might begin with the partition function so that the formation of a coalition structure at any date has a definite impact on payoffs, perhaps through the writing of binding agreements within coalitions in any period. But the important difference...is that such agreements would—by assumption—be up for grabs at the end of every period. There are no binding agreements that last for longer than a single date.*

A special case of our model is Bernheim and Slavov (2009), who extend the notion of a Condorcet Winner to an infinitely repeated game. They study history-dependent policy programs that at each stage are majority-preferred to paths generated by deviations. Specialized to their

[4] There is also growing interest in dynamic matching; see Corbae, Temzelides and Wright (2003), Damiano and Lam (2005), Du and Livne (2016), Kadam and Kotowski (2018), Doval (2018), Kotowski (2019), and Liu (2019).

[5] A different approach to these issues describes sets of outcomes that are immune to profitable coalitional deviations where each deviating coalition anticipates potential chains of subsequent deviations. See Harsanyi (1974), Chwe (1994), Jordan (2006), Ray and Vohra (2015b), Dutta and Vohra (2017), Kimya (2019), and Vohra ⓡ Ray (2019). While most of this approach studies history-independent behavior, Dutta and Vartiainen (2019) show that history-dependence facilitate existence.

[6] Also related are Gomes and Jehiel (2005) and Acemoglu, Egorov and Sonin (2012), who model real-time coalition-formation through the Markov Perfect Equilibria of a non-cooperative extensive-form model.

setting, our solution-concept coincides with their's. They study properties and applications of this solution-concept, but do not derive bounds on what it can enforce; our results establish that, because individuals are powerless on their own, all payoffs are supportable in their game (so long as players have non-equivalent utilities) as $\delta \to 1$.

Our results emphasize how coalitional deviations coupled with secret side-payments undermine dynamic incentives in the repeated game. Barron and Guo (2019) study a related issue in a relational contracting game between a long-run Principal and a sequence of short-run agents. They capture the friction that secret side-payments expose the Principal to extortion by shirking agents. More broadly, the challenge of secret side-payments is also an important theme in collusion in mechanism design; see Section 5 of Mookherjee (2006) for a survey.

Numerous papers adopt cooperative criteria to select equilibria in repeated games. Aumann (1959) and Rubinstein (1980) respectively study the Strong Nash and Strong Perfect Equilibria of an infinitely repeated game with limit-of-means and overtaking discounting criteria. Their solution-concepts assume that each coalition can commit to arbitrary long-run deviations off the path of play but not on-path. DeMarzo (1992) focuses on finite-horizon games and proposes an inductive solution-concept where behavior corresponds to a Strong Nash Equilibrium of the reduced normal-form game. He uses scapegoat strategies to prove a similar Folk Theorem as our NTU result for finitely repeated games.[7] Also related is the study of renegotiation-proofness (e.g. Pearce 1987; Bernheim and Ray 1989; Farrell and Maskin 1989), most of which focuses on deviations by the grand coalition to different behavior in the continuation game. By contrast, our focus is on short-term deviations by all coalitions where players cannot "re-wire" expectations about continuation behavior.

2 Examples

2.1 The Roommates Problem

We illustrate the role of history-dependence in a repeated version of the "roommates problem." Consider three players—Alice, Bob, and Carol—who are choosing between rooming together or remaining unmatched. The challenge is that only a pair can room together, and so at least one player is always alone. Table 1 describes their stage-game payoffs.

A matching specifies who rooms with whom, and a stable match is immune to profitable individual and coalitional deviations: there should be no pair of players who prefer to room with each other over their current match nor an individual player who prefers rooming alone to her match. A well-known challenge is that every match in this one-shot interaction is unstable.

[7] He also briefly studies infinite-horizon games, but because his solution-concept differs from ours, a similar folk theorem obtains only for two-player games.

	Alice	Bob	Carol
Alice	1	3	2
Bob	2	1	3
Carol	3	2	1

TABLE 1. Payoffs of Row Player from matching with Column Player (or remaining unmatched).

We model a setting where players match repeatedly, share a common discount factor δ, and the match today can condition on past outcomes. A coalition may choose to jointly deviate today, but coalitions cannot commit to future deviations; in other words, the matching convention has to be immune to profitable one-shot coalitional deviations on and off the path of play. We call such history-dependent matching processes ***stable conventions***.

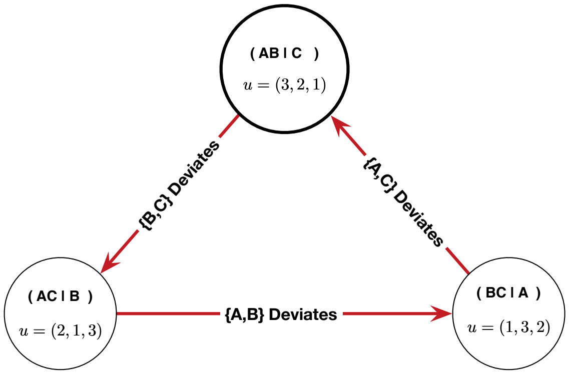


FIGURE 1. A stable convention for the roommates' problem if $\delta \geq 1/2$.

Figure 1 depicts a stable convention. In this stable convention, Alice and Bob are matched in every period on the path of play, and Carol remains unmatched. Bob and Carol each have a myopic incentive to deviate by matching with each other. But the history-dependent matching process ensures that Bob does not wish to deviate if he is sufficiently patient: should Bob and Carol deviate, then in every subsequent period, the process specifies that Bob remains unmatched. Given this punishment, Bob prefers to stay matched with Alice in each period if

$$\underbrace{(1-\delta)(3)}_{\text{Bob-Carol for a single period}} + \underbrace{\delta(1)}_{\text{Unmatched forever, discounted}} \leq \underbrace{2}_{\text{Alice-Bob Forever}},$$

which is satisfied whenever $\delta \geq \frac{1}{2}$.

The off-path behavior satisfies the same credibility as that on the path of play: when Alice and Carol are meant to match forever, Alice is punished in the future if she chooses to deviate with Bob. In this manner, the automaton depicted in Figure 1 guarantees that no coalition wishes to deviate when players are sufficiently patient. This example illustrates how a repeated matching environment has a stable convention even if the static one-shot environment lacks one.[8]

2.2 Dividing a Dollar with a Veto Player

Here, we illustrate our results using a ***simple game*** (Von Neumann and Morgenstern 1945): consider a divide-the-dollar game between three players—1, 2, and 3—where $\{1,2\}$ and $\{1,3\}$ can block and implement any division of the dollar, but the coalition of $\{2,3\}$ is powerless (as is any singleton).[9] Here, player 1 is an elite veto player who needs the support of one other (non-elite) player to capture the surplus. In the core of this stage game, player 1 captures the entire dollar; every other allocation guarantees that she and one other player has a profitable joint deviation.

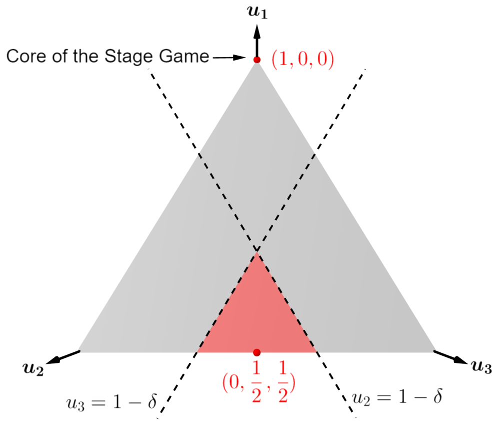


FIGURE 2. Supportable payoffs using core-reversion if $\delta \geq \frac{1}{2}$.

History-dependence can do more when this problem is repeated. Suppose that now, in every period, there is a dollar to be divided, and group behavior can condition on past allocations, whether any coalition blocked, etc. Similar to Nash-reversion equilibria of repeated non-cooperative games, we develop a stable convention that reverts to the core of the stage-game as a punishment. Consider a convention that recommends $\left(0,\frac{1}{2},\frac{1}{2}\right)$ every period so long as that has been the allocation in every prior period, and recommends the core of the stage game in any other history. Now, even if player 1 offers the entire dollar to either player 2 or 3, neither is willing to join her in blocking this outcome if $\delta \geq \frac{1}{2}$:

$$(1-\delta)(1)+\delta(0) \leq \frac{1}{2},$$

[8] Our rule shares similarities with previous dynamic resolutions. In a stochastic game where the state-variable is the previous period's chosen coalition structure, Konishi and Ray (2003) construct stable processes where the coalitional structure cycles stochastically when players are patient. Looking at a game without discounting, Vartiainen (2011) constructs an absorbing history-dependent process that shares a similar spirit to ours.

[9] We thank Elliot Lipnowski for suggesting this example to us. Ray and Vohra (2015b) and Dutta and Vohra (2017) also use this example to illustrate their approaches.

where the LHS is player 2's (or 3's) deviation payoff from being promised the entire surplus today and reverting to the core of the stage game from tomorrow onwards, and the RHS is her payoff from continuing on the path of play. Going further, core-reversion can support any allocation in the triangle formed by the vertices $\{(2\delta-1, 1-\delta, 1-\delta), (0, \delta, 1-\delta), (0, 1-\delta, \delta)\}$, which converges to the entire unit simplex as $\delta \to 1$. This is depicted in Figure 2.[10]

3 The Power of Conventions: Perfect Monitoring

This section describes our framework and results when monitoring is perfect. For expositional clarity, we first describe implications for non-transferable utility environments, and then introduce perfectly observed transfers.

3.1 A Non-Transferable Utility Environment

A set of players $N \equiv \{1, 2 \ldots, n\}$ interact repeatedly at $t = 0, 1, 2, \ldots$, and share a common discount factor $\delta < 1$. Players can make choices as individuals and as coalitions. The set of possible coalitions is the set of all nonempty subsets of N, denoted by $\mathcal{C}$.

The Stage Game: We consider a non-transferable utility (henceforth NTU) stage game using the language of cooperative game theory. Let A be the set of ***alternatives***, which is finite. An alternative a generates a payoff vector $v(a) \equiv (v_1(a), \ldots, v_n(a)) \in \mathsf{R}^n$, and we use $v : A \to \mathsf{R}^n$ to denote the generated payoff function. Using the language of Abreu, Dutta and Smith (1994), we sometimes focus on settings where no two players have perfectly aligned preferences and call these "games with nonequivalent utilities."

Definition 1. *The stage game satisfies **nonequivalent utilities** (NEU) if there is no pair of players $\{i, j\}$, and constants k and $\lambda > 0$ such that $v_i(a) = k + \lambda v_j(a)$ for all $a \in A$.*

In each period, the convention recommends an alternative, and feasible deviations for coalitions and individuals are defined relative to that recommendation. If a in A is recommended, then coalitions can decide whether to ***block*** the recommendation. If coalition C chooses to block the recommendation, it can deviate to any alternative in $E_C(a)$. If no coalition chooses to block, then the recommendation is implemented. The correspondence $E_C : A \rightrightarrows A$ is coalition C's ***effectivity correspondence***, as in Rosenthal (1972). This correspondence is "reflexive": $a \in E_C(a)$ for every alternative a and coalition C. Below are a few examples of commonly studied environments expressed in this language.

[10] In Section 6, we show how one can do more both in this example and more generally across a large class of simple games by using approaches from Abreu (1988) and Abreu, Pearce and Stacchetti (1990) to characterize the full set of supportable payoffs for fixed discount factors.

Example 1. Consider a strategic-form game in which player i's action set is A_i, the set of action profiles is $A \equiv \times_{i=1}^{n} A_i$. The effectivity correspondence is $E_C(a) \equiv \left\{a' \in A : a'_j = a_j \text{ for all } j \notin C\right\}$, modeling the possibility for a deviating coalition to choose action profiles in which players outside the coalition do not change their actions. This formulation extends the standard definition for individual deviations (used to define Nash equilibria) to a coalitional environment.

Example 2. Consider a general NTU characteristic function game (N, U) where the mapping $U(C) \subseteq \mathsf{R}^{|C|}$ specifies a set of feasible payoff vectors for coalition C if it forms. Let $\mathcal{P}$ be the set of all partitions of N and π be a generic partition. An alternative a is now a tuple (π, u), the effectivity correspondence $E_C(a)$ specifies the set of alternatives to which coalition C may move, and the payoff function is $v((\pi, u)) = u$.

Example 3. Suppose, as in Bernheim and Slavov (2009), that individuals vote in each period over a set of alternatives. Let $\mathcal{W}$ be the set of coalitions that have at least $\lceil \frac{N}{2} \rceil$ players. The effectivity correspondence specifies that for every a, $E_C(a) = A$ if $C \in \mathcal{W}$, and otherwise, $E_C(a) = \{a\}$.

Outcomes, Histories, and Paths: At the end of each period, the feasible outcome $o \equiv (a, C)$ specifies the chosen alternative and the identity of the blocking coalition (if any). We denote the set of feasible outcomes in this NTU environment by $\mathcal{O}^{NTU} \equiv A \times \mathcal{C}$. When referring to past outcomes, we denote the alternative chosen in period t by a^t and the blocking coalition in period t by C^t, where $C^t = \emptyset$ if the recommendation in period t was unblocked.[11]

A t-period history is a sequence $h^t \equiv (a^\tau, C^\tau)_{\tau=0,1,2,\ldots,t-1}$, that specifies alternatives and blocking coalitions for t periods. We denote the set of all feasible t-length histories by $\mathcal{H}^t$ for $t \geq 1$, and $\mathcal{H}^0 = \{\emptyset\}$ for the singleton comprising the initial null history. We denote by $\mathcal{H} \equiv \bigcup_{t=0}^{\infty} \mathcal{H}^t$ the set of all feasible histories. An ***outcome path*** is an infinite sequence $p \equiv (a^t, C^t)_{t=0,1,2,\ldots}$, specifying alternatives and blocking coalitions for each of infinitely many periods.

Plans and Conventions: A ***plan*** recommends an outcome following each history: a plan is a mapping $\sigma : \mathcal{H} \to \mathcal{O}^{NTU}$.[12] We denote the alternative and a blocking coalition recommended by a plan σ after history h by $a(h|\sigma)$ and $C(h|\sigma)$. A ***convention*** is a plan that recommends only outcomes that are unblocked: in other words, $\sigma : \mathcal{H} \to A \times \{\emptyset\}$.

Payoffs: For a path $p = (a^t, C^t)_{t=0,1,2,\ldots}$, $U_i(p) \equiv (1-\delta) \sum_{t=0}^{\infty} \delta^t v_i(a^t)$ denotes player i's normalized discounted continuation payoff from that path, where $0 \leq \delta < 1$ is the common discount

[11] Our model assumes that coalitional blocking is observable. If the stage game is a strategic-form game as in Example 1, then this assumption is unnecessary; instead, it suffices at every stage to punish someone from among those whose actions depart from the recommendation's. However, in a general partitional game (e.g., matching), the alternative itself may not code sufficient information about who deviated. We abstract from this monitoring imperfection, and as in the closely related papers (Hyndman and Ray 2007; Vartiainen 2011; Dutta and Vartiainen 2019), assume that the identity of the blocking coalition is directly observed.

[12] Because we focus on deterministic plans, there is no need to track a plan's past recommendation.

factor. For a plan σ and after history h, let $P(h|\sigma) \equiv (\sigma(h), \sigma(h, \sigma(h)), \ldots)$ denote the path generated recursively by σ after that history, and $U_i(h|\sigma)$ denote player i's payoff from that path.

3.2 A Definition of Stable Conventions

In this section, we define our notion of stability. For comparison, we begin with the notion for the stage game:

Definition 2. *An alternative a is a **core-alternative** if there exists no coalition C and alternative $a' \in E_C(a)$ such that for every i in C, $v_i(a') > v_i(a)$. A payoff vector $\widetilde{v}$ is in the **core** of the NTU stage game if there exists a core-alternative a such that $\widetilde{v} = v(a)$.*

The core focuses attention on alternatives where no coalition gains from blocking. Our dynamic solution-concept elaborates on the core in a straightforward way: we say that a convention is **stable** if after every history, no coalition unanimously prefers blocking the recommendation today, assuming that the future unfolds as anticipated by the convention.

Definition 3. *A convention σ is **stable in the NTU repeated game** if for every history h, there exists no coalition C and feasible deviation $a' \in E_C(a(h|\sigma))$ such that*

$$\text{For every } i \in C\text{:} \quad (1-\delta)v_i(a') + \delta U_i(h, a', C|\sigma) > U_i(h|\sigma). \tag{1}$$

In other words, no coalition has a profitable one-shot deviation.

The requirement for stability is that at every history and given future play, no coalition finds it profitable to block today. Coalitions anticipate that their choices today affect continuation play and a stable convention ensures that at least one member of each coalition finds the long-run cost of changing the path of play to outweigh her instantaneous gain from deviating.[13] Thus, players' shared understanding of the future—formalized through the convention—deters coalitional deviations today.

An alternative way to express the idea is that a stable convention recommends only core-alternatives of the *reduced normal-form game* at every history (whose payoffs are a convex combination of today's payoffs and continuation values). If $\delta = 0$, that reduced normal-form game collapses to the stage game and so stable conventions necessarily implement only core-alternatives of the stage game. One may proceed further with this connection. Suppose that a^* is a core-alternative, and consider a convention that prescribes a^* after every history. Such a convention is stable because behavior today does not impact continuation play, and in every period, no coalition gains myopically from deviating. The converse is also true: every "Markov" stable

[13] Our requirement for profitability is that every coalition member strictly gains from blocking. Alternatively, one could stipulate that every coalition member is weakly better off and at least one is strictly better off. Our main results are identical with this alternative definition.

convention—i.e., that in which the prescription does not depend on past play—can implement *only* core-alternatives. Thus, the relationship between a stable convention of the repeated game and the core of the stage game is analogous to that between sub-game perfect equilibria of the repeated game and the Nash equilibria of the corresponding stage game.

As mentioned before, our notion of a stable convention builds on important precursors. Konishi and Ray (2003) consider a similar recursive payoff in a stochastic game where players condition on the current coalitional structure (but not on past history), and this has been an important approach to studying farsighted stability in the subsequent literature. In the context of repeated elections, Bernheim and Slavov (2009) study *Dynamic Condorcet Winners*, which coincides with stable conventions when we when the stage game is specialized to their's.

3.3 What Can Be Enforced By Stable Conventions?

We turn to what stable conventions can enforce. We find that for NTU games, every feasible and "individually" rational payoff can be supported by a stable convention, if players are patient.

Let us define this payoff set. The set of feasible payoffs is the convex hull of the set of feasible payoffs denoted by $\mathcal{V}^\dagger \equiv \text{co}(\{\widetilde{v} \in \mathsf{R}^n : \exists a \in A \text{ such that } \widetilde{v} = v(a)\})$. Analogous to the (pure-action) minmax of noncooperative games, define each player's minmax payoff as the lowest payoff that she can be pushed down to by the convention when she can best-respond:

$$\underline{v}_i \equiv \min_{a \in A} \max_{a' \in E_{\{i\}}(a)} v_i(a'). \qquad \text{(Player } i\text{'s minmax)}$$

Thus, the subset of feasible payoffs that is strictly individually rational is

$$\mathcal{V}^\dagger_{IR} \equiv \left\{ v \in \mathcal{V}^\dagger : v_i > \underline{v}_i \text{ for every } i = 1, \ldots, n \right\}. \qquad \text{(NTU Feasible IR)}$$

With this in place, we state our first result.

Theorem 1. *For every $\delta \geq 0$, every stable convention gives each player i a payoff of at least $\underline{v}_i$. Moreover, if the stage game satisfies NEU, then for every $v \in \mathcal{V}^\dagger_{IR}$, there is a $\underline{\delta} < 1$ such that for every $\delta \in (\underline{\delta}, 1)$, there exists a stable convention with a discounted payoff equal to v.*

The statement of this folk theorem is nearly identical to that for sub-game perfect equilibria (Fudenberg and Maskin 1986; Abreu, Dutta and Smith 1994), except that we permit coalitional deviations, and do not limit our analysis to repeated play of a strategic-form game. Nevertheless, payoffs that are strictly *individually* rational can be sustained, and the possibility for coalitional deviations does not refine the set of sustainable outcomes. The key conceptual idea is that to deter coalitional deviations, it suffices to punish only a single constituent of each coalition—a "scapegoat"—as if she were a sole deviator.[14]

[14] The logic of Theorem 1 indicates that it would apply even if coalitions could commit to a sequence of deviations

We discuss the key steps. A convention is stable if no coalition, even those that are singletons (i.e., individuals), has profitable one-shot deviations. An implication of the standard one-shot deviation principle then is that no individual has a profitable multi-shot deviation. This property implies that no player can be pushed to below her individual minmax because otherwise she can profitably deviate. The second part of the result uses the NEU condition to construct player-specific punishments to deter individual deviations, and as mentioned above, identical punishments are used to deter coalitional deviations. Finally, because we have not augmented our model with a public correlation device, we use sequences of play (as in Sorin 1986 and Fudenberg and Maskin 1991) to achieve payoffs that are in the convex hull of generated payoffs.

3.4 Transferable Utility with Perfect Monitoring

This section augments the game with perfectly observed transfers.[15] We describe transfers using $T \equiv [T_{ij}]_{i,j\in N}$ where $T_{ij} \in [0, \infty)$ is the non-negative utility that is transferred to player j from player i. Let $\mathcal{T}$ denote the set of all $n \times n$ matrices with non-negative entries. We use $T_i = [T_{ij}]_{j\in N}$ to denote the vector of transfers paid by player i. Let $T_C = [T_i]_{i\in C}$ be the transfers paid by members of coalition C and $T_{-C} = [T_i]_{i\notin C}$ be transfers paid by members outside coalition C. Transfers modify payoffs in the usual way: a player's *experienced payoff* is the sum of her generated payoff and net transfers. That is $u_i(a, T) \equiv v_i(a) + \sum_{j\in N} T_{ji} - \sum_{j\in N} T_{ij}$.

A feasible outcome of the stage game now specifies the chosen alternative, the identity of a blocking coalition (if any), and the chosen transfers. We denote the set of feasible outcomes by $\mathcal{O}^{TU} \equiv \left\{o = (a, C, T) | a \in A, C \in \mathcal{C}, T \in \mathcal{T}\right\}$. Histories and paths are defined as in NTU stage games, with (a, C, T) replacing (a, C) whenever needed to account for transfers. A plan $\sigma : \mathcal{H} \to \mathcal{O}^{TU}$ specifies an outcome, including transfers, based on history. We continue to use $a(h|\sigma)$ and $C(h|\sigma)$ to denote the recommended alternative and blocking coalition in $\sigma(h)$, and in addition, we use $T(h|\sigma)$ to denote the transfers in $\sigma(h)$. As before, a convention recommends only outcomes that have empty blocking coalitions; in other words, $\sigma : \mathcal{H} \to A \times \{\emptyset\} \times \mathcal{T}$.

If coalition C blocks a recommended outcome $(a, \emptyset, T)$, it can choose any a' in $E_C(a)$, and change its transfer schedule to any T'_C so that the realized outcome is (a', C, T'_C, T_{-C}). This formulation assumes that when a coalition blocks, it still accept incoming transfers from outside the blocking coalition who do not know of the block at the time at which transfers are paid. This assumption is inessential to our results, and is assumed for notational convenience; identical results follow if one were to instead assume that blocking coalitions must achieve budget-balance.

Since the game has been augmented with transfers, we re-define the set of feasible and individually rational payoffs. Potential experienced payoff profiles after alternative a is chosen is

across histories, where the maximal number of deviations is bounded. We do not model this scenario explicitly because a profitable finite long-run deviation for a coalition must either involve a profitable one-shot deviation or call for an individual within the coalition to deviate even if that's not in her interest.

[15] We model transfers separately from alternatives to sharpen the contrast to the secret transfers case.

$\mathcal{U}(a) = \{u \in \mathsf{R}^n : \sum_{i\in N} u_i = \sum_{i\in N} v_i(a)\}$, its convex hull is $\mathcal{U}^\dagger \equiv \mathrm{co}(\cup_{a\in A}\mathcal{U}(a))$, and the set of feasible and strictly individually rational payoffs is

$$\mathcal{U}^\dagger_{IR} \equiv \{u \in \mathcal{U}^\dagger : u_i > \underline{v}_i \text{ for every } i = 1,\dots,n\}. \qquad \text{(TU Feasible IR)}$$

Players have preferences over the discounted stream of experienced payoffs. The definition of $U_i(p)$, $P(h|\sigma)$ and $U_i(h|\sigma)$ are modified in the obvious way to reflect the influence of transfers on experienced payoffs. To avoid Ponzi schemes, for all of our results, we restrict attention to conventions whose continuation values lie in a bounded set.

Assumption 1. We consider conventions σ such that continuation values are bounded across histories: $\{u \in \mathsf{R}^n : \exists h \in \mathcal{H} \text{ such that } U(h|\sigma) = u\}$ is a bounded subset of R^n.

We now extend the notion of a stable convention to allow for perfectly observed transfers.

Definition 4. *A convention σ is* ***stable in the TU repeated game*** *if for every history h, there exists no coalition C, alternative $a' \in E_C(a(h|\sigma))$, and transfers $T'_C = [T'_{ij}]_{i\in C, j\in N}$, such that*

$$\textit{For every } i \in C\textit{:} \quad (1-\delta)u_i(a', [T'_C, T_{-C}(h|\sigma)]) + \delta U_i(h, a', C, [T'_C, T_{-C}(h|\sigma)]|\sigma) > U_i(h|\sigma) \quad (2)$$

Because transfers are publicly observed, subsequent behavior may condition on these transfers when coalition C blocks the recommended outcome. A stable convention can use this information to sustain a large set of outcomes, as we prove below.

Theorem 2. *For every $\delta \geq 0$, every stable convention gives each player i a payoff of at least $\underline{v}_i$. For every $u \in \mathcal{U}^\dagger_{IR}$, there is a $\underline{\delta}$ such that for every $\delta \in (\underline{\delta}, 1)$, there exists a stable convention with a discounted payoff equal to u.*

The proof is similar to that of Theorem 1. Transfers ensure that players have opposed interests in the stage game, so payoffs necessarily satisfy NEU in this augmented game. The complication introduced by transfers is that if a deviating coalition anticipates a certain member to be punished, her coalition partners can use transfers today to compensate her. These transfers can potentially undermine the deterrence of future punishment, even as $\delta \to 1$. To overcome this problem, the convention targets the player who gains least from the deviation after transfers are made.

4 Secret Transfers Undermine Conventions

We see in Theorem 2 that even if coalitions can share their gains and losses through side-payments, a convention can deter coalitional deviations by punishing players for giving or receiving transfers. A different conclusion emerges if coalitions can make side-payments secretly. Section 4.1 describes this secret-transfers setting. Section 4.2 proves a one-shot coalitional deviation principle and

Section 4.3 uses this fact to show that each coalition can guarantee itself a coalitional minmax value. Section 4.4 characterizes the set of supportable payoffs, and connects that set to the β-core.

4.1 The Setup

We say that transfers are *secret* when the convention cannot condition on the amount of those payments: the future recommendation can depend on the identity of blocking coalitions as well as alternatives they've chosen but not on the amount of bribes and side-payments they have paid to one another.

Definition 5. *Two histories $h = (a^0, C^0, T^0, \ldots, a^t, C^t, T^t)$ and $\widetilde{h} = (\widetilde{a}^0, \widetilde{C}^0, \widetilde{T}^0, \ldots, \widetilde{a}^t, \widetilde{C}^t, \widetilde{T}^t)$ of the same length are **identical up to secret transfers** if for every $0 \leq \tau \leq t$,*

1. *the same alternative is chosen: $a^\tau = \widetilde{a}^\tau$,*
2. *the identify of the blocking coalition, if any, is the same: $C^\tau = \widetilde{C}^\tau$, and*
3. *the same transfers are made, except for those within the blocking coalition: $T^\tau_{ij} = \widetilde{T}^\tau_{ij}$ if $\{ij\} \nsubseteq C^\tau$.*

*A convention σ **respects secret transfers** if $\sigma(h) = \sigma(h')$ for any $h, h' \in \mathcal{H}$ that are identical up to secret transfers.*

A stable convention that respects secret transfers is one that satisfies both Definitions 4 and 5. A special case of a convention that respects secret transfers is one that ignores transfers altogether between any pair (blocking or otherwise). Our definition here assumes that all blocking coalitions can transfer utility secretly; as we elaborate in Section 5, our insights generalize to the case where some coalitions cannot do so.

Definition 5 entails that because players outside blocking coalitions do not observe transfers within a blocking coalition, their actions cannot condition on them. It also assumes that members of blocking coalitions do not condition their own subsequent *equilibrium* play on the transfers made within the blocking coalition. This measurability restriction may appear stronger than "secrecy," but we adopt this definition for two reasons.

The first reason is tractability. Our restriction is analogous to that of perfect public equilibria in repeated games with public monitoring (Mailath and Samuelson 2006) where all players condition their play on publicly observable variables. Secret transfers generate persistent private information, and it is beyond the scope of existing tools to characterize coalitional behavior that is *both* dynamic and where members of a coalition are asymmetrically informed.[16]

Second, one may envision that continuation play that attempts to elicit private information from deviators (about their transfers) might themselves be vulnerable to coalitional deviations.[17]

[16] Our approach is similar to that of collusion in mechanism design (Mookherjee 2006) where details of the side-contract are unobservable and cannot be conditioned on by the Principal.

[17] This issue emerges in one-shot environments, where as was noted by Maskin (1979), it may be impossible to implement any social choice correspondence that satisfies no-veto power in Strong Nash Equilibria.

This is an important issue but addressing it definitively is beyond our scope here, in particular because it would require a solution-concept that allows for both asymmetric information and dynamics. But as a preliminary result, we consider an expanded game in Supplementary Appendix B.7 where players can communicate about secret transfers. We show that an identical coalitional payoff guarantee result holds for semi-public equilibria of that game.

4.2 A One-Shot Coalitional Deviation Principle

We find that secret transfers undermine intertemporal incentives, and guarantees that each coalition obtains its coalitional minmax value. Central to this result is a one-shot *coalitional* deviation principle where we show that for conventions that respect secret transfers, a coalition has a profitable multi-shot coalitional deviation only if it has a profitable one-shot coalitional deviation. Therefore, any stable convention is immune to profitable multi-shot coalitional deviations.

Let us define multi-shot coalitional deviations. A multi-shot coalitional deviation is a plan that departs from the convention that is also feasible for the coalition: C is *solely* responsible for any deviations at any history, and the deviation (in terms of the alternative and transfers) at any history must be feasible for coalition C.

Definition 6. *A* ***multi-shot deviation by coalition*** $\mathbf{C}$ *from convention σ is a distinct plan $\sigma' : \mathcal{H} \to \mathcal{O}^{TU}$ such that for any history $h \in \mathcal{H}$ where $\sigma'(h) = (a', C', T') \neq \sigma(h)$, it must be that $C' = C$, $a' \in E_C(a(h|\sigma))$ and $T'_{-C} = T_{-C}(h|\sigma)$. A multi-shot deviation σ' by coalition C is* ***profitable*** *if there exists a history h such that $U_i(h|\sigma') > U_i(h|\sigma)$ for all $i \in C$.*

With these preliminaries defined, we prove the following result.

Lemma 1. (One-shot Coalitional Deviation Principle). Under secret transfers, a convention σ is stable if and only if it has no profitable multi-shot coalitional deviations.

Lemma 1 establishes that once coalitions can make secret transfers to each other, if the members of a coalition can gain from (committing to) a multi-shot deviation, then they can structure transfers so that they can also gain from a one-shot deviation. Players are effectively bribing others to join their coalition without worrying about being punished for these side-payments. This idea also does not require that every blocking coalition be able to make secret transfers; in Section 5, we show that if some but not all coalitions can make secret transfers, then the one-shot coalitional deviation principle applies for those coalitions that can make secret transfers.

Sketch of Proof: The "if" direction is true by definition. For the "only if" direction, suppose as a contrapositive that there is a profitable multi-shot deviation. Our steps below construct a profitable one-shot coalitional deviation using the following steps:

a. Since every member of C has a higher utility from that deviation path, it must be that the sum of the members' utilities is also higher.

b. Now treat the coalition C as a hypothetical player whose payoff is the sum of payoffs of members of coalition C. The standard argument establishes that this profitable multi-shot deviation is reducible to a profitable one-shot deviation for this hypothetical entity.

c. Under secret transfers, coalition C's gains in total value from that one-shot deviation can be freely distributed among its members using intra-coalition transfers when the coalition blocks, *without affecting continuation play*. Thus, there is a one-shot coalitional deviation that is profitable for every member of coalition C, and therefore, σ is unstable.

4.3 Coalitional Payoff Guarantees: An Anti-Folk Theorem

We use the one-shot coalitional deviation principle to show that in a stable convention, for every discount factor, each coalition can guarantee itself a total payoff below which it cannot be pushed down. We define this *coalitional minmax* as

$$\underline{v}_C \equiv \min_{a \in A} \max_{a' \in E_C(a)} \sum_{i \in C} v_i(a'). \tag{Coalition C's minmax}$$

This coalitional minmax adapts standard individual minmaxes in a natural way: it treats coalition C as a hypothetical entity that has a payoff that is the sum of the payoffs of its constituents, and can best-respond according to $E_C(\cdot)$. This minmax corresponds to the β-characteristic function proposed by Von Neumann and Morgenstern (1945) (see also Luce and Raiffa 1957 and Ray 2007) that assumes that those outside a blocking coalition act in ways to minimize the total value of those within it.[18] We argue that each coalition can guarantees itself at least this value.

Theorem 3. *Under secret transfers, for every $\delta \geq 0$, every stable convention gives each coalition C a total value of at least $\underline{v}_C$.*

Here is the argument for Theorem 3: if a convention σ could push a coalition down to a total value strictly less than $\underline{v}_C$, then we can construct a profitable multi-shot deviation by members of coalition C. By Lemma 1, there then exists a profitable one-shot coalitional deviation, which implies that σ is not stable.[19]

We view Theorem 3 as an Anti-Folk Theorem. In a general cooperative game without externalities, $\underline{v}_C$ corresponds to the value of coalition C given by its characteristic function. Thus, in

[18] A subtle difference is that the β-characteristic function is often used to convert a strategic-form game into a cooperative game. We are applying the same logic, but to an abstract transferable utility game, including those that lack a product-structure.

[19] We note that Theorem 3 applies even if the convention uses a public randomization device: for every realization of the public randomization device, coalition C can guarantee itself a total payoff of at least $\underline{v}_C$ by best-responding to the recommendation. Because Lemma 1 still applies, a stable convention then cannot push a coalition's value below this minmax.

such cases, stable conventions can implement payoffs only in the core of the stage game. More generally, when externalities are present, our result guarantees that payoffs supported by a stable convention are a subset of the β-core (i.e., the core when the characteristic function is the β-characteristic function defined above). If the conditions for the Bondareva-Shapley Theorem (Peleg and Sudhölter 2007) are satisfied, then the β-core is non-empty. However, if the conditions fail, then the β-core may be empty. We do not view this as a nihilistic conclusion, but rather as an illustration of how coalitional deviations coupled with secret side-payments severely erodes the power of conventions: every scheme of carrots and sticks is undermined by coalitional deviations.

4.4 The Efficient β-Core

Here, we develop a tighter characterization of supportable payoffs using an additional assumption that the grand coalition is omnipotent.[20]

Assumption 2. (Omnipotence of the Grand Coalition). For all $a \in A$, $E_N(a) = A$.

Assumption 2 yields an important implication when coupled with Theorem 3: if the grand coalition must achieve its minmax value and is omnipotent, then a stable convention must generate (utilitarian-)efficient continuation payoffs after every history. A simple logic implies that any stable convention must then use utilitarian-efficient *alternatives* after every history.[21] With this in mind, denote the set of *efficient alternatives* by $\overline{A} \equiv \arg\max_{a \in A} \sum_{i \in N} v_i(a)$. Define *efficient coalitional minmaxes* as the lowest payoff that a coalition can be pushed down to using only efficient alternatives:

$$\underline{v}_C^e \equiv \min_{a \in \overline{A}} \max_{a' \in E_C(a)} \sum_{i \in C} v_i(a'). \tag{3}$$

For every coalition C, $\underline{v}_C^e$ is weakly higher than $\underline{v}_C$ as the restriction to efficient alternatives limits how much coalitions can be punished. We use these minmaxes to define the "efficient β-core."

Definition 7. *The **efficient** β**-core** is the set*

$$\mathcal{B} \equiv \left\{ u \in \mathsf{R}^n : \sum_{i \in N} u_i = \max_{a \in A} \sum_{i \in N} v_i(a),\ \sum_{i \in C} u_i \geq \underline{v}_C^e \textit{ for all } C \in \mathcal{C} \backslash \{N\} \right\},$$

[20] This assumption, while commonly satisfied, is not innocuous: the grand coalition may be "too large" to be able to coordinate on joint deviations.

[21] Here is the argument. Suppose that a utilitarian-inefficient alternative is used in period t after some history h^t. The grand coalition's future payoff, from period $t+1$ onwards, cannot exceed the efficient level, and thus, it cannot recoup the efficiency loss incurred in period t. But then the grand coalition is not achieving its minmax value after history h^t, which is a contradiction.

*and the **strict efficient** β**-core** is the set*

$$\mathcal{B}^s \equiv \left\{ u \in \mathsf{R}^n : \sum_{i\in N} u_i = \max_{a\in A} \sum_{i\in N} v_i(a),\ \sum_{i\in C} u_i > \underline{v}^e_C \text{ for all } C \in \mathcal{C}\backslash\{N\} \right\}.$$

The efficient β-core guarantees that each coalition obtains at least its efficient coalitional minmax; the strict efficient β-core is in its relative interior guaranteeing that each "non-grand" coalition obtains strictly more than that minmax. Since we can treat these payoffs as that of a characteristic function, the set $\mathcal{B}$ is non-empty whenever that characteristic function satisfies the conditions of the Bondareva-Shapley Theorem. Moreover, as we prove in the Supplementary Appendix B.6, a mild strengthening of these conditions guarantee that $\mathcal{B}^s$ is also non-empty.

Theorem 4. *Under secret transfers, for every $\delta \geq 0$, every stable convention implements payoffs only within the efficient β-core. If the strict efficient β-core is non-empty, then for every payoff profile $u \in \mathcal{B}^s$, there is a $\underline{\delta} < 1$ such that for every $\delta \in (\underline{\delta}, 1)$, there exists a stable convention with a discounted payoff equal to u.*

The argument for the first part of Theorem 4 mirrors that of Theorem 3, but proves and embeds the idea that stable conventions can select only efficient alternatives. The second part of Theorem 4 is new. We treat each coalition—apart from the grand coalition—as a hypothetical player, and construct "player-specific" punishments for each such hypothetical player. While we do this step directly, one can see that this is feasible because the payoffs across coalitions satisfy the NEU condition in the game augmented with transfers. Using these "player-specific" punishments, a stable convention can support any payoff vector where each of these hypothetical players obtains strictly more than its efficient "individually rational" payoff.

5 A General Result on Public and Secret Transfers

For simplicity, we model transfers within blocking coalitions as being either all public (Theorem 2) or all secret (Theorems 3 and 4) but the insights generalize to settings that span these extremes. We model this more general setting here. We suppose that some but not all coalitions can transfer payoffs secretly. We show that only these coalitions are guaranteed a coalitional minmax value. The set of supportable payoffs then, if players are patient, are those that are feasible, strictly individually rational, and gives any coalition that can make secret transfers strictly more than its coalitional minmax.

Let $\mathcal{S} \subseteq \mathcal{C}$ denote the (non-singleton) set of coalitions that can make secret transfers.

Definition 5*. *Two histories $h = (a^0, C^0, T^0, \ldots, a^t, C^t, T^t)$ and $\widetilde{h} = (\widetilde{a}^0, \widetilde{C}^0, \widetilde{T}^0, \ldots, \widetilde{a}^t, \widetilde{C}^t, \widetilde{T}^t)$ of the same length are **identical up to $\mathcal{S}$-secret transfers** if for every $0 \leq \tau \leq t$,*

1. *the same alternative is chosen:* $a^\tau = \widetilde{a}^\tau$*,*
2. *the identity of the blocking coalition, if any, is the same:* $C^\tau = \widetilde{C}^\tau$*, and*
3. *the same transfers are made, except for those within a blocking coalition* that can make secret transfers*:* $T^\tau_{ij} = \widetilde{T}^\tau_{ij}$ *if either* $C^\tau \notin \mathcal{S}$ *or* $\{ij\} \nsubseteq C^\tau$*.*

A convention σ ***respects*** $\mathcal{S}$***-secret transfers*** *if* $\sigma(h) = \sigma(h')$ *for any* $h, h' \in \mathcal{H}$ *that are identical up to* $\mathcal{S}$*-secret transfers.*

Definition 5* is identical to Definition 5 except that only transfers within a blocking coalition in $\mathcal{S}$ cannot be conditioned upon by the convention.[22] We prove an analogoue of Lemma 1: once a coalition can make secret transfers, regardless of whether other coalitions can do so, it can always find a profitable one-shot deviation whenever it has a profitable multi-shot deviation.

Lemma 1*. Under $\mathcal{S}$-secret transfers, a convention σ is stable only if no coalition C in $\mathcal{S}$ has a profitable multi-shot deviation.

Lemma 1* implies that every coalition in $\mathcal{S}$ achieves at least its coalitional minmax. Below, we define the feasible and individually rational payoff set that also satisfies these coalitional minmaxes (we also define the payoff set where all minmax constraints hold strictly).

Definition 8. *The set of* $\mathcal{S}$***-rational payoffs*** *is*

$$\mathcal{D}(\mathcal{S}) \equiv \left\{ u \in \mathsf{R}^n : \sum_{i \in C} u_i \geq \underline{v}_C \text{ for all } C \in \mathcal{S} \cup N \right\},$$

and the set of ***strictly*** $\mathcal{S}$***-rational payoffs*** *is the set*

$$\mathcal{D}^s(\mathcal{S}) \equiv \left\{ u \in \mathsf{R}^n : \sum_{i \in C} u_i > \underline{v}_C \text{ for all } C \in \mathcal{S} \cup N \right\},$$

Theorem 5. *For every set of coalitions that can make secret transfers,* $\mathcal{S} \subseteq \mathcal{C}$*, and for every* $\delta \geq 0$*, every stable convention implements payoffs only within* $\mathcal{D}(\mathcal{S})$*. If* $\mathcal{D}^s(\mathcal{S})$ *is non-empty, then for every payoff profile* $u \in \mathcal{D}^s(\mathcal{S})$*, there is a* $\underline{\delta} < 1$ *such that for every* $\delta \in (\underline{\delta}, 1)$*, there exists a stable convention with a discounted payoff equal to* u*.*

The first conclusion of Theorem 5 follows from Lemma 1*. For the second conclusion, we use transfers to construct "player-specific" punishments where the set of hypothetical players is $\mathcal{S} \cup N$. Using these punishments, a stable convention can push the payoff of each hypothetical player arbitrarily close to the appropriate minmax values when players are sufficiently patient.[23]

[22] We do not impose any structure on $\mathcal{S}$, and study generally the consequences of secret transfers. An alternative approach would be to study settings where a directed graph specifies who can secretly transfer utility to whom, and that graph forms the basis of which transfers within blocking coalitions are secret.

[23] We note that $\mathcal{D}^s(\mathcal{S})$ is non-empty implies that the grand coalition cannot make secret transfers ($N \notin \mathcal{S}$). Once the grand coalition can make secret transfers, then the set of supportable payoffs reduces to those using only efficient alternatives, as in Section 4.4.

6 An Application to Simple Games

We apply our analysis to *simple games* (Von Neumann and Morgenstern 1945), re-visiting and generalizing the example of Section 2.2. Simple games are problems of pure division where certain *winning coalitions* have the rights to allocate a fixed surplus, and the question of interest is seeing how that surplus is divided. We study the class of simple games where no one is a dictator, but some players have veto power ("elites"). We study, for fixed discount factors, the degree to which history-dependence can support outcomes where elites share their resources with non-elites.

Let us first describe simple games in the language of our model. The set of alternatives is $A \equiv \{a \in \mathsf{R}^N_+ : \sum_{i \in N} a_i = 1\}$, where player i's generated payoff from alternative a is $v_i(a) \equiv a_i$. Let $\mathcal{W}$ be the set of *winning coalitions*, where each winning coalition C in $\mathcal{W}$ has the ability to choose how the dollar is divided, and each *losing coalition* $C \notin \mathcal{W}$ does not. In other words, for each a, $E_C(a) = A$ if $C \in \mathcal{W}$, and $E_C(a) = \{a\}$ otherwise. We assume that $\mathcal{W}$ is *monotonic* and *proper*.[24] Player i is an elite or *veto player* if she is a member of every winning coalition. The collection of all veto players is $D \equiv \cap_{C \in \mathcal{W}} C$, and a *collegial game* is that in which D is non-empty. We study non-dictatorial collegial games. These games are of interest because it models relatively common settings where there is at least one veto player, but that veto player does not have complete power (e.g. Winter 1996; McCarty 2000).[25]

For a non-dictatorial collegial game, the core of the stage game involves every non-veto player obtaining 0. We compare that allocation with those supported by stable conventions when players interact repeatedly. We have already seen in Section 2.2 how history-dependence can sustain a larger set of outcomes by using "core-reversion" as a punishment. Here, we consider a broader class of conventions and punishments, using approaches from Abreu (1988) and Abreu, Pearce and Stacchetti (1990).[26] (Below, Δ refers to the non-negative n-dimensional unit simplex.)

Theorem 6. *Suppose that the stage-game is non-dictatorial and collegial. With perfect monitoring, with or without transfers:*

a. *If there are at least two veto players, the set of supportable payoffs are those that give at least* $(1 - \delta)$ *to each winning coalition :*

$$U(\delta) \equiv \left\{ u \in \Delta : \sum_{i \in C} u_i \geq 1 - \delta \text{ for every } C \in \mathcal{W} \right\}.$$

[24] In other words, if $C \in \mathcal{W}$ and $C' \supseteq C$, then $C' \in \mathcal{W}$. Also, $C \in \mathcal{W}$ implies that $N \backslash C \notin \mathcal{W}$.

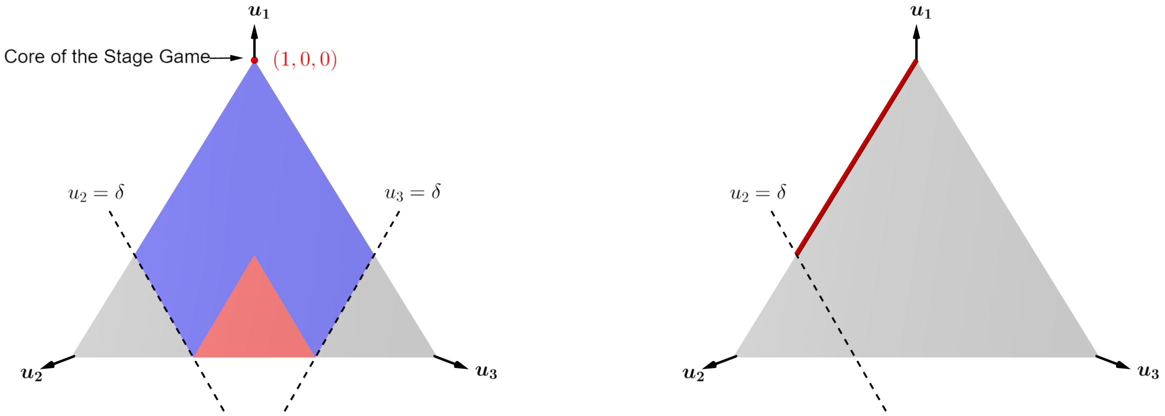
[25] One example is the interaction between a legislative body and an executive leader with veto power where neither body can pass a proposal on its own. Another example corresponds to organizations (e.g., the UN Security Council) where some members have veto power but the support of some non-veto players is also needed. Finally, power-sharing arrangements that resemble clientelism and patronage (Francois, Rainer and Trebbi 2015) often require the support of certain elites and sufficient support from non-elite citizens.

[26] To simplify exposition, we consider only those conventions that are stationary on the path of play. We conjecture that this is without loss of generality, particularly in our results for settings with transfers.

b. If there is only a single veto player, there exists $\underline{\delta}$ such that the set of supportable payo s is $U(\delta)$ for $\delta > \underline{\delta}$.

By contrast, when transfers are secret, the set of supportable outcomes, regardless of δ, is the core of the stage game: $K \equiv \left\{u \in \Delta : \sum_{i \in D} u_i = 1.\right\}$.

To interpret Theorem 6, suppose that there are at least two veto players. At $\delta = 0$, the set of supportable payoffs, $U(\delta)$, coincides with the core of the stage game. The set $U(\delta)$ strictly increases in δ in terms of set-inclusion, and it does so in the direction of sharing more surplus with non-veto players.[27] By contrast, with secret transfers, the set of supportable payoffs always coincides with the core of the stage game, regardless of δ. We illustrate these results in Figure 3 using our three-player example of Section 2.2.



(A) Perfect monitoring for $\delta > 1/2$

(B) Secret transfers between 1 and 2

FIGURE 3. (A) depicts the set of supportable outcomes with perfect monitoring. The red region depicts payoffs supported by core-reversion, and the blue region illustrates the gains that come from other stable conventions. (B) shows the set of supportable outcomes once coalition $\{1, 2\}$ can make secret transfers; player 3 then obtains 0. If all coalitions can make secret transfers, then the only supportable payoff is the core of the stage game.

These results illustrate a role of institutions that monitor bribes and side-payments. When all behavior is publicly observable, elite players can be motivated to share their surplus with non-elite players through a convention that punishes them if they deviate. However, that ability is lost once elite players can co-opt others with secret side-payments.

7 Conclusion

The primary contribution of this paper is to take the "repeated-games program"—of using carrots and sticks to discipline deviations—into a setting that has coalitional moves. We develop a frame-

[27] For the special case of voting rules like that in the UN Security Council, where every coalition that comprises the veto players and at least k of the $(n - |D|)$ non-veto players is a winning coalition, Theorem 6 implies that the wealthiest $n - |D| - k$ non-veto players cannot together obtain more than a δ fraction of the surplus.

work and solution-concept that tractably merges approaches in cooperative and repeated games. Because we model an abstract stage game, which includes both strategic-form and partitional games, our approach can be used to study repeated cooperative games as well as coalitional deviations in repeated noncooperative games.[28] The recursive nature of our solution-concept makes it feasible to analyze the set of supportable payoffs using standard techniques, such as self-generation.

We use this framework to investigate when and how coalitions can be credibly disciplined by carrots and sticks. The observability of transfers emerges as a critical feature: if transfers are perfectly observed, then stable conventions can support every feasible and strictly individually rational payoff vector. By contrast, if a coalition can make secret transfers, it can guarantee itself a minimal "coalitional minmax" value regardless of players' patience. When all coalitions can do so, then the set of supportable payoffs collapses to the core of the stage game (suitably defined).

We view this contrast to potentially speak to questions of enforcement and social order. An important consideration in enforcement is whether the elites' temptation to violate the law or abuse political power is disciplined by their expectations of future punishment.[29] But a challenge ubiquitous across time and space is that players can often evade sanctions by profitably bribing their punishers and partnering with them. Thus, we view credible enforcement to require immunity to coalitional deviations. Our results suggest that if transfers are observable, these coalitional deviations may not be so costly. But when players can secretly bribe others, they are less threatened by the prospect of future punishment even if deviating players cannot sign long-term contracts.

References

Abreu, Dilip (1988) "On the Theory of Infinitely Repeated Games with Discounting," *Econometrica*, Vol. 56, No. 2, pp. 383–396.

Abreu, Dilip, Prajit K. Dutta, and Lones Smith (1994) "The Folk Theorem for Repeated Games: A NEU Condition," *Econometrica*, Vol. 62, No. 4, pp. 939–948.

Abreu, Dilip, David Pearce, and Ennio Stacchetti (1990) "Toward A Theory of Discounted Repeated Games with Imperfect Monitoring," *Econometrica*, pp. 1041–1063.

Acemoglu, Daron, Georgy Egorov, and Konstantin Sonin (2012) "Dynamics and Stability of Constitutions, Coalitions, and Clubs," *American Economic Review*, Vol. 102, No. 4, pp. 1446–76.

Acemoglu, Daron and Alexander Wolitzky (2018) "A Theory of Equality Before the Law," Working Paper.

——— (2019) "Sustaining Cooperation: Community Enforcement vs. Specialized Enforcement," *Journal of the European Economic Association*.

[28] Liu (2019) applies a similar conceptual approach to study a repeated matching process where long-run players ("hospitals") may engage in coalitional deviations with short-run agents ("interns"). Because these short-run agents cannot be punished by continuation play, he finds limits to how much long-run players can be punished.

[29] A vast literature, dating back to Hume (1740) and recently including Basu (2000) and Mailath, Morris and Postlewaite (2017), identifies the law with history-dependent conventions that are credible and self-enforcing. See Acemoglu and Wolitzky (2018, 2019) for complementary analyses.

Ambrus, Attila (2006) “Coalitional Rationalizability,” *Quarterly Journal of Economics*, Vol. 121, No. 3, pp. 903–929.

——— (2009) “Theories of Coalitional Rationality,” *Journal of Economic Theory*, Vol. 144, No. 2, pp. 676–695.

Aumann, Robert J. (1959) “Acceptable Points in General Cooperative n-Person Games,” in Kuhn, H. W. and R. D. Luce eds. *Contributions to the Theory of Games IV*, Vol. 4, Princeton, NJ: Princeton University Press, p. 287.

Barron, Daniel and Yingni Guo (2019) “The Use and Misuse of Coordinated Punishments,” Working Paper.

Basu, Kaushik (2000) *Prelude to Political Economy: A Study of the Social and Political Foundations of Economics*, Oxford, UK: Oxford University Press.

Bernheim, B. Douglas and Debraj Ray (1989) “Collective Dynamic Consistency in Repeated Games,” *Games and Economic Behavior*, Vol. 1, No. 4, pp. 295–326.

Bernheim, B. Douglas and Sita N. Slavov (2009) “A Solution Concept for Majority Rule in Dynamic Settings,” *Review of Economic Studies*, Vol. 76, No. 1, pp. 33–62.

Blackwell, David (1965) “Discounted Dynamic Programming,” *Annals of Mathematical Statistics*, Vol. 36, No. 1, pp. 226–235.

Chwe, Michael (1994) “Farsighted Coalitional Stability,” *Journal of Economic theory*, Vol. 63, No. 2, pp. 299–325.

Compte, Olivier (1998) “Communication in repeated games with imperfect private monitoring,” *Econometrica*, pp. 597–626.

Corbae, Dean, Ted Temzelides, and Randall Wright (2003) “Directed Matching and Monetary Exchange,” *Econometrica*, Vol. 71, No. 3, pp. 731–756.

Damiano, Ettore and Ricky Lam (2005) “Stability in Dynamic Matching Markets,” *Games and Economic Behavior*, Vol. 52, No. 1, pp. 34–53.

DeMarzo, Peter M. (1992) “Coalitions, Leadership, and Social Norms: The Power of Suggestion in Games,” *Games and Economic Behavior*, Vol. 4, No. 1, pp. 72–100.

Doval, Laura (2018) “A Theory of Stability in Dynamic Matching Markets,” Working Paper.

Du, Songzi and Yair Livne (2016) “Rigidity of Transfers and Unraveling in Matching Markets,” Working Paper.

Dutta, Bhaskar and Hannu Vartiainen (2019) “Coalition Formation and History Dependence,” *Theoretical Economics*.

Dutta, Bhaskar and Rajiv Vohra (2017) “Rational Expectations and Farsighted Stability,” *Theoretical Economics*, Vol. 12, No. 3, pp. 1191–1227.

Farrell, Joseph and Eric Maskin (1989) “Renegotiation in Repeated Games,” *Games and Economic Behavior*, Vol. 1, No. 4, pp. 327–360.

Francois, Patrick, Ilia Rainer, and Francesco Trebbi (2015) “How Is Power Shared in Africa?” *Econometrica*, Vol. 83, No. 2, pp. 465–503.

Fudenberg, Drew and Eric Maskin (1986) “The Folk Theorem in Repeated Games with Discounting or with Incomplete Information,” *Econometrica*, pp. 533–554.

——— (1991) “On the Dispensability of Public Randomization in Discounted Repeated Games,” *Journal of Economic Theory*, Vol. 53, No. 2, pp. 428—438.

Gomes, Armando and Philippe Jehiel (2005) “Dynamic Processes of Social and Economic Interactions: On the Persistence of Inefficiencies,” *Journal of Political Economy*, Vol. 113, No. 3, pp. 626–667.

Harsanyi, John C. (1974) "An Equilibrium-Point Interpretation of Stable Sets and a Proposed Alternative Definition," *Management Science*, Vol. 20, No. 11, pp. 1472–1495.

Hume, David (1740) *A Treatise of Human Nature*, Oxford, UK: Oxford University Press.

Hyndman, Kyle and Debraj Ray (2007) "Coalition Formation with Binding Agreements," *Review of Economic Studies*, Vol. 74, No. 4, pp. 1125–1147.

Jordan, James S. (2006) "Pillage and Property," *Journal of Economic Theory*, Vol. 131, No. 1, pp. 26–44.

Kadam, Sangram V. and Maciej H. Kotowski (2018) "Multiperiod Matching," *International Economic Review*, Vol. 59, No. 4, pp. 1927–1947.

Kandori, Michihiro and Hitoshi Matsushima (1998) "Private observation, communication and collusion," *Econometrica*, Vol. 66, No. 3, p. 627.

Kimya, Mert (2019) "Equilibrium Coalitional Behavior," Working Paper.

Konishi, Hideo and Debraj Ray (2003) "Coalition Formation as A Dynamic Process," *Journal of Economic Theory*, Vol. 110, No. 1, pp. 1–41.

Kotowski, Maciej H. (2019) "A Perfectly Robust Approach to Multiperiod Matching Problems," Working Paper.

Lipnowski, Elliot and Evan Sadler (2019) "Peer-Confirming Equilibrium," *Econometrica*, Vol. 87, No. 2, pp. 567–591.

Liu, Ce (2019) "Stability in Repeated Matching Markets," Working Paper.

Liu, Qingmin (2018) "Rational Expectations, Stable Beliefs, and Stable Matching," Working Paper.

Liu, Qingmin, George J. Mailath, Andrew Postlewaite, and Larry Samuelson (2014) "Stable Matching with Incomplete Information," *Econometrica*, Vol. 82, No. 2, pp. 541–587.

Luce, R. Duncan and Howard Raiffa (1957) *Games and Decisions: Introduction and Critical Survey*: John Wiley and Sons, Inc.

Mailath, George J., Stephen Morris, and Andrew Postlewaite (2017) "Laws and Authority," *Research in Economics*, Vol. 71, No. 1, pp. 32–42.

Mailath, George and Larry Samuelson (2006) *Repeated Games and Reputations*, New York, NY: Oxford University Press.

Mangasarian, Olvi L (1994) *Nonlinear programming*, Philadelphia, PA: SIAM.

Maskin, Eric (1979) "Implementation and strong Nash equilibrium," in Laffont, J.J. ed. *Aggregation and Revelation of Preferences*: North-Holland, pp. 433–439.

McCarty, Nolan M. (2000) "Proposal Rights, Veto Rights, and Political Bargaining," *American Journal of Political Science*, pp. 506–522.

Mookherjee, Dilip (2006) "Decentralization, Hierarchies, and Incentives: A Mechanism Design Perspective," *Journal of Economic Literature*, Vol. 44, No. 2, pp. 367–390.

Pearce, David G (1987) "Renegotiation-Proof Equilibria: Collective Rationality and Intertemporal Cooperation," Working Paper.

Peleg, Bezalel and Peter Sudhölter (2007) *Introduction to the theory of cooperative games*, Vol. 34: Springer Science & Business Media.

Ray, Debraj (2007) *A Game-Theoretic Perspective on Coalition Formation*, New York, NY: Oxford University Press.

Ray, Debraj and Rajiv Vohra (2015a) "Coalition Formation," in Young, H. Peyton and Shmuel Zamir eds. *Handbook of Game Theory*, Vol. 4: Elsevier, pp. 239–326.

——— (2015b) "The Farsighted Stable Set," *Econometrica*, Vol. 83, No. 3, pp. 977–1011.

Rosenthal, Robert W. (1972) "Cooperative Games in Effectiveness Form," *Journal of Economic Theory*, Vol. 5, No. 1, pp. 88–101.

Rubinstein, Ariel (1980) "Strong Perfect Equilibrium in Supergames," *International Journal of Game Theory*, Vol. 9, No. 1, pp. 1–12.

Sorin, Sylvain (1986) "On Repeated Games with Complete Information," *Mathematics of Operations Research*, Vol. 11, No. 1, pp. 147–160.

Vartiainen, Hannu (2011) "Dynamic Coalitional Equilibrium," *Journal of Economic Theory*, Vol. 146, No. 2, pp. 672–698.

Vohra, Rajiv ⓡ Debraj Ray (2019) "Maximality in the Farsighted Stable Set," *Econometrica.*

Von Neumann, John and Oskar Morgenstern (1945) *Theory of Games and Economic Behavior*: Princeton University Press Princeton, NJ.

Winter, Eyal (1996) "Voting and Vetoing," *American Political Science Review*, pp. 813–823.

A Appendix

A.1 Outline and Preliminaries

This main appendix contains the proofs of the Folk Theorem for NTU Games (Theorem 1), the One-Shot Coalitional Deviation Principle for Secret Transfers (Lemma 1), and the Anti-Folk Theorem for Secret Transfers (Theorem 3).

The Supplementary Appendix contains proofs for our other results. Some of these arguments share a similar spirit to those of the above results, but with modifications that address important issues that arise. The proof of the Folk Theorem for TU Games with perfectly observed transfers (Theorem 2) mirrors that of Theorem 1 but addresses considerations that involve bounding the amount of transfers and selecting members of coalitions to punish in a way that cannot be undone through side-payments. The proof of the result identifying the connection with the efficient β-core (Theorem 4) iterates on the logic of Theorem 3, uses transfers to construct "coalition-specific" punishments, and then proves the bounds using an argument similar to Theorem 1.

Below, we exposit a notation and a result used throughout our proofs.

Let $BR_C(a) \equiv \arg\max_{a' \in E_C(a)} \sum_{i \in C} v_i(a')$ denote coalition C's best-response alternatives to a recommended alternative a.

Our analysis uses sequences of play to convexify payoffs, following standard arguments from Sorin (1986) and Fudenberg and Maskin (1991). Below, we reproduce the statement that we invoke in our arguments.

Lemma 2. (Lemma 2 of Fudenberg and Maskin 1991) Let X be a convex polytope in R^n with vertices $x^1, \ldots, x^K$. For all $\epsilon > 0$, there exists a $\underline{\delta} < 1$ such that for all $\underline{\delta} < \delta < 1$, and any $x \in X$, there exits a sequence $\{x_\tau\}_{\tau=0}^{\infty}$ drawn from $\{x^1, \ldots, x^K\}$, such that $(1-\delta)\sum_{\tau=0}^{\infty} \delta^\tau x_\tau = x$ and at any t, $\left|\left|x - (1-\delta)\sum_{\tau=t}^{\infty} \delta^{\tau-t} x_\tau\right|\right| < \epsilon$.

A.2 Proof of Theorem 1 on p. 12

Part 1: *For every $\delta \geq 0$, every stable convention gives each player i a payoff of at least $\underline{v}_i$.*

Consider any convention σ and player i such that $U_i(\emptyset|\sigma) < \underline{v}_i$. We first show that player i has a profitable multi-shot deviation from this convention and then use a one-shot deviation principle to show that there is a profitable one-shot deviation. Therefore σ cannot be stable.

A **multi-shot deviation for player** i from convention σ is a distinct plan $\sigma' : \mathcal{H} \to \mathcal{O}^{NTU}$ such that for any history $h \in \mathcal{H}$ where $\sigma'(h) = (a', C') \neq \sigma(h)$, it must be that $C' = \{i\}$ and $a' \in E_{\{i\}}(a(h|\sigma))$. A multi-shot deviation is **profitable** if there exists a history h such that $U_i(h|\sigma') > U_i(h|\sigma)$.

We consider the following multi-shot deviation: in every period, player i blocks and best-responds to the convention. Formally, this is a plan σ' where $C(h|\sigma') = \{i\}$ and $a(h|\sigma') \in BR_i\big(a(h|\sigma)\big)$ for every history $h \in \mathcal{H}$. By the definition of $\underline{v}_i$, the deviation σ' satisfies $v_i(a(h|\sigma')) \geq \underline{v}_i$ for all $h \in \mathcal{H}$, so player i's continuation value from period 0 must be higher: $U_i(\emptyset|\sigma') > U_i(\emptyset|\sigma)$.

We apply the standard one-shot deviation principle for individual decision making (Blackwell 1965) to this setting, which is now a simple decision tree.[30] Because stage-game payoffs are bounded for player i and there is discounting, the one-shot deviation principle implies that there exists a history $\overline{h} \in \mathcal{H}$ such that

$$(1-\delta)v_i(a(\overline{h}|\sigma')) + \delta U_i\Big(\overline{h}, a(\overline{h}|\sigma'), \{i\}\Big|\sigma\Big) > U_i(\overline{h}|\sigma),$$

which is a profitable one-shot deviation for coalition $\{i\}$. Therefore, σ is unstable.

Part 2: *If the stage game satisfies NEU, then for every $v \in \mathcal{V}^{\dagger}_{IR}$, there is a $\underline{\delta} < 1$ such that for every $\delta \in (\underline{\delta}, 1)$, there exists a stable convention with discounted payoff equal to v.*

Fix $v^0 \in \mathcal{V}^{\dagger}_{IR}$. We begin with preliminaries, defining payoffs and alternatives to support v^0.

First, since the game satisfies NEU, by Lemma 1 and Lemma 2 of Abreu, Dutta and Smith (1994), we can find *player-specific punishments* for v^0: there exist payoff vectors $\{v^i\}_{i=1}^n \subseteq \mathcal{V}^{\dagger}_{IR}$ such that $v^i_i < v^0_i$ for all $i \in N$, and $v^j_i > v^i_i$ for all $j \in N, j \neq i$. Second, let us define *minmaxing alternatives*: let $\underline{a}_i \in \arg\min_{a \in A} \max_{a' \in E_{\{i\}}} v_i(a')$ be an alternative that can be used to minmax player i. By construction, it follows that $v_i(\underline{a}_i) \leq \underline{v}_i$.

Given these payoffs and punishments, let $\kappa \in (0,1)$ be such that for every $\widetilde{\kappa} \in [\kappa, 1]$, the following is true for every i:

$$(1-\widetilde{\kappa})v_i(\underline{a}_i) + \widetilde{\kappa}v^i_i > \underline{v}_i \tag{4}$$

$$\text{For every } j \neq i: \quad (1-\widetilde{\kappa})v_j(\underline{a}_i) + \widetilde{\kappa}v^i_j > (1-\widetilde{\kappa})\underline{v}_j + \widetilde{\kappa}v^j_j \tag{5}$$

Inequality (4) implies that player i is willing to bear the cost of $v_i(\underline{a}_i)$ with the promise of transitioning into her player-specific punishment rather than staying at her minmax, where the promise is discounted at $\widetilde{\kappa}$. Similarly, inequality (5) implies that player j is willing to bear the cost of minmaxing player i

[30] For a statement of the one-shot deviation principle that applies in this context, see https://www.econ.nyu.edu/user/debraj/Courses/GameTheory2003/Notes/osdp.pdf

with the promise of transitioning into player i's specific punishment rather than her own, when the post-minmaxing phase payoffs are discounted at $\widetilde{\kappa}$. Each inequality holds at $\widetilde{\kappa} = 1$ for each i and $j \neq i$. Since the set of players is finite, there exists a value of $\kappa \in (0, 1)$ such that the inequality holds for all $\widetilde{\kappa} \in [\kappa, 1]$, $i \in N$ and $j \in N \backslash \{i\}$.

Let $L(\delta) \equiv \left\lceil \frac{\log \kappa}{\log \delta} \right\rceil$ where $\lceil \cdot \rceil$ is the ceiling function. Observe that $\delta^{L(\delta)} \in [\delta^{\frac{\log \kappa}{\log \delta}+1}, \delta^{\frac{\log \kappa}{\log \delta}}] = [\delta\kappa, \kappa]$. Therefore, $\lim_{\delta \to 1} \delta^{L(\delta)} = \kappa$.

Lemma 2 guarantees that for any $\epsilon > 0$, there exists $\underline{\delta} \in (0, 1)$ such that for all $\delta \in (\underline{\delta}, 1)$, there exist sequences $\left\{ \{a^{i,\tau}\}_{\tau=0}^{\infty} : i = 0, 1, \ldots, n \right\}$ such that for each i and t, $(1-\delta)\sum_{\tau=0}^{\infty} \delta^{\tau} v(a^{i,\tau}) = v^i$ and $\left|\left| v^i - (1-\delta)\sum_{\tau=t}^{\infty} \delta^{\tau} v(a^{i,\tau}) \right|\right| < \epsilon$. We fix an $\epsilon < (1-\kappa)\min\{\min_{i,j\neq i}(v_i^j - v_i^i), \min_i v_i^i - \underline{v}_i\}$, and given that ϵ, consider δ exceeding the appropriate $\underline{\delta}$.

We now describe the convention used to sustain v^0. Consider the automaton $(W, w(0,0), f, \gamma)$, where

- $W \equiv \left\{ w(d,\tau) | 0 \leq d \leq n, \tau \geq 0 \right\} \cup \{ \underline{w}(i,\tau) | 1 \leq i \leq n, 0 \leq \tau < L(\delta) \}$ is the set of possible states;
- $w(0,0)$ is the initial state;
- $f : W \to \mathcal{O}^{NTU}$ is the output function, where $f(w(d,\tau)) = (a^{d,\tau}, \emptyset)$ and $f(\underline{w}(i,\tau)) = (\underline{a}_i, \emptyset)$.
- $\gamma : W \times \mathcal{O}^{NTU} \to W$ is the transition function. For states of the form $w(d,\tau)$, the transition is

$$\gamma\big(w(d,\tau), (a, C)\big) = \begin{cases} \underline{w}(j^*, 0) & \text{if } C \neq \emptyset \text{ , } j^* = \min_{j \in C} j \\ w(d, \tau+1) & \text{otherwise} \end{cases}$$

For states in $\{\underline{w}(i,\tau) | 0 \leq \tau < L(\delta) - 1\}$,

$$\gamma\big(\underline{w}(i,\tau), (a, C)\big) = \begin{cases} \underline{w}(j^*, 0) & \text{if } C \notin \{\emptyset, \{i\}\} \text{ , } j^* = \min_{j \in C \backslash \{i\}} j \\ \underline{w}(i, 0) & \text{if } C = \{i\} \\ \underline{w}(i, \tau+1) & \text{otherwise} \end{cases}$$

For states of the form $\underline{w}(i, L(\delta)-1)$, the transition is

$$\gamma\big(\underline{w}(i, L(\delta)-1), (a, C)\big) = \begin{cases} \underline{w}(j^*, 0) & \text{if } C \notin \{\emptyset, \{i\}\} \text{ , } j^* = \min_{j \in C \backslash \{i\}} j \\ \underline{w}(i, 0) & \text{if } C = \{i\} \\ w(i, 0) & \text{otherwise} \end{cases}$$

The convention represented by the above automaton yields payoff profile v^0. By construction, the continuation values in different states, $V(\cdot)$, satisfy:

$$\begin{aligned} \left|\left| v^d - V(w(d,\tau)) \right|\right| &< \epsilon, & \tau &= 0, 1, \ldots \\ V(\underline{w}(i,\tau)) &= (1 - \delta^{L(\delta)-\tau}) v(\underline{a}_i) + \delta^{L(\delta)-\tau} V(w(i,0)), & \tau &= 0, \ldots, L(\delta) - 1 \end{aligned}$$

Below, we show that this convention is stable by showing that there is no profitable one-shot deviation in any state of this automaton.

Stability in states of the form $w(d,\tau)$: Set $B > \sup_{\{u \in \mathcal{V}^\dagger, i \in N\}} u_i$. Consider a one-shot deviation to (a, C) by coalition C. Let $j^* = \min\{j \in C\}$. For all τ, without the deviation j^* obtains a payoff greater than $v^d_{j^*} - \epsilon$. By deviating, j^* obtains a payoff less than

$$(1-\delta)B + \delta V_{j^*}(\underline{w}(j^*, 0)) = (1-\delta)B + \delta\left[(1-\delta^{L(\delta)})v_{j^*}(\underline{a}_{j^*}) + \delta^{L(\delta)} v^{j^*}_{j^*}\right]$$

For the deviation to be profitable, everyone in C, including player j^*, must be better off. So the one-shot deviation is unprofitable if the above term is no more than $v^d_{j^*} - \epsilon$. We prove that this is the case both for $j^* \neq d$ and $j^* = d$.

First consider $j^* \neq d$. Observe that

$$\lim_{\delta \to 1}(1-\delta)B + \delta\left[(1-\delta^{L(\delta)})v_{j^*}(\underline{a}_{j^*}) + \delta^{L(\delta)} v^{j^*}_{j^*}\right] = \lim_{\delta \to 1}\left[(1-\delta^{L(\delta)})v_{j^*}(\underline{a}_{j^*}) + \delta^{L(\delta)} v^{j^*}_{j^*}\right] < v^{j^*}_{j^*},$$

where the inequality follows from $v_{j^*}(\underline{a}_{j^*}) \leq \underline{v}_j < v^{j^*}_{j^*}$. Because ϵ by construction is strictly less than $v^d_{j^*} - v^{j^*}_{j^*}$, it follows that the deviation payoff is less than $v^d_{j^*} - \epsilon$ when δ is sufficiently large.

Now suppose that $j^* = d$. The deviation payoff being less than $v^d_{j^*} - \epsilon$ can be re-written as

$$(1-\delta)(B - v^{j^*}_{j^*}) + \epsilon \leq \delta(1-\delta^{L(\delta)})(v^{j^*}_{j^*} - v_{j^*}(\underline{a}_{j^*}))$$

As $\delta \to 1$, the LHS converges to ϵ. Because $\lim_{\delta\to 1}\delta^{L(\delta)} = \kappa$, the RHS converges to $(1-\kappa)(v^{j^*}_{j^*} - v_{j^*}(\underline{a}_{j^*}))$. By definition of ϵ, the above inequality holds, and therefore, there is no profitable one-shot deviation if δ is sufficiently high.

Stability in states of the form $\underline{w}(i,\tau)$: We prove that no coalition has a profitable one-shot deviation.

We first consider the case where $C = \{i\}$. Since player i is being minmaxed, her best possible deviation generates a payoff of $\underline{v}_i$ for her. She finds this deviation to be unprofitable if

$$(1-\delta^{L(\delta)-\tau})v_i(\underline{a}_i) + \delta^{L(\delta)-\tau}v^i_i \geq (1-\delta)\underline{v}_i + \delta(1-\delta^{L(\delta)})v_i(\underline{a}_i) + \delta^{L(\delta)+1}v^i_i. \tag{6}$$

Because $v^i_i > \underline{v}_i \geq v_i(\underline{a}_i)$, it suffices to show that

$$(1-\delta^{L(\delta)})v_i(\underline{a}_i) + \delta^{L(\delta)}v^i_i \geq (1-\delta)\underline{v}_i + \delta(1-\delta^{L(\delta)})v_i(\underline{a}_i) + \delta^{L(\delta)+1}v^i_i.$$

Re-arranging terms:

$$(1-\delta)(1-\delta^{L(\delta)})v_i(\underline{a}_i) + (1-\delta)\delta^{L(\delta)}v^i_i \geq (1-\delta)\underline{v}_i.$$

Dividing by $(1-\delta)$ yields:

$$(1-\delta^{L(\delta)})v_i(\underline{a}_i) + \delta^{L(\delta)}v^i_i \geq \underline{v}_i.$$

Let us verify that this inequality holds for sufficiently high δ. Taking $\delta \to 1$ yields Inequality (4), which is true. Hence Inequality (6) holds for sufficiently high δ.

If $C \neq \{i\}$, then j^* exists. Player j^* finds this one-shot deviation to be unprofitable if

$$(1-\delta^{L(\delta)-\tau})v_{j^*}(\underline{a}_i) + \delta^{L(\delta)-\tau}v^i_{j^*} \geq (1-\delta)B + \delta(1-\delta^{L(\delta)})v_{j^*}(\underline{a}_{j^*}) + \delta^{L(\delta)+1}v^{j^*}_{j^*}. \tag{7}$$

We prove that this inequality is satisfied if δ is sufficiently high. Examining the LHS, observe that for all τ such that $0 \leq \tau \leq L(\delta) - 1$,

$$\begin{aligned}\lim_{\delta\to1}\left[(1-\delta^{L(\delta)-\tau})v_{j^*}(\underline{a}_i) + \delta^{L(\delta)-\tau}v^i_{j^*}\right] &= \lim_{\delta\to1}\left[\left(1-\frac{\kappa}{\delta^\tau}\right)v_{j^*}(\underline{a}_i) + \frac{\kappa}{\delta^\tau}v^i_{j^*}\right]\\ &= (1-\widetilde{\kappa})v_{j^*}(\underline{a}_i) + \widetilde{\kappa}v^i_{j^*}\end{aligned}$$

for some $\widetilde{\kappa} \in [\kappa, 1]$. Examining the RHS of (7), observe that

$$\begin{aligned}&\lim_{\delta\to1}\left[(1-\delta)B + \delta(1-\delta^{L(\delta)})v_{j^*}(\underline{a}_{j^*}) + \delta^{L(\delta)+1}v^{j^*}_{j^*}\right] = \lim_{\delta\to1}\left[(1-\delta^{L(\delta)})v_{j^*}(\underline{a}_{j^*}) + \delta^{L(\delta)}v^{j^*}_{j^*}\right]\\ &= (1-\kappa)v_{j^*}(\underline{a}_{j^*}) + \kappa v^{j^*}_{j^*} \;\leq\; (1-\kappa)\underline{v}_{j^*} + \kappa v^{j^*}_{j^*} \;\leq\; (1-\widetilde{\kappa})\underline{v}_{j^*} + \widetilde{\kappa}v^{j^*}_{j^*},\end{aligned}$$

where the first equality follows from taking limits, the second from $\lim_{\delta\to1}\delta^{L(\delta)} = \kappa$, the first weak inequality follows from $v_{j^*}(\underline{a}_{j^*}) \leq \underline{v}_{j^*}$, the second weak inequality follows from $\widetilde{\kappa} \geq \kappa$ and $\underline{v}_{j^*} < v^{j^*}_{j^*}$. Since $\widetilde{\kappa} \in [\kappa, 1]$, (5) delivers that $(1-\widetilde{\kappa})v_{j^*}(\underline{a}_i) + \widetilde{\kappa}v^i_{j^*}$ is strictly higher than $(1-\widetilde{\kappa})\underline{v}_{j^*} + \widetilde{\kappa}v^{j^*}_{j^*}$. This term guarantees that (7) holds for sufficiently high δ.

A.3 Proof of Lemma 1 on p. 16

The "if" direction is true by definition. For the "only if" direction, consider a convention σ that respects secret transfers for which coalition C has a profitable multi-shot deviation, σ'. In other words, there exists a history $\overline{h} \in \mathcal{H}$ such that $U_i(\overline{h}|\sigma') > U_i(\overline{h}|\sigma)$ for every $i \in C$. We show that the convention σ has a profitable one-shot deviation, and therefore is not stable.

Since $U_i(\overline{h}|\sigma') > U_i(\overline{h}|\sigma)$ for every $i \in C$, it follows that $\sum_{i\in C} U_i(\overline{h}|\sigma') > \sum_{i\in C} U_i(\overline{h}|\sigma)$. Treat coalition C as a hypothetical player whose payoff is the sum of the payoffs of members of coalition C. Consider σ' as a multi-shot deviation by player C that increases its payoff.

By Assumption 1, the convention σ has bounded continuation value. We establish, in Lemma 4 in the Supplementary Appendix that if coalition C has a profitable multi-shot deviation, that it also has a profitable multi-shot deviation σ' in which $\left\{\sum_{i\in C} U_i(h|\sigma') : h \in \mathcal{H}\right\}$ is also bounded. Thus, the hypothetical player C faces a decision tree with bounded values and given discounting, the standard one-shot deviation principle applies. Therefore, there exists a history $\widehat{h} \in \mathcal{H}$ such that

$$(1-\delta)\sum_{i\in C} u_i\Big(a(\widehat{h}|\sigma'),\ T(\widehat{h}|\sigma')\Big) + \delta\sum_{i\in C} U_i\Big(\widehat{h},\ a(\widehat{h}|\sigma'),\ C,\ T(\widehat{h}|\sigma')\Big|\sigma\Big) > \sum_{i\in C} U_i(\widehat{h}|\sigma)$$

Thus, as a hypothetical player, C has a profitable one-shot deviation. We construct transfers to divide these gains so that each member of coalition C strictly profits from this one-shot deviation. Let T^* be the transfers matrix such that for all $(j,k) \notin C \times C$, $T^*_{jk} = T_{jk}(\widehat{h}|\sigma')$; but for $(j,k) \in C \times C$, T^*_{jk} satisfies for every $i \in C$,

$$(1-\delta)u_i\Big(a(\widehat{h}|\sigma'),\ T^*\Big) + \delta U_i\Big(\widehat{h},\ a(\widehat{h}|\sigma'),\ C,\ T(\widehat{h}|\sigma')\Big|\sigma\Big) > U_i(\widehat{h}|\sigma). \tag{8}$$

Consider the two histories

$$h_1 \equiv \Big(\widehat{h},\ a(\widehat{h}|\sigma'),\ C,\ T(\widehat{h}|\sigma')\Big) \text{ and } h_2 \equiv \Big(\widehat{h},\ a(\widehat{h}|\sigma'),\ C,\ T^*\Big).$$

By the construction of T^*, h_1 and h_2 are identical up to the transfers within coalition C. Since the convention σ respects secret transfers, it must be the case that for all $i \in N$,

$$U_i\Big(\widehat{h},\ a(\widehat{h}|\sigma'),\ C,\ T(\widehat{h}|\sigma')\Big|\sigma\Big) = U_i\Big(\widehat{h},\ a(\widehat{h}|\sigma'),\ C,\ T^*\Big|\sigma\Big).$$

Inequality (8) can therefore be re-written as, for every $i \in C$,

$$(1-\delta)u_i\Big(a(\widehat{h}|\sigma'),\ T^*\Big) + \delta U_i\Big(\widehat{h},\ a(\widehat{h}|\sigma'),\ C,\ T^*\Big|\sigma\Big) > U_i(\widehat{h}|\sigma). \tag{9}$$

According to Definition 4, inequality (9) implies that σ is not a stable convention.

A.4 Proof of Theorem 3 on p. 17

We prove a stronger statement: every stable convention σ guarantees that for every coalition C and every history $h \in \mathcal{H}$,

$$\sum_{i \in C} U_i(h|\sigma) \geq \underline{v}_C. \tag{10}$$

Consider a convention σ such that there exists a coalition C and history $\widehat{h}$ such that $\sum_{i\in C} U_i(\widehat{h}|\sigma) < \underline{v}_C$. We prove that σ must not be stable.

The convention σ recommends an alternative $a(h|\sigma)$ at every history $h \in \mathcal{H}$. We construct a profitable multi-shot deviation for coalition C. At every history $h \in \mathcal{H}$, let $d(h) \in BR_C(a(h|\sigma))$ be an alternative in coalition C's best-response to the recommended alternative. By the definition of $\underline{v}_C$ and $BR_C(.)$, it follows that $\sum_{i\in C} v_i(d(h)) \geq \underline{v}_C > \sum_{i\in C} U_i(\widehat{h}|\sigma)$. Since coalition C's total generated payoff from $d(h)$, $\sum_{i\in C} v_i(d(h))$, is higher than $\sum_{i\in C} U_i(\widehat{h}|\sigma)$, we can find transfers among players in C such that the payoff of each individual player $i \in C$ is higher than $U_i(\widehat{h}|\sigma)$. Formally, at every history h, there exist transfers $\widetilde{T}_C(h) \equiv [\widetilde{T}_{ij}(h)]_{i\in C, j\in N}$ such that $\widetilde{T}_{ij}(h) = 0$ for all $j \in N \backslash C$, and

$$v_i(d(h)) + \sum_{j\in C} \widetilde{T}_{ji}(h) - \sum_{j \in C} \widetilde{T}_{ij}(h) > U_i(\widehat{h}|\sigma)$$

for all $i \in C$. As a result, for each player $i \in C$, the experienced payoff from the stage-game outcome

$\Big(d(h),\ C,\ [\widetilde{T}_C(h), T_{-C}(h|\sigma)]\Big)$ satisfies

$$\begin{aligned} u_i\Big(d(h),\ [\widetilde{T}_C(h), T_{-C}(h|\sigma)]\Big) &= v_i(d(h)) + \sum_{j\in C}\widetilde{T}_{ji}(h) + \sum_{j\in N\backslash C} T_{ji}(h|\sigma) - \sum_{j\in N}\widetilde{T}_{ij}(h) \\ &\geq v_i(d(h)) + \sum_{j\in C}\widetilde{T}_{ji}(h) - \sum_{j\in C}\widetilde{T}_{ij}(h) \\ &> U_i(\widehat{h}|\sigma) \end{aligned}$$

where the weak inequality follows because $T_{ij}(h|\sigma) \geq 0$ for all $j \in N$, and $\widetilde{T}_{ij}(h) = 0$ for all $j \in N\backslash C$. Observe that the LHS concerns every history, including $\widehat{h}$ and those that follow. These steps prove that the multi-shot deviation σ' by coalition C, defined by $\sigma'(h) \equiv \Big(d(h),\ C,\ [\widetilde{T}_C(h), T_{-C}(h|\sigma)]\Big)$ for every history $h \in \mathcal{H}$, is profitable: $U_i(\widehat{h}|\sigma') > U_i(\widehat{h}|\sigma)$ for every $i \in C$. Lemma 1 then implies that σ is not stable.

B Supplementary Appendix (Not for Publication)

B.1 Preliminary Results

Below, we list two preliminary results used in our proofs.

Lemma 3. Suppose σ is a stable convention and let $diam(.)$ denote the diameter of a set. Then for any player i and any history $h \in \mathcal{H}$, the recommended transfers $\overline{T} = T(h|\sigma)$ from the convention must satisfy

$$\sum_{j \neq i} \overline{T}_{ji} \leq \frac{1+\delta}{1-\delta} \, diam(\{U(h|\sigma) : h \in \mathcal{H}\}) + diam(\mathcal{V}_{IR}^{\dagger})$$

Proof. At any history, the recommended alternative $\overline{a} = a(h|\sigma)$ and the recommended transfers $\overline{T} = T(h|\sigma)$ from the convention must satisfy

$$(1-\delta)[v_i(\overline{a}) + \sum_{j \neq i} \overline{T}_{ji}] + \delta \inf\{U_i(h|\sigma) : h \in \mathcal{H}\} \leq \sup\{U_i(h|\sigma) : h \in \mathcal{H}\}.$$

Otherwise, player i would have a profitable one-shot individual deviation from accepting all incoming transfers and reneging on all outgoing transfers. Rearranging terms, we have

$$\begin{aligned} \sum_{j \neq i} \overline{T}_{ji} &\leq \frac{\sup\{U_i(h|\sigma) : h \in \mathcal{H}\} - [(1-\delta)v_i(\overline{a}) + \delta \inf\{U_i(h|\sigma) : h \in \mathcal{H}\}}{(1-\delta)} \\ &= \frac{\sup\{U_i(h|\sigma) : h \in \mathcal{H}\}}{1-\delta} - \frac{\delta \inf\{U_i(h|\sigma) : h \in \mathcal{H}\}}{1-\delta} - v_i(\overline{a}) \end{aligned}$$

By the triangle inequality,

$$\sum_{j \neq i} \overline{T}_{ji} \leq \left| \frac{\sup\{U_i(h|\sigma) : h \in \mathcal{H}\}}{1-\delta} \right| + \left| \frac{\delta \inf\{U_i(h|\sigma) : h \in \mathcal{H}\}}{1-\delta} \right| + |v_i(\overline{a})| \,.$$

Since $|\sup\{U_i(h|\sigma) : h \in \mathcal{H}\}| \leq diam(\{U(h|\sigma) : h \in \mathcal{H}\})$, $|\inf\{U_i(h|\sigma) : h \in \mathcal{H}\}| \leq diam(\{U(h|\sigma) : h \in \mathcal{H}\})$, and $|v_i(\overline{a})| \leq diam(\mathcal{V}_{IR}^{\dagger})$, we have

$$\sum_{j \neq i} \overline{T}_{ji} \leq \frac{1+\delta}{1-\delta} \, diam(\{U(h|\sigma) : h \in \mathcal{H}\}) + diam(\mathcal{V}_{IR}^{\dagger})$$

□

Lemma 4. Suppose σ' is a profitable multi-shot deviation by coalition C from a stable convention σ, then there exists a profitable multi-shot coalitional deviation σ'' from σ, such that the set $\{\sum_{i \in C} U_i(h|\sigma'') : h \in \mathcal{H}\}$ is bounded.

Proof. We break this argument into two steps.

Step 1: We show that the set $\{\sum_{i \in C} U_i(h|\sigma') : h \in \mathcal{H}\}$ is bounded from above. It suffices to show that $\{\sum_{i \in C} u_i(\sigma'(h)) : h \in \mathcal{H}\}$ is bounded from above.

First we show that for player $i \notin C$, his stage-game values in σ' is bounded from below regardless of h. Since i is making the same outgoing transfers in $\sigma'(h)$ as in $\sigma(h)$, we have

$$u_i(\sigma'(h)) - u_i(\sigma(h)) = \Big[v_i(a(h|\sigma')) + \sum_{k \neq i} T_{ki}(h|\sigma')\Big] - \Big[v_i(a(h|\sigma)) + \sum_{k \neq i} T_{ki}(h|\sigma)\Big]$$

Rearranging terms, we have

$$\begin{aligned} u_i(\sigma'(h)) &= u_i(\sigma(h)) + \Big[v_i(a(h|\sigma')) - v_i(a(h|\sigma))\Big] - \sum_{k \neq i} T_{ki}(h|\sigma) + \sum_{k \neq i} T_{ki}(h|\sigma') \\ &\geq u_i(\sigma(h)) + \Big[v_i(a(h|\sigma')) - v_i(a(h|\sigma))\Big] - \sum_{k \neq i} T_{ki}(h|\sigma). \end{aligned} \tag{11}$$

By definition,

$$U_i(h|\sigma) = (1-\delta) u_i(\sigma(h)) + \delta U_i(h, \sigma(h)|\sigma),$$

or

$$u_i(\sigma(h)) = \frac{\delta U_i(h, \sigma(h)|\sigma) - U_i(h|\sigma)}{1-\delta}.$$

Plugging the above equation into inequality (11), we have

$$u_i(\sigma'(h)) \geq \frac{\delta U_i(h, \sigma(h)|\sigma) - U_i(h|\sigma)}{1-\delta} + \Big[v_i(a(h|\sigma')) - v_i(a(h|\sigma))\Big] - \sum_{k \neq i} T_{ki}(h|\sigma).$$

In the inequality above, $[\delta U_i(h, \sigma(h)|\sigma) - U_i(h|\sigma)]/(1-\delta)$ is bounded since σ has bounded continuation values; $\Big[v_i(a(h|\sigma')) - v_i(a(h|\sigma))\Big]$ is bounded because there are finite number of alternatives; and lastly, $\sum_{k \neq i} T_{ki}(h|\sigma)$ is bounded from above by Lemma 3. As a result, we can find number K such that $u_i(\sigma'(h)) \geq K$ for every history h and every player $i \notin C$.

After every history $h \in \mathcal{H}$, since the total experienced utility across all players must equal the total generated utility, and because $\overline{a}$ is a maximizer of $\sum_{i \in N} v_i(s)$,

$$\sum_{i \in C} u_i(\sigma'(h)) + \sum_{i \notin C} u_i(\sigma'(h)) \leq \sum_{i \in N} v_i(\overline{a}),$$

or

$$\sum_{i \in C} u_i(\sigma'(h)) \leq \sum_{i \in N} v_i(\overline{a}) - \sum_{i \notin C} u_i(\sigma'(h)).$$

After plugging in the bounds derived above, we have

$$\sum_{i \in C} u_i(\sigma'(h)) \leq \sum_{i \in N} v_i(\overline{a}) - (n - |C|) \times K \ \ \forall h \in \mathcal{H},$$

so the set $\{\sum_{i \in C} u_i(\sigma'(h)) : h \in \mathcal{H}\}$ is bounded from above.

Step 2: We show that $\{\sum_{i\in C} U_i(h|\sigma') : h \in \mathcal{H}\}$ is bounded from below. Suppose otherwise. We can construct another profitable deviation σ'' such that $\{\sum_{i\in C} U_i(h|\sigma'') : h \in \mathcal{H}\}$ is bounded: if $\sum_{i\in C} U_i(\widehat{h}|\sigma')$ falls below $\arg\min_{a\in A} \sum_{i\in C} v_i(a)$, at all histories following $\widehat{h}$ we ask C to block and refuse all outgoing transfers, while leaving the recommended alternative unchanged.

Formally, for a history $\widehat{h} \in \mathcal{H}$, let $F(\widehat{h}) \equiv \{h\widehat{h} : h \in \mathcal{H}\}$ denote the set of histories that can follow from $\widehat{h}$. Let $\underline{H}_C(\sigma') \equiv \{h \in \mathcal{H} : \sum_{i\in C} U_i(h|\sigma') < \min_{a\in A} \sum_{i\in C} v_i(a)\}$. Let $\mathbf{0}_C$ denote the vector of zero-valued transfers made from players in C. Define

$$\sigma''(h) = \begin{cases} \Big(a(h|\sigma), C, [\mathbf{0}_C, T_{-C}(h|\sigma)]\Big) & \forall h \in F(\widehat{h}) \text{ for some } \widehat{h} \in \underline{H}_C(\sigma') \\ \sigma'(h) & \text{otherwise} \end{cases}$$

By construction, the deviation σ'' has continuation values bounded below by $\arg\min_{a\in A} \sum_{i\in C} v_i(a)$, and is is still profitable. □

B.2 Proof of Theorem 2 on p. 14

Part 1: *For every $\delta \geq 0$, every stable convention gives each player i a payoff of at least $\underline{v}_i$.*

The proof mirrors that of the same part in Theorem 1, and so we elaborate on the necessary changes to the argument below. Consider any convention σ and player i such that $U_i(\emptyset|\sigma) < \underline{v}_i$. We first show that player i has a profitable multi-shot deviation from this convention.

We consider the following multi-shot deviation: in every period, player i blocks and best-responds to the convention, and refuses to make any outgoing transfers. Formally, this is a plan

$$\sigma'(h) = \Big((a(h|\sigma')), \{i\}, [\mathbf{0}_i, T_{-i}(h|\sigma)]\Big) \ \ \forall h \in \mathcal{H}$$

where $a(h|\sigma') \in BR_i\big(a(h|\sigma)\big)$ for every $h \in \mathcal{H}$. By the definition of $\underline{v}_i$, this multi-shot deviation gives i at least $\underline{v}_i$ after every history, so $U_i(\emptyset|\sigma') > U_i(\emptyset|\sigma)$.

By Assumption 1, the convention σ has bounded continuation value. Moreover, Assumption 1 implies that all incoming transfers player i receives on the path of the deviation σ' are also bounded (as proven in Lemma 3). As a result, player i faces a decision tree with bounded values in the deviation plan σ' and we can apply the standard one-shot deviation principle to prove that there exists a profitable one-shot deviation for $\{i\}$. Therefore, σ is not stable.

Part 2: *For every $u \in \mathcal{U}^{\dagger}_{IR}$, there is a $\underline{\delta} < 1$ such that for every $\delta \in (\underline{\delta}, 1)$, there exists a stable convention with a discounted payoff equal to u.*

Fix any $u^0 \in \mathcal{U}^{\dagger}_{IR}$. We argue below, using transfers, that we can find player-specific punishments for u^0: consider the vectors $\{u^i : i \in N\}$ defined by

$$u^i_j = \begin{cases} u_j - \epsilon & \text{if } j = i, \\ u_j + \frac{\epsilon}{n-1} & \text{if } j \neq i. \end{cases}$$

Observe that $\{u^i\}_{i=1}^n \subseteq \mathcal{U}_{IR}^\dagger$ when ϵ is sufficiently small, and that for all i, $u_i^i < u_i$ and for all $j \neq i$, $u_i^j > u_i^i$. Therefore, this is a vector of player-specific punishments.

Given these player-specific punishments, let $\kappa \in (0,1)$ be such that for every $\widetilde{\kappa} \in [\kappa, 1]$, the following is true for every i:

$$(1-\widetilde{\kappa}) v_i(\underline{a}_i) + \widetilde{\kappa} u_i^i > \underline{v}_i \tag{12}$$

$$\text{For every } j \neq i: \quad (1-\widetilde{\kappa}) v_j(\underline{a}_i) + \widetilde{\kappa} u_j^i > (1-\widetilde{\kappa}) \underline{v}_j + \widetilde{\kappa} u_j^j \tag{13}$$

By an argument identical to that which we saw in Theorem 1, there exists a value of $\kappa \in (0,1)$ such that the inequality holds for all $\widetilde{\kappa} \in [\kappa, 1]$, $i \in N$ and $j \in N \backslash \{i\}$. Let $L(\delta) \equiv \left\lceil \frac{\log \kappa}{\log \delta} \right\rceil$ where $\lceil \cdot \rceil$ is the ceiling function. As before, we use the property that $\lim_{\delta \to 1} \delta^{L(\delta)} = \kappa$.

Next we argue that there exists a finite set of payoff vectors whose convex hull contains the set $\mathcal{U}_{IR}^\dagger$.

Lemma 5. Let $\overline{a} \in \arg\max_{a \in A} \sum_{i \in N} v_i(a)$ and $\underline{a} \in \arg\min_{a \in A} \sum_{i \in N} v_i(a)$ two alternatives that maximize and minimize players' total generated payoffs, respectively. There exist payoff vectors $\{\widetilde{u}^1, \ldots, \widetilde{u}^M\} \subseteq \mathcal{U}(\overline{a}) \cup \mathcal{U}(\underline{a})$, such that $\mathcal{U}_{IR}^\dagger \subseteq \text{co}(\widetilde{u}^1, \ldots, \widetilde{u}^M)$.

Proof. By definition,

$$\mathcal{U}_{IR}^\dagger \subseteq \overline{\mathcal{U}}_{IR}^\dagger \equiv \left\{ u \in \mathsf{R}^n : \sum_{i \in N} v_i(\underline{a}) \leq \sum_{i \in N} u_i \leq \sum_{i \in N} v_i(\overline{a}) \text{ and } u_i \geq \underline{v}_i \forall i \in N \right\}.$$

Since $\overline{\mathcal{U}}_{IR}^\dagger$ is a bounded polyhedron, it is also a polytope. Let $x^1, \ldots, x^K$ be its vertices. Any point inside $\mathcal{U}_{IR}^\dagger$ can then be expressed as convex combinations of these vertices. Since $x^k \in \text{co}(\mathcal{U}(\overline{a}) \cup \mathcal{U}(\underline{a}))$ for all $1 \leq k \leq K$, for each k, there exist $\{\widetilde{u}^{k,1}, \ldots, \widetilde{u}^{k,m_k}\} \subseteq \mathcal{U}(\overline{a}) \cup \mathcal{U}(\underline{a})$ such that $x^k \subseteq \text{co}(\widetilde{u}^{k,1}, \ldots, \widetilde{u}^{k,m_k})$. As a result $\mathcal{U}_{IR}^\dagger \subseteq \text{co}(\cup_{1 \leq k \leq K} \{\widetilde{u}^{k,1}, \ldots, \widetilde{u}^{k,m_k}\})$. □

Lemma 5 implies that there exist payoff vectors $\{\widetilde{u}^1, \ldots, \widetilde{u}^M\} \subseteq \mathcal{U}(\overline{a}) \cup \mathcal{U}(\underline{a})$ such that $\mathcal{U}_{IR}^\dagger \subseteq \text{co}(\widetilde{u}^1, \ldots, \widetilde{u}^M)$, where $\widetilde{u}^m = u(\widetilde{a}^m, \widetilde{T}^m)$ for some alternative $\widetilde{a}^m \in \{\overline{a}, \underline{a}\}$ and transfers matrix $\widetilde{T}^m$ for each $m = 1, \ldots, M$. Lemma 2 then guarantees that for any $\epsilon > 0$, there exists $\underline{\delta} \in (0,1)$ such that for all $\delta \in (\underline{\delta}, 1)$, there exist sequences $\left\{ \{a^{i,\tau}, T^{i,\tau}\}_{\tau=0}^\infty : i = 0, 1, \ldots, n \right\}$ such that for each i and t, $(1-\delta) \sum_{\tau=0}^\infty \delta^\tau u(a^{i,\tau}, T^{i,\tau}) = u^i$ and $\left|\left| u^i - (1-\delta) \sum_{\tau=t}^\infty \delta^\tau u(a^{i,\tau}, T^{i,\tau}) \right|\right| < \epsilon$. We fix an $\epsilon < (1-\kappa) \min\{ \min_{i, j \neq i}(u_i^j - u_i^i), \min_i u_i^i - \underline{v}_i\}$, and given that ϵ, consider δ exceeding the appropriate $\underline{\delta}$.

Now we describe the convention that we use to sustain u^0. Let $\mathbf{0}$ denote the transfer matrix where all players make no transfers. Consider the convention represented by the automaton $(W, w(0,0), f, \gamma)$, where

- $W \equiv \{w(d,\tau) | 0 \leq d \leq n, \tau \geq 0\} \cup \{\underline{w}(i,\tau) | 1 \leq i \leq n, 0 \leq \tau < L(\delta)\}$ is the set of possible states;
- $w(0,0)$ is the initial state;
- $f : W \to \mathcal{O}^{TU}$ is the output function, where $f(w(d,\tau)) = (a^{d,\tau}, \emptyset, T^{d,\tau})$ and $f(\underline{w}(i,\tau)) = (\underline{a}_i, \emptyset, \mathbf{0})$;

- $\gamma : W \times \mathcal{O}^{TU} \to W$ is the transition function. For states of the form $w(d,\tau)$, the transition is

$$\gamma\big(w(d,\tau),(a,C,T)\big) = \begin{cases} \underline{w}(j^*,0) & \text{if } C \neq \emptyset,\ j^* = \arg\min_{j\in C}\{u_j(a,T) - u_j^{d,t}\} \\ w(d,\tau+1) & \text{otherwise} \end{cases}$$

For states in $\{\underline{w}(i,\tau) | 0 \leq \tau < L(\delta)-1\}$,

$$\gamma\big(\underline{w}(i,\tau),(a,C,T)\big) = \begin{cases} \underline{w}(j^*,0) & \text{if } \{C \neq \emptyset\} \cap \big(\{u_i(a,T) > \underline{v}_i\} \cup \{i \notin C\}\big) \\ & j^* = \arg\min_{C\setminus\{i\}}\{u_j(a,T) - v_j(\underline{a}_i)\} \\ \underline{w}(i,0) & \text{if } \{C \neq \emptyset\} \cap \{u_i(a,T) \leq \underline{v}_i\} \cap \{i \in C\} \\ \underline{w}(i,\tau+1) & \text{otherwise} \end{cases}$$

For states of the form $\underline{w}(i, L(\delta)-1)$, the transition is

$$\gamma\big(\underline{w}(i,L(\delta)-1),(a,C,T)\big) = \begin{cases} \underline{w}(j^*,0) & \text{if } \{C \neq \emptyset\} \cap \big(\{u_i(a,T) > \underline{v}_i\} \cup \{i \notin C\}\big) \\ & j^* = \arg\min_{C\setminus\{i\}}\{u_j(a,T) - v_j(\underline{a}_i)\} \\ \underline{w}(i,0) & \text{if } \{C \neq \emptyset\} \cap \{u_i(a,T) \leq \underline{v}_i\} \cap \{i \in C\} \\ w(i,0) & \text{otherwise} \end{cases}$$

The convention represented by the above automaton yields payoff profile u^0. By construction, the continuation values in different states, $V(\cdot)$, satisfy:

$$\Big|\Big|u^d - V(w(d,\tau))\Big|\Big| < \epsilon, \quad \tau = 0,1,\ldots$$

$$V(\underline{w}(i,\tau)) = (1-\delta^{L(\delta)-\tau})v(\underline{a}_i) + \delta^{L(\delta)-\tau}V(w(i,0)), \quad 0 \leq \tau \leq L(\delta)-1$$

In the NTU environment, since the feasible payoff set $\mathcal{V}^\dagger$ is bounded, whenever a coalition deviates, we can find number $B > 0$ that bounds every player's stage-game payoff. With transfers, however, players' stage-game payoffs are no longer bounded: in particular, we do not impose a priori bounds on the transfers made among members of the blocking coalition. This makes it more difficult to deter coalitional deviations, since players can use transfers to compensate each other.

Regardless, the *total* stage-game payoff of the deviating coalition is still bounded, so at least one member still has a bounded payoff. The definition of j^* in the automaton above ensures that the "scapegoat" selected by the convention can be effectively deterred as $\delta \to 1$. It remains to show that this convention is stable. This is the next step.

Stability in states of the form $w(d,\tau)$: If a coalition $C \neq \emptyset$ blocks in automaton state $w(d,\tau)$ and the outcome $(\widehat{a}, C, \widehat{T})$ is realized, the convention punishes $j^* = \arg\min_{j\in C}\{u_j(\widehat{a},\widehat{T}) - u_j^{d,\tau}\}$. It follows that

$$
\begin{aligned}
u_{j^*}(\widehat{a},\widehat{T}) - u_{j^*}^{d,\tau} &\leq \frac{1}{|C|}\Big[\sum_{j\in C} u_j(\widehat{a},\widehat{T}) - \sum_{j\in C} u_j^{d,\tau}\Big] \\
&\leq \frac{1}{|C|}\left[\max_{a\in A}\sum_{j\in C} v_j(a) - \min_{a\in A}\sum_{j\in C} v_j(a) + \sum_{j\in C}\sum_{k\notin C} T_{jk}^{d,\tau}\right] \\
&\leq \frac{1}{|C|}\left[\max_{a\in A}\sum_{j\in C} v_j(a) - \min_{a\in A}\sum_{j\in C} v_j(a) + \max_{1\leq m\leq M}\sum_{j\in C}\sum_{k\notin C} \widetilde{T}_{jk}^{m}\right],
\end{aligned}
$$

where the first inequality follows from the minimum among a set of numbers being less than their average; the second inequality follows from the difference between $\sum_{j\in C} u_j(\widehat{a},\widehat{T})$ and $\sum_{j\in C} u_j^{d,\tau}$ resulting from either differences in the generated payoffs from the realized alternative, or the outgoing transfers to players in $N\backslash C$; lastly, the third inequality follows because all $T^{d,\tau}$ are drawn from $\{\widetilde{T}^m\}_{m=1}^M$. Rearranging terms:

$$
u_{j^*}(\widehat{a},\widehat{T}) \leq \max_{j\in N, 1\leq m\leq M} \widetilde{u}_j^m + \max_{C\subseteq N, C\neq\emptyset} \frac{1}{|C|}\left[\max_{a\in A}\sum_{j\in C} v_j(a) - \min_{a\in A}\sum_{j\in C} v_j(a) + \max_{1\leq m\leq M}\sum_{j\in C}\sum_{k\notin C} \widetilde{T}_{jk}^{m}\right].
$$

In the inequality above, each term in the RHS is independent of δ and (d,τ). Thus, we can find a uniform bound B_1 such that $u_{j^*}(\widehat{a},\widehat{T}) < B_1$ for every δ and (d,τ).

Given this bound, we can use the analogue of the argument used in Theorem 1. For all τ, j^* obtains a payoff greater than $u_{j^*}^d - \epsilon$. By deviating, j^* obtains a payoff less than

$$
(1-\delta)B_1 + \delta V_{j^*}(\underline{w}(j^*,0)) = (1-\delta)B_1 + \delta\left[(1-\delta^{L(\delta)})v_j(\underline{a}_j) + \delta^{L(\delta)} u_{j^*}^{j^*}\right]
$$

By the exact same argument as in Theorem 1, this one-shot deviation is unprofitable for j^* and hence, for coalition C if δ is sufficiently high.

Stability in states of the form $\underline{w}(i,\tau)$: Suppose coalition $C \neq \emptyset$ blocks, leading to the outcome $(\widehat{a}, C, \widehat{T})$, We prove that at least one player in C does not find this one-shot deviation to be profitable. There are two cases to consider:

Case 1: $i \in C$ and $u_i(\widehat{a},\widehat{T}) \leq \underline{v}_i$. In this case, the convention selects player i to be the scapegoat. She finds this deviation to be unprofitable if

$$
(1-\delta^{L(\delta)-\tau})v_i(\underline{a}_i) + \delta^{L(\delta)-\tau}u_i^i \geq (1-\delta)\underline{v}_i + \delta(1-\delta^{L(\delta)})v_i(\underline{a}_i) + \delta^{L(\delta)+1}u_i^i. \tag{14}
$$

which follows from Inequality (12) for sufficiently high δ (using steps identical to the analogous argument in Theorem 1).

Case 2: Either $i \notin C$ or $u_i(\widehat{a},\widehat{T}) > \underline{v}_i$. In this case it cannot be that $C = \{i\}$ because otherwise $u_i(\widehat{a},\widehat{T}) \leq \underline{v}_i$. The convention then punishes $j^* = \arg\min_{j\in C\backslash\{i\}}\{u_j(\widehat{a},\widehat{T}) - v_j(\underline{a}_i)\}$. Denote $C\backslash\{i\}$ by

C'. It follows that

$$\begin{aligned} u_{j^*}(\widehat{a},\widehat{T}) - v_{j^*}(\underline{a}_i) &\leq \frac{1}{|C'|}\Big[\sum_{j\in C'} u_j(\widehat{a},\widehat{T}) - \sum_{j\in C'} v_j(\underline{a}_i)\Big] \\ &= \frac{1}{|C'|}\Big[\sum_{j\in C'\cup\{i\}} u_j(\widehat{a},\widehat{T}) - \sum_{j\in C'\cup\{i\}} v_j(\underline{a}_i) + v_i(\underline{a}_i) - u_i(\widehat{a},\widehat{T})\Big] \\ &= \frac{1}{|C'|}\Big[\sum_{j\in C'\cup\{i\}} u_j(\widehat{a},\widehat{T}) - \sum_{j\in C'\cup\{i\}} v_j(\underline{a}_i)\Big] + \frac{1}{|C'|}\Big[v_i(\underline{a}_i) - u_i(\widehat{a},\widehat{T})\Big]. \qquad (15) \end{aligned}$$

Furthermore,

$$\sum_{j\in C'\cup\{i\}} u_j(\widehat{a},\widehat{T}) - \sum_{j\in C'\cup\{i\}} v_j(\underline{a}_i) \leq \max_{a\in A} \sum_{j\in C'\cup\{i\}} v_j(a) - \min_{a\in A} \sum_{j\in C'\cup\{i\}} v_j(a). \qquad (16)$$

The inequality above follows since in the outcome $(\widehat{a}, C, \widehat{T})$, all players outside of $C' \cup \{i\}$ are following the recommendation from automaton state $\underline{w}(i,\tau)$ and making zero transfers.

Finally, if $i \notin C$ then $u_i(\widehat{a},\widehat{T}) \geq \min_{a\in A} v_i(a)$, since player i is following the recommendation from automaton state $\underline{w}(i,\tau)$ and makes zero outgoing transfers in the outcome $(\widehat{a}, C, \widehat{T})$; otherwise if $i \in C$ then $u_i(\widehat{a},\widehat{T}) > \underline{v}_i$. In either case,

$$v_i(\underline{a}_i) - u_i(\widehat{a},\widehat{T}) \leq v_i(\underline{a}_i) - \min\{\underline{v}_i, \min_{a\in A} v_i(a)\} \qquad (17)$$

Plugging inequalities (16) and (17) into inequality (15), we have

$$\begin{aligned} u_{j^*}(\widehat{a},\widehat{T}) - v_{j^*}(\underline{a}_i) &\leq \frac{1}{|C'|}\left[\max_{a\in A} \sum_{j\in C'\cup\{i\}} v_j(a) - \min_{a\in A} \sum_{j\in C'\cup\{i\}} v_j(a) - v_i(\underline{a}_i) + \min\{\underline{v}_i, \min_{a\in A} v_i(a)\}\right] \\ &\equiv b_2(i, C') \end{aligned}$$

As a result, across all states $\underline{w}(i,\tau)$ and all possible blocking coalitions $C \neq \emptyset$, we have

$$u_{j^*}(\widehat{a},\widehat{T}) \leq \max_{i\in N, C'\subseteq N\setminus\{i\}, C'\neq\emptyset} b_2(i, C')$$

In the inequality above, all the terms in the RHS are independent of δ. Therefore, we can find a uniform bound B_2 such that $u_{j^*}(\widehat{a},\widehat{T}) < B_2$ for every δ. We use these steps to show that player j^* finds this one-shot deviation to be unprofitable. Player j^* does not benefit from this deviation if

$$(1-\delta^{L(\delta)-\tau})v_{j^*}(\underline{a}_i) + \delta^{L(\delta)-\tau}u^i_{j^*} \geq (1-\delta)B_2 + \delta(1-\delta^{L(\delta)})v_{j^*}(\underline{a}_{j^*}) + \delta^{L(\delta)+1}u^{j^*}_{j^*}. \qquad (18)$$

This inequality is satisfied for sufficiently high δ, and the argument follows the same steps as that of the analogous part of Theorem 1.

B.3 Proof of Theorem 4 on p. 19

Part 1: *Under secret transfers, for every $\delta \geq 0$, every stable convention implements payoffs only within the efficient β-core.*

We first argue that for every stable convention σ, an efficient alternative must be chosen at every history: $a(h|\sigma) \in \overline{A}$ at every $h \in \mathcal{H}$. Suppose otherwise that $\widehat{a} \equiv a(\widehat{h}|\sigma) \notin \overline{A}$ for some history $\widehat{h}$, so that $\sum_{i\in N} v_i(\widehat{a}) < \max_{a\in A} \sum_{i\in N} v_i(a)$. It follows that

$$\begin{aligned}\sum_{i\in N} U_i(\widehat{h}|\sigma) &= (1-\delta)\sum_{i\in N} v_i(\widehat{a}) + \delta \sum_{i\in N} U_i(\widehat{h}, \widehat{a}, \emptyset, \widehat{T}|\sigma) \\ &< (1-\delta)\max_{a\in A}\sum_{i\in N} v_i(a) + \delta \sum_{i\in N} U_i(\widehat{h}, \widehat{a}, \emptyset, \widehat{T}|\sigma) \\ &\leq (1-\delta)\max_{a\in A}\sum_{i\in N} v_i(a) + \delta \max_{a\in A}\sum_{i\in N} v_i(a) \\ &= \max_{a\in A}\sum_{i\in N} v_i(a) \\ &= \underline{v}_N\end{aligned}$$

where the strict inequality follows from the definition of $\widehat{a}$, the weak inequality follows from the total experienced payoff being the total generated payoff in every period, and the final equality follows from Assumption 2. This strict inequality contradicts Inequality (10) established in the proof of Theorem 3.

Having argued that a stable convention must choose actions in $\overline{A}$ at every history, the remainder of the proof is identical, but replacing A with $\overline{A}$.

Part 2: *If the strict efficient β-core is non-empty, then for every payoff profile $u \in \mathcal{B}^s$, there is a $\underline{\delta} < 1$ such that for every $\delta \in (\underline{\delta}, 1)$, there exists a stable convention with a discounted payoff equal to u.*

Fix any payoff vector $u^N \in \mathcal{B}^s$. Below we construct "coalition-specific" punishments for all coalitions but the grand coalition.

Lemma 6. There exist coalition-specific punishments $\{u^C : C \in \mathcal{C}\backslash\{N\}\}$ in $\mathcal{B}^s$ such that

$$\sum_{i\in C} u_i^C < \sum_{i\in C} u_i^N \tag{19}$$

and for any coalition $C' \neq C$

$$\sum_{i\in C} u_i^C < \sum_{i\in C} u_i^{C'} \tag{20}$$

Proof. For any coalition $C \in \mathcal{C}\backslash\{N\}$, consider the vector u^C defined by

$$u_i^C = \begin{cases} u_i^N - \frac{\epsilon}{|C|} & i \in C \\ u_i^N + \frac{\epsilon}{|N\backslash C|} & i \notin C \end{cases}$$

Compared to the payoff vector u^N, in u^C every player in C is charged equally, with a total summing up to ϵ; by contrast, players outside of C are paid equally, with a total of amount also summing up to ϵ. The ϵ may be set sufficiently small to ensure all u^C's are in $\mathcal{B}^s$.

We show that these vectors satisfy inequalities (19) and (20). By construction, $\sum_{i\in C} u_i^C = \sum_{i\in C} u_i^N - \epsilon < \sum_{i\in C} u_i^N$, so Inequality (19) is satisfied. To verify (20), consider two coalitions $C, C' \in \mathcal{C}\backslash\{N\}$ with $C \neq C'$. Coalition C can be partitioned as the union of two components $C = (C\backslash C') \cup (C \cap C')$. So

$$\begin{aligned}
\sum_{i\in C} u_i^{C'} &= \sum_{i\in C\backslash C'} u_i^{C'} + \sum_{i\in C\cap C'} u_i^{C'} \\
&= \left[\sum_{i\in C\backslash C'} u_i^N + \frac{|C\backslash C'|}{|N\backslash C'|}\epsilon\right] + \left[\sum_{i\in C\cap C'} u_i^N - \frac{|C\cap C'|}{|C'|}\epsilon\right] && (21)\\
&= \sum_{i\in C} u_i^N - \left[\frac{|C\cap C'|}{|C'|} - \frac{|C\backslash C'|}{|N\backslash C'|}\right]\epsilon \\
&> \sum_{i\in C} u_i^N - \epsilon && (22)\\
&= \sum_{i\in C} u_i^C
\end{aligned}$$

Equality (21) follows since compared to u^N, $u^{C'}$ gives every player outside of C' an extra payoff of $\frac{\epsilon}{|N\backslash C|}$, while lowering the payoff of every player inside C' by $\frac{\epsilon}{|C'|}$. Since $C \neq C'$, either $C\backslash C' \neq \emptyset$ or $C \cap C' \neq C'$ must be true; in other words, either $\frac{|C\backslash C'|}{|N\backslash C'|} > 0$ or $\frac{|C\cap C'|}{|C'|} < 1$. In either cases, inequality (22) follows, which verifies (20). □

Using Lemma 6, let $\{u^C : C \in \mathcal{C}\backslash\{N\}\}$ be the vector of coalition-specific punishments for u^N. Fix an alternative $\overline{a} \in \overline{A}$. Since $\{u^C : C \in \mathcal{C}\} \subseteq \mathcal{U}(\overline{a})$, we can find transfer matrices $\{T^C : C \in \mathcal{C}\}$ such that $u(\overline{a}, T^C) = u^C$ for all $C \in \mathcal{C}$.

Let $\underline{a}_C^e \in \arg\min_{a\in\overline{A}} \max_{a'\in E_C} \sum_{i\in C} v_i(a')$ be an efficient alternative that can be used to minmax coalition C. Note that by construction, $\sum_{i\in C} v_i(\underline{a}_C^e) \leq \underline{v}_C^e$. Given the coalition-specific punishments, let $\kappa \in (0,1)$ be such that for every $\widetilde{\kappa} \in [\kappa, 1]$, the following is true for every coalition C:

$$(1-\widetilde{\kappa})\sum_{i\in C} v_i(\underline{a}_C^e) + \widetilde{\kappa}\sum_{i\in C} u_C^i > \underline{v}_C^e \tag{23}$$

$$\text{For every } C' \neq C: \quad (1-\widetilde{\kappa})\sum_{i\in C'} v_i(\underline{a}_C^e) + \widetilde{\kappa}\sum_{i\in C'} u_i^C > (1-\widetilde{\kappa})\underline{v}_{C'}^e + \widetilde{\kappa}\sum_{i\in C'} u_i^{C'}. \tag{24}$$

Inequality (23) implies that in terms of total value, coalition C is willing to bear the cost of $\sum_{i\in C} v_i(\underline{a}_i^e)$ with the promise of transitioning into its coalition-specific punishment rather than staying at its minmax. Inequality (24) implies that every coalition prefers punishing other coalitions than being punished itself. By an argument identical to that we saw in Theorem 1, there exists a value of $\kappa \in (0,1)$ such that all the inequalities above hold for all $\widetilde{\kappa} \in [\kappa, 1]$, $i \in N$ and $j \in N\backslash\{i\}$. Let $L(\delta) \equiv \left\lceil \frac{\log\kappa}{\log\delta} \right\rceil$ where $\lceil\cdot\rceil$ is the ceiling function. As before, we use the property that $\lim_{\delta\to 1} \delta^{L(\delta)} = \kappa$.

We describe the convention that we use to sustain u^N. Let $\mathbf{0}$ denote the transfer matrix where all player make zero transfers. Consider the convention represented by the automaton $(W, w(N), f, \gamma)$, where

- $W \equiv \big\{w(C) : C \in \mathcal{C}\big\} \cup \big\{\underline{w}(C, \tau) | C \in \mathcal{C}\backslash\{N\}, 0 \leq \tau < L(\delta)\big\}$ is the set of possible states;
- $w(N)$ is the initial state;
- $f : W \to \mathcal{O}^{TU}$ is the output function: for every $C \in \mathcal{C}$, $f(w(C)) = (\overline{a}, \emptyset, T^C)$; for every $C \in \mathcal{C}\backslash\{N\}$, $f(\underline{w}(C, \tau)) = (\underline{a}^e_C, \emptyset, \mathbf{0})$;
- $\gamma : W \times \mathcal{O}^{TU} \to W$ is the transition function. For states of the form $w(C)$, the transition is

$$\gamma\big(w(C), (a, C', T)\big) = \begin{cases} \underline{w}(C', 0) & \text{if } C' \notin \{N\} \\ w(C) & \text{otherwise} \end{cases}$$

For states in $\{\underline{w}(C, \tau) | 0 \leq \tau < L(\delta) - 1\}$, the transition is

$$\gamma(\underline{w}(C, \tau), (a, C', T)) = \begin{cases} \underline{w}(C', 0) & \text{if } C' \notin \{N\} \\ \underline{w}(C, \tau + 1) & \text{otherwise} \end{cases}$$

For states of the form $\underline{w}(C, L(\delta) - 1)$, the transition is

$$\gamma(\underline{w}(C, L(\delta) - 1), (a, C', T)) = \begin{cases} \underline{w}(C', 0) & \text{if } C' \notin \{N\} \\ w(C), & \text{otherwise} \end{cases}$$

The convention represented by the above automaton yields payoff profile u^0. By construction, the continuation values in different states, $V(\cdot)$, satisfy:

$$V(w(C)) = u^C, \ \ \forall C \in \mathcal{C}$$

$$V(\underline{w}(C, \tau)) = (1 - \delta^{L(\delta) - \tau}) v(\underline{a}^e_C) + \delta^{L(\delta) - \tau} V(w(C)), \ \ \forall C \in \mathcal{C}, 0 \leq \tau \leq L(\delta) - 1$$

Next, we check that this automaton representation has no profitable one-shot coalitional deviation for any $C \in \mathcal{C}$. To this end, it suffices to check that for each $C \in \mathcal{C}$, no deviation can result in higher *total* value for C: if this is true, then it is impossible to make every player $i \in C$ better off.

Since deviations by the grand coalition do not change continuation play, and the recommended alternatives are always efficient in all possible automaton states, the grand coalition N does not have profitable deviations. It remains to check that none of the other coalitions have profitable one-shot deviations. This is the next step.

Stability in states of the form $w(C)$: Suppose coalition C' blocks and the outcome (a', C', T') is realized. The total payoff of C' from this outcome satisfies

$$\begin{aligned}\sum_{i\in C'} u_i(a',T') &= \sum_{i\in C'} v_i(a') + \sum_{i\in C'}\sum_{j\in N\setminus C'} T'_{ji} - \sum_{i\in C'}\sum_{j\in N\setminus C'} T'_{ij} \\ &\leq \sum_{i\in C'} v_i(a') + \sum_{i\in C'}\sum_{j\in N\setminus C'} T'_{ji} \\ &\leq \max_{a\in A}\sum_{i\in C'} v_i(a) + \max_{C\in\mathcal{C}}\sum_{i\in C'}\sum_{j\in N\setminus C'} T^C_{ji} \equiv b_1(C').\end{aligned}$$

The final inequality follows from players outside of the blocking coalition C' making the same transfers as recommended by the convention, and T^C being the transfers that are recommended in automaton state $w(C)$. As a result, we can find number $B_1 \equiv \max_{C'\in\mathcal{C}\setminus\{N\}} b_1(C')$ that the total stage-game payoff for any deviation coalition from any automaton state is less than B_1. Crucially, B_1 does not depend on δ.

Consider a one-shot deviation to (a, C', T) by coalition $C' \in \mathcal{C}\setminus\{N\}$. Coalition C' has total payoff $\sum_{i\in C'} u_i^C$ without deviating. By deviating, C' obtains a total payoff less than

$$(1-\delta)B_1 + \delta\sum_{i\in C'} V_i(\underline{w}(C',0)) = (1-\delta)B_1 + \delta\left[(1-\delta^{L(\delta)})\sum_{i\in C'} v_i(\underline{a}^e_{C'}) + \delta^{L(\delta)}\sum_{i\in C'} u_i^{C'}\right]$$

For the deviation to be profitable, the total value for C' must be higher. So the one-shot deviation is unprofitable if the above term is no more than $\sum_{i\in C'} u_i^C$. We prove that this is the case both for $C' \neq C$ and $C' = C$.

First consider $C' \neq C$. Observe that

$$\begin{aligned}&\lim_{\delta\to 1}(1-\delta)B_1 + \delta\left[(1-\delta^{L(\delta)})\sum_{i\in C'} v_i(\underline{a}^e_{C'}) + \delta^{L(\delta)}\sum_{i\in C'} u_i^{C'}\right] \\ &\qquad = (1-\kappa)\sum_{i\in C'} v_i(\underline{a}^e_{C'}) + \kappa\sum_{i\in C'} u_i^{C'} < \sum_{i\in C'} u_i^{C'} < \sum_{i\in C'} u_i^C.\end{aligned}$$

It follows that the one-shot coalition deviation is not profitable.

Now suppose that $C' = C$. The deviation payoff being less than $\sum_{i\in C'} u_i^{C'}$ can be re-written as

$$(1-\delta)(B_1 - \sum_{i\in C'} u_i^{C'}) \leq \delta(1-\delta^{L(\delta)})(\sum_{i\in C'} u_i^{C'} - \sum_{i\in C'} v_i(\underline{a}^e_{C'}))$$

As $\delta \to 1$, the LHS converges to 0. Because $\lim_{\delta\to 1}\delta^{L(\delta)} = \kappa$, the RHS converges to $(1-\kappa)(\sum_{i\in C'} u_i^{C'} - \sum_{i\in C'} v_i(\underline{a}^e_{C'}))$. So the above inequality holds, and therefore, there is no profitable one-shot deviation if δ is sufficiently high.

Stability in states of the form $\underline{w}(C,\tau)$: Suppose coalition C' blocks and the outcome (a', C', T') is realized. Coalition C''s total payoff from this outcome satisfies

$$\begin{aligned}\sum_{i\in C'} u_i(a',T') &= \sum_{i\in C'} v_i(a') + \sum_{i\in C'}\sum_{j\in N\setminus C'} T'_{ji} - \sum_{i\in C'}\sum_{j\in N\setminus C'} T'_{ij}\\ &\le \sum_{i\in C'} v_i(a') + \sum_{i\in C'}\sum_{j\in N\setminus C'} T'_{ji}\\ &\le \max_{a\in A}\sum_{i\in C'} v_i(a) \equiv b_2(C').\end{aligned}$$

The inequality above follows because, in states $\underline{w}(C,\tau)$, the convention recommends players to make zero transfers, so there are no incoming transfers from players outside of the blocking coalition C'. As a result, we can find number $B_2 \equiv \max_{C'\in\mathcal{C}} b_2(C')$ that the total stage-game payoff for any deviating coalition from any automaton state is less than B_2. Note that B_2 does not depend on δ. We now prove that no coalition has a profitable one-shot deviation.

Case 1: $C' = C$. by the definition of $\underline{a}^e_C$, when coalition C blocks the outcome $(\underline{a}^e_C, \emptyset, \mathbf{0})$, its stage-game payoff cannot exceed $\underline{v}^e_C$. As a result, coalition C has no profitable deviation if

$$(1-\delta^{L(\delta)-\tau})\sum_{i\in C} v_i(\underline{a}^e_C) + \delta^{L(\delta)-\tau}\sum_{i\in C} u_i^C \ge (1-\delta)\underline{v}^e_C + \delta(1-\delta^{L(\delta)})\sum_{i\in C} v_i(\underline{a}^e_C) + \delta^{L(\delta)+1}\sum_{i\in C} u_i^C. \tag{25}$$

Because $\sum_{i\in C} u^i_C > \underline{v}^e_C \ge \sum_{i\in C} v_i(\underline{a}^e_C)$, it suffices to show that

$$(1-\delta^{L(\delta)})\sum_{i\in C} v_i(\underline{a}^e_C) + \delta^{L(\delta)}\sum_{i\in C} u_i^C \ge (1-\delta)\underline{v}^e_C + \delta(1-\delta^{L(\delta)})\sum_{i\in C} v_i(\underline{a}^e_C) + \delta^{L(\delta)+1}\sum_{i\in C} u_i^C.$$

Re-arranging terms:

$$(1-\delta)(1-\delta^{L(\delta)})\sum_{i\in C} v_i(\underline{a}^e_C) + (1-\delta)\delta^{L(\delta)}\sum_{i\in C} u_i^C \ge (1-\delta)\underline{v}^e_C.$$

Dividing by $(1-\delta)$ yields:

$$(1-\delta^{L(\delta)})\sum_{i\in C} v_i(\underline{a}^e_C) + \delta^{L(\delta)}\sum_{i\in C} u_i^C \ge \underline{v}^e_C.$$

Now taking the limit of the LHS as $\delta \to 1$ yields Inequality (23), and hence Inequality (25) is true for sufficiently high δ.

Case 2: $C' \neq C$. Coalition C' finds no profitable one-shot deviation to be unprofitable if

$$(1-\delta^{L(\delta)-\tau})\sum_{i\in C'} v_i(\underline{a}^e_C) + \delta^{L(\delta)-\tau}\sum_{i\in C'} u_i^C \ge (1-\delta)B_2 + \delta(1-\delta^{L(\delta)})\sum_{i\in C'} v_i(\underline{a}^e_{C'}) + \delta^{L(\delta)+1}\sum_{i\in C'} u_i^{C'}. \tag{26}$$

We prove that this inequality is satisfied if δ is sufficiently high. Examining the LHS, observe that for all

τ such that $0 \leq \tau \leq L(\delta) - 1$,

$$\begin{aligned}\lim_{\delta\to 1}\left[(1-\delta^{L(\delta)-\tau})\sum_{i\in C'} v_i(\underline{a}_C^e) + \delta^{L(\delta)-\tau}\sum_{i\in C'} u_i^C\right] &= \lim_{\delta\to 1}\left[\left(1-\frac{\kappa}{\delta^\tau}\right)\sum_{i\in C'} v_i(\underline{a}_C^e) + \frac{\kappa}{\delta^\tau}\sum_{i\in C'} u_i^C\right]\\ &= (1-\widetilde{\kappa})\sum_{i\in C'} v_i(\underline{a}_C^e) + \widetilde{\kappa}\sum_{i\in C'} u_i^C\end{aligned}$$

for some $\widetilde{\kappa} \in [\kappa, 1]$.[31]

Examining the RHS of (26), observe that

$$\begin{aligned}&\lim_{\delta\to 1}\left[(1-\delta)B_2 + \delta(1-\delta^{L(\delta)})\sum_{i\in C'} v_i(\underline{a}_{C'}^e) + \delta^{L(\delta)+1}\sum_{i\in C'} u_i^{C'}\right]\\ =&\lim_{\delta\to 1}\left[(1-\delta^{L(\delta)})\sum_{i\in C'} v_i(\underline{a}_{C'}^e) + \delta^{L(\delta)}\sum_{i\in C'} u_i^{C'}\right] = (1-\kappa)\sum_{i\in C'} v_i(\underline{a}_{C'}^e) + \kappa\sum_{i\in C'} u_i^{C'}\\ \leq&(1-\kappa)\underline{v}_{C'}^e + \kappa\sum_{i\in C'} u_i^{C'} \leq (1-\widetilde{\kappa})\underline{v}_{C'}^e + \widetilde{\kappa}\sum_{i\in C'} u_i^{C'},\end{aligned}$$

where the first equality follows from taking limits, the second from $\lim_{\delta\to 1}\delta^{L(\delta)} = \kappa$, the first weak inequality follows from $\sum_{i\in C'} v_i(\underline{a}_{C'}^e) \leq \underline{v}_{C'}^e < \sum_{i\in C'} u_i^C$, and the second weak inequality follows from $\widetilde{\kappa} \geq \kappa$ and $\underline{v}_{C'}^e < \sum_{i\in C'} u_i^{C'}$. Since $\widetilde{\kappa} \in [\kappa, 1]$, (24) delivers that $(1-\widetilde{\kappa})\sum_{i\in C'} v_i(\underline{a}_C^e) + \widetilde{\kappa}\sum_{i\in C'} u_i^C$ is strictly higher than $(1-\widetilde{\kappa})\underline{v}_{C'}^e + \widetilde{\kappa}\sum_{i\in C'} u_i^{C'}$. This term guarantees that (26) holds for sufficiently high δ.

B.4 Proof of Theorem 6 on p. 21

The argument comprises several steps. Throughout this argument, we restrict attention to *stationary conventions*, i.e., those in which the recommendation is identical across all on-path histories.

First, we construct punishments for each player. Lemmas 7 and 8 establish the existence of stable conventions σ^i that guarantee $U_i(\emptyset|\sigma^i) = 0$ for each player i. The case where there is a single veto player ($|D| = 1$), analyzed in Lemma 7, requires the discount factor to be sufficiently high. The case where there are two or more veto players ($|D| \geq 2$), analyzed in Lemma 8, applies for every discount factor.

Our second step compares the set of outcomes enforced using the above stable conventions as punishments with those enforced by punishments where every member of a deviating coalition simultaneously obtains 0. Lemma 9 proves that these two sets are identical.

The third step (Lemma 10) shows, given the earlier two steps, that a stationary convention is stable if and only if every winning coalition obtains at least $(1 - \delta)$.

The proof for the secret transfers component of our result follows immediately from Theorem 3. The proof for the single veto-player case, in both the NTU and perfectly monitored transfers settings, follows from combining Lemmas 7, 9 and 10. The proof for the multiple veto-player case, in both the NTU and perfectly monitored transfers settings, follows from combining Lemmas 8 to 10.

[31] In the second equality, we use $\widetilde{\kappa}$ rather than κ because τ is any integer between 0 and $L(\delta) - 1$.

Lemma 7. Suppose $|D| = 1$. When monitoring is perfect either with or without transfers, for every player $i \in N$, there is a stable convention σ^i such that $U_i(\emptyset|\sigma^i) = 0$ when $\delta > \frac{n-2}{n-1}$.

Proof. Without loss of generality, suppose the collegium D consists of player 1. Let $\widehat{a} \equiv (1, 0, \ldots, 0)$ denote the unique alternative in the core, and $\overline{a} \equiv (0, \frac{1}{n-1}, \ldots, \frac{1}{n-1})$ denote the alternative that equally divides the total payoff among all non-veto players.

Case 1: Non-Transferable Utility. Let σ^1 be the core-reversion convention that recommends $(\overline{a}, \emptyset)$ on path, and recommends $(\widehat{a}, \emptyset)$ indefinitely after any history where blocking has occurred. σ^1 gives player 1 zero payoff. We will verify that σ^1 is stable.

No coalition has profitable deviations once continuation play reverts back to the core. To check stability on path of play, consider a blocking coalition C. Since the game is non-dictatorial, if C is a winning coalition, it must be the case that $\{1\} \subseteq C$ but $C \neq \{1\}$. Let $j \neq 1$ be a player in C and consider any deviation (a', C) by C. Since $a'_j \leq 1$, we have

$$(1-\delta)a'_j + \delta 0 \leq 1 - \delta \leq \frac{1}{n-1}$$

so player j prefers following the convention over deviating and reverting to the core. As a result, no coalition C has profitable one-shot deviation after any history, so σ^1 is stable.

For $i \neq 1$, let σ^i be the convention that recommends $(\widehat{a}, \emptyset)$ after every history. The convention is stable, and gives each player $i \neq 1$ zero payoff.

Case 2: Perfectly Monitored Transfers. Let σ^1 be the core-reversion convention such that σ^1 recommends $(\overline{a}, \emptyset, \mathbf{0})$ on path; suppose blocking (a', C, T') has occurred, σ^1 recommends $(\widehat{a}, \emptyset, \mathbf{0})$ indefinitely afterwards if $u_1(a', T') \geq 0$, but ignores the blocking if instead $u_1(a', T') < 0$. σ^1 gives player 1 zero payoff. We will verify that σ^1 is stable.

No coalition has profitable deviations once continuation play reverts back to the core. To check stability on path of play, consider a blocking coalition C. Since the game is non-dictatorial, if C is a winning coalition it must be the case that $\{1\} \subseteq C$ and $C \neq \{1\}$.

Let C be a winning coalition. If $u_1(a', T') < 0$, since there is no change in continuation value, player 1 finds the deviation unprofitable. If $u_1(a', T') \geq 0$, then it must be the case that $\sum_{j \in C \setminus \{1\}} u_j(a', T') \leq 1$, so there must be a player $j \in C \setminus \{1\}$ such that $u_j(a', T') \leq 1$, and we have

$$(1-\delta)u_j(a', T') + \delta(0) \leq 1 - \delta \leq \frac{1}{n-1}$$

so player j prefers following the convention over deviating and reverting to the core. As a result, no coalition C has profitable one-shot deviation after any history, so σ^1 is stable.

For $i \neq 1$, let σ^i be the convention that recommends $(\widehat{a}, \emptyset, \mathbf{0})$ after every history. The convention is stable, and gives each player $i \neq 1$ zero payoff. □

Lemma 8. If $|D| \geq 2$, when monitoring is perfect either with or without transfers, for every player $i \in N$, there is a stable convention σ^i such that $U_i(\emptyset|\sigma^i) = 0$ for every δ.

Proof. Without loss of generality, suppose $\{1,2\} \subseteq D$. Let $a^1 \equiv (1,0,\dots,0)$ and $a^2 \equiv (0,1,0,\dots,0)$ be two alternatives that allocate all payoff to player 1 and 2, respectively. It follows that both a^1 and a^2 are in the core.

Case 1: Non-Transferable Utility. Let σ^1 be the convention that recommends $(a^2,\emptyset)$ regardless of history; for all $i \neq 1$, let σ^i be the convention that recommends $(a^1,\emptyset)$ regardless of history. Each σ^i is stable, and $U_i(\emptyset|\sigma^i) = 0$ for every $i \in N$.

Case 2: Perfectly Monitored Transfers. Let σ^1 be the convention that recommends $(a^2,\emptyset,\mathbf{0})$ regardless of history; for all $i \neq 1$, let σ^i be the convention that recommends $(a^1,\emptyset,\mathbf{0})$ regardless of history. Each σ^i is stable, and $U_i(\emptyset|\sigma^i) = 0$ for every $i \in N$.

□

Lemma 9. Suppose the set of payoff profiles from stable conventions is $\mathcal{U}$. For each player $i \in N$, let $\underline{u}_i \equiv \min_{u\in\mathcal{U}} u_i$ be player i's smallest possible payoff from stable conventions.
Non-Transferable Utility: let $(a,\emptyset)$ be a stage-game outcome. Then $(a,\emptyset)$ can be sustained as the outcome of a stationary stable convention if and only if for every coalition C and alternative $a' \in E_C(a)$, there is a player $i \in C$ such that

$$(1-\delta)v_i(a') + \delta\underline{u}_i \leq v_i(a) \tag{27}$$

Perfectly Monitored Transfers: let $(a,\emptyset,T)$ be a stage-game outcome. Then $(a,\emptyset,T)$ can be sustained as the outcome of a stationary stable convention if and only if for every coalition C, alternative $a' \in E_C(a)$, and transfers T'_C, there is a player $i \in C$ such that

$$(1-\delta)u_i(a',[T'_C,T_{-C}]) + \delta\underline{u}_i \leq u_i(a,T)$$

Proof. We prove the result for the case of non-transferable utility. The proof for perfectly monitored transfers uses a similar argument, the only difference being the augmentation of stage-game outcomes with transfers.

To see the "only if" direction, suppose there exists a coalition C and alternative a' such that inequalities (27) fails for every $i \in C$. Towards a contradiction, suppose also that there exists a stationary stable convention σ that sustains $(a,\emptyset)$. Since σ is a stable convention, it follows that $U_i(h|\sigma) \geq \underline{u}_i$ for every $i \in C$ and all $h \in \mathcal{H}$. As a result, for every $i \in C$,

$$(1-\delta)v_i(a') + \delta U_i(a',C|\sigma) \geq (1-\delta)v_i(a') + \delta\underline{u}_i > v_i(a),$$

which implies that (a',C) is a profitable deviation for coalition C, contradicting σ being a stable convention.

For the "if" direction, Inequality (27) implies that for every coalition C and alternative $a' \in E_C(a)$, there exits a player $i^*|_{(a',C)}$ and a *stable* convention $\sigma^{i^*|_{(a',C)}}$ such that

$$(1-\delta)v_{i^*|_{(a',C)}}(a') + \delta U_{i^*|_{(a',C)}}(a',C|\sigma^{i^*|_{(a',C)}}) \leq v_{i^*|_{(a',C)}}(a). \tag{28}$$

Consider a convention σ that recommends $(a, \emptyset)$ on path, but switches to $\sigma^{i^*|_{(a',C)}}$ if deviation (a', C) has occurred. Inequality (28) implies that on path, no coalition can find a deviation that makes every member better-off. In addition, the fact that $\sigma^{i^*|_{(a',C)}}$ is a stable convention for each $i^*|_{(a',C)}$ ensures that after any off-path history, no coalition can find deviation that makes every member better-off. Therefore σ is a stationary stable convention that sustains $(a, \emptyset)$.

□

Lemma 10. Suppose there exist stable conventions $\{\sigma^i : i \in N\}$ such that $U_i(\emptyset|\sigma^i) = 0$ for all $i \in N$. Then for every fixed δ, the set of payoff profiles sustainable by stationary stable conventions is $U_{PM}(\delta)$.

Proof. Since the game is non-dictatorial, no single player can form a winning coalition. It follows that $\underline{v}_i = 0$ for all $i \in N$. For each player i, 0 is i's smallest possible payoff from all stable conventions (achieved, in particular, by the stable convention σ^i).

Case 1: Non-Transferable Utility. By Lemma 9, in order for a payoff profile u to be sustainable by a stationary stable convention, it is necessary and sufficient that for every winning coalition $C \in \mathcal{W}$, there exist no alternative $a' \in E_C(a)$ such that for every $i \in C$

$$(1-\delta)a'_i + \delta \cdot 0 = (1-\delta)a'_i > u_i. \tag{29}$$

Note that for every winning coalition C, this is true if and only if

$$\sum_{i \in C} u_i \geq \sum_{i \in C} (1-\delta)a'_i = 1 - \delta$$

for all $a'_i \in E_C(a)$. To see why, note that $E_C(a)$ consists of all points on the unit simplex such that $\sum_{i \in C} a'_i = 1$, so if $\sum_{i \in C} u_i < (1-\delta) \cdot 1$, there must be a certain a', representing a division of total payoff 1 among players in C, such that inequality (29) holds for every $i \in C$.

It follows that a payoff profile u is sustainable by a stationary stable convention if and only if

$$\sum_{i \in C} u_i \geq 1 - \delta$$

for every $C \in \mathcal{W}$.

Case 2: Perfectly Monitored Transfers. Let $(a, \emptyset, T)$ be an outcome that can be sustained by a stationary stable convention, and $u \equiv u(a, T)$. By Lemma 9, this is true if and only if for every winning coalition $C \in \mathcal{W}$, there exist no alternative $a' \in E_C(a)$ and transfers T'_C such that for every $i \in C$,

$$(1-\delta)\Big[a'_i + \sum_{j \in C} T'_{ji} - \sum_{j \in C} T'_{ij}\Big] + (1-\delta) \sum_{j \notin C} T_{ji} + \delta \cdot 0 > u_i.$$

In the inequality above, it is without loss to focus on alternative a' such that $\sum_{i \in C} a'_i = 1$. Let $s_i^C(T) \equiv (1-\delta) \sum_{j \notin C} T_{ji}$ denote the total transfer player i receives from outside of coalition C. Note that $s_i^C(T) \geq$

0. Since $\sum_{i \in C}\left[a_i' + \sum_{j \in C} T_{ji}' - \sum_{j \in C} T_{ij}'\right] = \sum_{i \in C} a_i' = 1$, the above condition is satisfied if and only if there are no numbers $\{u_i'\}_{i \in C}$ such that $\sum_{i \in C} u_i' = 1$, and for every $i \in C$,

$$(1-\delta)u_i' + s_i^C(T) > u_i.$$

Following a similar argument as that in *Case 1*, this is satisfied if and only if for every winning coalition C,

$$\sum_{i \in C} u_i \geq 1 - \delta + \sum_{i \in C} s_i^C(T). \tag{30}$$

Now, since $s_i^C(T) \geq 0$ for all C, i and T, it follows that $u \in U_{PM}(\delta)$, so nothing outside of $U_{PM}(\delta)$ can be sustained.

To see everything in $U_{PM}(\delta)$ can be sustained, fix any $u \in U_{PM}(\delta)$ and let $a \equiv u$ be the alternative identified with u, we will show the outcome $(a, \emptyset, \mathbf{0})$ can be sustained by a stationary stable convention. Now, for every winning coalition C, since $s_i^C(\mathbf{0}) = 0$ for all C and i, it follows that inequality (30) is satisfied if and only if

$$\sum_{i \in C} u_i \geq 1 - \delta. \tag{31}$$

Since $u \in U_{PM}(\delta)$, inequality (31) indeed holds for every winning coalition, so u can be sustained by a stationary stable convention.

□

B.5 Proofs of Results in Section 5

B.5.1 Proof of Lemma 1*

The proof follows identical arguments as the "only if" direction in the proof of Lemma 1, except now we only focus on coalitions that are in $\mathcal{S}$.

B.5.2 Proof of Theorem 5

Part 1: *Given $\mathcal{S} \subseteq \mathcal{C}$, for every $\delta \geq 0$, every stable convention implements payoffs only within $\mathcal{D}(\mathcal{S})$.*

The proof follows identical arguments as in the proof of Theorem 3, except that now we only focus on multi-shot deviations by coalitions in $\mathcal{S}$, and make use of Lemma 1* instead of Lemma 1.

Part 2: *If $\mathcal{D}^s(\mathcal{S})$ is non-empty, then for every payoff profile $u \in \mathcal{D}^s(\mathcal{S})$, there is a $\underline{\delta} < 1$ such that for every $\delta \in (\underline{\delta}, 1)$, there exists a stable convention with a discounted payoff equal to u.*

Since $\mathcal{D}^s(\mathcal{S})$ is always empty when the grand coalition can make secret transfers, if $\mathcal{D}^s(\mathcal{S})$ is non-empty, it must be that $N \notin \mathcal{S}$. Fix any payoff vector $u^0 \in \mathcal{D}^s(\mathcal{S})$. Let $\widehat{\mathcal{S}} \equiv \mathcal{S} \cup N$, the first step is to construct "$\widehat{\mathcal{S}}$-specific" punishments for all coalitions (and individuals) in $\widehat{\mathcal{S}}$.

Lemma 6*. There exist $\widehat{\mathcal{S}}$-specific punishments $\{u^C : C \in \widehat{\mathcal{S}}\}$ in $\mathcal{D}^s(\mathcal{S})$ such that

$$\sum_{i \in C} u_i^C < \sum_{i \in C} u_i^0$$

and for any $C' \neq C$

$$\sum_{i \in C} u_i^C < \sum_{i \in C} u_i^{C'}$$

Proof. The argument uses the same construction as in Lemma 6, except here the payoff vectors are only constructed for coalitions and players in $\widehat{\mathcal{S}}$. $\square$

For every $C \in \widehat{\mathcal{S}}$, let $\underline{a}_C \in \arg\min_{a \in A} \max_{a' \in E_C} \sum_{i \in C} v_i(a')$ be an alternative that can be used to minmax C. Note that by construction, $\sum_{i \in C} v_i(\underline{a}_C) \leq \underline{v}_C$. Given the $\widehat{\mathcal{S}}$-specific punishments, let $\kappa \in (0,1)$ be such that for every $\widetilde{\kappa} \in [\kappa, 1]$, the following is true for every $C \in \widehat{\mathcal{S}}$:

$$(1-\widetilde{\kappa}) \sum_{i \in C} v_i(\underline{a}_C) + \widetilde{\kappa} \sum_{i \in C} u_i^C > \underline{v}_C \tag{32}$$

For every $C' \in \widehat{\mathcal{S}}$, $C' \neq C$: $\quad (1-\widetilde{\kappa}) \sum_{i \in C'} v_i(\underline{a}_C) + \widetilde{\kappa} \sum_{i \in C'} u_i^C > (1-\widetilde{\kappa}) \sum_{i \in C'} v_i(\underline{a}_{C'}) + \widetilde{\kappa} \sum_{i \in C'} u_i^{C'}. \quad (33)$

Inequality (32) implies that in terms of total value, coalition C is willing to bear the cost of $\sum_{i \in C} v_i(\underline{a}_C)$ with the promise of transitioning into its coalition-specific punishment rather than staying at its minmax. Inequality (33) implies that every coalition prefers punishing other coalitions than being punished itself. By an argument identical to that we saw in Theorem 1, there exists a value of $\kappa \in (0,1)$ such that all the inequalities above hold for all $\widetilde{\kappa} \in [\kappa, 1]$, $i \in N$ and $j \in N \backslash \{i\}$. Let $L(\delta) \equiv \left\lceil \frac{\log \kappa}{\log \delta} \right\rceil$ where $\lceil \cdot \rceil$ is the ceiling function. As before, we use the property that $\lim_{\delta \to 1} \delta^{L(\delta)} = \kappa$.

Since $\mathcal{D}^s(\mathcal{S}) \subseteq \mathcal{U}_{IR}^{\dagger}$, by Lemma 5, there exist payoff vectors $\{\widetilde{u}^1, \ldots, \widetilde{u}^M\} \subseteq \mathcal{U}(\overline{a}) \cup \mathcal{U}(\underline{a})$ such that $\mathcal{D}^s(\mathcal{S}) \subseteq \text{co}(\widetilde{u}^1, \ldots, \widetilde{u}^M)$, where $\widetilde{u}^m = u(\widetilde{a}^m, \widetilde{T}^m)$ for some alternative $\widetilde{a}^m \in \{\overline{a}, \underline{a}\}$ and transfers matrix $\widetilde{T}^m$ for each $m = 1, \ldots, M$. Lemma 2 then guarantees that for any $\epsilon > 0$, there exists $\underline{\delta} \in (0,1)$ such that for all $\delta \in (\underline{\delta}, 1)$, there exist sequences $\left\{\{a^{d,\tau}, T^{d,\tau}\}_{\tau=0}^{\infty} : d \in \widehat{\mathcal{S}} \cup \{0\}\right\}$ such that for each d and t, $(1-\delta)\sum_{\tau=0}^{\infty} \delta^{\tau} u(a^{d,\tau}, T^{d,\tau}) = u^d$ and $\left|\left|u^d - (1-\delta)\sum_{\tau=t}^{\infty} \delta^{\tau} u(a^{d,\tau}, T^{d,\tau})\right|\right| < \epsilon$. We fix an ϵ such that

$$\epsilon < (1-\kappa) \min \left\{ \min_{d \in \widehat{\mathcal{S}}, d' \in \widehat{\mathcal{S}} \cup \{0\}, d' \neq d} \Big(\sum_{i \in d} u_i^{d'} - \sum_{i \in d} u_i^d \Big), \min_{d \in \widehat{\mathcal{S}}} \sum_{i \in d} u_d^d - \underline{v}_d \right\},$$

and given that ϵ, consider δ exceeding the appropriate $\underline{\delta}$.

We now describe the convention that we use to sustain u^0. Let $\mathbf{0}$ denote the transfer matrix where all players make no transfers. Consider the convention represented by the automaton $(W, w(0,0), f, \gamma)$, where

- $W \equiv \left\{w(d,\tau) | d \in \widehat{\mathcal{S}} \cup \{0\}, \tau \geq 0\right\} \cup \{\underline{w}(C,\tau) | C \in \widehat{\mathcal{S}}, 0 \leq \tau < L(\delta)\}$ is the set of possible states;
- $w(0,0)$ is the initial state;
- $f : W \to \mathcal{O}^{TU}$ is the output function, where $f(w(d,\tau)) = (a^{d,\tau}, \emptyset, T^{d,\tau})$ and $f(\underline{w}(C,\tau)) = (\underline{a}_C, \emptyset, \mathbf{0})$;

- $\gamma : W \times \mathcal{O}^{TU} \to W$ is the transition function. For states of the form $w(d, \tau)$, the transition is

$$\gamma\big(w(d,\tau),(a,C',T)\big) = \begin{cases} \underline{w}(C',0) & \text{if } C' \in \widehat{\mathcal{S}} \\ \underline{w}(j^*,0) & \text{if } C' \in \mathcal{C}\backslash\widehat{\mathcal{S}},\ j^* = \arg\min_{j \in C'}\{u_j(a,T) - u_j^{d,t}\} \\ w(d,\tau+1) & \text{if } C' = \emptyset \end{cases}$$

For states of the form $\{\underline{w}(i,\tau) | 0 \leq \tau < L(\delta) - 1\}$ where $i \in N$, the transition is

$$\gamma(\underline{w}(i,\tau),(a,C',T)) = \begin{cases} \underline{w}(C',0) & \text{if } C' \in \widehat{\mathcal{S}} \\ \underline{w}(j^*,0) & \text{if } \{C' \in \mathcal{C}\backslash\widehat{\mathcal{S}}\} \cap \big(\{u_i(a,T) > \underline{v}_i\} \cup \{i \notin C'\}\big) \\ & j^* = \arg\min_{C'\backslash\{i\}}\{u_j(a,T) - v_j(\underline{a}_i)\} \\ \underline{w}(i,0) & \text{if } \{C \in \mathcal{C}\backslash\widehat{\mathcal{S}}\} \cap \{u_i(a,T) \leq \underline{v}_i\} \cap \{i \in C'\} \\ \underline{w}(C,\tau+1) & \text{if } C' = \emptyset \end{cases}$$

For states of the form $\{\underline{w}(C,\tau) | 0 \leq \tau < L(\delta) - 1\}$ where $C \in \mathcal{S}$, the transition is

$$\gamma(\underline{w}(C,\tau),(a,C',T)) = \begin{cases} \underline{w}(C',0) & \text{if } C' \in \widehat{\mathcal{S}} \\ \underline{w}(j^*,0) & \text{if } C' \in \mathcal{C}\backslash\widehat{\mathcal{S}},\ j^* = \arg\min_{j \in C'}\{u_j(a,T)\} \\ \underline{w}(C,\tau+1) & \text{if } C' = \emptyset \end{cases}$$

For states of the form $\underline{w}(C, L(\delta) - 1)$, the transition is

$$\gamma(\underline{w}(C,L(\delta)-1),(a,C',T)) = \begin{cases} \gamma(\underline{w}(C,0),(a,C',T)) & \text{if } C' \neq \emptyset \\ w(C), & \text{if } C' = \emptyset \end{cases}$$

The convention represented by the above automaton yields payoff profile u^0. By construction, the continuation values in different states, $V(\cdot)$, satisfy:

$$V(w(C)) = u^C, \ \ \forall C \in \widehat{\mathcal{S}}$$

$$V(\underline{w}(C,\tau)) = (1 - \delta^{L(\delta)-\tau})v(\underline{a}_C^e) + \delta^{L(\delta)-\tau}V(w(C)), \ \ \forall 0 \leq \tau \leq L(\delta) - 1, C \in \widehat{\mathcal{S}}$$

Stability in states of the form $w(d, \tau)$: Depending on whether or not the blocking coalition can make secret transfers, there are two cases to consider.
Case 1: $C' \in \mathcal{C}\backslash\widehat{\mathcal{S}}$. Suppose the outcome $(\widehat{a}, C', \widehat{T})$ is realized, then the convention punishes player $j^* = \arg\min_{j \in C'}\{u_j(\widehat{a}, \widehat{T}) - u_j^{d,\tau}\}$. Following the same argument as in the proof of Theorem 2, this one-shot deviation is unprofitable for j^* and hence, for coalition C' if δ is sufficiently high.

Case 2: $C' \in \widehat{\mathcal{S}}$. The convention punishes coalition C'. Since all $T^{d,\tau}$ are drawn from $\{\widetilde{T}^m\}_{m=1}^M$, we have

$$\sum_{i \in C'} u_i(\widehat{a}, \widehat{T}) \leq \max_{a \in A} \sum_{j \in C} v_j(a) + \max_{1 \leq m \leq M} \sum_{j \in C} \sum_{k \notin C} \widetilde{T}^m_{jk}.$$

In the inequality above, each term in the RHS is independent of δ and (d, τ). Thus, we can find a uniform bound B_1 such that the total payoff from deviation for coalition C', $\sum_{i \in C'} u_i(\widehat{a}, \widehat{T})$ is less than B_1 for every C', δ and (d, τ). Coalition C' has total payoff of at least $\sum_{i \in C'} u_i^C - \epsilon$ without deviating. By deviating, C' obtains a total payoff less than

$$(1-\delta)B_1 + \delta \sum_{i \in C'} V_i(\underline{w}(C', 0)) = (1-\delta)B_1 + \delta \left[(1 - \delta^{L(\delta)}) \sum_{i \in C'} v_i(\underline{a}_{C'}) + \delta^{L(\delta)} \sum_{i \in C'} u_i^{C'} \right]$$

For the deviation to be profitable, the total value for C' must be higher. So the one-shot deviation is unprofitable if the above term is no more than $\sum_{i \in C'} u_i^C - \epsilon$. We prove that this is the case both for $C' \neq C$ and $C' = C$.

First consider $C' \neq C$. Observe that

$$\lim_{\delta \to 1} (1-\delta)B_1 + \delta \left[(1 - \delta^{L(\delta)}) \sum_{i \in C'} v_i(\underline{a}_{C'}) + \delta^{L(\delta)} \sum_{i \in C'} u_i^{C'} \right]$$
$$= (1-\kappa) \sum_{i \in C'} v_i(\underline{a}_{C'}) + \kappa \sum_{i \in C'} u_i^{C'} < \sum_{i \in C'} u_i^{C'} < \sum_{i \in C'} u_i^C - \epsilon$$

The last inequality above follows from the definition of ϵ. It follows that the one-shot coalition deviation is not profitable for C'.

Now suppose that $C' = C$. The deviation payoff being less than $\sum_{i \in C'} u_i^{C'}$ can be re-written as

$$(1-\delta)(B_1 - \sum_{i \in C'} u_i^{C'}) + \epsilon \leq \delta(1 - \delta^{L(\delta)})(\sum_{i \in C'} u_i^{C'} - \sum_{i \in C'} v_i(\underline{a}_{C'}))$$

As $\delta \to 1$, the LHS converges to ϵ. Because $\lim_{\delta \to 1} \delta^{L(\delta)} = \kappa$, the RHS converges to $(1-\kappa)(\sum_{i \in C'} u_i^{C'} - \sum_{i \in C'} v_i(\underline{a}_{C'}))$. By the definition of ϵ, the above inequality holds, and therefore, there is no profitable one-shot deviation if δ is sufficiently high.

Stability in states of the form $\underline{w}(i, \tau)$ where $i \in N$: If $C' \in \mathcal{C} \backslash \widehat{\mathcal{S}}$, we can verify that such deviations are not profitable using the same arguments as in the proof of Theorem 2. If $C' \in \widehat{\mathcal{S}}$, then the same arguments as in the proof of Theorem 4 apply, except that instead of $\underline{a}_C^e$ (or $\underline{a}_{C'}^e$), coalitions are minmaxed using $\underline{a}_C$ (or $\underline{a}_{C'}$).

Stability in states of the form $\underline{w}(C, \tau)$ where $C \in \mathcal{S}$: Again, there are two cases to consider.
Case 1: $C' \in \widehat{\mathcal{S}}$. The same arguments as in the proof of Theorem 4 applies, except that instead of $\underline{a}_C^e$ (or $\underline{a}_{C'}^e$), coalitions are minmaxed using $\underline{a}_C$ (or $\underline{a}_{C'}$).
Case 2: $C' \in \mathcal{C} \backslash \widehat{\mathcal{S}}$. Suppose the outcome $(\widehat{a}, C', \widehat{T})$ is realized. The convention punishes $j^* = \arg\min_{j \in C'} \{u_j(\widehat{a}, \widehat{T})\}$.

It follows that

$$u_{j^*}(\widehat{a},\widehat{T}) \le \frac{1}{|C'|}\Big[\sum_{j\in C'} u_j(\widehat{a},\widehat{T})\Big] \le \frac{1}{|C'|}\Big[\sum_{j\in C'} v_j(\widehat{a})\Big] \le \frac{1}{|C'|}\Big[\max_{a\in A}\sum_{j\in C'} v_j(a)\Big] \equiv b_2(C').$$

The first inequality above follows since the minimum among any numbers is less than their average; the second inequality follows because when C' blocks, all players outside of C' are following the recommendation from the convention and making zero transfers, so the total value of C' cannot be higher than the total utility generated from $\widehat{a}$.

Define $B_2 \equiv \max_{C'\in\mathcal{C}} b_2(C')$, so $u_{j^*}(\widehat{a},\widehat{T}) < B_2$ for all C' and δ. B_2 provides a bound on the payoff obtained by j^* when C' deviates. Player j^* does not benefit from this deviation if

$$(1-\delta^{L(\delta)-\tau})v_{j^*}(\underline{a}_C) + \delta^{L(\delta)-\tau}u^C_{j^*} \ge (1-\delta)B_2 + \delta(1-\delta^{L(\delta)})v_{j^*}(\underline{a}_{j^*}) + \delta^{L(\delta)+1}u^{j^*}_{j^*}.$$

In addition, since $C \in \mathcal{S}$ and $j^* \in N$, it must be that $C \neq j^*$. The inequality above is satisfied for sufficiently high δ, and the argument follows the same steps as that of the analogous part of Theorem 1.

B.6 Existence of Payoff Set for Efficient β-Core

The following result helps provide necessary and sufficient conditions for the non-emptiness of $\mathcal{B}^s$ and $\mathcal{D}^s(\mathcal{S})$.

Proposition 1. *Let (N,ϕ) be a characteristic function game. The set*

$$\Lambda \equiv \left\{ u \in \mathsf{R}^n : \sum_{i\in N} u_i = \phi(N), \sum_{i\in C} u_i > \phi(C) \text{ for all } C \in \mathcal{C}\backslash\{N\} \right\}$$

is non-empty if and only if for every set of weights $\left\{0 \le \lambda_C \le 1 : C \in \mathcal{C}\backslash\{N\}\right\}$ such that

$$\forall i : \sum_{C\in\mathcal{C}\backslash\{N\}, i\in C} \lambda_C = 1$$

the following condition holds:

$$\sum_{C\in\mathcal{C}\backslash\{N\}} \lambda_C\,\phi(C) < \phi(N)$$

Proposition 1 is analogous to the Bondareva-Shapley Theorem (Peleg and Sudhölter 2007) but uses strict rather than weak inequalities. It allows us to determine whether Λ is empty by solving the following linear programming problem:

$$\begin{aligned} &\max \sum_{C\in\mathcal{C}\backslash\{N\}} \lambda_C\phi(C) \\ \text{s.t.}\quad & 0 \le \lambda_C \le 1 \;\; \forall C \in \mathcal{C}\backslash\{N\} \\ & \sum_{C\in\mathcal{C}\backslash\{N\}, i\in C} \lambda_C = 1 \;\; \forall i \in N \end{aligned}$$

The set Λ is nonempty if and only the value of the program above is strictly less than $\phi(N)$.

To determine whether $\mathcal{B}^s$ is empty, one can apply Proposition 1 by setting $\phi(N) = \max_A \sum_{i\in N} v_i(a)$, and $\phi(C) = \underline{v}^e_C$ for every $C \in \mathcal{C}\backslash\{N\}$. Similarly for $\mathcal{D}^s(\mathcal{S})$, we can set $\phi(N) = \max_A \sum_{i\in N} v_i(a)$, $\phi(C) = \underline{v}_C$ for all $C \in \mathcal{S}$, and $\phi(C) = \sum_{i\in C} \underline{v}_i$ for all $C \notin \mathcal{S} \cup \{N\}$.

B.6.1 Preliminaries

To prove Proposition 1, we use a variant of Motzkin's Theorem of the Alternative. We first state the original result in Lemma 11 (see Chapter 2.4 of Mangasarian 1994) and then prove the variant that we need in Lemma 11* below. For notational clarity, we use bold capital letters ($\boldsymbol{A}$) to denote matrices, bold lowercase letters ($\boldsymbol{x}$) for vectors, and plain lowercase letters (x) for numbers. We use $\boldsymbol{x} > 0$ to denote $x_i > 0$ for all i, and $\boldsymbol{x} \geq 0$ to denote $x_i \geq 0$ for all i.

Lemma 11. (Motzkin's Theorem of the Alternative) Let $\boldsymbol{A} \neq 0$, $\boldsymbol{B}$, and $\boldsymbol{D}$ be given matrices. Then either

$$\boldsymbol{Ax} > 0, \quad \boldsymbol{Bx} \geq 0, \quad \boldsymbol{Dx} = 0 \text{ has a solution } \boldsymbol{x}$$

or

$$\begin{cases} \boldsymbol{A}^T\boldsymbol{w} + \boldsymbol{B}^T\boldsymbol{y} + \boldsymbol{D}^T\boldsymbol{z} = 0 \\ \boldsymbol{w} \geq 0, \quad \boldsymbol{w} \neq 0, \quad \boldsymbol{z} \geq 0 \end{cases} \text{has a solution } \boldsymbol{w}, \boldsymbol{y}, \boldsymbol{z}$$

but never both.

While Lemma 11 deals with the solvability of systems of *homogeneous* linear inequalities, for our purpose, Lemma 11* deals with the solvability of systems of *non-homogeneous* linear inequalities.

Lemma 11*. Let $\boldsymbol{A} \neq 0$ and $\boldsymbol{D}$ be given matrices. Then either

$$\boldsymbol{Ax} > \boldsymbol{b}, \quad \boldsymbol{Dx} = \boldsymbol{d} \text{ has a solution } \boldsymbol{x} \tag{I}$$

or

$$\begin{cases} \boldsymbol{A}^T\boldsymbol{y} + \boldsymbol{D}^T\boldsymbol{z} = 0, \boldsymbol{y} \geq 0 \text{ with} \\ \boldsymbol{b}^T\boldsymbol{y} + \boldsymbol{d}^T\boldsymbol{z} > 0 \text{ or } \left(\boldsymbol{b}^T\boldsymbol{y} + \boldsymbol{d}^T\boldsymbol{z} = 0 \text{ and } \boldsymbol{y} \neq 0\right) \end{cases} \text{has a solution } \boldsymbol{y}, \boldsymbol{z} \tag{II}$$

but never both.

Proof. Letting $\boldsymbol{x} = \boldsymbol{u}/t$ where $t > 0$, statement (I) above is equivalent to

$$\begin{cases} \boldsymbol{Au} - t\boldsymbol{b} > 0 \\ t > 0 \\ \boldsymbol{Du} - t\boldsymbol{d} = 0 \end{cases} \text{has a solution } \boldsymbol{u}, t$$

or

$$\begin{cases} \begin{bmatrix} \boldsymbol{A} & -\boldsymbol{b} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ t \end{bmatrix} > 0 \\ \begin{bmatrix} \boldsymbol{D} & -\boldsymbol{d} \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ t \end{bmatrix} = 0 \end{cases} \text{has a solution } \begin{bmatrix} \boldsymbol{u} \\ t \end{bmatrix}$$

Applying Lemma 11 to the system above, we have that statement (I) is mutually exclusive to

$$\begin{cases} \begin{bmatrix} \boldsymbol{A}^T & 0 \\ -\boldsymbol{b}^T & 1 \end{bmatrix} \widetilde{\boldsymbol{y}} + \begin{bmatrix} \boldsymbol{D} \\ -\boldsymbol{d}^T \end{bmatrix} \boldsymbol{z} = 0 \\ \widetilde{\boldsymbol{y}} \geq 0, \widetilde{\boldsymbol{y}} \neq 0 \end{cases} \qquad \text{has a solution } \widetilde{\boldsymbol{y}}, \boldsymbol{z}$$

where $\widetilde{\boldsymbol{y}} = (\boldsymbol{y}, y^*)$ (note that y^* is a number). This is equivalent to

$$\begin{cases} \boldsymbol{A}^T \boldsymbol{y} + \boldsymbol{D}^T \boldsymbol{z} = 0 \\ \boldsymbol{b}^T \boldsymbol{y} + \boldsymbol{d}^T \boldsymbol{z} = y^* \\ \boldsymbol{y} \geq 0; y^* \geq 0; \text{ and } \boldsymbol{y}, y^* \text{ not both equal to } 0 \end{cases} \qquad \text{has a solution } \boldsymbol{y}, y^*, \boldsymbol{z}$$

which can be further simplified to

$$\begin{cases} \boldsymbol{A}^T \boldsymbol{y} + \boldsymbol{D}^T \boldsymbol{z} = 0, \boldsymbol{y} \geq 0 \text{ with} \\ \boldsymbol{b}^T \boldsymbol{y} + \boldsymbol{d}^T \boldsymbol{z} > 0 \text{ or } \left(\boldsymbol{b}^T \boldsymbol{y} + \boldsymbol{d}^T \boldsymbol{z} = 0 \text{ and } \boldsymbol{y} \neq 0 \right) \end{cases} \qquad \text{has a solution } \boldsymbol{y}, \boldsymbol{z}$$

This is statement (II), and thus completes the proof. □

B.6.2 Proof of Proposition 1

For every $C \in \mathcal{C} \backslash \{N\}$, let χ_C denote the n-dimensional row vector consisting of 0's and 1's, with the i-th entry being 1 if and only if $i \in C$. The non-emptyness of Λ is equivalent to the solvability of the following system of linear inequalities:

$$\begin{bmatrix} \vdots \\ \chi_C \\ \vdots \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} > \begin{bmatrix} \vdots \\ \phi(C) \\ \vdots \end{bmatrix} \tag{34}$$

and

$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \phi(N) \tag{35}$$

By Lemma 11*, it follows that the solvability of inequalities (34) and (35) is mutually exclusive to the existence of numbers $\mu \in \mathsf{R}$ and $\{\gamma_C \geq 0 : C \in \mathcal{C} \backslash \{N\}\}$ that satisfy (36) and (37) below:

$$\forall i : \sum_{C \in \mathcal{C} \backslash \{N\}, i \in C} \gamma_C = \mu \tag{36}$$

$$\begin{cases} \displaystyle\sum_{C \in \mathcal{C} \setminus \{N\}} \gamma_C \phi(C) > \mu \phi(N) \\ \text{or} \\ \displaystyle\sum_{C \in \mathcal{C} \setminus \{N\}} \gamma_C \phi(C) = \mu \phi(N) \text{ with } \gamma_C \neq 0 \text{ for some } C \end{cases} \tag{37}$$

In the equations above, (36) implies that $\mu \geq 0$. However μ cannot be 0, for otherwise (36) would imply that all γ_C's are zero, making (37) impossible. Define $\lambda_C \equiv \gamma_C / \mu$. It follows that Λ is non-empty if and only if there does not exist numbers $\{\lambda_C \geq 0 : C \in \mathcal{C} \setminus \{N\}\}$ that satisfy

$$\forall i : \sum_{C \in \mathcal{C} \setminus \{N\}, i \in C} \lambda_C = 1 \tag{38}$$

and

$$\begin{cases} \displaystyle\sum_{C \in \mathcal{C} \setminus \{N\}} \lambda_C \phi(C) > \phi(N) \\ \text{or} \\ \displaystyle\sum_{C \in \mathcal{C} \setminus \{N\}} \lambda_C \phi(C) = \phi(N) \text{ with } \lambda_C \neq 0 \text{ for some } C \end{cases} \tag{39}$$

Since (38) already implies $\lambda_C \neq 0$ for some C, (39) can be simplified, and Λ is non-empty if and only if there are no numbers $\{\lambda_C \geq 0 : C \in \mathcal{C} \setminus \{N\}\}$ that satisfy

$$\begin{cases} \forall i : \displaystyle\sum_{C \in \mathcal{C} \setminus \{N\}, i \in C} \lambda_C = 1 \\ \displaystyle\sum_{C \in \mathcal{C} \setminus \{N\}} \lambda_C \phi(C) \geq \phi(N) \end{cases}$$

In other words, any numbers $\{\lambda_C \geq 0 : C \in \mathcal{C} \setminus \{N\}\}$ that satisfy $\sum_{C \in \mathcal{C} \setminus \{N\}, i \in C} \lambda_C = 1$ for all i must at the same time satisfy $\sum_{C \in \mathcal{C} \setminus \{N\}} \lambda_C \phi(C) < \phi(N)$, which completes the proof.

B.7 Communicating about Secret Transfers

In this section, we investigate whether communication rounds can be added to the game to elicit information about secret transfers from members of a blocking coalition and use that information to deter deviations. We consider a tractable variant of this problem, following approaches to study communication in repeated games with private monitoring (Compte 1998; Kandori and Matsushima 1998). Suppose that each player can privately observe only the transfers that are sent or received by herself, but can communicate publicly at the end of each period. We consider "semi-public conventions" where behavior across periods conditions only on the publicly observable history, but that within a period can condition on private information. We show that the analogue of a one-shot deviation principle (Lemma 1) applies, which generates the same coalitional payoff guarantee.

Let us describe the game. At each period t, players first choose alternatives and transfers, as in our baseline game. Each player then chooses a publicly observed message $m_i \in \mathcal{M}$, where the set of messages is sufficiently rich to describe all potential transfers ($\mathcal{M} \supseteq \mathcal{T}$). Let $\mathcal{M}^n$ denote the space of all feasible profiles of messages $m = (m_1, \dots, m_n)$. Messages are communicated simultaneously.

The set of public outcomes at the end of each period is $\mathcal{O}_p^{TU} \equiv \Big\{o = (a, C, m) | a \in A, C \in \mathcal{C}, m \in \mathcal{M}^n\Big\}$. The set of public histories is the set of finite sequences of public outcomes $\mathcal{H}_p \equiv \cup_{t=0}^{\infty} (\mathcal{O}_p^{TU})^t$. For each player i and transfers matrix T, let $I_i(T) \equiv \{T_{jk} | j = i \text{ or } k = i\}$ denote the transfers in T that are paid out or received by player i, and let $\mathcal{I}_i \equiv \big\{I_i(T) | T \in \mathcal{T}\big\}$ denote the set of all possible $I_i(T)$. Player i's semi-public histories are elements in $\mathcal{H}_i \equiv \mathcal{H}_p \times (A \times \mathcal{C} \times \mathcal{I}_i)$.

Definition 9. *A semi-public convention σ is a collection of mappings $\{\sigma_p\} \cup \{\sigma_i\}_{i=1}^n$, where $\sigma_p : \mathcal{H}_p \to A \times \mathcal{C} \times \mathcal{T}$ is the public component of the convention, and $\sigma_i : \mathcal{H}_i \to \mathcal{M}$ is the private reporting strategy for player i.*

When defining coalitional deviations, a coalition C when blocking can choose any a' in $E_C(a)$, change its transfer schedule to any T'_C, and send any profile of messages $m_C = \times_{i \in C} m_i$. Given players' private reporting strategies $\{\sigma_i\}_{i=1}^n$, if a coalition C blocks after public history $h \in \mathcal{H}_p$ choosing alternative a' and transfers $T'_C = [T'_{ij}]_{i \in C, j \in N}$, we use $m_{-C}(h, a, T'_C | \sigma_{-C}) \equiv \times_{i \notin C} \sigma_i \Big(h, a', C, I_i \big[T'_C, T_{-C}(h|\sigma_p)\big]\Big)$ to denote the resulting public messages from players outside of the blocking coalition.

Definition 10. *A semi-public convention $\sigma = \{\sigma_p\} \cup \{\sigma_i\}_{i=1}^n$ is stable if for every public history $h \in \mathcal{H}_p$, there exists no coalition C, alternative $a' \in E_C(a(h|\sigma))$, transfers $T'_C = [T'_{ij}]_{i \in C, j \in N}$, and messages $m'_C = \times_{i \in C} m'_i$, such that*

$$\text{For every } i \in C\text{:} \quad (1-\delta) u_i\Big(a', \big[T'_C, T_{-C}(h|\sigma_p)\big]\Big) + \delta U_i\Big(h, a', C, \big[m'_C, m_{-C}(h, a, T'_C | \sigma_{-C})\big] \,\Big|\, \sigma_p\Big) > U_i(h|\sigma_p)$$

We can analogously define a coalition's multi-shot deviation from a semi-public convention as we did in Section 4.

Definition 11. *A multi-shot deviation by coalition C from a semi-public convention $\sigma = \{\sigma_p\} \cup \{\sigma_i\}_{i=1}^n$ is a distinct collection of mappings $\sigma' = \{\sigma'_p\} \cup \{\sigma'_i\}_{i=1}^n$ such that*

1. *For any public history $h \in \mathcal{H}_p$ where $\sigma'(h) = (a', C', T') \neq \sigma(h)$, it must be that $C' = C$, $a' \in E_C(a(h|\sigma))$ and $T'_{-C} = T_{-C}(h|\sigma)$.*
2. *For all $i \notin C$, $\sigma'_i = \sigma_i$: players outside C follow their original reporting strategy.*

*A multi-shot deviation σ' by coalition C is **profitable** if there exists a public history h such that $U_i(h|\sigma') > U_i(h|\sigma)$ for all $i \in C$.*

Note that the analogue of Lemma 1 applies in this setting: if a coalition has a profitable multi-shot deviation (including at the communication stage), then it has a profitable one-shot deviation.

Lemma 12. A semi-public convention σ is stable if and only if no coalition has a profitable multi-shot devaition.

For brevity, we do not write the proof below but the argument is identical to that of Lemma 1. Therefore, Theorem 3 continues to apply: in any stable semi-public convention, every coalition must achieve payoffs of at least its coalitional minmax.