Conventions and Coalitions in Repeated Games*

S. Nageeb Ali† Ce Liu‡

June 2, 2019

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 (on and off the path), and anticipate that today's actions influence tomorrow's behavior. We evaluate the degree to which history-dependence can ward off 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, reducing the set of supportable payoffs to the core of the stage game.

*We thank (in random order) Daniel Barron, Federico Echenique, Joel Sobel, Mike Powell, Alex Wolitzky, Debraj Ray, Elliot Lipnowski, Joel Watson, Ben Golub, Vijay Krishna, Laura Doval, Matt Elliott, Hideo Konishi, Ron Siegel, and Maciej Kotowski. Ali gratefully acknowledges financial support from NSF grants SES-1127643 and SES-1530639.

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

‡Department of Economics, Michigan State University. Email: celiu0402@gmail.com.

Contents

1 Introduction

The theory of repeated games models self-enforcing conventions where players share an understanding of how the future unfolds in response to choices made today and in the past, and given that shared understanding, no individual wishes to deviate. This theory is central to our understanding of dynamic incentives and has been applied across a range of settings.

The main approach for the study of repeated games is non-cooperative, relying on individual optimization, and without any 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 all of these settings, one could in principle study the non-cooperative extensive-form that permits players to jointly deviate by modeling how players form alliances through a process of public and private offers with acceptance and rejection decisions. But our solutions for modeling how and which alliances form are sensitive to the extensive-form. Because it is difficult to assess which extensive-form is appropriate and infeasible to study them all, many studies of collective behavior follow the approach of cooperative game theory in taking a bird's eye view to how exactly alliances form and instead focusing on when alliances are profitable.

Our objective is to combine this cooperative approach with the repeated-games understanding of dynamic incentives. When such cooperative environments are repeated, what is the appropriate notion of stability? To what degree and when does the power of expectations 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. Effectivity correspondences specify the moves that each coalition can make. 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

[1] We believe that answering these questions is useful not only for repeated *cooperative* games but also for investigating 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.

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 the prospect of these coalitional deviations.

Because there is no "off-the-shelf" solution-concept for this coalitional repeated game, we develop one that is consistent with this motivation by building on pioneering approaches to farsighted stability in cooperative games (surveyed in Ray 2007 and 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 deviation at any history; in other words, a stable convention lacks profitable one-shot deviations for all coalitions. We then ask the following 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. The first observation is that history-dependence is a source of stability.[2] By making behavior 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 force using simple examples in Section 2: first, we show that in a repeated roommates problem, every efficient allocation can be supported by a stable convention even if the one-shot interaction has no stable match. Second, we illustrate in a repeated division problem, if the core of the stage game is non-empty, a convention can use "core-reversion" 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 sustainable outcomes beyond individual deviations when players are patient; dynamic incentives effectively ward off coalitional deviations. To put it differently, a shared understanding of the future—and its associated carrots and sticks—removes the incentives for coalitions to deviate today if players are sufficiently patient. This result has a simple intuition: to ward off coalitional deviations, it suffices to punish an individual member

[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.

of each coalition as if she were an individual deviator. This logic applies even when players can transfer utility to each other to "bribe" others to join their coalition because the convention can then punish players for paying or receiving bribes.

Secret Transfers: This possibility result leans heavily on the observability of side-payments. But in many contexts, the power of bribes comes from their secrecy and the inability to condition future play on them. Our second set of results, exposited in Section 4, finds a sharp contrast when coalitions can use secret side-payments.

Specifically, suppose that when a coalition 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; while 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 payoff (adding across its constituents) that coalition C can be pushed down to by others when it can best-respond. This is a coalitional payoff guarantee that 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 given 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.*

The above result states that when transfers are secret, payoffs supported by stable conventions gives 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.[3]

[3] Von Neumann and Morgenstern (1945) also suggest the α-characteristic function, which assumes a maxmin

In general, the set of payoffs where each coalition is guaranteed at least its coalitional minmax is smaller than the set of feasible and strictly individually rational payoffs. Indeed, for some games, this set is empty.[4] Our result implies that in these games, no convention is stable, regardless of the patience of players. We do not view this conclusion as being nihilistic but as a stark illustration of how short-term coalitional deviations coupled with secret transfers undermine the dynamic incentives of a convention.

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 deviation, 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* applies: a coalition lacks a profitable one-shot deviation from a convention if and only if it lacks a profitable multi-shot deviation.

This result (Lemma 1) 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.

Iterating this logic yields a tighter bound. Since the grand coalition can also guarantee itself a coalitional minmax, payoffs must be on the efficiency frontier at every history. Define the *efficient β-core* to be the set of payoffs that are both (i) efficient, and (ii) give each coalition above the coalitional minmaxes 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 as players become arbitrarily patient, every payoff within the relative interior of that set can be sustained.

Thus, we see that once coalitions can make secret transfers, they do not need long-term commitments to effectively deviate and guarantee a coalitional minmax. The appropriate analysis treats each non-singleton coalition as a fictitious entity, expanding the number of players from n to $2^n - 1$, and 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 try to minimize the payoffs of those within the blocking coalition; for example, see Chapter 2 of Ray (2007). That criticism is exactly right when the concept is applied to one-shot interactions where those outside a blocking coalition have no reason to hurt themselves to punish

procedure. In settings where each coalition can use a private correlation device, the Minmax Theorem implies that the two are identical.

[4]The set is non-empty if and only if the induced characteristic function satisfies the conditions of the Bondareva-Shapley Theorem.

deviators. In a repeated game, however, coalitions can be rewarded for punishing others. But this can be done only to a limited extent and must use efficient alternatives. If transfers can be made secretly within blocking coalitions, the efficient β-core may be an appropriate description of the set of sustainable outcomes.

Understanding Laws and Norms: In addition to investigating the role of dynamics and history-dependence in coalitional behavior, we view these results as speaking to the role of expectations in legal, community, and political enforcement. One perspective of laws and norms—dating back at least to the work of Hume (1740) if not earlier—treats them as shared understandings that individuals have of each other with the proviso that that understanding be credible and self-enforcing. Basu (2000) summarizes this idea beautifully:

> *In the end, all are caught in a web of self-reinforcing sanctions...law's empire, tangible and all-encompassing as it may seem, is founded on little else than beliefs.*

This perspective is a recurring theme of academic and popular discourse, reflected both in work that assesses the strength of institutions by the degree to which they are self-enforcing, and in concerns expressed about how the actions of some political elites erode institutional norms.[5] It appears to us that for a shared understanding to be self-enforcing, it must be immune to both individual and coalitional deviations. As we know from countless cases of corruption and coups, individuals who are supposed to be punished often are able to evade sanctions by profitably bribing their punishers and partnering with them. Analogously, when the protest of multiple citizen groups is needed to oust a political leader, that leader may offer patronage to some of those groups to retain power. With this issue in mind, we study when a shared understanding of sanctions and rewards is credible from the perspective of not only individual but also coalitional deviations. In other words, if all players share an understanding of future behavior, when is it that no coalition of players finds it profitable to deviate today?

Our results suggest that the observability of transfers plays an important role in enforcement. If transfers are observable, laws may successfully implement a large range of outcomes without encouraging any coalition to deviate. But once parties can make secret side-payments, there is less that laws can implement that are immune to coalitional deviations.

We illustrate some of these ideas in Section 5 where we study pure division problems in which a group of players choose how to divide resources. These games are studied as *simple games* (Von Neumann and Morgenstern 1945) in cooperative game theory, and feature in the study of legislative bargaining (Baron and Ferejohn 1989). We study the degree to which history-dependent interactions can motivate political elites to share resources with those who are not elites. We show

[5]This is a vast literature across the social sciences, some examples of which are Posner (1997), Weingast (1997), Przeworski and Maravall (2003), Aghion, Alesina and Trebbi (2004), Acemoglu, Egorov and Sonin (2010, 2012), Fearon (2011), Bidner and Francois (2013), Francois, Rainer and Trebbi (2015), Acemoglu and Jackson (2017), Mailath, Morris and Postlewaite (2017), and Acemoglu and Wolitzky (2018, 2019).

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 literature that combines elements from both cooperative and non-cooperative game theory to understand stable social arrangments; 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).[6] 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 is closely related to the study of farsighted stability in coalitional games, surveyed in Ray (2007) and Ray and Vohra (2015a). One 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), Xue (1998), Diamantoudi and Xue (2003), Jordan (2006), Ray and Vohra (2015b), Dutta and Vohra (2017), Kimya (2019), and Vohra ⓡ Ray (2019). While most of this literature considers chains of deviations in a way that is history-independent, Dutta and Vartiainen (2019) illustrate how history-dependence can guarantee existence across a general class of games.

A more closely related strand, initiated by Konishi and Ray (2003), studies real-time coalition-formation processes, and our solution-concept builds on their's. Their framework is dynamic and cooperative: coalitional structures generate payoffs 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.[7] Behavior in this setting is "Markov," where coalitions condition their behavior only on the 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 notable 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, and we have not seen analogues of our

[6] Analogously, there is growing interest in modeling dynamic reasoning in matching; see Corbae, Temzelides and Wright (2003), Damiano and Lam (2005), Du and Livne (2016), Kadam and Kotowski (2018a,b), Doval (2018), Liu (2019), and Kotowski (2019).

[7] 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.

folk and anti-folk theorems in this prior literature. Incidentally, the direction in which we proceed is suggested in the conclusion of Ray (2007, pp. 301) as being a potentially important direction for further research on coalitional games:

> *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 the innovative model of 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 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. Since individuals have no individual actions in their model, our results establish that all payoffs are sustainable (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 closely related issue in the context of relational contracting between a long-run Principal and a sequence of short-run agents. They capture a beautiful and realistic friction: secret side-payments exposes the Principal to extortion by shirking agents. Our results are complementary in that the strategic issue of our paper is not that of extortion but of being able to structure transfers in a way that allows coalitions to deviate while shielding its members from excessive punishment. 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 in repeated games adopt cooperative criteria to select equilibria. 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.[8] Also related is the important work on 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

[8]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.

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 our ideas 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. The table below describes their stage-game payoffs:

	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).

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.

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. We call such history-dependent matching processes *stable conventions*.

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 $\delta \geq \frac{1}{2}$.

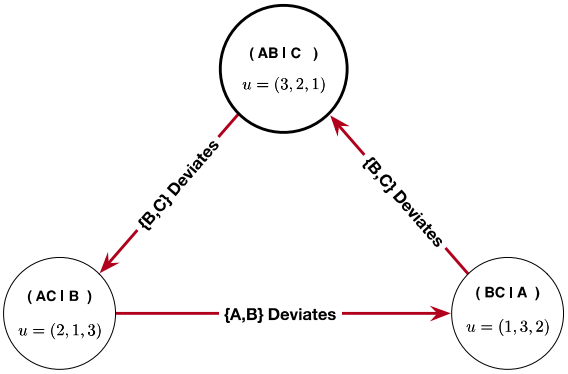


FIGURE 1. A stable convention for the roommates' problem if $\delta \geq 1/2$.

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, where the match in each period is enforced through future history-dependent matches.[9] The rule specified above is not the only stable convention; in fact, every feasible payoff vector in the above game (if players are sufficiently patient) may be enforced through some configuration of carrots and sticks. However, if players can make side-payments that cannot be conditioned upon in future play, every convention is unstable and undermined by some scheme of coalitional deviations and secret side-payments.

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 pass any division of the dollar, but the coalition of $\{2, 3\}$ is powerless (as is any singleton).[10] Here, player 1 is an elite veto player who needs the support of one other (non-elite) player to capture the surplus. The core of this stage game involves player 1 capturing the entire dollar; every other allocation guarantees that she and one other player has a profitable joint deviation.

History-dependence in a repeated bargaining problem can do more. Suppose that now, in

[9] 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.

[10] Ray and Vohra (2015b) and Dutta and Vohra (2017) also use this example to illustrate their approaches; we thank Elliot Lipnowski for suggesting that we do so.

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 use a "core-reversion" convention to enforce more here.

Consider a convention that recommends the allocation $\left(0, \frac{1}{2}, \frac{1}{2}\right)$ every period so long as that has been the division 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 $\delta \geq \frac{1}{2}$:

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

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 below. In Section 5, 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.

By contrast, our anti-folk theorem result implies that once coalitions can make secret transfers, then the only supportable outcome is the core of the stage game, where the elite veto player captures the entire surplus.

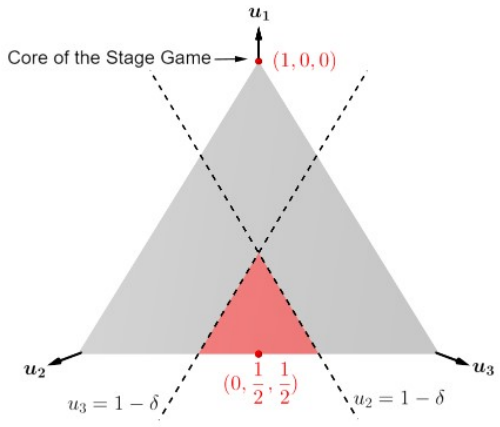


FIGURE 2. Supportable payoffs using core-reversion if $\delta \geq \frac{1}{2}$.

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 $v(a) \equiv (v_1(a), \ldots, v_n(a)) \in \mathbb{R}^n$, and we use $v : A \to \mathbb{R}^n$ to denote the 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). We assume the following about these correspondences:

Assumption 1. ***(Reflexivity)*** *For every coalition C and alternative a, $a \in E_C(a)$.*

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

Assumption 1 guarantees that a coalition can block an outcome without necessarily changing the alternative; this assumption is used when we later augment the game with transfers. Assumption 2 guarantees that it is feasible for the grand coalition to deviate and choose any alternative. All of these assumptions are satisfied in a number of settings commonly studied in the literature, as we illustrate below.

Example 1. Consider a strategic-form game in which each player's action set is A_i, the set of action profiles is $A \equiv A_1 \times \ldots \times A_n$. 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 coalitional or characteristic function game (N, U) where the mapping $U(C) \subseteq \mathbb{R}^{|C|}$ 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. Now let an alternative $a = (\pi, u)$ where π is a partition and u is a feasible payoff vector given that partition. The effectivity correspondence $E_C(a)$ specifies the set of alternatives to which coalition C may move,[11] 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.[12]

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 $\sigma : \mathcal{H} \to \mathcal{O}^{NTU}$. 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] It is natural to impose restrictions on this effectivity correspondence. For example, one may require in the spirit of *coalitional sovereignty* (Ray and Vohra 2015b) that (i) $(\pi', u') \in A$ is an element of $E_C((\pi, u))$ only if C is a union of members of π'; and (ii) π' has as a member any C' such that $C' \cap C = \emptyset$ and $C' \in \pi$.

[12] 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.

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 conventional 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

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

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 recursive payoff similar to that in Definition 3 in a stochastic game where players condition on the current coalitional structure (but not on past history). Vartiainen (2011) augments this solution-concept to allow for history-dependence and studies a setting without discounting. In the context of repeated elections, Bernheim and Slavov (2009) study *Dynamic Condorcet Winners*, which coincides with stable conventions when we specialize our stage game to their's.

3.3 What Can Be Enforced By Stable Conventions?

The previous section defined stable conventions. Here, we turn to the limits of their enforceability. We establish that for NTU games, every payoff that is feasible and "individually" rational can be implemented in a stable convention, if players are sufficiently patient.

Analogous to the (pure-action) minmax of noncooperative repeated games, let us define each player's minmax payoff as the lowest payoff that she attains when she has the opportunity to best-respond to the recommendation:

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

Based on this minmax payoff, let us define the set of feasible and strictly individually rational payoffs. The set of feasible payoffs is $\mathcal{V}^\dagger \equiv \text{co}(\{\widetilde{v} \in \mathbb{R}^n : \exists a \in A \text{ such that } \widetilde{v} = v(a)\})$ where $\text{co}(S)$ is the convex hull of a set of payoff profiles S, and the subset of these 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 set of results.

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 the folk theorem is nearly identical to that for sub-game perfect equilibria (Fudenberg and Maskin 1986; Abreu, Dutta and Smith 1994), with the differences being 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]

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. We model transfers separately from collective choices to sharpen the contrast to the secret transfers case. We begin with preliminaries to define the game before stating our results.

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 $\mathcal{O}^{TU} \equiv \Big\{o = (a,C,T) | a \in A, C \in \mathcal{C}, T \in \mathcal{T}\Big\}$. 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

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

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

$$\mathcal{U}^\dagger_{IR} \equiv \left\{u \in \mathcal{U}^\dagger : u_i > \underline{v}_i \text{ for every } i = 1, \ldots, n\right\}. \tag{TU Feasible IR}$$

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 payoff. To avoid Ponzi schemes, for all of our results, we restrict attention to conventions whose continuation values lie in a bounded set.

Assumption 3. *We consider conventions σ such that continuation values are bounded across histories: $\{u \in \mathbb{R}^n : \exists h \in \mathcal{H}$ such that $U(h|\sigma) = u\}$ is a bounded subset of $\mathbb{R}^n$.*

With these preliminaries defined, we can 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 for every i in C,

$$(1-\delta)u_i(a', [T'_C, T_{-C}(h|\sigma)]) + \delta U_i(h, a', C, [T'_C, T_{-C}(h|\sigma)]|\sigma) > U_i(h|\sigma) \tag{2}$$

Because transfers are publicly observed, subsequent behavior may be conditioned on these transfers when coalition C blocks the recommended outcome. We use this tool to prove a folk theorem analogous to Theorem 1.

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 for Theorem 2 is similar to that of Theorem 1. Transfers ensure that players have opposed interests in the stage game, so NEU in this augmented game is automatically satisfied. The complication introduced by transfers is that if the deviating coalition C anticipates a certain member to be punished, other members can transfer utility to her to compensate for the subsequent punishment. These transfers can potentially undermine the deterrence effect of future continuation, even as $\delta \to 1$. To overcome this problem, the convention targets the player who gained least from the deviation after transfers are made.

4 Secret Transfers Undermine Conventions

We see in Section 3.4 that side-payments alone do not undermine the power of expectations: by punishing players for giving or receiving transfers, the convention ensures that coalitions do not deviate. In this section, we see a different conclusion emerges once coalitions can make secret side-payments. Section 4.1 describes the secret-transfers setting that we analyze. Section 4.2 proves a one-shot coalitional deviation principle. Section 4.3 proves our result that each coalition can guarantee itself a coalitional minmax value. Section 4.4 establishes our formal result connecting behavior to the efficient β-core.

4.1 The Setup

We say that transfers are *secret* when the convention cannot condition on the amount of those payments. In other words, the future 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. Our analysis isolates this monitoring imperfection as being critical in a repeated coalitional setting.

This form of secrecy is a measurability restriction. Consider two $(t+1)$-length histories $h = (a^0, C^0, T^0, \ldots, a^t, C^t, T^t)$ and $\widetilde{h} = (\widetilde{a}^0, \widetilde{C}^0, \widetilde{T}^0, \ldots, \widetilde{a}^t, \widetilde{C}^t, \widetilde{T}^t)$. We say that h and $\widetilde{h}$ are **identical up to transfers within blocking coalitions** if they are of the same length, and the alternative chosen, the identity of the blocking coalition if any, and transfers made outside the blocking coalition are all identical across these two histories:

$$\text{For every } 0 \leq \tau \leq t: \quad a^\tau = \widetilde{a}^\tau,\ C^\tau = \widetilde{C}^\tau,\ T_{-C_\tau} = \widetilde{T}_{-\widetilde{C}_\tau}.$$

In other words, the only potential difference between histories h and $\widetilde{h}$ is in the transfers made within blocking coalitions. This is the information that we model as being secret from the convention.

Definition 5. A convention σ **respects secret transfers** if $\sigma(h) = \sigma(h')$ for any $h, h' \in \mathcal{H}$ that are identical up to transfers within blocking coalitions.

A stable convention that respects secret transfers is one that satisfies both Definitions 4 and 5.[15] To be transparent about what Definition 5 entails: because players outside blocking coalitions do not observe transfers within a blocking coalition, their actions cannot condition on them. Definition 5 also assumes that members of blocking coalitions do not condition their subsequent *equilibrium* play on the transfers made within the blocking coalition. This measurability restriction may be stronger than secrecy, but we view there to be several rationales for it. First,

[15] A special case of a convention that respects secret transfers is one that ignores transfers altogether between any pair (blocking or otherwise).

this 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. Second, one may envision that mechanisms or continuation play that attempt to elicit private information from deviators (about their transfers) might themselves be vulnerable to coalitional deviations, so it's unclear that information about transfers can be credibly elicited from members of a blocking coalition. Third, secret transfers generate persistent private information, and it is beyond the scope of existing tools to characterize coalitional behavior that is both dynamic and conditions on persistent private information. Given all of these reasons, we view this to be a useful starting point to investigate how coalitions may use secret side-payments.[16]

4.2 A One-Shot Coalitional Deviation Principle

Our central result is that secret transfers are a destabilizing force that undermines intertemporal incentives, and guarantees that each coalition obtains its coalitional minmax value. At the core of this result is a one-shot *coalitional* deviation principle: we prove that for conventions that respect secret transfers, the existence of a profitable multi-shot coalitional deviation implies that of a profitable one-shot coalitional deviation. Hence, any stable convention in this setting must also be immune to profitable multi-shot coalitional deviations.

We begin by defining 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 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.*

This result has important implications. A challenge central to coalitional behavior is that they cannot commit to long-term deviations, and thus, can be potentially defeated by the power of expectations. Lemma 1 establishes that once coalitions can make secret transfers to each other, such long-term commitments are no longer needed: if a coalition can jointly gain from a long-term commitment to a multi-shot deviation, they can structure their short-term deviations alongside

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

transfers to obtain those gains.[17] Being able to bribe other players to join one's coalition and to not be punished for it is an important tool that can protect coalitions from the power of intertemporal incentives.[18]

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 prove that in a stable convention, for every discount factor, each coalition can guarantee itself a total payoff below which it cannot be pushed down to by the convention. We define this *coalitional minmax* as follows:

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

The coalitional minmax builds on standard individual minmaxes in a natural way, treating coalition C as a hypothetical entity whose total value is the sum of the payoffs of its constituents, and with an ability to best-respond represented by $E_C(\cdot)$. 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.[19] We argue that each coalition can guarantees itself

[17] We have described Lemma 1 in a setting where every blocking coalition can make secret transfers. An analogous coalition-specific result holds in a more general setting where only some coalitions can make secret transfers; for those coalitions, the one-shot coalitional deviation principle applies.

[18] In closely related work, Barron and Guo (2019) illustrate how secret transfers can destroy the power of relational contracting between a long-run agent and a sequence of short-run opponents, and cooperation may be restored when those transfers are observable. Our results are complementary in that the economic channel here is that of side-payments that evade punishment whereas they study the issue of extortion.

[19] One subtle difference is that the β-characteristic function is often used to convert a strategic-form game into a cooperative game. We are preserving the same logic, but applying it to an abstract transferable utility game,

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 $\underline{v}_C$.*

The argument for Theorem 3 has a straightforward conceptual structure. 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.[20]

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 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). In certain cases, the set of payoffs where each coalition obtains at least its minmax value is empty. We do not view this as a negative conclusion, but rather as a stark illustration of how short-term coalitional deviations coupled with the ability to make side-payments secretly severely erodes the power of conventions.[21]

Because the grand coalition must also achieve its coalitional minmax, and is omnipotent (Assumption 2), Theorem 3 also implies that stable conventions implement only efficient alternatives. We hesitate to interpret this as a positive result for the (usual) reason that the active players modeled in a game may not include all those whose utilities are relevant for welfare-evaluation. For example, if all of the active players are firms (or political leaders)—and not consumers (or citizens)—one may not be sanguine about those active players being able to effectively collude to maximize their payoffs.

4.4 The Efficient β-Core

We use the idea that only efficient alternatives are selected to derive a tighter result. Let us denote the *efficient alternatives* by $\overline{A} \equiv \arg\max_{a \in A} \sum_{i \in N} v_i(a)$. We define *efficient coalitional minmaxes* as the lowest payoff that a coalition can be pushed down to using only efficient alternatives:

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

including those that lack a product-structure.

[20] 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 $\underline{v}_C$ with a multi-shot deviation that best-responds to the recommendation. Because Lemma 1 still applies, a stable convention then cannot push a coalition's value below this minmax.

[21] This issue would not arise in the absence of the possibility for coalitional deviations. In a standard repeated strategic-form game, augmenting the game with transfers that are made simultaneously with actions could only expand and not reduce the set of supportable payoffs.

Naturally, for every coalition C, $\underline{v}^e_C$ is weakly higher than $\underline{v}_C$, since the restriction to using efficient alternatives diminishes the capability of the convention to punish coalitions. With this in mind, let us define the efficient β-core:

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

$$\mathcal{B} \equiv \left\{ u \in \mathbb{R}^N : \sum_{i \in N} u_i = \max_{a \in A} \sum_{i \in N} v_i(a),\ \sum_{i \in C} u_i \geq \underline{v}^e_C \text{ for all } C \neq N \right\},$$

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

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

The efficient β-core is the set of efficient payoffs that gives each coalition at least its efficient coalitional minmax. The strict efficient β-core is in the relative interior of the efficient β-core, where each non-grand coalition obtains strictly more than its efficient coalitional minmax. We prove the following result below.

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.*

This result is a tight folk theorem for the setting with secret transfers: the set of payoff profiles supportable by stable conventions is a subset of the efficient β-core; moreover, any payoff profile within the strict efficient β-core is supportable so long as players are sufficiently patient. Accordingly, our result offers a connection between the efficient β-core and payoffs sustained by stable conventions.

The argument for the first part of the result mirrors Theorem 3, but now embedding the restriction that stable conventions can select only efficient alternatives. The second part of the result treats each coalition—apart from the grand coalition—as a hypothetical player, and constructs "player-specific" punishments for these hypothetical players.[22] Once these "player-specific" punishments are generated, we use an argument analogous to that of Fudenberg and Maskin (1986) (and Theorems 1 and 2) to push each of these hypothetical player arbitrarily close to its efficient coalitional minmax.

[22]While we do this step directly, one can see that this is feasible because the payoffs of these coalitions satisfy the NEU condition in the game augmented with transfers.

5 An Application to Simple Games

Here, we specialize our analysis to *simple games* (Von Neumann and Morgenstern 1945), revisiting and generalizing our analysis of the example in 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. Simple games are relevant both for cooperative games (Ray and Vohra 2015b; Dutta and Vohra 2017) and are extensively studied in the vast literature on legislative bargaining that builds on Baron and Ferejohn (1989) (see Eraslan and Evdokimov 2019 for a survey).

We study a particular sub-class of simple games: namely those games where no single player is a dictator, but some are political elites. We ask the following questions. First, for fixed discount factors, to what degree can history-dependence support outcomes where political elites share their resources with non-elites? Second, what happens when coalitions can make secret transfers?

To apply our approach, let us re-formulate simple games in the language of our model. The set of alternatives is the division of a dollar among n players: $A \equiv \{a \in \mathbb{R}^N_+ : \sum_{i \in N} a_i = 1\}$, where player i's generated payoff from alternative a is $v_i(a) \equiv a_i$. Some coalitions have the ability to affect the outcome whereas others do not: 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. Formally, for each a, $E_C(a) = A$ if $C \in \mathcal{W}$, and $E_C(a) = \{a\}$ otherwise. As is standard, we assume that $\mathcal{W}$ is *monotonic* and *proper*.[23]

To define political elites, we call player i a *veto player* if she is a member of every winning coalition. The collection of all veto players—referred to as the *collegium*—is $D \equiv \cap_{C \in \mathcal{W}} C$, and a *collegial game* is that in which D is non-empty. Our analysis studies collegial games that are non-dictatorial; in other words, every winning coalition has at least two members.

Non-dictatorial collegial games are of interest to political economy 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 2000a,b). 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.

While non-elites may have *de jure* power in all of these cases—e.g., if the set of veto players, D, is not itself a winning coalition—they have no *de facto* power when interacting only once. Indeed, the core of the stage game involves elites sharing all of the surplus among themselves, and not at all with non-veto players. Our interest is in comparing that outcome with what can be sustained through stable conventions when the players interact repeatedly.

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

We have already seen in Section 2.2 how history-dependence can sustain a larger set of outcomes by using reversion to the stage-game core as a punishment. Here, we consider a broader class of conventions and punishments, using approaches from Abreu (1988) and Abreu, Pearce and Stacchetti (1990).[24] We prove the following result. (Below, Δ refers to the non-negative n-dimensional unit simplex.)

Theorem 5. *Under 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 :

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

b. If there is only a single veto player, there exists $\underline{\delta}$ such that the set of supportable payoffs 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: $K \equiv \left\{ u \in \Delta : \sum_{i \in D} u_i = 1. \right\}$.

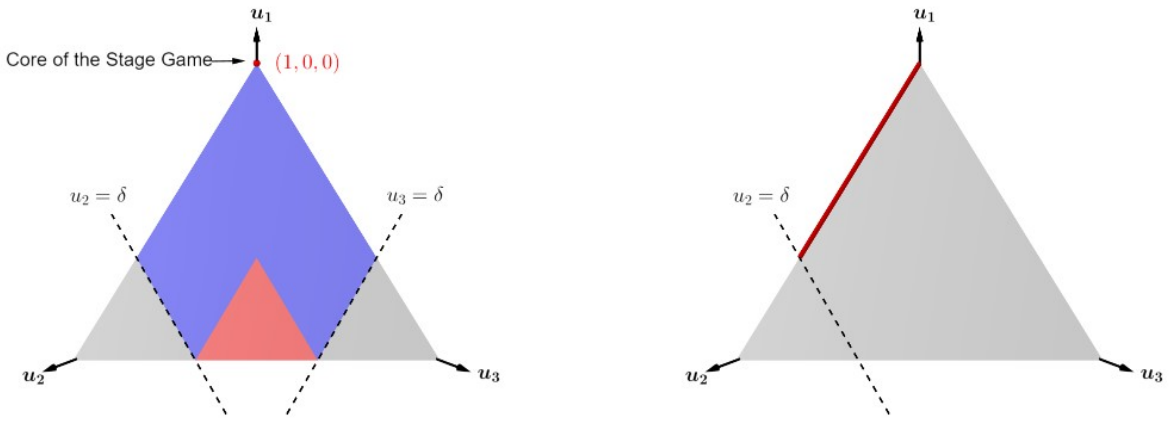
Theorem 5 describes the set of supportable payoffs when monitoring is perfect, with and without transfers. To interpret this result, consider the case of there being at least two veto players. At $\delta = 0$, this set coincides with the core of the stage game. As players become more patient, a larger range of payoffs is supportable: the set of supportable payoffs, $U(\delta)$, is strictly increasing in δ (in terms of set-inclusion). The only way for $U(\delta)$ to expand is by allowing non-veto players to potentially capture larger shares of the surplus (K is a subset of $U(\delta)$ for all δ).[25] By contrast, with secret transfers, the set of supportable payoffs shrinks to the core of the stage game, giving veto players the entire surplus.

These results illustrate the importance 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. However, that ability is lost once elite players can co-opt others with secret side-payments.

To prove Theorem 5, we consider optimal penal codes that feature the worst possible punishment for players, namely giving them 0. However, our coalitional setting introduces a subtlety absent in standard repeated games. In standard repeated games, after a player deviates, continuation play can condition only on the *identity* of the deviator without being sensitive to how she

[24] 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.

[25] 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 5 implies that the wealthiest $n - |D| - k$ non-veto players cannot together obtain more than a δ fraction of the surplus.



(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. (B) depicts how secret transfers reduces the set of supportable outcomes once coalition $\{1, 2\}$ can make secret transfers.

deviated (Abreu 1988). By contrast, in our setting, if a coalition deviates, being sensitive to how it does so is important for constructing the optimal penal code.

Figure 3 illustrates these results by revisiting our example in Section 2.2: this is a 3-player divide-the-dollar game where player 1 is a veto player. Figure 3(A) illustrates the set of supportable payoffs for perfect monitoring, with and without transfers. The red region depicts payoffs supported by core-reversion, and the blue region illustrates the gains that come from using the approach here. We see that the region of supportable payoffs expands to all those that give each of players 2 and 3 less than δ. Figure 3(B) illustrates how once coalition $\{1, 2\}$ can make secret transfers, the set of supportable payoffs reduces to those where player 3 is completely excluded and obtains 0.[26] If all coalitions can make secret transfers, then the only supportable payoff is the core of the stage game.

6 Conclusion

This paper models self-enforcing conventions for repeated games in which coalitions can commit to short-term deviations but not long-term behavior. We investigate the degree to which the motive to deviate as coalitions is disciplined by players' shared understanding of the future. We find that when all behavior is perfectly observed, then stable conventions can support every feasible and strictly individually rational payoff vector, so the possibility for coalitional deviations comes at little cost when players are patient. By contrast, if coalitions can make secret transfers to each other, then they can guarantee themselves a minimal "coalitional minmax" value regardless of

[26] While our formal results do not cover this case, it is straightforward to extend Lemma 1 and Theorem 3 to argue that if a coalition can make secret transfers, then it obtains its coalitional minmax regardless of whether other coalitions can make secret transfers.

players' patience. In cooperative games without externalities, the set of supportable payoffs is then reduced to the core of the stage game; without externalities, the set is a subset of the β-core.

We view our results to have both theoretical and applied import. On the theoretical side, our framework and solution-concept offer a tractable merger of important ideas 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 think both about repeated cooperative games as well as coalitional deviations in repeated noncooperative games. The recursive nature of our solution-concept makes it feasible to analyze the set of supportable payoffs in applications using standard self-generation techniques.

On the applied side, we view our results as speaking to important issues of enforcing law, power, and social order. An important consideration in the design of legal and political institutions is the degree to which the temptation to violate the law or abuse political power is disciplined by players' expectations of how these actions affect the future. Our results suggest that monitoring side-payments is critical for the credibility of enforcement: in settings where players can secretly bribe others, they are less threatened by the prospect of future punishment.

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 (2010) "Political Selection and Persistence of Bad Governments," *Quarterly Journal of Economics*, Vol. 125, No. 4, pp. 1511–1575.

——— (2012) "Dynamics and Stability of Constitutions, Coalitions, and Clubs," *American Economic Review*, Vol. 102, No. 4, pp. 1446–76.

Acemoglu, Daron and Matthew O. Jackson (2017) "Social Norms and the Enforcement of Laws," *Journal of the European Economic Association*, Vol. 15, No. 2, pp. 245–295.

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.*

Aghion, Philippe, Alberto Alesina, and Francesco Trebbi (2004) "Endogenous Political Institutions," *Quarterly Journal of Economics*, Vol. 119, No. 2, pp. 565–611.

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.

Baron, David P. and John A. Ferejohn (1989) "Bargaining in Legislatures," *American Political Science Review*, Vol. 83, No. 4, pp. 1181–1206.

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.

Bidner, Chris and Patrick Francois (2013) "The Emergence of Political Accountability," *Quarterly Journal of Economics*, Vol. 128, No. 3, pp. 1397–1448.

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.

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.

Diamantoudi, Effrosyni and Licun Xue (2003) "Farsighted Stability in Hedonic Games," *Social Choice and Welfare*, Vol. 21, No. 1, pp. 39–61.

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.

Eraslan, Hülya and Kirill S. Evdokimov (2019) "Legislative and Multilateral Bargaining," *Annual Review of Economics*, Vol. 11, No. 1.

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

Fearon, James D. (2011) "Self-Enforcing Democracy," *Quarterly Journal of Economics*, Vol. 126, No. 4, pp. 1661–1708.

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 (2018a) "Multiperiod Matching," *International Economic Review*, Vol. 59, No. 4, pp. 1927–1947.

——— (2018b) "Time Horizons, Lattice Structures, and Welfare in Multi-Period Matching Markets," *Games and Economic Behavior*, Vol. 112, pp. 1–20.

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.

McCarty, Nolan M. (2000a) "Presidential Pork: Executive Veto Power and Distributive Politics," *American Political Science Review*, Vol. 94, No. 1, pp. 117–129.

——— (2000b) "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.

Posner, Richard A. (1997) "Social Norms and the Law: An Economic Approach," *American Economic Review*, Vol. 87, No. 2, pp. 365–369.

Przeworski, Adam and José M. Maravall (2003) *Democracy and the Rule of Law*, Vol. 5: Cambridge University Press.

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.

Weingast, Barry R. (1997) "The Political Foundations of Democracy and the Rule of Law," *American Political Science Review*, Vol. 91, No. 2, pp. 245–263.

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

Xue, Licun (1998) "Coalitional Stability under Perfect Foresight," *Economic Theory*, Vol. 11, No. 3, pp. 603–627.

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 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-responses 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 $\mathbb{R}^N$ 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. 14

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.[27] 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

[27] 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

true for every i:

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

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

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 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 $\delta^{L(\delta)}\in[\delta^{\frac{\log\kappa}{\log\delta}},\delta^{\frac{\log\kappa}{\log\delta}+1}]=[\delta\kappa,\kappa]$. Therefore, $\lim_{\delta\to1}\delta^{L(\delta)}=\kappa$.

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

- $W\equiv\left\{w(d,\tau)|0\leq d\leq n,\tau\geq0\right\}\cup\{\underline{w}(i,\tau)|1\leq i\leq n,0\leq\tau<L(\delta)\}$ 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

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

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

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

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

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

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

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

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, 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

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

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

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

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

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

where the first equality follows from taking limits, the second from $\lim_{\delta\to 1}\delta^{L(\delta)} = \kappa$, 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 $(1-\widetilde{\kappa})v_{j^*}(\underline{a}_i) + \widetilde{\kappa}v_{j^*}^i$ 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. 18

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 3, 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. 20

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. Consider an alternative $d(h) \in BR_C(a(h|\sigma))$ 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

B.1 Preliminary Results

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

Lemma 3. *Suppose σ is a stable convention. 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*

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

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 $|v_i(\overline{a})| \leq diam(\mathcal{V}^{\dagger}_{IR})$, we have

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

□

Lemma 4. *Suppose σ' is a profitable multi-shot coalitional deviation 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 is bounded from below regardless of h. Since j 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 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

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

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

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

The deviation σ'' is still profitable. □

B.2 Proof of Theorem 2 on p. 16

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 3, the convention σ has bounded continuation value. Moreover, Assumption 3 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.*

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

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 $\mathcal{U}^{\dagger}_{IR} \subseteq co(\widetilde{u}^1, \ldots, \widetilde{u}^M)$.*

Proof. By definition,

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

Since $\overline{\mathcal{U}}^{\dagger}_{IR}$ is a bounded polyhedron, it is also a polytope. Let $x^1, \ldots, x^K$ be its vertices. Any point inside $\mathcal{U}^{\dagger}_{IR}$ 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 $\mathcal{U}^{\dagger}_{IR} \subseteq \text{co}(\cup_{1 \leq k \leq K} \{\widetilde{u}^{k,1}, \ldots, \widetilde{u}^{k,m_k}\})$. □

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 $\mathcal{U}^{\dagger}_{IR} \subseteq \text{co}(\widetilde{u}^1, \ldots, \widetilde{u}^M)$, 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 $\epsilon < (1-\kappa)\min\{\min_{i,j \neq i}(u^j_i - u^i_i), \min_i u^i_i - \underline{v}_i\}$, and given that ϵ, consider δ exceeding the appropriate $\underline{\delta}$.

Now 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 $\{u^i\}_{i=1}^n \subseteq \mathcal{U}^{\dagger}_{IR}$ when ϵ is sufficiently small, and that for all i, $u^i_i < u_i$ and for all $j \neq i$, $u^j_i > u^i_i$. 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:

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

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

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 $\lim_{\delta \to 1} \delta^{L(\delta)} = \kappa$.

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

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

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

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

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

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

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

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

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

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

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

Furthermore,

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

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,

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

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

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

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. 21

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 fee summing

up to ϵ; by contrast, players outside of C are paid equally, with a total of amount also summing up to ϵ. This fee 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

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

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 $\underline{a}^e_C \in \arg\min_{a\in\overline{A}} \max_{a'\in E_C} \sum_{i\in C} v_i(a')$ be an efficient alternative that can be used to minmax player i. Note that by construction, $\sum_{i\in C} v_i(\underline{a}^e_C) \leq \underline{v}^e_C$. 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:

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

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

Inequality (23) implies that in terms of total value, coalition C is willing to bear the cost of $\sum_{i\in C} v_i(\underline{a}^e_i)$ with the promise of transitioning into its coalition-specific punishment rather than staying at its minmax. By an argument identical to that 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 $\lim_{\delta\to 1} \delta^{L(\delta)} = \kappa$.

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

- $W \equiv \{w(C) : C \in \mathcal{C}\} \cup \{\underline{w}(C,\tau) | C \in \mathcal{C}\backslash\{N\}, 0 \leq \tau < L(\delta)\}$ 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Re-arranging terms:

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

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

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

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

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

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

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

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

for some $\widetilde{\kappa} \in [\kappa, 1]$.[28]

Examining the RHS of (26), observe that

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

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 $\sum_{i\in C'} v_i(\underline{a}^e_{C'}) \leq \underline{v}^e_{C'} < \sum_{i\in C'} u_i^C$, and the second weak inequality follows from $\widetilde{\kappa} \geq \kappa$ and $\underline{v}^e_{C'} < \sum_{i\in C'} u_i^{C'}$. Since $\widetilde{\kappa} \in [\kappa, 1]$, (24) delivers that $(1-\widetilde{\kappa})\sum_{i\in C'} v_i(\underline{a}^e_C) + \widetilde{\kappa}\sum_{i\in C'} u_i^C$ is strictly higher than $(1-\widetilde{\kappa})\underline{v}^e_{C'} + \widetilde{\kappa}\sum_{i\in C'} u_i^{C'}$. This term guarantees that (26) holds for sufficiently high δ.

B.4 Proof of Theorem 5 on p. 23

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.

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

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 $a^1 \equiv (1,0,\ldots,0)$ and $a^2 \equiv (0,1,0,\ldots,0)$ 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 $(a^1, \emptyset)$ regardless of history; for all $i \neq 1$, let σ^i be the convention that recommends $(a^2, \emptyset)$ 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 $(a^1, \emptyset, \mathbf{0})$ regardless of history; for all $i \neq 1$, let σ^i be the convention that recommends $(a^2, \emptyset, \mathbf{0})$ 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 $i \in C$,

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

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.

□