Buck-passing Dumping in a Pure Exchange Game of Bads

Takaaki Abe*

Abstract

We study stable strategy profiles in a pure exchange game of bads, where each player dumps his/her bads such as garbage onto someone else. Hirai et al. (2006) show that cycle dumping, in which each player follows an ordering and dumps his/her bads onto the next player, is a strong Nash equilibrium and that self-disposal is α-stable for some initial distributions of bads. In this paper, we show that a strategy profile of bullying, in which all players dump their bads onto a single player, becomes α-stable for every exchange game of bads. We also provide a necessary and sufficient condition for a strategy profile to be α-stable in an exchange game of bads. Moreover, we show that cycle dumping is the only dumping behavior that generates a strong Nash equilibrium. In addition, we show that repeating an exchange after the first exchange makes self-disposal stationary.

Keywords: bads; dumping; exchange; stability

JEL Classification: C72; C71

1 Introduction

A bad is a commodity or an object that causes disutility to its owner. Typical examples of bads include garbage, industrial waste, and pollutants. This paper seeks to address the following question: why does buck-passing dumping behavior exist everywhere? More specifically, why do a small number of individuals or nations receive and dispose of a large quantity of bads? We attempt to answer this question in terms of stability and dumping strategy. Therefore, we first need a model for people's dumping behavior. One of the pioneering models that formally deal with bads is the garbage disposal game introduced by Shapley and Shubik (1969). They assume that each player has a bag of garbage, namely, an initial endowment of bads. Each player dumps his/her bads into someone's yard. Shapley and Shubik (1969) model this situation as a cooperative game with transferable utility, in which if a coalition S of players is formed, then the players outside S dump all of their bads to coalition S, and the members of S similarly dump their bads to the outside players. Therefore, the quantity of bads the players in S have to dispose of is the sum of all bads dumped by the outside players to S. If the coalition of all players is formed, they dispose of all bads by themselves. Shapley and Shubik (1969) show that the core of this game is empty.

* School of Political Science and Economics, Waseda University. 1-6-1, Nishi-waseda, Shinjuku-ku, Tokyo 169-8050, Japan. Email: takatomo3639@asagi.waseda.jp

The author is grateful to Yukihiko Funaki and the participants of CREST Workshop 2019 at Ecole Polytechnique for their comments. The author gratefully acknowledges financial support from JSPS Grant-in-Aid for Research Activity Start-up (No. 19K23206) and Waseda University Grant-in-Aid for Research Base Creation (2019C-486).

Hirai et al. (2006) focus on strategic dumping. They replace goods by bads in the pure exchange game of goods analyzed by Scarf (1971). Therefore, a strategy of each player is to distribute his/her bads over the players. In a pure exchange game of goods, keeping all initial endowments is a dominant strategy for every player. However, if players are endowed with bads, keeping bads is not a rational strategy, and dumping all of one's bads to someone else is a dominant strategy. Hirai et al. (2006) show that if the number of types of bads is one (*e.g.*, garbage), then cycle dumping in which each player dumps his/her bads to the next player is a strong Nash equilibrium for any ordering of players. Moreover, the authors offer a sufficient condition that an initial distribution of bads should satisfy for self-disposal to be an α-core element.

Given that we are interested in players' dumping behavior, we should focus on "*who* dumps bads to *whom*" in a model. Therefore, we use the model proposed by Hirai et al. (2006) and consider a strategy to be a distribution of one's initial bads. Moreover, we formally introduce the notion of a *dumping function*. A dumping function describes a dumping behavior/policy that a player consistently follows for all initial distributions of bads: a dumping function assigns a strategy profile to every initial distribution of bads. Introducing this notion enables us to regard players' dumping behavior as a class of dumping functions and formally analyze its properties.

Besides the model, what stability notion should we use to analyze a strategy profile generated by each dumping function? There are six stability notions that have been widely accepted in the literature: $\alpha, \beta, \gamma, \delta$-cores, strong Nash equilibrium (SNE), and coalition-proof Nash equilibrium (CPNE). The α-core (C^α) and the β-core (C^β) concepts were proposed by Aumann and Peleg (1960) to model the payoffs that deviating players can achieve independent of the reaction of non-deviating players. The γ-core (C^γ) and the δ-core (C^δ) concepts were proposed by Chander and Tulkens (1997) and Currarini and Marini (2004), respectively. The notion of coalition-proof Nash equilibrium was introduced by Bernheim et al. (1987). Fortunately, the following equivalences hold for every pure exchange game of bads:

- $C^\alpha = C^\beta$,
- $C^\gamma = C^\delta = \emptyset$,
- $SNE = CPNE$.

Therefore, we focus on the concepts of α-core and strong Nash equilibrium. In this paper, a stable profile or an equilibrium does not necessarily mean a desirable profile because if a dumping profile in which a small number of players suffer a large quantity of bads is stable, then stability prevents individuals and coalitions from splitting off from the dumping profile. In contrast, if a dumping profile refers to an acceptable profile such as self-disposal, then we might consider stability to be a desirable property.

The rest of this paper is organized as follows. In Section 2, we introduce the model of a pure exchange game of bads and the definitions of α-stability and strong Nash equilibrium. The proposition of Hirai et al. (2006) is also discussed. In Section 3, we discuss the dumping functions that generate an α-stable strategy profile for every exchange game of bads. In Section 4, we show that cycle dumping is the only dumping function that generates a strong Nash equilibrium for every exchange game of bads. In Section 5, we show that introducing the second exchange may facilitate self-disposal. We conclude this paper by summarizing our results and proposing a direction for future research in Section 6. All proofs are

provided in the appendix.

2 Preliminaries

2.1 Pure exchange game of bads

Let $N = \{1, ..., n\}$ be a set of players. We assume $n \geq 3$. A coalition S is a nonempty subset of the player set: $S \subseteq N$. We denote by $|S|$ the number of members in coalition S. Every player $i \in N$ has an initial endowment of bads, given by $b^i > 0$. We assume that bads are homogeneous and divisible and the number of types of bads is one. Let $b = (b^1, ..., b^n) \in \mathbb{R}^N_{++}$ and $b^1 \leq ... \leq b^n$ without loss of generality. Let $B^N := \{b \in \mathbb{R}^N_{++} | b^1 \leq ... \leq b^n\}$.

Player i's strategy is a distribution of bads over players, given by a n-dimensional nonnegative vector $x^i = (x^{i1}, ..., x^{in}) \in \mathbb{R}^N_+$, where x^{ij} means the quantity of bads dumped by i to j. We define $X^i := \mathbb{R}^N_+$ and $X^S := \times_{i \in S} X^i$ for each coalition $S \subseteq N$. Once $b \in B^N$ is given, the strategy of player i is a distribution of b^i over all players: $x^i = (x^{i1}, ..., x^{in}) \in \mathbb{R}^N_+$ with $\sum_{j \in N} x^{ij} = b^i$. Let $X^i_b := \{x^i \in X^i | \sum_{j \in N} x^{ij} = b^i\}$ and $X^S_b := \times_{i \in S} X^i_b$ for each coalition $S \subseteq N$. Let x denote a strategy profile, $x = (x^1, ..., x^n)$.

For each player $i \in N$, define $v^i : X^N \to \mathbb{R}$ as follows: for any $x \in X^N$,

$$v^i(x) := u^i \left(\sum_{j \in N} x^{ji} \right),$$

where u^i is a strictly decreasing utility function, the input $\sum_{j \in N} x^{ji}$ of which represents the total quantity of bads player i receives after exchange x. We call $G_b = (N, \{X^i_b\}_{i \in N}, \{v^i\}_{i \in N})$ a pure exchange game of bads or simply an exchange game of bads. Note that an initial distribution $b \in B^N$ of bads yields a game G_b. Therefore, we can identify G_b with b. A dumping function $x : B^N \to X^N$ assigns a strategy profile $x(b) \in X^N_b$ to every initial distribution $b \in B^N$ of bads.

2.2 Stability

In this subsection, we introduce stability concepts. In the same manner as Shapley and Shubik (1969) and Hirai et al. (2006), we assume that players can form a coalition $S \subseteq N$ and have a joint strategy $x^S \in X^S$.

Definition 2.1. A coalition $S \subseteq N$ *deviates* from $x \in X^N$ if there is $y^S \in X^S$ such that

$$v^i(y^S, x^{N \setminus S}) > v^i(x) \text{ for all } i \in S.$$

A strategy profile x is a *strong Nash equilibrium* if no coalition deviates from x. A strategy profile x is said to be a *Nash equilibrium* if no one-person coalition deviates from x.

The concept of strong Nash equilibrium was defined by Aumann (1959). If a strategy profile x is a strong Nash equilibrium, no coalition has an incentive to deviate from profile x.

Aumann and Peleg (1960) define α-effectiveness to formulate the payoff the members of coalition S can achieve independently from the players outside S.

Definition 2.2. Let $x \in X^N$. A coalition $S \subseteq N$ is α-*effective* for x if there is $y^S \in X^S$ such that for any $z^{N\setminus S} \in X^{N\setminus S}$,

$$v^i(y^S, z^{N\setminus S}) > v^i(x) \text{ for all } i \in S.$$

A strategy profile x is α-*stable* if no coalition is α-effective for x.

The α-core of a game is the set of α-stable strategy profiles in the game. If coalition S is α-effective for x, the members of S can find a strategy y^S by which all members improve their payoffs regardless of strategy $z^{N\setminus S}$ chosen by other players. Therefore, an α-effective coalition represents a cautious attitude of players who try to split off from profile x: they take into account all possible reactions of outside players and decide to split off from profile x if their payoffs strictly increase even in the worst-case scenario.

2.3 Cycle dumping and self-disposal

We now introduce two dumping functions proposed by Hirai et al. (2006). For any coalition $S \subseteq N$, let σ_S be an ordering of all members of S and $\sigma_S(k)$ denote the kth player in ordering σ_S. For convenience, set $\sigma_S(|S|+1) := \sigma_S(1)$ and $\sigma_S(1-1) := \sigma_S(|S|)$ for any $S \subseteq N$. Moreover, let $\sigma_S(1) = \arg\min_{i\in S} b^i$ without loss of generality. If two or more players have the same minimal quantity of initial bads, then $\sigma_S(1)$ refers to the player whose player index is the smallest among those players. For each coalition $S \subseteq N$, let Ψ^S be the set of such orderings σ^S. For simplicity, for $S = N$, we omit N and write $\sigma := \sigma_N$ and $\sigma \in \Psi^N$.

For every ordering $\sigma \in \Psi^N$ and every player $i \in N$, let $\lambda^\sigma(i)$ denote the **predecessor** of i in ordering σ: for some index k such that $i = \sigma(k)$, $\lambda^\sigma(i) := \sigma(k-1)$. Similarly, let $\eta^\sigma(i)$ denote the **successor** of i in ordering σ; namely, $\eta^\sigma(i) := \sigma(k+1)$. If ordering $\sigma \in \Psi^N$ is fixed, we omit σ and write $\lambda(i)$ and $\eta(i)$.

Definition 2.3. Let $\sigma \in \Psi^N$. σ-*Cycle dumping* $x^\sigma : B^N \to X^N$ is a dumping function defined as follows: for any $b \in B^N$ and any $i \in N$,

$$x^\sigma(b)^{i\eta(i)} = b^i.$$

σ-Cycle dumping describes that every player follows ordering σ and dumps all his/her bads to the next player.

Definition 2.4. *Self-disposal dumping* $x^* : B^N \to X^N$ is a dumping function defined as follows: for any $b \in B^N$ and any $i \in N$,

$$x^*(b)^{ii} = b^i.$$

Self-disposal dumping means that every player does not dump bads to anyone else and disposes of his own bads by himself.

Proposition 2.5 (Hirai et al., 2006)**.**

i. For any $b \in B^N$ and any $\sigma \in \Psi^N$, strategy profile $x^\sigma(b)$ is a strong Nash equilibrium.

ii. If $\sum_{j=1}^m b^j \geq b^{m+1}$ for all $m = 1, ..., n-1$, then strategy profile $x^*(b)$ is α-stable.

The first statement means that every σ-cycle dumping generates a strong Nash equilibrium for any initial distribution of bads. In this paper, we show that cycle dumping is the only dumping function that

yields a strong Nash equilibrium for every exchange game of bads. We elaborate on this in Section 4.

The second statement shows that if an initial distribution of bads satisfies a condition, the self-disposal profile becomes α-stable. The sufficient condition requires that there is no "very big player m^*" such that $b^{m^*} > \sum_{j=1}^{m^*-1} b^j$. Since dumping all bads to someone else is a dominant strategy, self-disposal generates neither a Nash equilibrium nor a strong Nash equilibrium. Their condition suggests that the possibility of a "counterattack" incorporated in α-stability makes self-disposal stable for some distributions of bads because the strategies of the other players are fixed in the definition of a strong Nash equilibrium, while they may vary in that of α-stability. However, the self-disposal profile is not the only α-stable profile. On the contrary, there are many α-stable profiles in an exchange game of bads. To see this, in the following section we first provide a necessary and sufficient condition for a strategy profile to be α-stable.

3 Dumping functions generating α-stable strategy profiles

3.1 Necessary and sufficient condition for a strategy profile to be α-stable

Let $b \in B^N$. For any $x \in X_b^N$ and $i \in N$, let $r_x^i := \sum_{j \in N} x^{ji}$ denote the quantity of bads player i receives after exchange x. Therefore, $r_x = (r_x^1, ..., r_x^n)$ represents the distribution of bads that results from x.

Proposition 3.1. Let $b \in B^N$. A strategy profile $x \in X_b^N$ is α-stable if and only if for any $S \subseteq N$ there is $i \in S$ such that $\sum_{j \in N \setminus S} b^j \geq r_x^i$.

Once coalition S splits off from strategy profile x, each member of S may be counterattacked by outside players. The left-hand side of the inequality, $\sum_{j \in N \setminus S} b^j$, represents the maximum quantity of bads that may be dumped by players outside S to a member of S when player i splits off from x together with the other members of S. The right-hand side, r_x^i, is the quantity of bads player i receives when player i accepts profile x. Therefore, the inequality means that player i prefers to accept the result of exchange x rather than possibly cause the worst-case scenario by splitting off from x.

Proposition 3.1 is a useful result because it shows that r_x is the only information we need to determine whether x is α-stable. Strategy profile $x \in \mathbb{R}_+^{N \times N}$ contains the information on *who* dumps bads to *whom*, while $r_x \in \mathbb{R}_+^N$ is an n-dimensional vector that describes the result of exchange x, in which the first *who* is removed. In this sense, r_x is a reduction of x. Proposition 3.1 shows that α-stability of x depends on only r_x.

Moreover, since $\sum_{i \in N} r_x^i = \sum_{i \in N} b^i$, the two distributions r_x and b are on the same hyperplane. To see this, we fix, *e.g.*, $b = (3, 12, 15)$. The shaded area in Figure 1 is the set of distributions r_x satisfying the condition of Proposition 3.1, namely, $\{r_x | x \text{ is } \alpha\text{-stable}\}$. This set is clearly not convex. Since this example does not satisfy the "no big player" condition of Hirai et al. (2006), the self-disposal profile $(3, 12, 15)$ is not α-stable, as illustrated in the figure. In the following subsection, we introduce some new dumping functions that generate an α-stable strategy profile.

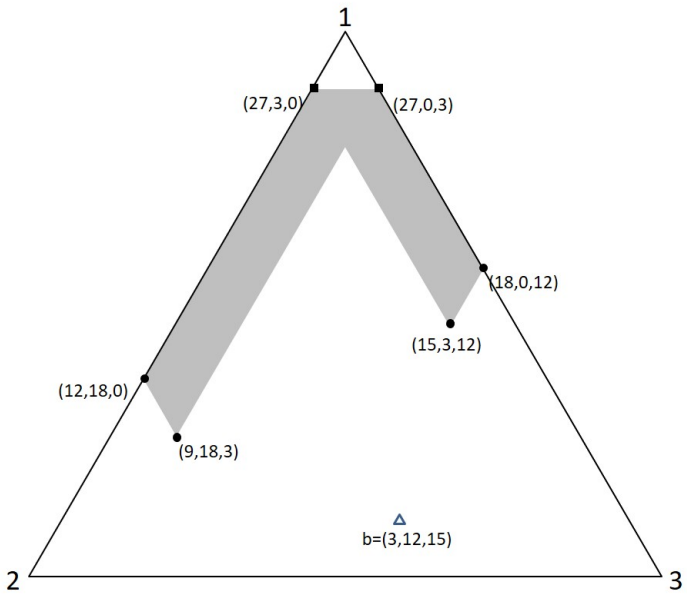

Figure 1 The set of distributions generated by α-stable strategy profiles.

3.2 Dumping functions with a focus and a cycle

Once an initial distribution $b \in B^N$ is fixed, Proposition 3.1 specifies the condition for a strategy profile x to be α-stable in the fixed game G_b. Then, is there a dumping function that always assigns an α-stable strategy profile to every $b \in B^N$? In other words, what dumping function always returns strategy profile x so that r_x can be inside the shaded area illustrated in Figure 1 for every $b \in B^N$? To answer this question, below we introduce two dumping functions, both of which model buck-passing dumping behavior. The following dumping function describes that all players except for player 1 dump all of their bads to player 1, and player 1 dumps his/her bads to an arbitrary player i.

Definition 3.2. Let $i \in N \setminus \{1\}$. *Focus dumping on* 1 *against* i, $\hat{x}^i : B^N \to X^N$, is a dumping function defined as follows: for any $b \in B^N$,

$$\hat{x}^i(b)^{j1} = b^j \text{ for all } j \in N \setminus \{1\},$$
$$\hat{x}^i(b)^{1i} = b^1.$$

Focus dumping describes bullying, in which all players pass the buck by dumping all bads to player 1, whose power of counterattack is the smallest, as $b^1 = \min_{j \in N} b^j$. Note that i denotes the player to whom player 1 dumps his/her bads. Since dumping function $\hat{x}^i$ is defined for every $i \in N \setminus \{1\}$, there are $n-1$ focus dumping functions, in each of which player 1 is the target of focus dumping.

The following dumping function is an extension of cycle dumping.

Definition 3.3. Let $\sigma \in \Psi^N$ and $i \in N \setminus \{\lambda(1)\}$. *$i$-Incomplete cycle dumping* $x^{\sigma i} : B^N \to X^N$ is a

dumping function defined as follows: for any $b \in B^N$,

$$x^{\sigma i}(b)^{j\eta(j)} = b^j \text{ for all } j \in N \setminus \{1\},$$
$$x^{\sigma i}(b)^{i1} = b^i.$$

This dumping function describes that player i violates cycle σ and dumps his bads b^i to player 1 instead of his successor $\eta(i)$. All players except for i follow cycle σ. For each ordering $\sigma \in \Psi^N$, all players except for the predecessor of player 1 can be such a cycle-breaking player. Therefore, there are $n-1$ incomplete cycle profiles for each ordering. Note that player 1 can also be a cycle-breaking player: in this case, player 1 retains his/her bads.

The following lemma shows that the dumping functions defined above do not overlap each other in the sense of convex combination. Moreover, Proposition 3.5 shows that every combination of these dumping functions yields an α-stable strategy profile for every $b \in B^N$.

Lemma 3.4. Let $n \geq 4$. For any $b \in B^N$ and any $\sigma \in \Psi^N$, no strategy profile of the following $2n-1$ profiles can be defined as a nonnegative convex combination of the other $2n-2$ profiles:

$$x^\sigma(b),\ \hat{x}^i(b) \text{ for all } i \in N \setminus \{1\},\ x^{\sigma i}(b) \text{ for all } i \in N \setminus \{\lambda(1)\}.$$

If $n = 3$, then the same holds for the following four profiles:

$$x^\sigma(b),\ \hat{x}^2(b),\ \hat{x}^3(b),\ x^{\sigma 1}(b).$$

Lemma 3.4 shows that for each $b \in B^N$, the above $2n-1$ dumping functions yield a convex hull without any slack. A combination of the dumping functions is distributive dumping: a player divides her initial bads into some parts and dumps them to some players, while in each extreme point of the convex hull, every player dumps all her initial bads to a single player. Note that for $n = 3$, the number of dumping functions reduces to four because the following two functions coincide for each ordering: $\hat{x}^{\sigma(2)}(b) = x^{\sigma 2}(b)$ for ordering (123) and $\hat{x}^{\sigma(2)}(b) = x^{\sigma 3}(b)$ for ordering (132).

For simplicity, we use $E(n, b, \sigma)$ to denote the set of the above $2n-1$ profiles for all $n \geq 4$ and that of the four profiles for $n = 3$.

Proposition 3.5. For any $n \geq 3$, any $b \in B^N$, and any $\sigma \in \Psi^N$, all nonnegative convex combinations of $E(n, b, \sigma)$ are α-stable.

We first consider an ordering σ and σ-cycle dumping x^σ, in which all players dump their bads according to cycle σ. Now, every player i splits $\epsilon \cdot b^i$ off from b^i with $0 \leq \epsilon \leq 1$ and dumps it to player 1, dumping the remaining $(1-\epsilon) \cdot b^i$ bads to his successor. Let x^σ_ϵ denote such a dumping. As ϵ increases, the quantity of bads dumped to player 1 increases, and the dumping profile x^σ_ϵ gets closer to focus dumping and finally reaches it at $\epsilon = 1$. In the same manner, we can consider all intermediate dumping functions between each incomplete cycle dumping and focus dumping, and those between cycle dumping and each incomplete cycle dumping. This result suggests that although cycle dumping yields a strong Nash equilibrium, even a small disturbance ϵ of cycle dumping leads players to focus dumping, in which player 1 ends up suffering all bads dumped by the other players.

Figure 2 illustrates the distributions obtained as a combination of the above dumping functions for $b = (3, 12, 15)$. In view of Proposition 3.5, ordering (123) yields convex combinations of $(15, 3, 12)$, $(18, 0, 12)$, $(27, 0, 3)$, and $(27, 3, 0)$. Similarly, ordering (132) generates convex combinations of $(12, 15, 3)$, $(15, 15, 0)$, $(27, 3, 0)$, and $(27, 0, 3)$. This proposition does not cover the convex hull of $(15, 15, 0)$, $(12, 18, 0)$, $(9, 18, 3)$, and $(12, 15, 3)$, which implies that there are other dumping functions that provide α-stable strategy profiles. In the next subsection, we introduce another class of dumping functions.

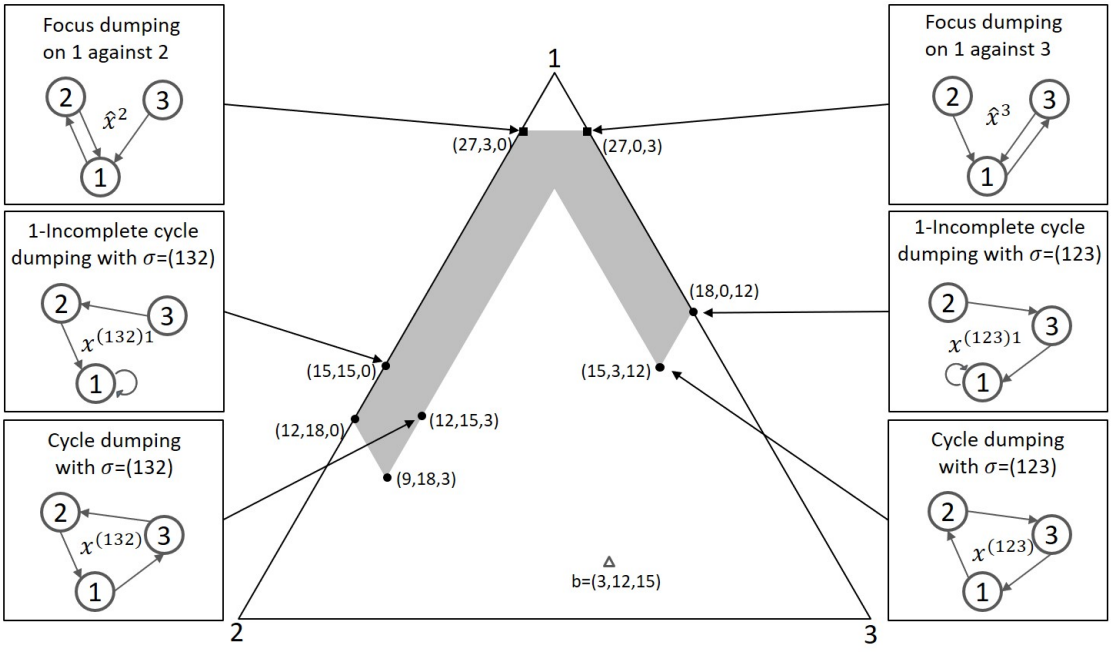

Figure 2 Proposition 3.5 and dumping functions

3.3 Dumping functions with a partition and cycles

In the previous subsection, players form a complete/incomplete cycle that consists of all players, and player 1 is the target of focus dumping. In what follows, players are partitioned into some coalitions, and each coalition has its own coalitional cycle.

For any nonempty $S \subseteq N$, let $\Pi(S)$ be the set of all partitions of S and $\Pi^*(S) := \Pi(S) \setminus \{\{S\}\}$. For any $i \in S$ and $\mathcal{P} \in \Pi(S)$, let $\mathcal{P}(i)$ denote the coalition to which player i belongs in partition $\mathcal{P}$. Let $\Psi^{\mathcal{P}} = \times_{S \in \mathcal{P}} \Psi^S$ and $\sigma_{\mathcal{P}} \in \Psi^{\mathcal{P}}$. Similarly to the previous section, we use $\lambda^{\sigma_{\mathcal{P}}}(i)$ and $\eta^{\sigma_{\mathcal{P}}}(i)$ to denote i's predecessor and successor in ordering $\sigma_{\mathcal{P}}$.[*1] When an ordering $\sigma_{\mathcal{P}} \in \Psi^{\mathcal{P}}$ is fixed, we omit $\sigma_{\mathcal{P}}$ and write $\lambda(i)$ and $\eta(i)$.

Definition 3.6. Let $\mathcal{P} \in \Pi(N)$ and $\sigma_{\mathcal{P}} \in \Psi^{\mathcal{P}}$. *$\sigma_{\mathcal{P}}$-Cycle dumping* $x^{\sigma_{\mathcal{P}}} : B^N \to X^N$ is a dumping

[*1] Formally, assuming that player i is the kth player in ordering $\sigma_{\mathcal{P}(i)}$ of coalition $\mathcal{P}(i)$, let $\lambda^{\sigma_{\mathcal{P}}}(i) := \sigma_{\mathcal{P}(i)}(k-1)$ and $\eta^{\sigma_{\mathcal{P}}}(i) := \sigma_{\mathcal{P}(i)}(k+1)$. Note that if i belongs to that player's one-person coalition $\mathcal{P}(i) = \{i\}$, then $\lambda^{\sigma_{\mathcal{P}}}(i) = \eta^{\sigma_{\mathcal{P}}}(i) = i$.

function defined as follows: for any $b \in B^N$ and any $i \in N$,

$$x^{\sigma_{\mathcal{P}}}(b)^{i\eta(i)} = b^i.$$

$\sigma_{\mathcal{P}}$-Cycle dumping is an extension of σ-cycle dumping. If $\mathcal{P} = \{N\}$, then each $\sigma_{\mathcal{P}}$-cycle dumping function coincides with a σ-cycle dumping function. Therefore, we focus on $\Pi^*(N)$ below.

Lemma 3.7. Let $\mathcal{P} \in \Pi^*(N)$, $\sigma_{\mathcal{P}} \in \Psi^{\mathcal{P}}$, and $b \in B^N$. If for any $\mathcal{Q} \subsetneq \mathcal{P}$ with $\mathcal{P}(n) \in \mathcal{Q}$ there is $j \in \cup_{T \in \mathcal{Q}} T$ such that $b^j \leq \sum_{i \in N \setminus (\cup_{T \in \mathcal{Q}} T)} b^i$, then strategy profile $x^{\sigma_{\mathcal{P}}}(b)$ is α-stable.

Lemma 3.7 implies that once $b \in B^N$ is given, we can partition the player set N so that the partition yields an α-stable profile. The following lemma provides the construction of such a partition.

Lemma 3.8. Let $\mathcal{P} \in \Pi^*(N)$ with $\mathcal{P}(n) = \mathcal{P}(1)$. For any $b \in B^N$ and any $\sigma_{\mathcal{P}} \in \Psi^{\mathcal{P}}$, strategy profile $x^{\sigma_{\mathcal{P}}}(b)$ is α-stable.

Although σ-cycle dumping always generates an α-stable profile for all $\sigma \in \Psi^N$, $\sigma_{\mathcal{P}}$-cycle dumping does not for some $\sigma_{\mathcal{P}} \in \Psi^{\mathcal{P}}$. Lemma 3.8 shows that as long as the smallest player 1 and the largest player n belong to the same coalition in partition $\mathcal{P}$, every $\sigma_{\mathcal{P}}$-cycle dumping generates an α-stable profile for all $b \in B^N$. For example, let $N = \{1, 2, 3, 4, 5, 6, 7\}$ and $\mathcal{P} = \{\{1, 4, 7\}, \{2, 3, 5\}, \{6\}\}$. Then, $x^{\sigma_{\mathcal{P}}}(b)$ is α-stable for all $b \in B^N$ and all $\sigma_{\mathcal{P}} \in \Psi^{\mathcal{P}}$. Therefore, by grouping players so that the partition can satisfy the condition $\mathcal{P}(n) = \mathcal{P}(1)$, every ordering on the partition generates an α-stable profile.

For any $\mathcal{P} \in \Pi^*(N)$ with $\mathcal{P}(n) = \mathcal{P}(1)$, let $T_{\mathcal{P}}$ be the coalition to which both players 1 and n belong, namely, $T_{\mathcal{P}} := \mathcal{P}(n) = \mathcal{P}(1)$. The following dumping function signifies that players in a coalition dump their bads to a certain player i who may be a member of another coalition.

Definition 3.9. Let $\mathcal{P} \in \Pi^*(N)$ with $\mathcal{P}(n) = \mathcal{P}(1)$ and $\sigma_{\mathcal{P}} \in \Psi^{\mathcal{P}}$. Let $t \in \mathbb{R}_+$ and $i \in N \setminus T_{\mathcal{P}}$. *$\sigma_{\mathcal{P}}$-Cycle dumping with t-focus on i*, $x_{ti}^{\sigma_{\mathcal{P}}} : B^N \to X^N$, is a dumping function defined as follows: for any $b \in B^N$, there is $(t^1, ..., t^n) \in \mathbb{R}_+^N$ such that

- for every $j \in N$, $0 \leq t^j \leq b^j$ and $t^{\lambda(i)} = 0$,
- $\sum_{j \in N} t^j = t$,
- for every $j \in N$, $x_{ti}^{\sigma_{\mathcal{P}}}(b)^{j\eta(j)} = b^j - t^j$,
- for every $j \in N$, $x_{ti}^{\sigma_{\mathcal{P}}}(b)^{ji} = t^j$.

Each player, *e.g.*, player j, divides her initial bads b^j into two parts: $b^j - t^j$ and t^j. The former is dumped to her successor $\eta(j)$, and the latter to the fixed player i. Therefore, the dumping function $x_{ti}^{\sigma_{\mathcal{P}}}$ can be regarded as an intermediate one between $\sigma_{\mathcal{P}}$-cycle dumping and focus dumping on i. The nonnegative real number t represents the total quantity of bads dumped to player i by players other than $\lambda(i)$. Therefore, t satisfies $0 \leq t \leq \sum_{j \in N} b^j - b^{\lambda(i)}$. The following proposition shows that t has a threshold for $x_{ti}^{\sigma_{\mathcal{P}}}$ to be α-stable.

Proposition 3.10. Let $\mathcal{P} \in \Pi^*(N)$ with $\mathcal{P}(n) = \mathcal{P}(1)$, $\sigma_{\mathcal{P}} \in \Psi^{\mathcal{P}}$, $t \in \mathbb{R}_+$, and $i \in N \setminus T_{\mathcal{P}}$. For any $b \in B^N$, the following two statements are equivalent:

i. $x_{ti}^{\sigma_{\mathcal{P}}}(b)$ is α-stable,

ii. $t \leq \sum_{j \in N} b^j - (b^i + b^{\lambda(i)})$.

If $t = 0$, then dumping function $x_{ti}^{\sigma_{\mathcal{P}}}$ coincides with $\sigma_{\mathcal{P}}$-cycle dumping $x^{\sigma_{\mathcal{P}}}$ since every player dumps all his bads to the respective next player. As t increases, the fixed player i receives more bads, and $x_{ti}^{\sigma_{\mathcal{P}}}$ gets closer to focus dumping on i. If t is less than or equal to the threshold $\sum_{j \in N} b^j - (b^i + b^{\lambda}(i))$, then profile $x_{ti}^{\sigma_{\mathcal{P}}}(b)$ is α-stable, while if t exceeds the threshold, then the fixed player i has an incentive to split off from such an unfavorable profile. This proposition not only provides a necessary and sufficient condition for $x_{ti}^{\sigma_{\mathcal{P}}}$ to be α-stable but also shows that if player i belongs to a coalitional cycle that consists of two or more players, it is difficult for that player to split off from $x_{ti}^{\sigma_{\mathcal{P}}}$. The reason is that as long as player i dumps all his bads to another player, t does not exceed the threshold. In other words, t exceeds the threshold only if $t^i > 0$. Therefore, Proposition 3.10 states that if player i follows a rational strategy, namely, $t^i = 0$, then player i cannot split off from $x_{ti}^{\sigma_{\mathcal{P}}}$ in the sense of α-stability.

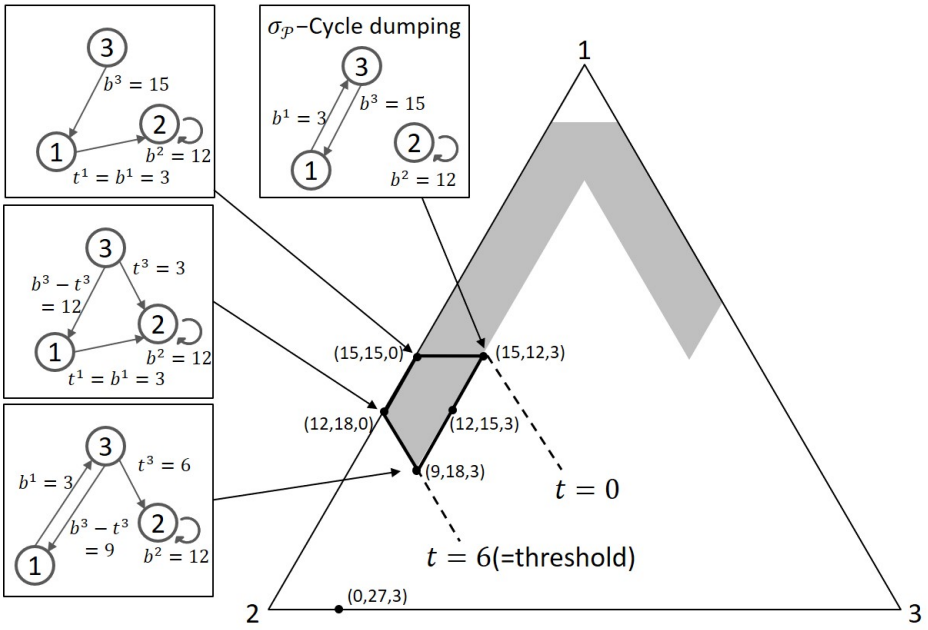

Figure 3 The threshold and the transition of $x_{ti}^{\sigma_{\mathcal{P}}}$

Figure 3 illustrates the distributions that result from α-stable profiles $x_{ti}^{\sigma_{\mathcal{P}}}$ for the initial distribution $b = (3, 12, 15)$. We focus on the threshold and the transition of $x_{ti}^{\sigma_{\mathcal{P}}}$ with respect to t. Consider $\mathcal{P} = \{\{1, 3\}, \{2\}\}$ and $\sigma_{\mathcal{P}} = ((13), (2))$. Since $i \in N \setminus T_{\mathcal{P}}$, let $i = 2$. The threshold is $\sum_{j \in N} b^j - (b^i + b^{\lambda(i)}) = 3 + 12 + 15 - (12 + 12) = 6$, where $\lambda(2) = 2$, as player 2 forms her one-person coalition. Note that $0 \leq t^1 \leq 3$, $0 \leq t^3 \leq 15$ and $t = t^1 + t^3$. If $t = 0$, profile $x_{ti}^{\sigma_{\mathcal{P}}}(b)$ coincides with $x^{\sigma_{\mathcal{P}}}(b)$ and results in $(15, 12, 3)$. As t^1 increases, profile $x_{ti}^{\sigma_{\mathcal{P}}}(b)$ gets closer to $(15, 15, 0)$. Since $0 \leq t^1 \leq 3$, profile $x_{ti}^{\sigma_{\mathcal{P}}}(b)$ stops before it reaches the threshold of 6. Now, we fix $t^1 = 0$ and increase t^3. Starting from $(15, 12, 3)$, profile $x_{ti}^{\sigma_{\mathcal{P}}}(b)$ gets closer to $(0, 27, 3)$ as t^3 increases. Although strategy profile $x_{ti}^{\sigma_{\mathcal{P}}}(b)$ is α-stable before reaching the threshold of 6, it is not so after exceeding the threshold. Therefore, the threshold depicts a

boundary of the set of α-stable strategy profiles.

4 Strong Nash equilibrium

In this section, we analyze strong Nash equilibria in an exchange game of bads. The main purpose is to show that the concepts of strong Nash equilibrium and cycle dumping are equivalent in an exchange game of bads.

Let $b \in B^N$, $x \in X_b^N$, and $S \subseteq N$. Coalition S has *no isolated player* for $x \in X_b^N$ if for every $i \in S$, there is $j \in S$ such that $x^{ji} > 0$. Moreover, coalition S is said to be *cyclic* for x if there is an ordering $\sigma_S \in \Psi^S$ such that $x^{i\eta(i)} > 0$ for every $i \in N$.

Lemma 4.1. Let $b \in B^N$, $x \in X_b^N$, and $S \subseteq N$. Coalition S deviates from x if and only if S has no isolated player for x.

Lemma 4.1 shows that a coalition S has an incentive to deviate if and only if every member of S receives a positive quantity of bads dumped by a member of S. A player i has an incentive to form a deviating coalition only with players who dump a positive quantity of bads to the player i because he can persuade them to avoid dumping bads to him, which decreases his bads; hence, player i has no incentive to invite into a deviating coalition any player who dumps no bads to him. Therefore, if coalition S has an isolated player, the isolated player has no incentive to deviate together with the other members of S.

The following proposition updates the first statement of Proposition 2.5 offered by Hirai et al. (2006).

Proposition 4.2. Let $b \in B^N$ and $x \in X_b^N$. The following statements are equivalent:

i. x is a strong Nash equilibrium,
ii. every proper coalition of N has no isolated player for x,
iii. every proper coalition of N is not cyclic for x,
iv. x is a σ-cycle dumping profile for some $\sigma \in \Psi^N$.

Proposition 4.2 shows that cycle dumping is the only class of dumping functions that generate a strong Nash equilibrium for every exchange game of bads. In other words, Hirai et al (2006) show that for any $b \in B^N$,

$$\{x \in X_b^N | x \text{ is a strong Nash equilibrium}\} \supseteq \{x^\sigma(b) | \text{for all } \sigma \in \Psi^N\},$$

while we show that for any $b \in B^N$,

$$\{x \in X_b^N | x \text{ is a strong Nash equilibrium}\} = \{x^\sigma(b) | \text{for all } \sigma \in \Psi^N\}.$$

Given that the set of coalition-proof Nash equilibria also coincides with that of strong Nash equilibria, these three concepts are equivalent for all exchange games of bads.

5 Self-disposal and the second stage

In this section, we show that adding another stage facilitates self-disposal: players exchange their bads again after the first exchange. We fix an arbitrary initial distribution of bads, $b \in B^N$. The first

stage is the same as the exchange we discussed in the previous sections, $G_b = (N, \{X_b^i\}_{i\in N}, \{v^i\}_{i\in N})$. After players exchange their bads according to an action profile $x \in X_b^N$, the n-dimensional vector r_x represents the resulting distribution of bads. We now consider r_x to be the initial distribution of bads in the second stage, and the players play an exchange game again in the same manner as in the first stage. Since in the beginning of the second stage player i possesses r_x^i bads, player i's set of actions in the second stage is given by $X_{r_x}^i := \{z^i \in X^i | \sum_{j\in N} z^{ij} = r_x^i\}$. For simplicity, we omit r and write

$$X_x^i := X_{r_x}^i = \{z^i \in X^i | \sum_{j\in N} z^{ij} = r_x^i\}.$$

The utility of each player i is defined for the final quantity of bads player i possesses after the second stage.

What stability notion should we now use to analyze an exchange game of bads with two stages? The concept of subgame perfect Nash equilibrium (SPNE) might not be an effective approach because of the following two reasons. One is that there are infinitely many SPNEs. In each stage, every action profile in which each player keeps no bads becomes a Nash equilibrium. Since the set of such actions is infinite, there are infinitely many Nash equilibria in each stage, which results in infinitely many SPNEs. The other is that the concept of SPNE does not take into account coalitional actions, which is incompatible with the results we provided in the preceding sections. Therefore, in this section we introduce a new stability concept by extending the notion of strong Nash equilibrium.

The difficulty in defining a coalitional stability notion in a game with multiple stages lies in the fact that a coalition formed in the first stage does not necessarily last in the second stage. We incorporate this point into our stability notion by considering each player's maximin payoff in the second stage. The following is the maximin payoff of player i in the second stage if action profile x is played in the first stage:

$$m^i(x) := \max_{z^i \in X_x^i} \min_{z^{-i} \in X_x^{-i}} v^i(z^i, z^{-i}),$$

where $z^{-i} := z^{N\setminus\{i\}}$ and $X_x^{-i} := X_x^{N\setminus\{i\}}$. Player i can achieve $m^i(x)$ by himself if x is played in the first stage. Now, we define our stability concept similarly to the definition of strong Nash equilibrium.

Definition 5.1. A coalition $S \subseteq N$ *m-deviates* from $x \in X^N$ if there is $y^S \in X^S$ such that

$$m^i(y^S, x^{N\setminus S}) > m^i(x) \text{ for all } i \in S.$$

An action profile $x \in X^N$ is *m-stable* if no coalition m-deviates from x.

Function m is the only difference between Definitions 2.1 (strong Nash equilibrium) and 5.1. When coalition S m-deviates from x, the members of S have a joint action y^S by which all members improve their maximin payoffs in the second stage.

The concept of m-stability has the following two features. One is that we do not have to assume that a coalition formed in the first stage lasts in the second stage. The members of a coalition S agree that playing y^S in the first stage gives them higher maximin payoffs than does playing x^S and do not necessarily agree that they should cooperate with each other again in the second stage. Therefore, by choosing y^S the members of an m-deviating coalition can improve their final payoffs even if their coalition

splits up after the first stage. The second feature is that an m-stable strategy profile is a stationary profile. If a strategy profile is m-stable, then no coalition has an incentive to change its action in the first stage as long as there is the second stage. Since this holds for every pair of stages t and $t+1$, an m-stable strategy profile is stationary.

In an exchange game of bads, the definition of m-stability has a more straightforward form. For any $b \in B^N$, any $x \in X_b^N$, and any $i \in N$, we have

$$m^i(x) = \max_{z^i \in X_x^i} \min_{z^{-i} \in X_x^{-i}} v^i(z^i, z^{-i}) = u^i\left(\sum_{j \in N \setminus \{i\}} r_x^j\right).$$

Hence, in an exchange game of bads, a coalition $S \subseteq N$ m-deviates from $x \in X_b^N$ if there is $y^S \in X_b^S$ such that for every $i \in S$,

$$\sum_{j \in N \setminus \{i\}} r^j_{(y^S, x^{N \setminus S})} < \sum_{j \in N \setminus \{i\}} r_x^j.$$

We obtain the following result:

Proposition 5.2. For any $b \in B^N$, the self-disposal profile $x^*(b)$ is the only m-stable profile in X_b^N.

The importance of the second stage is that if player i dumps his/her bads to another player j in the first stage, then player j can dump the bads back to player i in the second stage. Proposition 5.2 shows that the possibility of counterattack in the second stage may reduce outside dumping because keeping bads weakens future counterattacks.

6 Conclusion

In this paper, a stable profile or an equilibrium is not necessarily a desirable profile since if a dumping profile in which a single individual suffers a large quantity of bads is stable, then the individual cannot split off from the dumping profile. The main purpose of this paper is to show that such an undesirable dumping profile becomes stable. This objective is achieved through Proposition 3.5. This proposition shows that focus dumping where all players dump all of their bads to player 1 yields an α-stable strategy profile for every exchange game of bads. Moreover, the proposition also shows that every dumping function defined as a combination of cycle dumping, focus dumping, and incomplete cycle dumping generates an α-stable strategy profile.

Player 1 is not the only target of focus dumping: an arbitrary player i may also become a target. We define a dumping function in which all players are partitioned into some coalitions and dump all of their bads to player i across coalitions. We show that such a dumping function generates an α-stable profile as long as the quantity of bads dumped to player i is less than or equal to a certain threshold. In addition, we show that cycle dumping is the only dumping function that generates a strong Nash equilibrium. We also show that allowing another stage after the first exchange makes self-disposal stationary.

In this paper, the number of types of bads is assumed to be one to clarify "who dumps bads to whom." One might be more interested in the model with two or more types of bads. This generalization does not change the coincidence among the stability notions mentioned in Section 1. Therefore, we can focus

on the α-core and strong Nash equilibrium. However, the extended version of cycle dumping, in which each player dumps his/her all types of bads to the next player, does not necessarily yield a strong Nash equilibrium. Moreover, a strong assumption on utility functions is needed to hold the equivalence of Proposition 3.1 between α-stability and a distribution of bads. We conjecture that *additive separability* discussed by Konishi et al. (2001) in the Shapley-Scarf economy with multiple types of goods may weaken complementarity between different types of bads. However, further studies are needed to clarify the relationship between complementarity and stable dumping profiles.

In addition, we analyze the transition of a distribution of bads by varying parameter ϵ in Proposition 3.5 and t in Proposition 3.10; however, repeating an exchange of bads similarly to Section 5 should be another approach to analyzing a transition. As mentioned in Section 2, an exchange x is an $n \times n$ matrix that assigns a distribution of bads to another distribution. Therefore, by using x as a transition matrix, we may define a Markov chain over the distributions of bads. Future studies could examine transition x that converges to the self-disposal profile.

Appendix

Proof of Proposition 3.1

Proof. Let $x \in X_b^N$ be α-stable. By the definition of α-stability, for any $S \subseteq N$ and any $y^S \in X_b^S$, there is $z^{N\setminus S} \in X_b^{N\setminus S}$ such that

$$v^i(y^S, z^{N\setminus S}) \leq v^i(x) \text{ for some } i \in S.$$

Since v^i is represented by u^i, the inequality is equivalent to $u^i(r^i_{(y^S,z^{N\setminus S})}) \leq u^i(r^i_x)$. Moreover, since u^i is a strictly decreasing function, this is equivalent to $r^i_{(y^S,z^{N\setminus S})} \geq r^i_x$. Hence, x is α-stable if and only if for any $S \subseteq N$ and any $y^S \in X_b^S$, there is $z^{N\setminus S} \in X_b^{N\setminus S}$ such that

$$r^i_{(y^S,z^{N\setminus S})} \geq r^i_x \text{ for some } i \in S.$$

Below, fixing $S \subseteq N$, we show that the following two statements are equivalent:

(i) For any $y^S \in X_b^S$, there is $z^{N\setminus S} \in X_b^{N\setminus S}$ such that $r^i_{(y^S,z^{N\setminus S})} \geq r^i_x$ for some $i \in S$.
(ii) There is $i \in S$ such that $\sum_{j\in N\setminus S} b^j \geq r^i_x$.

We first show (ii) $\Rightarrow$ (i). Let i be the player satisfying the condition of (ii). Set $\bar{z}^{ji} = b^j$ for all $j \in N \setminus S$. It readily follows that $\sum_{j\in N\setminus S} \bar{z}^{ji} = \sum_{j\in N\setminus S} b^j$. Hence, for any $y^S \in X_b^S$,

$$r^i_{(y^S,z^{N\setminus S})} = \sum_{j\in S} y^{ji} + \sum_{j\in N\setminus S} \bar{z}^{ji} \geq \sum_{j\in N\setminus S} b^j \overset{\text{(ii)}}{\geq} r^i_x$$

We now show (i) $\Rightarrow$ (ii). Assume that for any $i \in S$,

$$\sum_{j\in N\setminus S} b^j < r^i_x. \tag{A.1}$$

Let $\bar{y}^S$ be a profile satisfying $\bar{y}^{ji} = 0$ for all $j, i \in S$. It holds that for any $z^{N\setminus S} \in X_b^{N\setminus S}$ and any $i \in S$,

$$r^i_{(\bar{y}^S, z^{N\setminus S})} = \sum_{j\in S} \bar{y}^{ji} + \sum_{j\in N\setminus S} z^{ji} \leq \sum_{j\in N\setminus S} b^j \overset{\text{(A.1)}}{<} r^i_x.$$

This contradicts (i). □

Proof of Lemma 3.4

Proof. We prove the statement for $n \geq 4$. Let $b \in B^N$ and $\sigma \in \Psi^N$.

We begin with $x^\sigma(b)$. We assume that there is a collection of non-negative real numbers that sum to 1, $\hat{c}^i$ with $i \in N \setminus \{1\}$ and $c^{\sigma i}$ with $i \in N \setminus \{\lambda(1)\}$, such that

$$x^\sigma(b) = \sum_{i\in N\setminus\{1\}} \hat{c}^i \hat{x}^i(b) + \sum_{i\in N\setminus\{\lambda(1)\}} c^{\sigma i} x^{\sigma i}(b).$$

If there is $i \in N \setminus \{1\}$ such that $\hat{c}^i > 0$, then for any player $j \in N \setminus \{1\}$, we have $x^\sigma(b)^{j1} \geq \hat{c}^i \hat{x}^i(b)^{j1} = \hat{c}^i b^j > 0$. However, in the σ-cycle profile $x^\sigma(b)$, there is a player j^* in $N \setminus \{1\}$ who dumps no bads to player 1, namely, $x^\sigma(b)^{j^*1} = 0$. This is a contradiction. If there is $i \in N \setminus \{\lambda(1)\}$ such that $c^{\sigma i} > 0$, then $x^\sigma(b)^{i1} \geq c^{\sigma i} x^{\sigma i}(b)^{i1} = c^{\sigma i} b^i > 0$. However, since $\lambda(1)$ is the only player who dumps bads to player 1 in $x^\sigma(b)$, we have $x^\sigma(b)^{i'1} = 0$ for every $i' \in N \setminus \{\lambda(1)\}$. This is a contradiction. Hence all the coefficients above are zero, while $x^\sigma(b)$ contains positive entries, which implies that $x^\sigma(b)$ can not be written as a non-negative convex combination of the other profiles.

Now, let $i^* \in N \setminus \{1\}$. We suppose that

$$\hat{x}^{i^*}(b) = c^\sigma x^\sigma(b) + \sum_{i\in N\setminus\{1,i^*\}} \hat{c}^i \hat{x}^i(b) + \sum_{i\in N\setminus\{\lambda(1)\}} c^{\sigma i} x^{\sigma i}(b)$$

holds for some collection of non-negative coefficients that sum to 1. Assume $c^\sigma > 0$. For any $i \in N \setminus \{1, \lambda(1)\}$, $\hat{x}^{i^*}(b)^{i\eta(i)} = 0$. However, $x^\sigma(b)^{h\eta(h)} = b^h > 0$ for some $h \in N \setminus \{1, \lambda(1)\}$. This is a contradiction. If there is $i \in N \setminus \{1, i^*\}$ such that $\hat{c}^i > 0$, then $\hat{x}^{i^*}(b)^{1i} = 0 < b^1 = \hat{x}^i(b)^{1i}$, which is a contradiction. If there is $i \in N \setminus \{\lambda(1)\}$ such that $c^{\sigma i} > 0$, then since $n \geq 4$ we obtain a contradiction in the same manner as the case $c^\sigma > 0$ by replacing $x^\sigma(b)$ with $x^{\sigma i}(b)$. Thus, $\hat{x}^{i^*}(b)$ can not be written as a non-negative convex combination of the other profiles.

Let $i^* \in N \setminus \{\lambda(1)\}$. We assume that

$$x^{\sigma i^*}(b) = c^\sigma x^\sigma(b) + \sum_{i\in N\setminus\{1\}} \hat{c}^i \hat{x}^i(b) + \sum_{i\in N\setminus\{\lambda(1),i^*\}} c^{\sigma i} x^{\sigma i}(b)$$

holds for some collection of non-negative coefficients that sum to 1. If $c^\sigma > 0$ then we have $x^{\sigma i^*}(b)^{i^*\eta(i^*)} = 0 < b^{i^*} = x^\sigma(b)^{i^*\eta(i^*)}$, which is a contradiction. If there is $i \in N \setminus \{1\}$ such that $\hat{c}^i > 0$, then at least one player $j^* \in N \setminus \{1, i^*, \lambda(1)\}$ satisfies $x^{\sigma i^*}(b)^{j^*1} = 0$, while $\hat{x}^i(b)^{j1} = 1$ for every $j \in N \setminus \{1\}$. This is a contradiction. If there is $i \in N \setminus \{\lambda(1), i^*\}$ such that $c^{\sigma i} > 0$, then $x^{\sigma i^*}(b)^{i^*\eta(i^*)} = 0 < b^{i^*} = x^{\sigma i}(b)^{i^*\eta(i^*)}$, which is a contradiction. Thus, $x^{\sigma i^*}(b)$ can not be written as a non-negative convex combination of the other profiles. □

Proof of Proposition 3.5

Proof. In view of Lemma 3.4, the statement for $n = 3$ is straightforward to show since the number of the profiles is four. Below, we prove the statement for $n \geq 4$. Let $b \in B^N$ and $\sigma \in \Psi^N$. Let $y(b)$ be a non-negative convex combination of the profiles in $E(n, b, \sigma)$:

$$y(b) = c^{\sigma} x^{\sigma}(b) + \sum_{i \in N \setminus \{1\}} \hat{c}^i \hat{x}^i(b) + \sum_{i \in N \setminus \{\lambda(1)\}} c^{\sigma i} x^{\sigma i}(b).$$

In view of Proposition 3.1, we show that for every coalition $S \subseteq N$, there is $i \in N$ such that $\sum_{j \in N \setminus S} b^j \geq r_y^i$.

We have

$$r_y = c^{\sigma} r_{x^{\sigma}} + \sum_{i \in N \setminus \{1\}} \hat{c}^i r_{\hat{x}^i} + \sum_{i \in N \setminus \{\lambda(1)\}} c^{\sigma i} r_{x^{\sigma i}}.$$

Let $h = N \setminus \{1\}$. We have

$$\begin{aligned}
r_y^h &= c^{\sigma} r_{x^{\sigma}}^h + \sum_{i \in N \setminus \{1\}} \hat{c}^i r_{\hat{x}^i}^h + \sum_{i \in N \setminus \{\lambda(1)\}} c^{\sigma i} r_{x^{\sigma i}}^h \\
&= c^{\sigma} r_{x^{\sigma}}^h + \sum_{i \in N \setminus \{1\}} \hat{c}^i r_{\hat{x}^i}^h + \sum_{i \in N \setminus \{\lambda(1), \lambda(h)\}} c^{\sigma i} r_{x^{\sigma i}}^h + c^{\sigma \lambda(h)} r_{x^{\sigma \lambda(h)}}^h \\
&= c^{\sigma} b^{\lambda(h)} + \sum_{i \in N \setminus \{1\}} \hat{c}^i r_{\hat{x}^i}^h + \sum_{i \in N \setminus \{\lambda(1), \lambda(h)\}} c^{\sigma i} b^{\lambda(h)} + c^{\sigma \lambda(h)} 0 \\
&= \left(c^{\sigma} + \sum_{i \in N \setminus \{\lambda(1), \lambda(h)\}} c^{\sigma i} \right) b^{\lambda(h)} + \sum_{i \in N \setminus \{1\}} \hat{c}^i r_{\hat{x}^i}^h. \qquad \text{(A.2)}
\end{aligned}$$

The third equality holds because (i) in the profiles x^{σ} and $x^{\sigma i}$, the bads player h receives are equal to the initial endowment of player $\lambda(h)$, and (ii) in the profile $x^{\sigma \lambda(h)}$, player h receives no bads as player $\lambda(h)$ dumps his bads all to player 1. Moreover, we have

$$\begin{aligned}
\text{(A.2)} &= \left(c^{\sigma} + \sum_{i \in N \setminus \{\lambda(1), \lambda(h)\}} c^{\sigma i} \right) b^{\lambda(h)} + \hat{c}^h b^1 \\
&\leq \left(c^{\sigma} + \sum_{i \in N \setminus \{\lambda(1), \lambda(h)\}} c^{\sigma i} \right) b^{\lambda(h)} + \hat{c}^h b^{\lambda(h)} \\
&\leq b^{\lambda(h)}. \qquad \text{(A.3)}
\end{aligned}$$

The first equality holds because $r_{\hat{x}^i}^h = b^1$ holds only for $i = h$. The second inequality holds as $b^1 \leq ... \leq b^n$. Since the coefficients sum to 1, we obtain the third inequality. Thus, (A.3) shows that $r_y^h \leq b^{\lambda(h)}$ for every $h \in N \setminus \{1\}$, which implies that if a coalition S contains a player $h \in N \setminus \{1\}$ but not $\lambda(h)$, then, in view of Proposition 3.1, the coalition is not α-effective. Therefore, below, we focus on a coalition defined as $\{\sigma(1), ..., \sigma(s)\}$ for some $1 \leq s \leq n-1$.

For any $s \geq 2$, coalition $S := \{\sigma(1), ..., \sigma(s)\}$ contains player $\sigma(2)$. From (A.3), it follows that $r_y^{\sigma(2)} \leq b^{\sigma(1)} = b^1 = \min_{\emptyset \neq T \subseteq N} \sum_{j \in T} b^j \leq \sum_{j \in N \setminus S} b^j$. Hence, for any $s \geq 2$, coalition $S = \{\sigma(1), ..., \sigma(s)\}$ is

not α-effective. We now consider the coalition $S = \{1\}$. We have

$$\begin{aligned} r_y^1 &= c^\sigma r_{x^\sigma}^1 + \sum_{i\in N\setminus\{1\}} \hat{c}^i r_{\hat{x}^i}^1 + \sum_{i\in N\setminus\{\lambda(1)\}} c^{\sigma i} r_{x^{\sigma i}}^1 \\ &= c^\sigma b^{\lambda(1)} + \sum_{i\in N\setminus\{1\}} \hat{c}^i \left(\sum_{j\in N\setminus\{1\}} b^j \right) + \sum_{i\in N\setminus\{\lambda(1)\}} c^{\sigma i} \left(b^{\lambda(1)} + b^i \right) \\ &= c^\sigma b^{\lambda(1)} + \sum_{j\in N\setminus\{1\}} b^j \left(\sum_{i\in N\setminus\{1\}} \hat{c}^i \right) + \sum_{i\in N\setminus\{\lambda(1)\}} c^{\sigma i} b^{\lambda(1)} + \sum_{i\in N\setminus\{\lambda(1)\}} c^{\sigma i} b^i. \end{aligned} \tag{A.4}$$

For simplicity, let $\Sigma\hat{c} := \sum_{i\in N\setminus\{1\}} \hat{c}^i$. We have

$$\begin{aligned} \text{(A.4)} &= c^\sigma b^{\lambda(1)} + \sum_{j\in N\setminus\{1\}} b^j \Sigma\hat{c} + \sum_{i\in N\setminus\{\lambda(1)\}} c^{\sigma i} b^{\lambda(1)} + \sum_{i\in N\setminus\{\lambda(1)\}} c^{\sigma i} b^i \\ &= c^\sigma b^{\lambda(1)} + \left[b^{\lambda(1)} \Sigma\hat{c} + \sum_{j\in N\setminus\{1,\lambda(1)\}} b^j \Sigma\hat{c} \right] + \sum_{i\in N\setminus\{\lambda(1)\}} c^{\sigma i} b^{\lambda(1)} + \left[c^{\sigma 1} b^1 + \sum_{i\in N\setminus\{1,\lambda(1)\}} c^{\sigma i} b^i \right] \\ &= \left[\sum_{i\in N\setminus\{1,\lambda(1)\}} (c^{\sigma i} + \Sigma\hat{c}) b^i \right] + \left[c^\sigma + \Sigma\hat{c} + \sum_{i\in N\setminus\{\lambda(1)\}} c^{\sigma i} \right] b^{\lambda(1)} + c^{\sigma 1} b^1 \\ &= \left[\sum_{i\in N\setminus\{1,\lambda(1)\}} (c^{\sigma i} + \Sigma\hat{c}) b^i \right] + b^{\lambda(1)} + c^{\sigma 1} b^1 \\ &\leq \left[\sum_{i\in N\setminus\{1,\lambda(1)\}} (c^{\sigma i} + \Sigma\hat{c}) b^i \right] + b^{\lambda(1)} + \sum_{i\in N\setminus\{1,\lambda(1)\}} c^{\sigma 1} b^i \\ &= \left[\sum_{i\in N\setminus\{1,\lambda(1)\}} (c^{\sigma 1} + c^{\sigma i} + \Sigma\hat{c}) b^i \right] + b^{\lambda(1)} \\ &\leq \sum_{i\in N\setminus\{1,\lambda(1)\}} b^i + b^{\lambda(1)} \\ &= \sum_{i\in N\setminus\{1\}} b^i. \end{aligned}$$

Hence, coalition $\{1\}$ is not α-effective. Thus, no coalition is α-effective, and $y(b)$ is α-stable. □

Proof of Lemma 3.7

Proof. Since N and $\emptyset$ are not α-effective, we consider coalition S that satisfies $\emptyset \neq S \subsetneq N$. If there is a coalition $T \in \mathcal{P}$ such that $T \cap S \neq \emptyset$ and $T \setminus S \neq \emptyset$, then then for some $h \in S$, $\lambda(h) \in T \setminus S$ and $h \in T \cap S$. From the definition of $x^{\sigma_\mathcal{P}}$, we have

$$r_{x^{\sigma_\mathcal{P}}}^h = b^{\lambda(h)} \overset{\lambda(h)\notin S}{\leq} \sum_{j\in N\setminus S} b^j,$$

and S is not α-effective. Therefore, for coalition S to be α-effective, S needs to satisfy $S = \cup_{T\in\mathcal{Q}} T$ for some $\mathcal{Q} \subsetneq \mathcal{P}$. If $\mathcal{P}(n) \notin \mathcal{Q}$, then $n \in N \setminus S$. Hence, for some $h \in S$, $r_{x^{\sigma_\mathcal{P}}}^h = b^{\lambda(h)} \leq b^n \leq \sum_{j\in N\setminus S} b^j$,

which means that S is not α-effective. If $\mathcal{P}(n) \in \mathcal{Q}$, then in view of the assumption of the claim, there is $j \in \cup_{T\in\mathcal{Q}} T$ such that $b^j \leq \sum_{i\in N\setminus(\cup_{T\in\mathcal{Q}}T)} b^i$. We fix this player j and consider $\eta(j)$. It holds that $r^{\eta(j)}_{x^{\sigma_\mathcal{P}}} = b^j \leq \sum_{i\in N\setminus S} b^i$, which means that S is not α-effective. Therefore, no coalition is α-effective for $x^{\sigma_\mathcal{P}}(b)$. □

Proof of Lemma 3.8

Proof. If there is a coalition $T \in \mathcal{P}$ such that $T \cap S \neq \emptyset$ and $T \setminus S \neq \emptyset$, then S is not α-effective in the same manner as Lemma 3.7. Hence, to be α-effective, S must satisfy $S = \cup_{T\in\mathcal{Q}} T$ for some $\mathcal{Q} \subsetneq \mathcal{P}$. If $\mathcal{P}(n) \notin \mathcal{Q}$, then player n is in $N \setminus S$, and for some $h \in S$, $b^h \leq b^n \leq \sum_{j\in N\setminus S} b^j$. If $\mathcal{P}(n) \in \mathcal{Q}$, then player 1 is in S, and for some $h \in N \setminus S$, $b^1 \leq b^h \leq \sum_{j\in N\setminus S} b^j$. □

Proof of Proposition 3.10

Proof. In this proof, we fix $\mathcal{P}$, $\sigma_\mathcal{P}$, t, i, and b as mentioned in the proposition and write $r_{**} := r_{x^{\sigma_\mathcal{P}}_{ti}}$ and $r_* := r_{x^{\sigma_\mathcal{P}}}$ for simplicity. We begin with [(ii)⇒(i)]. Below, we consider that a coalition S satisfies $\emptyset \neq S \subsetneq N$, because N and $\emptyset$ are not α-effective. Claim 1 is used for Claims 2-5. Claims 2-5 are used for the α-effectiveness of coalition S.

Claim 1. For any $j \in N \setminus \{i\}$, $r^j_{**} \leq r^j_*$. For the player i, $r^i_{**} \geq r^i_*$.

Proof of Claim 1. For every player $j \in N \setminus \{i\}$, $r^j_{**} = b^{\lambda(j)} - x^{\sigma_\mathcal{P}}_{ti}(b)^{\lambda(j)i} \leq b^{\lambda(j)} = r^j_*$. For the fixed player i, $r^i_{**} = b^{\lambda(i)} + t \geq b^{\lambda(i)} = r^i_*$. //

Claim 2. For any S with $\emptyset \neq S \subsetneq N$, if $i \in N \setminus S$, then S is not α-effective.

Proof of Claim 2. In view of lemma 3.8, $x^{\sigma_\mathcal{P}}$ is α-stable. Hence, for the coalition S, there is $h \in S$ such that $r^h_* \leq \sum_{j\in N\setminus S} b^j$. Since $h \neq i$, it follows from Claim 1 that

$$r^h_{**} \overset{\text{Claim 1}}{\leq} r^h_* \leq \sum_{j\in N\setminus S} b^j,$$

which means that S is not α-effective. //

Claim 3. For any S with $\emptyset \neq S \subsetneq N$, if $T_\mathcal{P} \subseteq S$, then S is not α-effective.

Proof of Claim 3. Since $\{1, n\} \subseteq T_\mathcal{P}$, $\sigma_{T_\mathcal{P}}(1) = 1$. We write $h := \sigma_{T_\mathcal{P}}(2)$. Since $i \in N \setminus T_\mathcal{P}$, $h \neq i$. We have

$$r^h_{**} \overset{\text{Claim 1}}{\leq} r^h_* = b^{\lambda(h)} = b^1 \leq \min_{j\in N} b^j.$$

Since S is a proper subset of N, for some $h' \in N \setminus S$, $\min_{j\in N} b^j \leq b^{h'} \leq \sum_{j\in N\setminus S} b^j$. Hence, S is not α-effective. //

Claim 4. For any S with $\emptyset \neq S \subsetneq N$, if S satisfies the following condition, then S is not α-effective:

$$\text{there is } T \in \mathcal{P} \setminus \{\mathcal{P}(i)\} \text{ such that } T \cap S \neq \emptyset \text{ and } T \setminus S \neq \emptyset.$$

Proof of Claim 4. Since $T \cap S \neq \emptyset$ and $T \setminus S \neq \emptyset$, there is a player $h \in T \cap S$ such that $\lambda(h) \in T \setminus S$. Considering that $h \in T \neq \mathcal{P}(i)$ implies $h \neq i$, we have

$$r^h_{**} \overset{\text{Claim 1}}{\leq} r^h_* = b^{\lambda(h)} \overset{\lambda(h) \notin S}{\leq} \sum_{j \in N \setminus S} b^j.$$

Hence, S is not α-effective. //

Claim 5. For any S with $\emptyset \neq S \subsetneq N$ and $S \cap T_{\mathcal{P}} = \emptyset$, if S satisfies the following condition, then S is not α-effective:

$$\text{there is } T \in \mathcal{P} \setminus \{\mathcal{P}(i)\} \text{ such that } T \cap S \neq \emptyset.$$

Proof of Claim 5. Since $T \cap S \neq \emptyset$, we consider a player $h \in T \cap S$. Moreover, since $h \in T \neq \mathcal{P}(i)$, we obtain $h \neq i$. We have

$$r^h_{**} \overset{\text{Claim 1}}{\leq} r^h_* = b^{\lambda(h)} \leq b^n$$

Since $n \in T_{\mathcal{P}}$ and $S \cap T_{\mathcal{P}} = \emptyset$, we have $n \in N \setminus S$, which implies

$$b^n \leq \sum_{j \in N \setminus S} b^j.$$

Hence, S is not α-effective. //

In view of Claim 4, for a coalition S to be α-effective, it must obey the following condition: for some partition $\mathcal{Q} \subseteq \mathcal{P} \setminus \{\mathcal{P}(i)\}$, $S = (\cup_{T \in \mathcal{Q}} T) \cup (S \cap \mathcal{P}(i))$, where $\mathcal{Q}$ may be $\emptyset$. Moreover, by Claim 3 and the fact that $T_{\mathcal{P}}$ is an element of $\mathcal{P}$, the previous condition must become the following form: for some partition $\mathcal{Q} \subseteq \mathcal{P} \setminus \{\mathcal{P}(i), T_{\mathcal{P}}\}$, $S = (\cup_{T \in \mathcal{Q}} T) \cup (S \cap \mathcal{P}(i))$. Hence, $S \cap T_{\mathcal{P}} = \emptyset$. By Claim 5, $(\cup_{T \in \mathcal{Q}} T)$ must be empty. Hence, $S = S \cap \mathcal{P}(i)$, which is equivalent to the condition $S \subseteq \mathcal{P}(i)$. By Claim 2, S satisfies $i \in S \subseteq \mathcal{P}(i)$.

If there is $h \in S$ such that $h \neq i$, then we have $r^h_{**} \overset{\text{Claim 1}}{\leq} r^h_* = b^{\lambda(h)} \leq b^n \overset{n \in N \setminus S}{\leq} \sum_{j \in N \setminus S} b^j$. Hence, S is not α-effective. Therefore, it suffices to show $S = \{i\}$ is not α-effective. By the condition of t, we obtain $r^i_{**} = b^{\lambda(i)} + t \leq b^{\lambda(i)} + \sum_{j \in N \setminus \{i, \lambda(i)\}} b^j = \sum_{j \in N \setminus \{i\}} b^j$, which means that S is not α-effective. Thus, for every $S \subseteq N$, S is not α-effective for $x^{\sigma_{\mathcal{P}}}_{ti}(b)$: $x^{\sigma_{\mathcal{P}}}_{ti}(b)$ is α-stable.

Now, we assume that the opposite inequality holds for the condition of t, $t > \sum_{j \in N \setminus \{i, \lambda(i)\}} b^j$. Then, in the same manner as above, $r^i_{**} > \sum_{j \in N \setminus \{i\}} b^j$. Given Proposition 3.1 and the fact that i is the only member of S, $S = \{i\}$ becomes α-effective. Hence, $x^{\sigma_{\mathcal{P}}}_{ti}(b)$ is not α-stable. This establishes [not (ii) $\Rightarrow$ not (i)]. □

Proof of Lemma 4.1

Proof. We first prove the if-part. Let $y^S \in X^S_b$ satisfy the following: for every $i, j \in S$, $y^{ji} = 0$ and for every $j \in S$ there is $h \in N \setminus S$ such that $y^{jh} = b^j$. Now, fixing $j \in S$, we have $v^j(y^S, x^{N \setminus S}) = u^j(\sum_{h \in N \setminus S} x^{hj}) > u^j(\sum_{i \in S} x^{ij} + \sum_{h \in N \setminus S} x^{hj}) = v^j(x)$, where the inequality holds because for every

$i \in S$ there is $j \in S$ such that $x^{ji} > 0$, and u^j is strictly decreasing. Since this holds for every $h \in S$, S deviates from x.

We now prove the only-if-part. Suppose that there is $i^* \in S$ such that for every $j \in S$, $x^{ji} = 0$. In strategy profile x, such a player i^* has $v^{i^*}(x) = u^{i^*}(\sum_{h \in N \setminus S} x^{hi^*}) = u^{i^*}(\sum_{h \in N \setminus S} x^{hi^*} + \min_{y^S \in X_b^S}[\sum_{i \in S} y^{ii^*}]) \geq v^{i^*}(y^S, x^{N \setminus S})$ for any $y^S \in X_b^S$. Hence, S does not deviate from x. □

Proof of Proposition 4.2

Proof. Claim [(i)⇒(ii)] follows from Lemma 4.1. Now consider [(ii)⇒(iii)]. If a coalition is cyclic then the coalition has no isolated player. Therefore, if a coalition has an isolated player, then it is not cyclic. Now, we show [(iii)⇒(iv)]. Suppose the negation of (iv): for any $\sigma \in \Psi^N$, there is a player $i \in N$ such that $x^{ij} > 0$ for some $j \neq \eta(i)$. Coalition $S := \{1, ..., i, j, \eta(j), ..., \lambda(1)\}$ is proper subset of N since $\eta(i)$ is not in S and is cyclic for x. This contradicts (iii). Claim [(iv)⇒(i)] follows from the first statement of Proposition 2.5. □

Proof of Proposition 5.2

Proof. Let $b \in b^N$. We first show that if profile $x \in X_b^N$ is not the self-disposal profile $x^*(b)$, then x is not m-stable. Assume that there is a player $i \in N$ such that $x^{ih} > 0$ for some $h \in N \setminus \{i\}$. We define $y^i \in X_b^i$ as follows: $y^{ii} = b^i$ and $y^{ij} = 0$ for every $j \in N \setminus \{i\}$. We have

$$\begin{aligned}
\sum_{j \in N \setminus \{i\}} r_x^j - \sum_{j \in N \setminus \{i\}} r_{y^i, x^{-i}}^j &= \sum_{j \in N \setminus \{i\}} \sum_{l \in N} x^{lj} - \sum_{j \in N \setminus \{i\}} \left(y^{ij} + \sum_{l \in N \setminus \{i\}} x^{lj} \right) \\
&= \sum_{j \in N \setminus \{i\}} (x^{ij} - y^{ij}) \\
&= \sum_{j \in N \setminus \{i\}} x^{ij} \\
&> 0.
\end{aligned}$$

Hence, $\{i\}$ m-deviates from x, and profile x is not m-stable.

Now, we show that $x^*(b)$ is m-stable. We write $x^* := x^*(b)$ for simplicity. Assume that a coalition S m-deviates from x^*. By the definition of m-deviation, there is $y^S \in X_b^S$ such that for every $i \in S$,

$$\sum_{j \in N \setminus \{i\}} r_{y^S, x^{*N \setminus S}}^j < \sum_{j \in N \setminus \{i\}} r_{x^*}^j. \tag{A.5}$$

Let $T := \{i \in S | y^{ij} > 0 \text{ for some } j \in N \setminus \{i\}\}$. Note that T is nonempty because if $y^{ij} = 0$ for every $j \in N \setminus \{i\}$ and every $i \in S$, then $y^S = x^{*S}$, which contradicts (A.5). For every $i \in T$, define

$$I^i(y^S) := \sum_{h \in S \setminus \{i\}} y^{hi} \text{ and } O^i(y^S) := \sum_{h \in N \setminus \{i\}} y^{ih}.$$

For every $i \in T$, we have

$$\sum_{j \in N \setminus \{i\}} r^j_{y^S, x^{*N \setminus S}} = \left(\sum_{j \in N \setminus \{i\}} b^j \right) + O^i(y^S) - I^i(y^S). \tag{A.6}$$

Since $\sum_{j \in N \setminus \{i\}} r^j_{x^*} = \sum_{j \in N \setminus \{i\}} b^j$, (A.5) implies

$$\sum_{j \in N \setminus \{i\}} r^j_{y^S, x^{*N \setminus S}} < \sum_{j \in N \setminus \{i\}} b^j. \tag{A.7}$$

It follows from (A.6) and (A.7) that for every $i \in T$,

$$O^i(y^S) < I^i(y^S). \tag{A.8}$$

In view of the definition of T, $y^{ii} = b^i$ for every $i \in S \setminus T$. Hence, we have $I^i(y^S) \stackrel{\text{def}}{=} \sum_{h \in S \setminus \{i\}} y^{hi} = \sum_{h \in T \setminus \{i\}} y^{hi}$ and $O^i(y^S) \stackrel{\text{def}}{=} \sum_{h \in N \setminus \{i\}} y^{ih} = \sum_{h \in T \setminus \{i\}} y^{ih} + \sum_{h \in N \setminus T} y^{ih}$. We obtain

$$\begin{aligned}
\sum_{i \in T} I^i(y^S) &= \sum_{i \in T} \sum_{h \in T \setminus \{i\}} y^{hi} \\
&= \sum_{i \in T} \sum_{h \in T \setminus \{i\}} y^{ih} \\
&\leq \sum_{i \in T} \left(\sum_{h \in T \setminus \{i\}} y^{ih} + \sum_{h \in N \setminus T} y^{ih} \right) \\
&= \sum_{i \in T} O^i(y^S).
\end{aligned}$$

This contradicts (A.8). □

References

[1] Aumann, R.J. (1959). Acceptable points in general cooperative n-person games. *Contributions to the Theory of Games IV*, Eds. Tucker, A. and Luce, R., Princeton University Press, Princeton.

[2] Aumann, R.J., and Peleg, B. (1960). Von Neumann-Morgenstern solutions to cooperative games without side payments. *Bulletin of the American Mathematical Society* 66(3), 173-179.

[3] Bernheim, B.D., Peleg, B., and Whinston, M. D. (1987). Coalition-proof nash equilibria i. concepts. *Journal of Economic Theory* 42(1), 1-12.

[4] Chander, P., and Tulkens, H. (2006). The core of an economy with multilateral environmental externalities. *International Journal of Game Theory* 26(3), 379-401.

[5] Currarini, S., and Marini, M. (2004). A conjectural cooperative equilibrium for strategic form games. *Game Practice and the Environment*, Eds. Carraro, C. and Fragnelli, V., Edward Elgar, Cheltenham.

[6] Hirai, T., Masuzawa, T., and Nakayama, M. (2006). Coalition-proof Nash equilibria and cores in a strategic pure exchange game of bads. *Mathematical Social Sciences* 51(2), 162-170.

[7] Konishi, H., Quint, T., and Wako, J. (2001). On the Shapley-Scarf economy: the case of multiple types of indivisible goods. *Journal of Mathematical Economics* 35(1), 1-15.

[8] Scarf, H.E. (1971). On the existence of a coopertive solution for a general class of N-person games. *Journal of Economic Theory* 3(2), 169-181.

[9] Shapley, L.S., and Shubik, M. (1969). On the core of an economic system with externalities. *American Economic Review* 59(4), 678-684.