WARWICK
ECONOMICS

CRETA
Centre for Research in Economic Theory and its Applications

Discussion Paper Series

# Cooperation in a State of Anarchy

Abhinay Muthoo

(This paper also appears in
Warwick Economics Research Papers series No: 1296)

August 2020 No: 63

CRETA
Centre for Research in Economic Theory and its Applications
Department of Economics
University of Warwick, Coventry,
CV4 7AL, United Kingdom

warwick.ac.uk/fac/soc/economics/research/centres/creta/papers

# Cooperation in a State of Anarchy

Abhinay Muthoo†

July 31, 2020

**Abstract**

We lay down a simple (game-theoretic) model of a state-of-anarchy involving three players. We focus attention on the following question: Which subset of players (if any) will agree to cooperate amongst each other? Will all three players agree to do so, or only two of the three players (and if so, which two players)? Or will no player agree to cooperate with any other player? We show that the socially optimal outcome is for all three players to agree to cooperate with each other. We also show that due to the presence of positive externalities, in equilibrium, cooperation may only be established between two of the three players (which is sub-optimal).

*JEL Classification Numbers*: D62, D74, F02.

*Acknowledgements.* I thank Stephen Ansolabere, Cathy Hafer, Maggie Penn, Ken Shepsle, Jim Synder, Ariel Rubinstein and various seminar participants for their helpful comments.

† Department of Economics, University of Warwick, UK. Email: a.muthoo@warwick.ac.uk

"... there is no such thing as perpetual tranquillity while we live here because life is but motion and can never be without desire or without fear, no more than without sense ... there can be no content but in proceeding." THOMAS HOBBES, *Leviathan, 1651.*

# 1 Introduction

Cooperation amongst countries on a variety of matters - especially during (and post) the pandemic - is critical for peace and prosperity. In a context of a world that can be characterised as being in a (sort of) "state of anarchy", it is hard to secure such cooperation. We make three contributions motivated, in part, by this issue.

First, we lay down a simple (game-theoretic) model of a state of anarchy, with three players. Each player is endowed with some productive and fighting skills. Their interaction commences with a game of "coalition formation", in which they discuss whether or not to cooperate with each other. An outcome of this game generates a *coalitional structure*.[1] Once a coalitional structure is determined, the players then simultaneously and independently choose how much costly resource (time) to allocate to productive work. Then, after observing the output levels produced by each player (and given the coalitional structure that has previously been determined), each player decides whether or not to "fight" any other player who does not belong to the coalition to which he belongs (if any), in order to steal his output.

Much of the large literature in cooperative game theory and, with the exception of a few papers, the small literature on noncooperative games of coalition formation are concerned with the analyses of coalitional games in *characteristic function* form. A basic property of such games is that the payoff to a coalition or to a player in a coalition — depending on whether the game allows or does not allow for transferable utility — does not depend on what coalitions the other players do or do not form. In contrast, the coalitional game that is derived from the primitives of my model does not share this property: the payoff to a player when the other two players form a coalition is strictly greater than his payoff when they do not form such a coalition. Indeed, my coalitional game is a game with "positive spillovers". It is a game in *partition function* form, where the payoff to a player in a coalition depends on the entire coalitional structure. Key contributions to such games include Bloch (1996), Ray and Vohra (1997, 1999) and Maskin (2003).

In the context of our simple model, we establish two main results. First, we show that the socially optimal (or first-best) outcome is one in which the grand coalitional structure

[1]There are five possible coalitional structures. The grand coalition is one possible outcome, in which cooperation is established amongst all three players. At the other extreme is the degenerate coalitional structure, in which cooperation is not established at all. Finally, one of the three possible partial coalitional structures could emerge, in which a pair of the players agree to cooperate between themselves, while the third player is not party to any such agreement.

emerges. Second, applying Ray and Vohra's (1997) "Equilibrium Binding Agreements" solution concept, we show that under some circumstances, the equilibrium coalitional structure will be a partial coalitional structure. In particular, the grand coalition is not a stable coalitional structure. This arises because for such parameter values, each player has an individual incentive to break away from the grand coalition, and free-ride on the other two players as they would continue to have an incentive to maintain the cooperative agreement between the two of them.

# 2 A Simple Model of a State-of-Anarchy

The model has three stages, operating at three dates.

**Date 1 (Coalition Formation):** Three players, 1, 2 and 3, engage in a process of coalition formation. There are five possible outcomes of this process corresponding to the five possible coalitional structures: $G$ denotes the grand coalitional structure (i.e., $G = \{\{1,2,3\}\}$), $D$ the degenerate coalitional structure (i.e., $D = \{\{1\},\{2\},\{3\}\}$), and $ij$ the partial coalitional structure in which $i$ and $j$ form a coalition, with $k$ in a coalition on his own (i.e., $ij = \{\{i,j\},\{k\}\}$ where $i \neq j \neq k$ and $i,j,k = 1,2,3$).

After the coalitional structure is determined at date 1 - which defines which player has agreed to cooperate with whom - we move to:

**Date 2 (Production):** The players simultaneously and independently choose the quantities of time that they respectively engage in productive work. If $i$ works for $L_i$ units of time, where $0 \leq L_i \leq T$ (with $T$ being the total time endowment), then it produces $f_i(L_i)$ units of output. The twice differentiable production function $f_i$ satisfies the following (standard) conditions: $f_i(0) = 0$, $f_i'(0) = +\infty$, $f_i' > 0$ and $f_i'' \leq 0$.

All players now observe the quantities of output produced by each player, and then, we move to:

**Date 3 (Fights?):** Depending on the coalitional structure that was determined at date 1, a "fighting" process may take place in which players fight each other in order to steal each others' outputs.[2] If the grand coalition formed, then no fight occurs and each player consumes what he produced; that is, $c_i^G = f_i(L_i)$ (for all $i = 1,2,3$). Otherwise, three fights take place, one over each player's output. A player will not be involved in the fight over another player's output if and only if they belong to the same coalition.

If the degenerate coalitional structure is in place, then each player is involved in all three fights. The outcome of each fight depends on the players' respective fighting skills. Let $z_i > 0$ denote $i$'s fighting skill. The outcome of the fight over each player's output is such that the share of that output obtained by $i$ is $z_i/(z_1+z_2+z_3)$. Hence, if the degenerate

[2] "Fighting" may also be interpreted as non-violent, predatory behaviour, such as trade wars, which indirectly can lead to extraction of some of the other player's output.

coalitional structure is in place, then $i$'s consumption level is:

$$c_i^D = \frac{z_i}{z_1 + z_2 + z_3}\Big[f_1(L_1) + f_2(L_2) + f_3(L_3)\Big].$$

If the $ij$ coalitional structure is in place, then while $k$ will be involved in all three fights, $i$ and $j$ will be involved in only two fights (one over $k$'s output and one when they fight with $k$ over their own output). Hence, if the $ij$ coalitonal structure is in place, then the consumption levels of $i$, $j$ and $k$ ($i \neq j \neq k$, with $i, j, k = 1, 2, 3$) are respectively:[3]

$$\begin{aligned}
c_i^{ij} &= \frac{z_i}{z_i + z_k} f_i(L_i) + \frac{z_i}{z_i + z_j + z_k} f_k(L_k) \\
c_j^{ij} &= \frac{z_j}{z_j + z_k} f_j(L_j) + \frac{z_j}{z_i + z_j + z_k} f_k(L_k) \\
c_k^{ij} &= \frac{z_k}{z_i + z_k} f_i(L_i) + \frac{z_k}{z_j + z_k} f_j(L_j) + \frac{z_k}{z_i + z_j + z_k} f_k(L_k).
\end{aligned}$$

The (von Neumann-Morgenstern) utility to $i$ is $U_i(c, l)$, where $c$ and $l$ are respectively his levels of consumption and leisure. I assume that $U_i$ takes the following (quasi-linear) form: $U_i(c, l) = c + v_i(l)$, where $v_i$ satisfies the following (standard) conditions: $v_i(0) = 0$, $v_i'(0) = +\infty$, $v_i' > 0$ and $v_i'' < 0$.[4]

# 3 The Socially Optimal Outcome

Before establishing the socially optimal (i.e., the first-best) outcome, we first determine the equilibrium levels of work chosen by each player under each of the five possible coalitional structures. This is straightforward to establish, and the following lemma states the desired results:

**Lemma 1.** *Letting $L_i^G$, $L_i^D$, $L_i^{ij}$, $L_i^{ik}$ and $L_i^{jk}$ respectively denote $i$'s equilibrium work level under the five possible coalitional structures (where $i \neq j \neq k$, with $i, j, k = 1, 2, 3$), we have: $L_i^G$ is the unique solution to the first-order condition*

$$f_i'(L_i) = v_i'(T - L_i),$$

*$L_i^D$ is the unique solution to the first-order condition*

$$\frac{z_i}{z_1 + z_2 + z_3} f_i'(L_i) = v_i'(T - L_i),$$

[3]Notice that I am implicitly assuming that fighting does not lead to any loss (or destruction) of a player's output. I adopt this assumption partly because it makes cooperation (i.e., the formation of coalitions) that much harder. I should also emphasize that a key cost of fighting is that it affects the players' ex-ante incentives to work. This cost of fighting — which is indirect but fundamental — is a key element of my model and analysis.

[4]I adopt this particular utility function partly to simplify the analysis (the additive separability feature), and partly to capture the assumption that each player has risk-neutral preferences over consumption.

*$L_i^{ij}$ is the unique solution to the first-order condition*

$$\frac{z_i}{z_i + z_k} f_i'(L_i) = v_i'(T - L_i),$$

*$L_i^{ik}$ is the unique solution to the first-order condition*

$$\frac{z_i}{z_i + z_j} f_i'(L_i) = v_i'(T - L_i),$$

*and $L_i^{jk}$ is the unique solution to the first-order condition*

$$\frac{z_i}{z_i + z_j + z_k} f_i'(L_i) = v_i'(T - L_i).$$

Notice that

$$L_i^G > L_i^{ij}, L_i^{ik} > L_i^{jk} = L_i^D.$$

Thus, for each player, the amount of work done when the grand coalition is in place strictly exceeds the amount of work done under any partial coalitional structure, which, in turn, exceeds the amount of work done under the degenerate coalitional structure. This, of course, results from the fact that a player's marginal return from work — which comprise the left-hand sides of the above first-order conditions (the right-hand sides being the marginal cost of work) — is increasing in the extent to which his output is protected from theft. Under the grand coalitional structure it's protected from theft from both players, while under two of the three partial coalitional structures it's protected from theft from only one player; under the third partial coalitional structure and the degenerate coalitional structure both of the other players will try to steal his output. Not surprisingly, "productive efficiency" is achieved if and only if the grand coalitional structure is in place.

It will be helpful to state the unique equilibrium payoff to each player under each coalitional structure. Letting $\Pi_i^G$, $\Pi_i^D$, $\Pi_i^{ij}$, $\Pi_i^{ik}$ and $\Pi_i^{jk}$ respectively denote $i$'s unique equilibrium payoff under the five possible coalitional structures (where $i \neq j \neq k$ with $j, k = 1, 2, 3$), we obtain (given the results stated above and noting that $L_i^{jk} = L_i^D$):

$$\Pi_i^G = v_i(T - L_i^G) + f_i(L_i^G) \tag{1}$$

$$\Pi_i^D = v_i(T - L_i^D) + \frac{z_i}{z_1 + z_2 + z_3}\left[f_i(L_i^D) + f_j(L_j^D) + f_k(L_k^D)\right] \tag{2}$$

$$\Pi_i^{ij} = v_i(T - L_i^{ij}) + \frac{z_i}{z_i + z_k} f_i(L_i^{ij}) + \frac{z_i}{z_i + z_j + z_k} f_k(L_k^D) \tag{3}$$

$$\Pi_i^{ik} = v_i(T - L_i^{ik}) + \frac{z_i}{z_i + z_j} f_i(L_i^{ik}) + \frac{z_i}{z_i + z_j + z_k} f_j(L_j^D) \tag{4}$$

$$\Pi_i^{jk} = v_i(T - L_i^D) + \frac{z_i}{z_i + z_j} f_i(L_j^{jk}) + \frac{z_i}{z_i + z_k} f_k(L_k^{jk}) + \frac{z_i}{z_i + z_j + z_k} f_i(L_i^D). \tag{5}$$

These equilibrium payoffs are the players' payoffs in the absence of any intra-coalitional (or inter-coalitional) transfers of output. The solution concept defined below is aimed

for an analysis of coalition formation under this non-transferable utility (NTU) setting.[5] Notice that since $\Pi_i^{jk} > \Pi_i^D$, my coalitional game is a game with positive spillovers (or externalities). That is, $i$ strictly prefers the other two players ($j$ and $k$) to form a coalition than not to form one. This is because $i$ can then effectively "free-ride" on the agreement reached by $j$ and $k$ not to steal each others' outputs — since $i$ would then be the only one trying to steal both $j$'s and $k$'s outputs. The following proposition establishes that the grand coalitional structure is the socially optimal coalitional structure:

**Proposition 1** (**The Grand Coalition is the Socially Optimal Outcome).** *For any $i, j, k$ (where $i \neq j \neq k$ and $i, j, k = 1, 2, 3$):*

$$\Pi_1^G + \Pi_2^G + \Pi_3^G > \Pi_i^{ij} + \Pi_j^{ij} + \Pi_k^{ij} > \Pi_1^D + \Pi_2^D + \Pi_3^D.$$

*Proof.* I first establish the second inequality. After substituting for these expressions using (2)-(5), and then re-arranging terms, the difference

$$[\Pi_i^{ij} + \Pi_j^{ij} + \Pi_k^{ij}] - [\Pi_i^D + \Pi_j^D + \Pi_k^D]$$

can be written as the *sum* of the following three terms:

$$[v_i(T - L_i^{ij}) + f_i(L_i^{ij})] - [v_i(T - L_i^D) + f_i(L_i^D)]$$
$$[v_j(T - L_j^{ij}) + f_j(L_j^{ij})] - [v_j(T - L_j^D) + f_j(L_j^D)]$$
$$[v_k(T - L_k^{ij}) + f_k(L_k^{ij})] - [v_k(T - L_k^D) + f_k(L_k^D)].$$

Since $L_k^{ij} = L_k^D$, the last of these terms is zero. Furthermore, since $L_i^{ij} > L_i^D$ and $L_j^{ij} > L_j^D$, each of the first two terms is strictly positive. Hence, we have established the second inequality in the proposition. The first inequality can be established by a similar algebraic manipulation (and then appealing to the fact that each player works more under the grand coalitional structure than under the $ij$ partial coalitional structure). □

Proposition 1 implies that the players' collective incentive is to establish the grand coalitional structure over any partial coalitional structure, and any partial coalitional structure over the degenerate coalitional structure.

# 4 The Equilibrium Coalitional Structure

The solution concept which we adopt is essentially the concept of "Equilibrium Binding Agreements" introduced and studied in Ray and Vohra (1997). Definition 1 (below) states

[5] I will however later on briefly discuss coalition formation on the assumption that intra-coalitional transfers of output are possible; this additional instrument seems natural and may enhance the likelihood of the stability of the grand coalitional structure, which, as established below, maximizes the sum of the players utility payoffs. The idea is that players may agree to cooperate (join a coalition) only if they are bribed to do so (by having some other player transfer some of his output).

the conditions that are required to be satisfied in order for the grand coalition to be a stable coalitional structure. The first condition, G.1, states that no single player can unilaterally and profitably deviate (by breaking away from the grand coalition) *anticipating the rational responses of the other two players to such a deviation.* The second condition, G.2, states that a joint deviation by a pair of players is not profitable to at least one of them (anticipating their rational responses after they have conducted the joint deviation). The final condition, G.3, states that the three players cannot all (collectively) benefit from collectively agreeing to break up the grand coalition).[6]

**Definition 1** (Stability of the Grand Coalitional Structure)**.** *The grand coalitional structure is stable if and only if the following three conditions hold:*

*G.1 [Unilateral deviations]. For each* $i = 1, 2, 3$ *and* $i \neq j \neq k$ *(*$j, k = 1, 2, 3$*)*

$$\Pi_i^G \geq \begin{cases} \Pi_i^{jk} & \text{if } \Pi_j^{jk} \geq \Pi_j^D \text{ and } \Pi_k^{jk} \geq \Pi_k^D, \\ \Pi_i^D & \text{otherwise.} \end{cases}$$

*G.2 [Pairwise Deviations]. For any pair of players* $i$ *and* $j$ *(*$i \neq j$ *and* $i, j = 1, 2, 3$*), either* $\Pi_i^G \geq \Pi_i^{ij}$ *or* $\Pi_j^G \geq \Pi_j^{ij}$ *if* $\Pi_i^{ij} \geq \Pi_i^D$ *and* $\Pi_j^{ij} \geq \Pi_j^D$*; otherwise, either* $\Pi_i^G \geq \Pi_i^D$ *or* $\Pi_j^G \geq \Pi_j^D$*.*

*G.3 [The Collective Deviation]. Either* $\Pi_1^G \geq \Pi_1^D$ *or* $\Pi_2^G \geq \Pi_2^D$ *or* $\Pi_3^G \geq \Pi_3^D$*.*

The notion of stability that underlies Definition 1 — that a coalitional structure is stable if it is immune to "credible fragmentation" — implies that the degenerate coalitional structure is (trivially) stable, by definition. I now turn to the conditions that are required to be satisfied in order for an arbitrary partial coalitional structure to be stable. These conditions are stated in Definition 2. They state that neither of the two players in a partial coalition can unilaterally and profitably deviate.

**Definition 2** (Stability of the three Partial Coalitional Structures)**.** *The partial coalitional structure* $ij = \{\{i, j\}, \{k\}\}$ *(where* $i \neq j \neq k$ *and* $i, j, k = 1, 2, 3$*) is stable if and only if the following condition holds:*

*P.1 [Unilateral Deviations].* $\Pi_i^{ij} \geq \Pi_i^D$ *and* $\Pi_j^{ij} \geq \Pi_j^D$*.*

The above definitions provide the conditions under which each of the five possible coalitional structures is stable. The question then is which one of these coalitional structures can emerge in equilibrium of the (unspecified) coalition formation process at date 1. We

[6]Notice that this definition of stability of the grand coalition allows for any subset of players to consider breaking away from the grand coalition, but it does not allow for players to break up and then subsequently form new coalitions. It is also based on the notion that players when considering to deviate anticipate the subsequent rational responses by themselves and by the other players (on further potential break-ups). Implicit in this view is that splitting up or fragmentation (i.e., breaking away from a coalition) is costless relative to forming new coalitions.

adopt the following viewpoint. The three players first consider forming the grand coalition. If it's stable, then it is the unique equilibrium coalitional structure. If, on the other hand, the grand coalitional structure is not stable, then the players consider forming one of the three partial coalitional structures. If at least one of them is stable, then the equilibrium coalitional structure will be a partial coalitional structure. Of course, if more than one of the partial coalitional structures is stable, then which one is the equilibrium coalitional structure is left unspecified. Finally, if neither the grand coalitional structure nor any one of the partial coalitional structures are stable, then the unique equilibrium coalitional structure is the degenerate coalitional structure (which, as noted above, is stable, by definition).

I now derive the equilibrium coalitional structure in the case when the players are identical: that is, they have identical productive skills, identical fighting skills and identical preferences. Formally, for all $i = 1, 2, 3$, $f_i = f$, $z_i = z$ and $v_i = v$.

It follows from Lemma 1 that for all $i = 1, 2, 3$, $L_i^G = L^G$ and $L_i^D = L^D$, where $f'(L^G) = v'(T - L^G)$ and $f'(L^D)/3 = v'(T - L^D)$. Furthermore, for all $i \neq j$, $L_i^{ij} = L^P$, where $f'(L^P)/2 = v'(T - L^P)$. Finally, note that for all $i \neq j \neq k$, $L_i^{jk} = L^D$. Notice, as is expected, that $L^G > L^P > L^D$. It might also be convenient to state here the unique equilibrium payoffs to each player under each coalitional structure, which follow from a straightforward application of (1)-(5). The equilibrium payoffs to each player under the grand coalitional structure and the degenerate coalitional structure are respectively:

$$\Pi^G = v(T - L^G) + f(L^G) \tag{6}$$

$$\Pi^D = v(T - L^D) + f(L^D) \tag{7}$$

Under any partial coalitional structure, the equilibrium payoff to the player who is in a coalition on his own is

$$\Pi^{NP} = v(T - L^D) + f(L^P) + \frac{f(L^D)}{3}, \tag{8}$$

while the equilibrium payoff to each of the two coalitional partners is

$$\Pi^P = v(T - L^P) + \frac{f(L^P)}{2} + \frac{f(L^D)}{3}. \tag{9}$$

The following observations will prove helpful in studying the stability or otherwise of each of the possible coalitional structures. Note that $\Pi^G > \Pi^D$, by definition. Furthermore, it is straightforward to establish that $\Pi^G > \Pi^P$ and that $\Pi^{NP} > \Pi^P$ — where, it may be noted, that the latter inequality means that in any partial coalitional structure, the player who is not party to any agreement obtains a higher payoff than that obtained by the players who have agreed to be part of a coalition between themselves.[7] The following proposition characterizes the equilibrium coalitional structure in this case of identical players:

[7]First, I establish that $\Pi^G > \Pi^P$. Since $L^P > L^D$, it follows that $\eta(L^P) > \Pi^P$, where $\eta(L) = v(T - L) + f(L)$. Hence, since $\eta(L^G) > \eta(L^P)$ — by definition, since $\eta$ is maximized at $L = L^G$ — it follows that $\Pi^G > \Pi^P$. Now I establish that $\Pi^{NP} > \Pi^P$. By definition, $L^D$ is the unique maximizer

**Proposition 2** (**Equilibrium Coaltional Structure).** *Assume the players are identical.*
*(a)* **Inefficient Equilibria.** *If* $\Pi^P \geq \Pi^D$ *and* $\Pi^G < \Pi^{NP}$*, then there are three equilibrium coalitional structures, namely the three partial coalitional structures.*
*(b)* **Unique Efficient Equilibrium.** *If either* $\Pi^P < \Pi^D$ *or* $\Pi^G \geq \Pi^{NP}$*, then the unique equilibrium coalitional structure is the grand coalitional structure.*

*Proof.* In order to establish this proposition, I begin by checking whether the grand coalitional structure is stable; that is, whether or not conditions G.1-G.3 are satisfied when the players are identical. G.3, which is satisfied if and only if $\Pi^G \geq \Pi^D$, is satisfied, by definition. G.2 is satisfied if and only if

$$\Pi^G \geq \begin{cases} \Pi^P & \text{if } \Pi^P \geq \Pi^D, \\ \Pi^D & \text{if } \Pi^P < \Pi^D. \end{cases}$$

This is satisfied, since (as noted above) $\Pi^G > \Pi^P$ (and since $\Pi^G > \Pi^D$). Finally, note that G.1 is satisfied if and only if

$$\Pi^G \geq \begin{cases} \Pi^{NP} & \text{if } \Pi^P \geq \Pi^D, \\ \Pi^D & \text{if } \Pi^P < \Pi^D. \end{cases}$$

By definition $\Pi^G > \Pi^D$, and so I need to check that whenever $\Pi^P \geq \Pi^D$, it must be the case that $\Pi^G \geq \Pi^{NP}$. This statement is not always correct; and hence, in summary, the grand coalitional structure is stable if and only if either $\Pi^P < \Pi^D$ or $\Pi^G \geq \Pi^{NP}$.

I now turn to examine the potential stability of an arbitrary partial coalitional structure; since the players are identical, the three possible partial coalitional structures are essentially "identical". Condition P.1 is satisfied if and only if $\Pi^P \geq \Pi^D$.

The proposition is now an immediate consequence of these results, and the definition of what constitutes an equilibrium coalitional structure. □

The intuition for why the grand coalitional structure is not stable if $\Pi^G < \Pi^{NP}$ and $\Pi^D \leq \Pi^P$ is as follows. Under these circumstances (when these two inequalities hold), each player has a unilateral incentive to deviate and break away from the grand coalition. This is because by doing so a player obtains a payoff of $\Pi^{NP}$, which, given the supposition that $\Pi^G < \Pi^{NP}$, strictly exceeds his payoff $\Pi^G$ that he obtains under the grand coalitional structure. It should be noted that his payoff from this unilateral deviation is $\Pi^{NP}$ and *not* $\Pi^D$, since he rationally and correctly anticipates that if he unilaterally deviates and breaks away from the grand coalition then (because, by supposition, $\Pi^D \leq \Pi^P$) the other two players have an incentive to form a partial coalition.

of $v(T-L) + [f(L)/3]$, and hence, $\Pi^{NP} \geq v(T-L^P) + [4f(L^P)/3]$. The right-hand side of this inequality exceeds $\Pi^P$ (since $L^P > L^D$), and hence the left-hand side of this inequality exceeds $\Pi^P$, as desired.

Since, by Proposition 1, the grand coalitional structure maximizes aggregate (or social) surplus, the question is whether or not it could emerge in equilibrium if intra-coalitional transfers of output are entertained? In addressing such a question, the basic point to note is that $\Pi^{NP} > \Pi^{G}$ implies that even when intra-coalitional transfers are allowed, it will not be possible to make the grand coalitional structure stable — since at least some player will receive a payoff that is less than $\Pi^{NP}$, who would have an incentive to break away. Essentially, the instability of the grand coalition under such circumstances is made possible because under any partial coalitional structure the player who is not party to any agreement free-rides on the other two players who are in a coalition and hence have agreed not to steal each others' outputs. This means that there is more for the third player to steal from these two players.

The message that is emerging here is that in a world with three identical players, a partial coalitional structure entails free-riding by one player over the good behaviour of the other two. This makes sense. But it implies that under some circumstances the grand coalition would not emerge even if intra-coalitional transfers are entertained. We leave it for future research to characterize the equilibrium coalitional structure when the players can differ in some respects such as in their productive skills and/or in their fighting skills. There is furthermore much scope to extend the model in various interesting ways such as by allowing for players to inflict damage on each others' productive and/or fighting skills.

# References


Bloch, Francis (1996), "Sequential Formation of Coalitions in Games with Externalities and Fixed Payoff Division", *Games and Economic Behavior*, **14**, 90-123.

Hobbes, Thomas (1651), *Leviathan.* Pengiun classics, London 1986 (edited by C.B. Macpherson).

Maskin, Eric (2003), "Bargaining, Coalitions and Externalities", mimeo, Princeton.

Ray, Debraj and Rajiv Vohra (1997), "Equilibrium Binding Agreements", *Journal of Economic Theory*, **73**, 30-78.

Ray, Debraj and Rajiv Vohra (1999), "A Theory of Endogenous Coalition Structures", *Games and Economic Behavior*, **26**, 286-336.