Coalitions and Collective Action © Physica-Verlag, Wuerzburg (Germany) 1984

Stable Coalition Formation: Aspects of a Dynamic Theory

A.E. Roth, Pittsburgh

Abstract: This paper considers how coalitions gain and lose members and ultimately stabilize, under certain assumptions.

1. Introduction

Any comprehensive theory of coalition formation in economic environments must simultaneously deal with several closely related questions which each pose serious conceptual difficulties, since it must resolve both the question of which coalitions will form, and how the benefits of each coalition will be distributed. To the extent that economic agents are motivated to form coalitions precisely in order to enjoy these benefits, they select coalitions based on the benefits offered, and so the two questions are inseparable.

Nevertheless, it can be useful to consider certain limited aspects of coalition formation which abstract away from the strategic elements of the process, in an effort to identify phenomena common to a broad class of coalition formation processes. The model proposed in this essay is directed at the question of how coalitions gain and lose members and ultimately stabilize. Since the model abstracts away from particular strategic considerations, it will be useful for studying phenomena which are potentially common to a broad range of coalition forming processes, of the kind which occur not only in economic environments, but also in political, social, and ecological contexts. To emphasize this point of view, the discussion will be phrased in terms of the following ecological metaphor.

2. An Ecological Metaphor

Imagine a stretch of seashore immediately following an exceptionally violent storm. Along this stretch of coast are a number of tide-pools; isolated environments of rock and sand and sea water, swept clean of life by the exceptional violence of the recent storm. All of these tide pools are virtually identical as potential habitats for various forms of coastal life.

Now imagine the same stretch of shore after some suitably long time has elapsed. Each tide pool now contains a diverse population of living organisms, co-existing with each other in the same environment. Furthermore, it is likely that some, if not all, of the organisms in a particular tide pool can be classified as permanent residents; that is, over a period of many tides, these organisms remain part of the population of that tide pool.

Looking at another tide pool, there is no reason to expect that the population of organisms which it supports will be identical to that of the first. On the contrary, it is likely that two different tide pools, while sharing the same physical characteristics, will support populations of markedly different composition.

This difference in population among physically identical tide pools is due, presumably, to the different sequences in which organisms are introduced to each tide pool by the random action of the tides. The introduction of an organism to a tide pool changes the characteristics of that environment as a potential habitat for other organisms; and different organisms change the environment differently. Thus the environment associated with each tide pool undergoes a process of evolution as new organisms are introduced by the tide.

This is not to imply that the first organisms to be introduced to a particular tide pool will of necessity become permanent residents of that pool; they may be displaced by later arrivals. The dynamics of this process is one of the things which we hope to study by means of a formal model. We will also want to characterize the degree and manner in which this dynamic process becomes stable over time.

3. The Model

Let X be the (finite) universe of organisms, and let H be the set of available habitats. For most of what follows, we will consider only a single habitat h in H.

Let R be a binary relation defined on X such that for all organisms x in X the statement xRx is false (i.e., $\sim xRx$). The relation R is called "*prevents*", and if for organisms x, y in X, xRy then we say that (the presence of) x *prevents* y (from occupying the same habitat). The relation R need not be symmetric. That is, it may be that for some organisms x and y in X, xRy but $\sim yRx$.

A collection of organisms $x, y, \ldots, z$ such that $xRyR \ldots RzRx$ is called a cycle. A cycle is called even or odd, depending on whether the number of organisms in it is even or odd.

Time is divided into *periods*, and in each period at most one organism can be introduced to each habitat. (Allowing many organisms to be introduced simultaneously would not change the results.) The population of a habitat h at the end of period n will be written $P_n(h)$, or, when no confusion will result, P_n. We assume that $P_0 = \phi$, the empty set, and for all periods n, P_n is of course a subset of X. We further assume that the length of each period is short compared to the life-span of the organisms in question.

For each organism x in X define $D(x) = \{ y \in X \mid xRy \}$. $D(x)$ is thus the set of all organisms y which are prevented from occupying the same habitat as the organism x. For each population P (that is, for each subset of X) define $D(P) \equiv \bigcup_{x \in P} D(x)$. $D(P)$ is

the set of organisms y which are prevented from occupying the same habitat as some organism x in the population P. Finally, let $U(P) \equiv X - D(P)$. $U(P)$ is the set of all organisms which are *not* prevented from occupying the same habitat as any organism in the population P.

The population P_n of a habitat h at the end of period n evolves in the following way. If no new organism is introduced into the habitat at the start of period $n+1$, then $P_{n+1} = P_n$, that is, the population is unchanged.

If, at the start of period $n+1$, an organism y is introduced into the habitat such that y is in $D(P_n)$, then $P_{n+1} = P_n$. That is, if a new organism y is introduced into the habitat, such that y is prevented from occupying the same habitat as one of the organisms already in the population of that habitat, then the organism y is eliminated, and the population remains unchanged.

If, at the start of period $n+1$, an organism y is introduced such that y is in $U(P_n)$ (i.e., y is not in $D(P_n)$), then $P_{n+1} = P_n \cup \{y\} - D(y)$. That is, if the new organism y is not prevented from occupying the habitat by any member of the existing population of that habitat, then y occupies a place in the habitat. Any organism x in P_n which is prevented from occupying the same habitat as y is then eliminated from the population.

4. Analysis of the Model

The first question which we must answer is which populations are *feasible*, that is, for which subsets P of X is it possible that $P = P_n(h)$ for some habitat h at the end of some period n?

Proposition 1: A population P is feasible if and only if $P \subseteq U(P)$.

Proof: To see that P is not feasible if $P \not\subseteq U(P)$ note that, in this case, there must be x and y in P such that xRy. If x is a member of the population when y enters the habitat, then y is eliminated. If y is a member of the population when x enters, then y is eliminated if there is no z in the population such that zRx, otherwise x is eliminated.

To see that P is feasible if $P \subseteq U(P)$, suppose P had k elements which enter the habitat in periods $1, \ldots, k$. Then $P = P_k(h)$.

We also want to consider which populations are *permanent*; i.e., which populations P have the property that, if $P \subseteq P_n$, then for all $m \geq n$, $P \subseteq P_m$, no matter what the sequence by which new organisms enter the habitat. It will be convenient to denote the set $U(U(P))$ by $U^2(P)$.

Proposition 2: A feasible population P is permanent if and only if $P \subseteq U^2(P)$.

Proof: To see that P is permanent if $P \subseteq U(U(P))$ note that the only way in which some x in P could be eliminated from the population would be if some y such that yRx were to become a member of the population P_m for $m > n$. But if yRx, then y is

not in $U(P)$, since $P \subseteq U(U(P))$. Therefore y is in $D(P)$, so y cannot become a member of any population P_m such that $P \subseteq P_m$. So if $P \subseteq P_n$, then $P \subseteq P_{n+1}$, and $P \subseteq P_m$ for $m \geq n$ by induction.

To see that P is not permanent if $P \not\subseteq U(U(P))$, we need to show that some sequence of events leads to a population P_m such that $P \not\subseteq P_m$, and $m > n$. Let y be an element of $U(P)$ such that yRx for some x in P. If y is introduced to the habitat at period $n+1$, then y is an element of P_{n+1} and x is not, so $P \not\subseteq P_{n+1}$.

The set $U^2(P)$ is the set *protected* by the population P. This terminology is meant to reflect the fact, used in the above proof, that any organism which prevents a member of $U^2(P)$ is in turn prevented by some member of P. A population P such that $P \subseteq U^2(P)$ will be called *self protecting*, and, as shown above, self protecting populations are permanent.

Consider now a self protecting population P, an organism x in $U^2(P) - P$, and some other organism z such that zRx. Since x is protected by P, there is a member of P which prevents z. Thus no organism z which prevents x can ever become part of some population which contains the population P. Therefore, if at any period the organism x is introduced to a habitat with a population P_n which contains P, then the organism x occupies a place in that habitat. Furthermore, it is not difficult to see that $P \cup \{x\} \subseteq U^2(\{P \cup \{x\}\})$, so every population $P_m (m \geq n)$ contains $P \cup \{x\}$; that is, the population $P \cup \{x\}$ is a self-protecting permanent part of the population.

We say, therefore, that a feasible, self-protecting population is *stable* when it has grown to the point where it includes all organisms which it protects. Thus a stable population P is one such that $P = U^2(P) \subseteq U(P)$.

This definition of a stable population leaves open questions of existence and non-emptiness. That is, for an arbitrary universe of organisms X and binary relation R, must it always be the case that some stable population P exists? And under what circumstances might the empty set be stable, so that over time the population of some habitat could contain no permanent members, but only transients? The following two propositions answer these questions.

Proposition 3: There exists a stable population for every universe X and binary relation R. That is, there exists some $P \subseteq X$ such that $P = U^2(P) \subseteq U(P)$.

This proposition does not in fact depend on the finiteness of X. The mathematical result was first established, in a very different context, as a corollary of a general theorem about functions defined on lattices, in *Roth* [1975]. [In *Blair/Roth*, a close relationship was established between this result and the famous fixed-point theorem of *Tarski*].

Together with Proposition 3, the following result provides a sufficient (but not a necessary) condition for the existence of non-empty stable populations.

Proposition 4: The empty set $\varnothing$ is a stable population if and only if $\varnothing = U(X)$. (So a non-empty stable population always exists if $U(X) \neq \varnothing$.)

Proof: Since $\varnothing \subseteq X = U(\varnothing)$, the empty set is stable if and only if $\varnothing = U^2(\varnothing) = U(X)$.

The fact that circumstances exist under which even the empty set is a stable population serves to emphasize that the kind of stability we are talking about here is dynamic rather than static. Before we go on to consider what manner of static equilibrium can occur, let us describe the dynamics associated with a stable population.

Consider a habitat containing a stable population $P = P_n$. We may think of the set of all organisms as being partitioned into three sets: P, $D(P)$, and $U(P) - P$. For convenience we shall call the third set P'.

If a member of $D(P)$ is introduced into the habitat, it is, of course, immediately eliminated. If a member x of P' is introduced into the habitat, however, it becomes part of the population P_{n+1}, since it is contained in $U(P_n)$. However, x is unprotected by P, which means that there is an organism y in $U(P)$ such that yRx. Since this organism y is not in P, it must be in P'. Thus, unless xRy, x will be eliminated from the population if y is introduced into the habitat. Therefore the population of this habitat over subsequent periods will consist of the permanent population P, augmented by some transient organisms from P'.

Since stable populations are permanent, the population of any given habitat will tend towards the largest stable population compatible with its present population. Under suitable circumstances, this can lead to a stable population P such that the set P' is empty. In this case $P = U(P)$, and we say that P is *completely stable.* Any organism introduced into a habitat containing a completely stable population is immediately eliminated, since any organism outside of P is in $D(P)$. (It is clear that a completely stable population is stable, since if $P = U(P)$, then $P = U^2(P)$.) Note that a completely stable population is never empty.

The following sufficient condition for at least one completely stable population to exist is due to *Richardson* [1953].

Proposition 5: If there are no odd cycles (for a given X and R) then there exists a completely stable population.

As the references associated with Propositions 3 and 5 make clear, the mathematical structure associated with stable and completely stable populations has been studied in other contexts. In cooperative game theory, when R represents the domination relation and X is the set of imputations, the structure of stable populations occurs in sets of imputations called *subsolutions* [cf. *Roth*, 1976], while *solutions* are sets of outcomes which have the structure of completely stable populations [cf. *von Neumann/Morgenstern*]. Sets with the same structure as completely stable populations are also called *kernals* in graph theory, where X is the set of nodes of a graph, and R represents the arcs [cf. *Behzad/Harary*]. In certain kinds of games on graphs, in which players take turns moving along arcs, and the first player to have no legal move loses, the set of nodes P from which a win can be assured has the same structure as a stable population. Furthermore, the set of nodes from which a draw can be assured is the set

$P' = U(P) - P$, and the set of nodes from which a loss cannot be prevented is the set $D(P)$ [cf. *Roth*, 1978]. Finally, completely stable populations are studied by *Wilson* [1972], in a context closely related to the ecological metaphor employed here.

Examples

In the accompanying diagrams, organisms are indicated by letters and the relation R by arrows. An arrow pointing from one organism to another indicates that the first organism prevents the second from occupying the same habitat.

Example 1: (See Figure 1). In this example, the following populations are stable: $\varnothing$, $\{a, c\}$; $\{a, c, e, g\}$; $\{a, c, f, h\}$; and $\{b, d, f, h\}$. Of these the last three are completely stable.

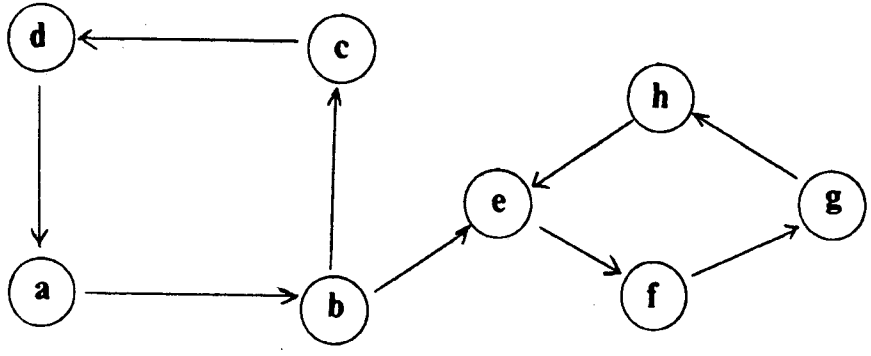

Figure 1

Example 2: (See Figure 2), In this example, the sole stable population is $\{a\}$, which happens to be equal to $U(X)$. No completely stable populations exist.

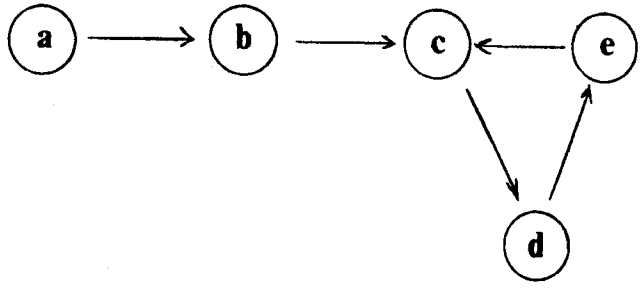

Figure 2

Example 3: (See Figure 3). In this example, the sole stable set is $\{a, c\}$, which is completely stable. This demonstrates that the sufficient condition of Proposition 5 is not necessary to insure the existence of completely stable sets.

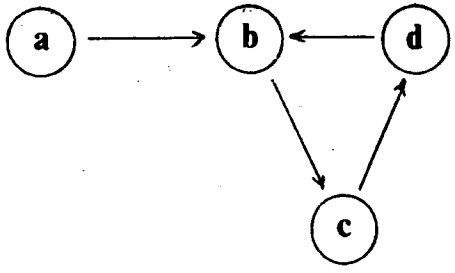

Figure 3

Example 4: (See Figure 4). In this example, the empty set is the sole stable population. Naturally, it is not completely stable.

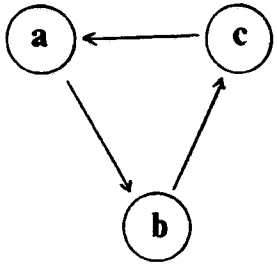

Figure 4

References

Behzad, M., and *F. Harary*: Which Directed Graphs Have a Solution? Mathematica Slovaca **27**, 1977, 37–42.

Blair, C., and *A.E. Roth*: An Extension and Simple Proof of a Constrained Lattice Fixed Point Theorem. Algebra Universalis **9**, 1979, 131–132.

von Neumann, J., and *O. Morgenstern*: The Theory of Games and Economic Behavior. Princeton, N.J., 1944.

Richardson, M.: Solutions of Irreflexive Relations. Annals of Mathematics **58**, 1953, 537–590.

Roth, A.E.: A Lattice Fixed-Point Theorem With Constraints. Bulletin of the American Mathematical Society **81**, 1975, 136–138.

– : Subsolutions and the Supercore of Cooperative Games. Mathematics of Operations Research **1**, 1976, 43–49.

– : Two-Person Games on Graphs. Journal of Combinational Theory, Series B **24**, 1978, 238–241.

Tarski, A.: A Lattice-Theoretical Fixpoint Theorem and Its Applications. Pacific Journal of Mathematics **5**, 1955, 285–309.

Wilson, R.: Consistent Modes of Behavior. Technical Report No. 55, Institute for Mathematical Studies in the Social Sciences, Stanford University 1972.

Coalitions and Collective Action © Physica-Verlag, Wuerzburg (Germany) 1984

Stable Coalition Structures[1])

S. Hart, Tel Aviv, and *M. Kurz*, Stanford

By a "coalition" one means a set of players who decide to act together, as one group, relative to the rest of the players. By the term "coalition structure" we mean a partition of the set of players into a number of coalitions each aiming to enhance the interests of its members. The typical coalition structure assumed in the literature is the collection of singletons. However, in many real situations individuals act through social organizations like political parties, unions, trade groups and others. Thus one notes that at any moment of time society organizes itself into a coalition structure and the outcome of any game calls for a division of gains among coalitions as well as among the members of each coalition. This means that the existence of coalition structures implies that the interactions among the players are conducted on two levels: First, *among* the coalitions, and second, *within* each coalition. In most of game theory it is assumed that the coalition structure is given exogenously. In contrast, the theory which we presented in *Hart/Kurz* [1983] addresses the problem of why do coalition structures form, and predicts, as an *endogenous* outcome, which one will indeed form.

This theory is based on two concepts. First, a *coalition structure value* (*CS*-value, for short) is defined; it is an evaluation of the players' prospects for any coalition structure. Second, based on this value, one finds which coalition structure is *stable*, in the sense that no player or group of players can change the coalition structure to their advantage.

One of the main properties which we postulate the *CS*-value to have is overall efficiency. This means that our analysis does not aim to characterize that organization of society which is needed in order to achieve social efficiency. Rather, we consider the formation of coalitions as one among many strategic acts used by the players, within the bargaining process, in order to increase as much as possible their share of the total social "pie". More specifically, we do not assume that each one of the groups receives what it can assure itself (in technical terms – its worth) but rather, that all the coalitions bargain for the division of the total, which is the worth of the grand coalition.[2])

[1]) This work was supported by National Science Foundation Grant SES80-06654 at the Institute for Mathematical Studies in the Social Sciences, Stanford University. The authors thank R.J. Aumann and L.S. Shapley for helpful comments.

[2]) Since one of the possible cooperative decisions is to act separately, cooperation can never decrease the total available resources (i.e., one has "super-additivity").