Games and Economic Behavior **38,** 318–346 (2002)
doi:10.1006/game.2001.0878, available online at http://www.idealibrary.com on IDEAL®

Stability and Segregation in Group Formation

Igal Milchtaich

Department of Economics, Bar-Ilan University, Ramat-Gan 52900, Israel; and Managerial Economics and Decision Sciences, J. L. Kellogg Graduate School of Management, Northwestern University, Evanston, Illinois 60208
E-mail: milchti@mail.biu.ac.il

and

Eyal Winter[1]

Department of Economics, The Hebrew University of Jerusalem, Jerusalem 91905, Israel; and Department of Economics, Washington University, St. Louis, Missouri 63130
E-mail: mseyal@pluto.mscc.huji.ac.il

Received June 3, 1998

This paper presents a model of group formation based on the assumption that individuals prefer to associate with people similar to them. It is shown that, in general, if the number of groups that can be formed is bounded, then a stable partition of the society into groups may not exist. (A partition is defined as stable if none of the individuals would prefer be in a different group than the one he is in.) However, if individuals' characteristics are one-dimensional, then a stable partition always exists. We give sufficient conditions for stable partitions to be segregating (in the sense that, for example, low-characteristic individuals are in one group and high-characteristic ones are in another) and Pareto efficient. In addition, we propose a dynamic model of individual myopic behavior describing the evolution of group formation to an eventual stable, segregating, and Pareto efficient partition. *Journal of Economic Literature* Classification Numbers: C72, H41. © 2002 Elsevier Science

Key Words: group formation; coalition structure; local public goods; segregation; myopic optimization; weak acyclicity.

1. INTRODUCTION

Group formation refers to environments in which an individual's payoff from performing a particular action only depends on the set of other

[1] The second author thanks the German–Israeli Foundation and the European Commission for financial support through the GIF and TMR grants.

0899-8256/02 $35.00

individuals taking that action. The motivation behind studying group formation is that, in a wide range of social and economic interactions, individuals take independent decisions that, among other things, determine the identities of the people with whom they associate. In such situations, the individuals' preferences over the available alternatives are often strongly dependent on their preferences for particular kinds of people. For example, in deciding about a residential neighborhood, recreational activity, fashion, or affiliation to a religious congregation, the question of who are the other people making the same choice is often of major importance. The objective of this paper is to study group formation in environments in which individuals seek to be in a group in which the other people, or at least the "average person," are similar to them.[2]

There is more than one way the notion of a similarity between an individual and a *group* of people may be interpreted and modeled. In this paper, we suggest three such models. Our first group formation model is very simple. It consists of a finite set of individuals and a metric which, for each pair of individuals, quantifies the relevant differences between them. For every possible partition of the set of individuals into groups, each person's satisfaction with the partition is determined by the average distance between himself and the other members of his group: the greater the distance, the lower the satisfaction. Thus, the highest degree of satisfaction is attained when a group is homogeneous, i.e., the distance between any two individuals in the group is zero. Therefore, if there were no bound on the number of groups, only homogeneous groups would be formed. To avoid such a triviality, we assume there *is* an upper bound to the possible number of groups. For example, each group of individuals may be associated with a distinct site or jurisdiction, so that the number of groups cannot exceed the number of available locales. The identification of groups with sites has the additional advantage of providing a natural labeling of groups. Thus, an individual's choice of site determines his group. Since our concern here is only with group formation, uncompounded by any other effects, we assume that there is no intrinsic difference between sites. However, there are clearly situations in which it would be more realistic to assume that the sites are not identical.

There is more than one sense in which a partition of a society into groups may be stable (see, for example, Bogomolnaia and Jackson, 1998). The definition of stability used in this paper is as follows: A partition is stable if none of the individuals would be better off leaving his current group and joining another group. This definition is rather demanding. An

[2] Needless to say, there are many situations in which people prefer the company of others *not* similar to them. They may, for example, prefer to associate with people of a higher socio-economic class. In such cases, our models do not apply.

individual is not required to obtain the consent of the members of his current group or the group he joins. Nor does he incur any travel costs. The number of stable partitions is, therefore, likely to be smaller than it would be if free mobility were not assumed. We show, in fact, in Section 2 that, in the model described above, stable partitions might not exist.

Our counterexample involves a society of five individuals who are represented by points in the plane: The distance between any two individuals is equal to the distance between the corresponding points. In this example, if only two groups are allowed to form, stable partitions do not exist. We show, however, in Section 2 that a similar problem does not arise when people can be represented by points on the *line*. In this case, at least one stable partition always exists. The argument, for the special case of only two groups, was already given by Schelling (1978). As he showed, it is sufficient to consider segregating partitions; in the two-group case, this means that all the individuals to the right of some imaginary point on the line are in one group and all the individuals to the left of that point are in another group. In each group, if the individual closest to the other group is satisfied with the proposed partition, nobody in the group objects to it. If not, that individual needs to be moved to the other group. After a finite number of iterations, this procedure gives a partition to which no one objects.

This algorithm for finding a stable partition (under the assumption that people's characteristics are unidimensional) yields a partition that is not only stable but also segregating. This is significant since, as we show in Section 3, all stable and segregating partitions are weakly Pareto efficient. There are, however, stable partitions which are not weakly Pareto efficient, and therefore not segregating. The reason for this potential lack of efficiency has to do with the assumption that individuals only care about the absolute value of the difference between their own characteristics and the characteristic of each of the other individuals in their group. They are not concerned whether the other members of the group are all to the left or to the right of themselves, or there are some on both sides. Consequently, an individual may be unwilling to leave a group even when his characteristic is closer to the average in another group than in his own group. This observation suggests that the inefficiency that accompanies lack of segregation may not arise when individuals, rather than looking at the *average distance* from the other people in their group, seek to minimize the *distance from the average* in the group. In Section 4, we show that, in such a case, all stable partitions are, indeed, segregating and weakly Pareto efficient. In fact, a partition is stable in the new, "distance from the average," model if and only if it is stable in the old, "average distance," model and is segregating. Such a partition is therefore efficient in *both* models.

One conceptual difference between the "average distance" and the "distance from the average" models is that, in the latter, people do not necessarily have to know the identity or the characteristic of every member of each group. Knowing the attributes of a "representative individual" in each group suffices. People's preferences are such that they seek to be the group in which the representative individual is most similar to them. Unlike in the "average distance" model, they do not necessarily object to being in a heterogeneous group. Therefore, the result that, under such preferences, stable partitions are always segregating, rather than arbitrarily mixed, is not immediately obvious from the assumptions. Depending on the application concerned, these preferences may be taken either as a primitive or as derived from some other, more basic, ones. Suppose, for example, that the members of each group are making collective decisions on issues affecting their welfare, e.g., decide on the level of expenditure on some public good. Individuals differ in their preferences, and hence also in their ideas about the optimal level of expenditure. An individual whose preferred level of expenditure is close to the average in his group is likely to be less dissatisfied with the collective decision than one whose preferred level is, say, higher than everybody else's. Another possibility is that group membership serves as a signal to the outside world. The idea is that outside observers categorize a person in a certain way when they learn he is a member of a particular group. To minimize errors, they ascribe the average characteristic of the people in the group to that person. The categorization is then most accurate for people close to the average. It should be emphasized that the issue here is *honest* signaling. People *want* to be recognized for what they are. Hence, they will try to be in a group in which the average person is as close as possible to themselves.

These examples of the "distance from the average" model may raise two objections. First, when individuals join a group, they change it. For example, if the group is going to take a decision, a new member has as much power as everybody else does to affect it. Correspondingly, people should consider not the average characteristic of the other members of their group but the average characteristic of *all* the individuals in the group, including themselves. This may very well affect their decisions on which group to join. In particular, they would generally prefer to join smaller rather than larger groups, everything else being equal. This is because the average characteristic tends to shift more towards that of a newcomer in a small rather than in a large group. In Section 5, we study the implications of such a "self-effect" on group formation. Because of the (indirect) effect of group size on individuals' behavior, this case is more difficult to study analytically than the previous one, in which self-effect is absent. In particular, we are only able to prove the existence of stable partitions when there are two groups. However, in other aspects, at least,

this model is similar to the "distance from the average" model without self-effect.

The second objection that can be raised against the above examples is that other statistics—in particular the median—may be more suitable than the average characteristic for predicting the decisions reached by the group or for describing its "representative individual." In certain situations, this is certainly the case (see Westhoff, 1977, 1979). We show, however, in Section 6 that if the average characteristic is replaced by a *weighted* average, then stable partitions do not necessarily exist. The existence of stable partitions is also not generally guaranteed if individuals' preferences over groups are allowed to depend in a more direct way on the number of group members: If congestion effects are strong, all partitions may be unstable. In the same section, we show that, in all the three models considered in this paper (namely, the "average distance" model and the "distance from the average" model with or without self-effect), there may not be a partition which is *strongly* stable in the sense that there are no profitable deviations, not just for individuals but also for coalitions (i.e., subsets of individuals). In fact, we give an example in which there is always a profitable deviation for a coalition consisting of identical individuals, who are members of the same group and move together to join some other group.

In the last section (Section 7), we describe a dynamic model of group formation. In this model, individuals behave myopically by moving, one at a time, to a group which they prefer to their current one. The order in which individuals move is random, at least to some degree. We show that, in both variants of the "distance from the average" model, this process converges almost surely to a stable and segregating partition within a finite number of periods, regardless of the initial partition. (As before, with self-effect we are able to prove convergence only if there are just two groups.) A small amount of randomness is indispensable since, without it, the process may be cyclical. We show, however, that there is always a way out of a cycle, and therefore the process will escape it eventually, provided that every order of movers occurs with positive probability.

The issue of group formation has been discussed in a variety of contexts in the past. Schelling's early analysis (1978) is particularly relevant to our own. Schelling provides several examples illustrating the strategic considerations related to group formation when individuals' preferences depend on such factors as the age, income, or race of those with whom they associate. Our analysis offers a more complete and rigorous treatment of some of the issues addressed by Schelling. Karni and Schmeidler (1990) give an example of an overlapping generation model in which two types of individuals choose one of three possible colors (or fashions). A consumer's payoff from choosing a color depends on the proportion of individuals of his type

who decide on the same color. In this framework, Karni and Schmeidler examine fashion cycles and the conditions for segregation. Group formation in a different context was discussed by Kaneko and Kimura (1992). These authors use a game theoretic model to explain racial discrimination. In their framework, the population consists of two types, blacks and whites, with the latter forming the majority. Players are engaged in a "festival game" generating a positive payoff for participants if and only if a majority of the population attends the festival. They show that a stable convention may emerge in which only the white population attends the festival. Ellickson *et al*. (1999) developed a general-equilibrium theory of clubs based on the view that clubs are involved in a variety of public activities, rather than only in the production of public good, and obtained results about the relation between the competitive equilibrium and the core in this framework. Our paper is related to this work in so far as clubs are a form of group formation. However, apart from the fact that our approach here is game theoretic, our notion of group formation differs from the notion of clubs as defined by Ellickson *et al*. (1999). In particular, in our framework there is no consumption and the individuals' payoffs are derived directly from their association with a group. The literature on core stability in games with effective coalitions (see, for example, Kaneko and Wooders, 1982; Le Breton *et al*., 1992) is also somewhat related to our paper. An effective coalition is characterized by a particular combination of types of individuals. For example, in a two-sided market, an effective coalition consists of a buyer and a seller. The existence of types in games with effective coalitions induces a particular form of preferences of individuals over coalitions. Even more closely related to our approach are the recent papers of Bogomolnaia and Jackson (1998) and Banerjee *et al*. (2001). These studies focus on several stability notions (in particular, core, individual, and Nash stability) of hedonic coalition formation, in which a player's payoff is only determined by the identity of the other members of his group. None of these notions is directly applicable to our framework, however, since they all imply that individuals would not like to deviate and form their own groups. In our model, individuals normally *would* like to form new groups, and it is only the exogenous constraint on the number of groups which prevents them from doing so. Correspondingly, the sufficient conditions for stability obtained by Bogomolnaia and Jackson (1998) are not satisfied by any of our models.

Finally, although our paper does not draw its motivation directly from the literature on local public goods, it is somewhat related to it. This literature aims to explain how consumers partition themselves into jurisdictions to maximize their utility from the consumption of public good (see, for example, Tiebout, 1956; Wooders, 1980; Bewley, 1981; Greenberg and Weber, 1993; Jehiel and Scotchmer, 1997; Konishi *et al*., 1998).

Conceptually, our model is similar. As indicated above, groups in our model can be interpreted as jurisdictions, and individuals' characteristics as their optimal levels of public good (with respect to a particular tax scheme). Perhaps the most significant departure of our paper from the relevant literature on local public goods is the dynamic process of movement between groups. Conceivably, this process may be applied to model the way jurisdictions evolve to their final structure. We are not aware of similar models in the current literature on local public goods.

2. THE SETUP

A *society* consists of a finite number n of *individuals*, indexed from 1 to n; it can thus be identified with the index set $I = \{1, 2, \ldots, n\}$. A partition of the society is a specification of m groups of individuals, such that each individual belongs to one, and only one, group. The groups are indexed from 1 to m, and the set of groups can thus be identified with the index set $K = \{1, 2, \ldots, m\}$. Group formation is viewed as endogenous. However, the *number* m of groups is an exogenous parameter (with $m \leq n$). The formation of new groups is viewed as being either physically impossible or prohibitively expensive. Moving between existing groups, on the other hand, is always possible and does not involve any cost. Formally, an (ordered) *partition* (of I) is an element $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ of K^n such that, for all $k \in K$, $\sigma_i = k$ for some i.[3] Here, σ_i denotes the (index of the) group to which individual i belongs. It is convenient to define a partial order $\leq$ on the set of partitions, as follows: $\sigma \leq \sigma'$ if and only if $\sigma_i \leq \sigma_i'$ for all i. The notation $\sigma < \sigma'$ means $\sigma \leq \sigma'$ and $\sigma \neq \sigma'$. A partition σ is *minimal* (with respect to $\leq$) in a given set of partitions if there is no partition σ' in that set such that $\sigma' < \sigma$. It is *maximal* in a given set if there is no partition σ' in that set such that $\sigma < \sigma'$. It is the *smallest* (*greatest*) partition in a set if $\sigma \leq \sigma'$ (respectively, $\sigma' \leq \sigma$) for every partition σ' in the set. A partition σ is *nondecreasing* if $\sigma_i \leq \sigma_j$ when $i < j$.

The *distance* d_{ij} between two individuals i and j is a measure of how different they are from one another. For a given partition σ, the average distance between individual i and the other individuals in his group is denoted $d_i(\sigma)$. If i is the only member of the group, then $d_i(\sigma)$ is defined as 0. Individuals prefer the company of other individuals who are similar to them. A partition is considered to be *unstable* if there is at least one

[3] The assumption that all m groups are nonempty is made for technical convenience only. None of the results in this paper would be affected if groups were allowed to remain empty, so that m would become an *upper bound* on the number of groups.

individual who would like to move to a different group, because the average distance between him and the members of that group is smaller than in his current group. Formally, a partition σ is unstable if there is a partition τ differing from σ in only one coordinate, say the ith one, such that $d_i(\tau) < d_i(\sigma)$. Otherwise, the partition is *stable*. A partition is stable if and only if it is a (pure-strategy) Nash equilibrium in the game in which the set of players is I, the set of actions available to each player is K, and the payoff function of player i is given by some strictly decreasing function of d_i. The following example shows that, even in the special case in which the d_{ij}'s are distances in a Euclidean space, an equilibrium does not always exist.

EXAMPLE 2.1. There are five individuals in the society and two groups. Each individual is represented by a point in the plane, such that the distance between any two individuals is equal to the Euclidean distance between the respective points. The points are $(0, 0)$, $(0, 47)$, $(55, 62)$, $(101, 55)$, and $(74, -34)$ (see Fig. 1). A straightforward, though somewhat tedious, computation shows that none of the 30 possible partitions in this example is stable.

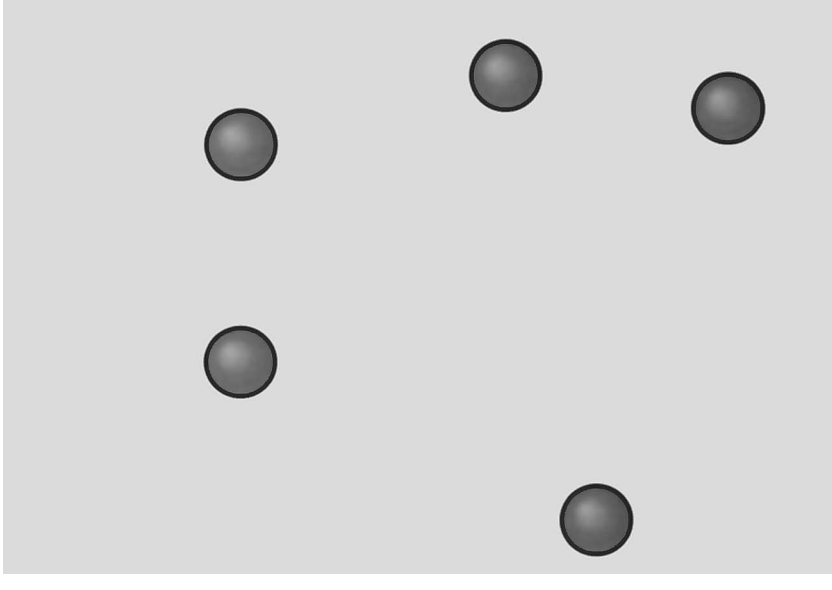

FIG. 1. A society in which all the partitions into two groups are unstable.

If, however, individuals can be represented by points on the *line*, a stable partition always exists. The following proposition follows as an immediate corollary from a stronger result (Theorem 2.4) that we prove below.

PROPOSITION 2.2. *If for every individual i there is a real number c_i, such that $d_{ij} = |c_i - c_j|$ for all i and j, then a stable partition exists.*

In the rest of this paper, we assume that distances can indeed be represented as in Proposition 2.2. We will refer to c_i as the *characteristic* of individual i and assume, without loss of generality, that $c_1 \le c_2 \le \cdots \le c_n$. A natural question arising in this context is whether stability implies that individuals are separated, or segregated, according to their characteristics, so that, for example, low-characteristic individuals are in one group and high-characteristic individuals are in another. Formally, a partition is *segregating* if, for every pair of individuals j and j' such that $c_j < c_{j'}$, if j and j' are members of the same group k then *all* the individuals i such that $c_j \le c_i \le c_{j'}$ are in k. Perhaps somewhat surprisingly, stable partitions need *not* be segregating.

EXAMPLE 2.3. The following partition is stable, but not segregating:

$$0 \quad 5 \quad 5 \quad 5 \qquad 4 \quad 4 \quad 4 \quad 9.$$

In this example, and elsewhere in this paper, the characteristics of individuals belonging to the same group are written alongside one another.

While a stable partition, as Example 2.3 shows, need not be segregating, it always has the following weaker property (which holds for every segregating partition): If two individuals in two different groups have the same characteristic c, then *all* the individuals in these two groups have characteristic c. The proof of this is rather simple. If σ is a partition which does not have this property, then there are two individuals i and j, which are not in the same group, and a third individual j' in j's group, such that $c_i = c_j \neq c_{j'}$. The average distance between i and the individuals in j's group is less than $d_j(\sigma)$. (Since $d_{ij} = |c_i - c_j| = 0$, the average distance is $(1/n_k)(n_k - 1)d_j(\sigma)$, where n_k is the number of individuals in j's group. Since not all the individuals in that group have the same characteristic, $d_j(\sigma) > 0$.) The average distance between j and the individuals in i's group is less than or equal to $d_i(\sigma)$. Therefore, the partition σ is unstable both if $d_j(\sigma) \le d_i(\sigma)$ or if $d_i(\sigma) < d_j(\sigma)$. We thus conclude that all stable partitions have the property described above. It is easy to see that a *nondecreasing* partition that has this property is segregating. Therefore, every nondecreasing stable partition is segregating. The next theorem shows that at least one nondecreasing stable partition always exists. In the theorem, the following terminology is used. Individual i *u-objects* to a

nondecreasing partition σ if $\sigma_i < m$ and $d_i(\sigma) > d_i(\tau)$, where τ is the partition differing from σ only in that i is a member of group $\sigma_i + 1$. Individual i *d-objects* to σ if $\sigma_i > 1$ and $d_i(\sigma) > d_i(\tau')$, where τ' is the partition differing from σ only in that i is a member of group $\sigma_i - 1$.

THEOREM 2.4. *Every minimal element* (*with respect to the partial order* $\leq$) *in the* (*nonempty*) *set of all nondecreasing partitions to which none of the individuals u-objects is stable and segregating, and the same is true for every maximal element in the set of all nondecreasing partitions to which none of the individuals d-objects.*

The proof of Theorem 2.4 uses the following lemma.

LEMMA 2.5. *Let* σ *be a nondecreasing partition. If the highest-* (*lowest-*) *index individual in a group does not u-object* (*respectively, d-object*) *to* σ, *then no one in that group does so. If there is no individual that u-objects or d-objects to* σ, *then this partition is stable and segregating.*

Proof. Let i be the highest-index individual in some group k, and j another member of k. Assuming that $k < m$, let τ be the partition differing from σ only in that individual i is in group $k + 1$, and τ' the partition differing from σ only in that individual j is in group $k + 1$. Since σ is nondecreasing and $c_i \geq c_j$, $d_j(\tau') = d_i(\tau) + (c_i - c_j)$. On the other hand, $d_j(\sigma) \leq d_i(\sigma) + (c_i - c_j)$. Therefore, $d_i(\sigma) - d_i(\tau) \geq d_j(\sigma) - d_j(\tau')$. It follows that if j u-objects to σ, then so does i. The proof for d-objections is similar. If no one u-objects or d-objects to σ, then it is clearly stable. As remarked above, a stable partition that is nondecreasing is also segregating. ■

Proof of Theorem 2.4. In view of Lemma 2.5, and of the symmetry between u- and d-objections, it is sufficient to prove the following: If a given partition is maximal in the set of all nondecreasing partitions to which none of the individuals d-objects, then no one u-objects to it either. We will show, in fact, that if σ is a nondecreasing partition to which none of the individuals d-objects, and if i is the highest-index individual who u-objects to σ (assuming there is such an individual), then no one d-objects to the partition $\tau > \sigma$ defined by $\tau_i = \sigma_i + 1$ and $\tau_j = \sigma_j$ for all $j \neq i$. (Note that τ is a legitimate partition. Since i u-objects to σ, there must be at least one more individual in i's group σ_i.) This partition is nondecreasing, for it follows from Lemma 2.5 that i must be the highest-index individual in his group.

Clearly, individual i does not d-object to τ. It therefore follows from Lemma 2.5 that none of the individuals j such that $\tau_j = \tau_i$ d-objects to τ. If j is an individual such that $\tau_j > \tau_i + 1$ or $\tau_j < \sigma_i$ then, since j did not d-object to σ, he does not d-object to τ either. The same is true if

$\tau_j = \tau_i + 1$: Since $c_i \le c_{j'} \le c_j$ for all individuals j' in group τ_i, the average distance between j and the individuals in group τ_i could only have increased when i joined that group. It remains to examine the case in which $\tau_j = \sigma_i$. If j is the lowest-index individual in group σ_i (in which the highest-index individual was i), the average distance between j and the other individuals in group σ_i could only have decreased when i left that group. Since individual j did not d-object to σ, a fortiori he does not d-object to τ. It follows, by Lemma 2.5, that none of the individuals j such that $\tau_j = \sigma_i$ d-objects to τ. ■

3. EFFICIENCY

A partition σ is *weakly Pareto efficient* if there is no partition τ such that $d_i(\tau) < d_i(\sigma)$ for all individuals i. A partition σ is *Pareto efficient* if there is no partition τ such that $d_i(\tau) \le d_i(\sigma)$ for all individuals i and strict inequality holds for at least one individual. As Example 2.3 demonstrates, a stable partition need not be even weakly Pareto efficient. (In that example, if individuals 1 and 8 switch places, everybody is better off.) Note, however, that the partition in Example 2.3 is not segregating. In fact, the next theorem shows that a stable partition is not weakly Pareto efficient *only* if it is not segregating. Note that, in a segregating partition, there may be two or more groups in which all the individuals have the same characteristic c. Such a segregating partition will be said to be *detaching*. Clearly, in the *generic* case, in which $c_i \ne c_j$ when $i \ne j$, all segregating partitions are nondetaching.

THEOREM 3.1. *A stable and segregating partition is weakly Pareto efficient. A stable and nondetaching segregating partition is Pareto efficient.*

The proof of Theorem 3.1 is given in the Appendix. This proof shows, in fact, even more than the theorem asserts. In the generic case, for example, not only is it true that every deviation from a stable and segregating partition which is advantageous to some individuals must be disadvantageous to at least one of the others, but *any* change that is not tantamount to re-labeling of groups must leave at least one individual worse off.

In Theorem 3.1, the condition of a stable partition cannot be dropped. In fact, a nondetaching segregating partition that is not stable need not be even weakly Pareto efficient. For example, all individuals are worse off in

0 3 4 5 6 9

than in

0 3 4 5 6 9.

In the second part of the theorem, the condition of a nondetaching partition cannot be dropped; stability alone is not a sufficient condition for Pareto efficiency of a segregating partition (but only for weak Pareto efficiency). For example, the partitions

$$\begin{array}{cccc} 0 & 1 & 3 & 3 \\ 0 & 1 & 3 & 3 \end{array}$$

are both stable. However, the first two individuals are better off in the second partition, while the other two are equally well off in both partitions.

4. DISTANCE FROM THE AVERAGE

In the model presented in Section 2, stable partitions may not be even weakly Pareto efficient. As shown, this potential lack of efficiency has to do with the fact that, in that model, some stable partitions are not segregating. This section presents a variant of the above model in which all the stable partitions are segregating and weakly Pareto efficient. In fact, in the new model, a partition is stable if and only if it is stable in the old model and is segregating. The difference between the two models is that, rather than seeking to minimize the *average distance* between their own characteristic and those of the other members of their group, individuals like the distance between their characteristic and the *average characteristic* of the other people in the group to be as small as possible. Formally, for a given partition σ, and for an individual i in group k, $d_i(\sigma)$ is redefined as $|\hat{C}_k - c_i|$, where $\hat{C}_k$ is the average characteristic of all individuals other than i in group k. If i is the only member of his group, then $d_i(\sigma) = 0$. As before, the stability of a partition σ is defined by the condition that, for every individual i, there is no partition τ differing from σ in the ith coordinate only such that $d_i(\tau) < d_i(\sigma)$. The old model will hereafter be referred to as the "average distance" model, and the new one as the "distance from the average" model.

One interpretation of the difference between the two models is that, in the "distance from the average" model, groups are viewed as significant entities, and not merely as collections of individuals. This model may be more suitable than the "average distance" model in cases in which the properties of the "average person" in a group are reasonably reliable predictors of the group's collective acts or decisions (see the Introduction). Note that there is one important case in which these two models coincide. When there are only two types of individuals (e.g., two races), people's preferences are the same in both models. Namely, they want to be with people of their own kind, and rank groups according to the proportion of

such people in them. In this special case, a partition is stable (in either model) if and only if it is segregating, which means that all groups are homogeneous.

THEOREM 4.1. *In the "distance from the average" model, all stable partitions are segregating and weakly Pareto efficient. Moreover, all stable partitions that are nondetaching are Pareto efficient.*

The proof of Theorem 4.1 is given in the Appendix. As it shows, the reason that stable partitions in the "distance from the average" model are always segregating is that individuals always want to move to a group in which they are closer to the current average characteristic than in their own group. This is not the case in the "average distance" model (cf. Example 2.3).

The existence of stable partitions in the "distance from the average" model can be proved in exactly the same way as in the "average distance" model. In fact, Lemma 2.5 and Theorem 2.4, as well as the proofs given for these results above, hold in both models. However, with the aid of Theorem 4.1, it is possible to prove more than mere existence. As the following theorem shows, the set of all stable partitions in the former model is, in fact, a subset of the corresponding set in the latter one.

THEOREM 4.2. *The set of all stable partitions in the "distance from the average" model is nonempty. It consists of all the stable partitions in the "average distance" model which are segregating.*

Proof. Consider a segregating partition σ that is unstable in one of these models. Without loss of generality, we may assume that σ is nondecreasing. By Lemma 2.5, there is some individual i, who is the highest- or lowest-index individual in his group, that (in the model in which σ is unstable) u-objects or d-objects, respectively, to σ. Clearly, the value of $d_i(\sigma)$ in the "distance from the average" and the "average distance" models is the same. Moreover, the value of $d_i(\tau)$ is also the same in both models, where τ is the nondecreasing partition that differs from σ in the ith coordinate only (i.e., differs only in that individual i is in the neighboring group). It follows that $d_i(\tau) < d_i(\sigma)$, and hence the partition σ is unstable, in *both* models. ■

5. SELF-EFFECT

In the "distance from the average" model presented in the previous section, individuals do not take into account the fact that, by joining a group, they change it: the "average person" in the group becomes more like themselves. This assumption is implicit in the definition that sets

$d_i(\sigma)$ as the distance between i's characteristic and the average characteristic of the *other* individuals in his group. This section explores the alternative possibility, that people contemplating a move take into consideration the effect it will have on the group they join. Formally, $d_i(\sigma)$ is redefined as the absolute value $|C_k - c_i|$ of the difference between the characteristic c_i of individual i and the average characteristic C_k of *all* the individuals in i's group k (including i himself). The new model is termed the "distance from the average" model with *self-effect*. The model presented in Section 4 will hereafter be referred to as the "distance from the average" model without self-effect. Note that, even with self-effect, it is possible to express $d_i(\sigma)$ in terms of the average characteristic $\hat{C}_k$ of the *other* individuals in i's group k. Specifically, if the number of people n_k in group k is at least two, $d_i(\sigma) = (1 - 1/n_k)|\hat{C}_k - c_i|$. The factor $1 - 1/x$ can be interpreted as quantifying the (negative) effect of group size on people's desire to join a group. Note that, although group size is not assumed to have a direct effect on the well-being of group members, congestion effects do come in indirectly because small groups are more affected by the arrival of new members.

As shown in the Appendix, Theorem 4.1 holds for the "distance from the average" model with self-effect as well as for the model without self-effect. In fact, a single proof holds for both models. The (indirect) effect of group size on individuals' willingness to join the group does, however, complicate things when one tries to establish the existence of stable partitions in the "distance from the average" model with self-effect. We can prove that such a partition always exists for two groups, but are unable to prove a similar result for more than two groups. However, we do not have a counter-example, either. Thus, the existence of stable partitions in the "distance from the average" model with self-effect and more than two groups is unresolved. Note that, if $m = 2$, there are precisely $n - 1$ nondecreasing partitions, which are linearly ordered by $\leq$. The proof of the following result is given in the Appendix.

THEOREM 5.1. *In the "distance from the average" model with self-effect, if the number of groups is two, then a stable partition exists. Specifically, the smallest (with respect to the partial order $\leq$) nondecreasing partition to which none of the individuals u-objects,* or *the greatest nondecreasing partition to which none of the individuals d-objects, is stable (and hence segregating).*

In contrast with the parallel result for the "distance from the average" model without self-effect (Theorem 2.4), "or" in Theorem 5.1 cannot be replaced by "and." For example, with self-effect, individual 1 u-objects to the nondecreasing partition

$$-8 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \qquad 5,$$

while no one u-objects to

$$-8 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 5.$$

The latter partition is therefore the smallest element in the set of all nondecreasing partitions to which none of the individuals u-objects. However, since this partition is not segregating, it cannot be stable.

6. OTHER VARIATIONS

There are a number of additional ways in which the three models presented in this paper, or our notion of stability, could be further modified or generalized. This section considers three possible generalizations. All of them are shown to have the undesirable property that the set of stable partitions may be empty.

Weighted averages. Even with only two groups, the existence of stable partitions is not generally guaranteed if individuals are seeking to minimize, not the distance from the average characteristic of the people in their group (as in the "distance from the average" model with self-effect) but the distance from a *weighted* average of these people's characteristics. Specifically, suppose that each individual is assigned some fixed positive weight, and that a partition is defined as stable if, for every individual i, the absolute value of the difference between c_i and the weighted average (with respect to these weights) of the characteristics of all the people in i's group (including i himself) is less than or equal to the absolute value of the difference between c_i and what would have been the weighted average of the characteristics in any other group if individual i joined that group. Then, stable partitions may not exist. For example, it is straightforward to check that stable partitions do not exist when there are two groups and four individuals, with characteristics -2, $1/8$, 1, and 5 and weights 3, 8, 2, and 1, respectively.

Strongly stable partitions. The last example can be modified to show that, even with equal weights, a *strongly* stable partition does not always exist. By a strongly stable partition we mean a partition that is a strong equilibrium in the corresponding game (Aumann, 1959): there is no subset of individuals that could make themselves better off by moving simultaneously from their current groups to some other groups. This is true in all of our models: the "average distance" model and both "distance from the average" models. For example, let s be a positive integer and suppose that there are two groups and $14s$ individuals: $3s$ individuals with characteristic -2, $8s$ individuals with characteristic $1/8$, $2s$ individuals with characteris-

tic 1, and s individuals with characteristic 5. Then, in all three models, there are precisely six stable partitions—all the segregating ones. However, for large enough s ($s \geq 17$ in the case of the "average distance" model and the "distance from the average" model without self-effect, and *all* values of s in the case of the "distance from the average" model with self-effect), none of these six partitions is strongly stable. In fact, none of them is even a coalition-proof equilibrium (Bernheim *et al.*, 1987) of the corresponding game. Indeed, for each of the six partitions there is a characteristic c such that (in all three models) the individuals with characteristic c would be better off moving *together* from their (common) group to the other one.

Congestion effects. One may wonder whether results similar to those we obtained for the "distance from the average" model with self-effect also hold for more general models in which individuals are (either directly or indirectly) affected by the size of their group, as well as by the characteristics of its members. One class of such models includes "distance from the average" models in which, for every partition σ, every group k in which the number of individuals n_k is at least two, and every individual i in group k, $d_i(\sigma)$ can be written as $\psi(n_k)|\hat{C}_k - c_i|$, where $\hat{C}_k$ is the average characteristic of all individuals other than i in group k. The factor $\psi(x)$ expresses the congestion effects. In the "distance from the average" model without self-effect it is equal to 1. In the model with self-effect it is $1 - 1/x$. Do the results in Sections 4 and 5 also hold for other functional forms of $\psi(x)$? In general, they do not. For example, if $\psi(x) = (1 - 1/x)^2$, $m = 2$, and $n = 5$ then there are no stable partitions when $c_1 = -1$, $c_2 = c_3 = 0$, $0 < c_4 < 1/13$, and $c_5 = 1$. In addition, even when a stable partition does exist, it might not be weakly Pareto efficient. For example, the non-detaching segregating partitions

$$0 \quad 3 \quad 3 \quad 4 \quad 4 \qquad 7$$
$$0 \quad 3 \quad 3 \qquad 4 \quad 4 \quad 7$$

are both stable if $\psi(2) = 0.74$, $\psi(3) = 0.94$, and $\psi(x) = 1 - 1/x$ for $x > 3$. However, everybody is better off in the first partition. We thus see that it is not possible to maintain existence and efficiency of stable partitions when people's desire to join a group may depend in an arbitrary way on the number of group members.

7. MYOPIC OPTIMIZATION AND CONVERGENCE TO A STABLE PARTITION

In the previous sections, the existence and efficiency of stable and segregating partitions in our three models were studied. This section is

concerned with how such partitions form. To study the evolution of group formation, we present a simple, and rather general, model in which individuals behave myopically, and possibly nondeterministically, in moving from one group to another. We show that, in both "distance from the average" models (but only for two groups, in the model with self-effect), this process converges to a stable partition almost surely, regardless of the initial partition. This may be interpreted as implying that social migration between groups is a temporary phenomenon in a trajectory leading to eventual stability. As in Section 5, we do not know whether the restriction on the number of groups in the model with self-effect is substantial, or is only due to the limitations of our techniques. We also do not know whether convergence to a stable partition occurs in the "average distance" model. (For a partial result, see Lemma 7.3 below.)

Starting with an initial partition σ, consider the following stochastic process: In each period t ($t = 1, 2, \ldots$), an individual i and a group k are randomly selected according to some probability measure on $I \times K$. The probability that a particular pair is selected may depend on t, on the current partition, or on history. However, this probability is greater than some fixed $\varepsilon > 0$ for every i and k such that, at time t, it is better for individual i to join group k than to stay with his current one, and there is no other group that would be even better for him. (Note that we do not specify whether or not other pairs are selected with positive probability.) If the individual i selected is, indeed, better off moving to the group k selected than staying with his current group, he moves there. Otherwise, there are no changes. The process then proceeds to period $t + 1$.

THEOREM 7.1. *In the "distance from the average" model without self-effect, as well as in the model with self-effect and only two groups, the stochastic process described above converges to a stable—and hence segregating and weakly Pareto efficient—partition with probability* 1, *regardless of the initial partition* σ.

To prove Theorem 7.1, the following definitions are needed. A finite sequence of partitions, $\sigma(0), \sigma(1), \ldots, \sigma(T)$, is called a *path* of length T if, for every $1 \le t \le T$, there is exactly one individual i_t, called the *mover* in $\sigma(t)$, such that $\sigma_{i_t}(t) \neq \sigma_{i_t}(t-1)$. It is an *improvement path* if, for every $1 \le t \le T$, $d_{i_t}(\sigma(t)) < d_{i_t}(\sigma(t-1))$, and a *best-reply path* if, for every $1 \le t \le T$, $d_{i_t}(\sigma(t)) \le d_{i_t}(\tau)$ for every partition τ such that $\tau_j = \sigma_j(t)$ for all $j \neq i_t$. A *best-reply improvement path* is one that is both an improvement and a best-reply path. (A path of length zero, consisting of a single partition, is also considered a best-reply improvement path.) For the stochastic process described above, if the partition at time t is equal to $\sigma(0)$, the initial point of a given best-reply improvement path of length T, then the probability that at time $t + T$ the partition is $\sigma(T)$, the terminal

point of the path, is at least ε^T. Therefore, to prove Theorem 7.1 it suffices to show that, under the conditions of the theorem, every partition is the initial point of some best-reply improvement path leading to a stable partition. The following two lemmas, the proofs of which are given in the Appendix, together establish this. The first lemma shows that it is always possible to reach a segregating partition. The second lemma shows that it is always possible to go from a segregating partition to a stable one. In all, it takes no more than $4(m-1)(n-m)$ moves to reach a stable partition.

LEMMA 7.2. *In both "distance from the average" models, for every partition σ there is an improvement path whose initial point is σ, whose terminal point is segregating, and whose length does not exceed $2(m-1)(n-m)$. In the model without self-effect, there is moreover a best-reply improvement path with these properties.*

LEMMA 7.3. *In the "average distance" model, the "distance from the average" model without self-effect, and the "distance from the average" model with self-effect and only two groups, for every segregating partition σ there is a best-reply improvement path whose initial point is σ, whose terminal point is stable and segregating, and whose length does not exceed $2(m-1)(n-m)$.*

Lemmas 7.2 and 7.3 are related to the notion of weak acyclicity of games (Young, 1993). A normal-form game is *weakly acyclic* if for every pure-strategy profile there is a best-reply path leading from that strategy profile to a strict Nash equilibrium. A game is *acyclic* if in every best-reply path all strategy profiles are distinct. Lemmas 7.2 and 7.3 show that if the number of groups is two, the games defined by our two "distance from the average" models are weakly acyclic (at least if there are no ties, so that all pure-strategy equilibria are strict). In general, however, these games are not *acyclic*; there are best-reply improvement paths (of lengths greater than zero) in which the initial and terminal points coincide. For example, in the "distance from the average" model with self-effect,

$$\begin{matrix} -1 & \varepsilon & 1 \\ \varepsilon & 1 & \\ 1 & & \end{matrix} \qquad \begin{matrix} & & 1 \\ & -1 & 1 \\ -1 & \varepsilon & 1 \end{matrix}$$

is an improvement path for $0 < \varepsilon < 3/7$. In the "distance from the average" model without self-effect, as well as in the "average distance" model,

$$\begin{matrix} 0 & & \\ 0 & \varepsilon & \\ 0 & \varepsilon & 1 \end{matrix} \qquad \begin{matrix} 0 & \varepsilon & 1 \\ & 0 & 1 \\ & & 0 \end{matrix}$$

is an improvement path for $0 < \varepsilon < 1/4$. Since, in both cases, the first and last partitions are essentially the same, these paths can be extended indefinitely. Thus, although our dynamic model converges to a stable partition almost surely in both "distance from the average" models (at least if $m = 2$), cycles may nevertheless occur. This shows that the corresponding games do not generally have any kind of potential (Monderer and Shapley, 1996). These examples also demonstrate the general point that, even in situations in which each individual's best reply is always unique, a weakly acyclic game with more than two individuals need not be acyclic. For another class of weakly acyclic, but generally not acyclic, games see Milchtaich (1996, 1998).

APPENDIX

Formally, at least, the three models presented in this paper may be viewed as special cases of a single, general model. This makes it possible to unify the proofs of some of the results stated above. The common features of the three models are as follows. If the number n_k of individuals in group k is at least two, individual i is a member of that group, and the average characteristic of all the individuals other than i in group k is $\hat{C}_k$, then, in both "distance from the average" models,

$$d_i(\sigma) = \psi(n_k)|\hat{C}_k - c_i|, \tag{1}$$

where $\psi(n_k) = 1$ in the model without self-effect and $\psi(n_k) = 1 - 1/n_k$ in the model with self-effect. Alternatively, $d_i(\sigma)$ can be expressed in terms of the average characteristic C_k of *all* the individuals in group k, as

$$d_i(\sigma) = \phi(n_k)|C_k - c_i|, \tag{2}$$

where $\phi(n_k) = n_k/(n_k - 1)$ in the model without self-effect and $\phi(n_k) = 1$ in the model with self-effect. Equations (1) and (2), with $\psi(n_k) = 1$ and $\phi(n_k) = n_k/(n_k - 1)$, also hold in the average distance model, provided that individual i has *the minimum or the maximum characteristic* in group k. If this is not the case, then the expressions on the right-hand side of (1) and (2) (which are equal to one another) are *less* than $d_i(\sigma)$.

It turns out that the exact functional forms of the *congestion factors* $\psi(x)$ and $\phi(x)$, which quantify the effect of group size on people's desire to join the group, are not very important. What is important about these functions is that they be defined for all positive integers x, be related by

$$\psi(x) = (1 - 1/x)\,\phi(x) \qquad \text{for all } x \tag{3}$$

(which implies $\psi(1) = 0$, but does not constrain $\phi(1)$ in any way), and satisfy the following conditions:

$$\psi(x) < \phi(y) \qquad \text{for all } x \text{ and } y, \tag{4}$$

and

$$\psi(x) \leq \psi(y) \qquad \text{when } x \leq y. \tag{5}$$

Note, however, that condition (4), and both conditions in conjunction, stringently limit the effect that group size can have on people's desire to join a group (especially one which is at least moderately large). These limitations cannot be removed. Indeed, the examples in Section 6 show that in a variant of the "distance from the average" model in which only (5) (but not (4)) holds stable partitions may not exist, and in a variant in which only (4) (but not (5)) holds they may not be weakly Pareto efficient.

A unified proof of Theorem 3.1, which holds in all three models, is now presented.

Proof of Theorem 3.1. Let σ and τ be two partitions such that σ is stable and segregating and $d_i(\tau) \leq d_i(\sigma)$ for all i. We claim that $d_i(\tau) = d_i(\sigma)$ for at least one individual i, and if σ is nondetaching, then there is a permutation π of K such that $\pi(\sigma_i) = \tau_i$ for all i (and hence $d_i(\tau) = d_i(\sigma)$ for *all* i). This claim is trivial if $n = 1$. Suppose, then, that $n \geq 2$, and that we have already proved this claim for all societies in which the number of individuals is smaller than n (and for any number of groups).

For $k \in K$, let n_k be the number of individuals, and C_k the average characteristic, in group k when the partition is σ. Let n'_k and C'_k be the number of individuals and the average characteristic, respectively, in group k when the partition is τ. Without loss of generality, it may be assumed that $C_k \leq C_l$ and $C'_k \leq C'_l$ for all k and $l > k$. (Under this additional assumption, the above permutation π is, in fact, the identity.)

Suppose, first, that $n'_1 \geq n_1$. Let i be the lowest-index individual such that $\sigma_i = 1$. If $n_1 = 1$ then $d_i(\sigma) = 0$, and hence also $d_i(\tau) = 0$. If, in addition, the segregating partition σ is nondetaching, then (since $C_1 \leq C_k$ for all k) $c_j > c_i$ for all $j \neq i$. Since $d_i(\tau) = 0$ and $C'_1 \leq C'_k$ for all k, this implies that $\tau_i = 1$ and $\tau_j \neq 1$ for all $j \neq i$. It then follows from our induction hypothesis (which we apply to the $(n-1)$-member society $I \setminus \{i\}$) that there is a permutation π of K such that $\pi(1) = 1$ and $\pi(\sigma_j) = \tau_j$ for all $j \neq i$. This completes the proof of our claim in the case in which $n_1 = 1$. Suppose now that $n'_1 \geq n_1 > 1$. Since σ is segregating and $C_1 \leq C_k$ for all k, the characteristic of every individual j such that $\sigma_j = 1$ is less than or equal to the characteristic of every individual j' such that $\sigma_{j'} > 1$, and is (strictly) less than it if σ is nondetaching. Since $n'_1 \geq n_1$, this implies $\hat{C}_1 \leq \hat{C}'_1$, where $\hat{C}_1$ is the average characteristic of all

individuals $j \neq i$ such that $\sigma_j = 1$ and $\hat{C}'_1$ is the average characteristic of all individuals $j \neq i$ such that $\tau_j = 1$. Furthermore, if σ is nondetaching, then $\hat{C}_1 = \hat{C}'_1$ only if, for every individual j, $\sigma_j = 1 \Leftrightarrow \tau_j = 1$. Therefore, $\psi(n_1)(\hat{C}'_1 - c_i) \geq \psi(n_1)(\hat{C}_1 - c_i) = d_i(\sigma) \geq d_i(\tau) = \phi(n'_{\tau_i})(C'_{\tau_i} - c_i) \geq \psi(n'_1)(\hat{C}'_1 - c_i)$, where the last inequality follows from (4) and the assumption $C'_{\tau_i} \geq C'_1$ if $\tau_i \neq 1$, and it is an equality if $\tau_i = 1$. But, by (5), $\psi(n_1) \leq \psi(n'_1)$, and therefore all the inequalities in the previous sentence must, in fact, be equalities. In particular, $d_i(\sigma) = d_i(\tau)$ and $\hat{C}'_1 = \hat{C}_1$. If σ is nondetaching then, as pointed out above, the latter equality implies that the set of all individuals j such that $\sigma_j = 1$ coincides with the set of all individuals j such that $\tau_j = 1$. We may therefore complete the proof of our claim by applying the induction hypothesis to the society whose members are all individuals j with $\sigma_j \neq 1$.

Very similar arguments prove our claim in the case in which $n'_m \geq n_m$. (In this case, it is the highest-index individual i such that $\sigma_i = m$ which has to be considered.) Therefore, if $m \leq 2$ then the proof is complete. Assume, then, that $m \geq 3$.

If k is a group such that $C_k \leq C'_k$ and i is an individual such that $\sigma_i < k \leq \tau_i$, then $d_i(\sigma) \geq d_i(\tau) \geq \phi(n'_{\tau_i})(C'_{\tau_i} - c_i) \geq \phi(n'_{\tau_i})(C'_k - c_i) \geq \psi(n_k + 1)(C_k - c_i)$ by (4), and the last inequality is strict if $C_k \neq c_i$. On the other hand, since σ is stable, $\psi(n_k + 1)(C_k - c_i) \geq d_i(\sigma)$. Therefore, $d_i(\sigma) = d_i(\tau)$ and $C_k = c_i$. Since $\sigma_i \neq k$, the latter equality implies that σ cannot be nondetaching, and the proof is complete. Similarly, if there is a group k and an individual i such that $C_k \geq C'_k$ and $\sigma_i > k \geq \tau_i$, then we are done. Assume, then, that for every individual i and every group k, if $C_k \leq C'_k$ and $\sigma_i < k$ then $\tau_i < k$, and if $C_k \geq C'_k$ and $\sigma_i > k$ then $\tau_i > k$.

If $C_2 \leq C'_2$ or $C_{m-1} \geq C'_{m-1}$, this assumption implies that $n'_1 \geq n_1$ or $n'_m \geq n_m$, respectively. By our previous results, this completes the proof. Conversely, if $C_2 > C'_2$ and $C_{m-1} < C'_{m-1}$, then there is some $k > 1$ such that $C_{k-1} \geq C'_{k-1}$ and $C_k \leq C'_k$ (recall that we are assuming that $m \geq 3$). Then, for every individual i, $\sigma_i < k \Leftrightarrow \tau_i < k$. We may therefore apply our induction hypothesis to the two complementary societies, those individuals i with $\sigma_i < k$ and those individuals j with $\sigma_j \geq k$, and thus complete the proof. ■

The proof of Theorem 4.1 can now be completed. It only has to be shown that, in both "distance from the average" models, all stable partitions are segregating.

Proof of Theorem 4.1. Let σ be a stable partition. For a given pair of distinct groups k and l such that $C_k \leq C_l$, set $C = (C_k + C_l)/2$. To prove that σ is segregating, it suffices to show that $c_i \leq C$ for all the individuals i in group k; $C \leq c_i$ for all the individuals i in group l; and if $c_i = C$ for

some individual i in group k or in group l, then a similar equality holds for all individuals in both groups.

If $\sigma_i = k$ then, by the assumed stability of σ, $\phi(n_k)|C_k - c_i| \leq \psi(n_l + 1)|C_l - c_i|$. It follows, by (4), that $|C_k - c_i| \leq |C_l - c_i|$, and equality holds only if both sides are zero, that is, only if $C_k = C_l = c_i$. Therefore, $c_i \leq C$. Similarly, if $\sigma_i = l$ then $C \leq c_i$. If $c_i = C$, and hence $|C_k - c_i| = |C_l - c_i|$, for some individual i in group k or in group l, then, as just shown, $C_k = C_l$. In this case, $|C_k - c_i| = |C_l - c_i|$, and therefore $c_i = C$, for *all* the individuals i in groups k and l. ■

To prove Theorem 5.1, we will need three lemmas. All lemmas refer to the "distance from the average" model with self-effect. The following notation is used. The average characteristic in the society, $(c_1 + c_2 + \cdots + c_n)/n$, is denoted $\bar{c}$. If the number of groups is two, then $\sigma^{(j)}$ $(j = 1, 2, \ldots, n-1)$ denotes the nondecreasing partition defined by $\sigma_i^{(j)} = 1$ for $i \leq j$ and $\sigma_i^{(j)} = 2$ for $i > j$.

LEMMA A.1. *For a given nondecreasing partition σ, let i and j be two individuals in the same group k such that $C_k \leq c_i \leq c_j$ or $c_j \leq c_i \leq C_k$. If i u-objects to σ, then j also u-objects to it. If i d-objects to σ, then j also d-objects to it.*

Proof. Since σ is nondecreasing, $C_k \leq c_i \leq c_j \leq C_{k+1}$ or $c_j \leq c_i \leq C_k \leq C_{k+1}$. Therefore, $\phi(n_k)|C_k - c_j| - \psi(n_{k+1} + 1)|C_{k+1} - c_j| = \phi(n_k)|C_k - c_i| - \psi(n_{k+1} + 1)|C_{k+1} - c_i| + [\phi(n_k)|c_i - c_j| - \psi(n_{k+1} + 1)(c_i - c_j)]$. By (4), the expression in square brackets is nonnegative. Therefore, if i u-objects to σ, then so does j. The proof for d-objections is similar. ■

LEMMA A.2. *Suppose that $m = 2$. If $c_1 + c_n \leq 2\bar{c}$, then individual n does not d-object to any nondecreasing partition. If $c_1 + c_n \geq 2\bar{c}$, then individual 1 does not u-object to any nondecreasing partition.*

Proof. By symmetry, it suffices to prove the first assertion. Suppose, then, that there is some $1 \leq j < n - 1$ such that individual n d-objects to the partition $\sigma^{(j)}$. Then, $\phi(j+1)(c_n - (c_1 + c_2 + \cdots + c_j + c_n)/(j+1)) < \psi(n-j)(c_n - (c_{j+1} + c_{j+2} + \cdots + c_{n-1})/(n-j-1))$, and therefore $(c_1 + c_2 + \cdots + c_j + c_n)/(j+1) > (c_{j+1} + c_{j+2} + \cdots + c_{n-1})/(n-j-1)$ by (4). Since $(c_2 + \cdots + c_j)/(j-1) \leq (c_{j+1} + c_{j+2} + \cdots + c_{n-1})/(n-j-1)$ if $j \geq 2$, the last inequality implies $(c_1 + c_n)/2 > (c_{j+1} + c_{j+2} + \cdots + c_{n-1})/(n-j-1) \geq (c_2 + \cdots + c_{n-1})/(n-2)$, and therefore $(c_1 + c_n)/2 > \bar{c}$. ■

LEMMA A.3. *Suppose that $m = 2$. If $j < n$ is such that there is some individual who u-objects to the partition $\sigma^{(j)}$, then there is an individual $2 \leq i \leq j$ who u-objects to the partition $\sigma^{(i)}$, and hence does not d-object to the partition $\sigma^{(i-1)}$.*

Proof. Suppose the contrary, that such an i does not exist. Then, for every $2 \le i \le j$,

$$\phi(i)\big(c_i - (c_1 + c_2 + \cdots + c_i)/i\big) \le \phi(n+1-i)\big((c_i + c_{i+1} + \cdots + c_n)/(n+1-i) - c_i\big).$$

In particular, individual j does not u-object to $\sigma^{(j)}$. However, by assumption, *someone* u-objects to $\sigma^{(j)}$, and it therefore follows from Lemma A.1 that individual 1 u-objects to it. Hence,

$$\phi(j)\big((c_1 + c_2 + \cdots + c_j)/j - c_1\big) > \psi(n+1-j)\big((c_{j+1} + c_{j+2} + \cdots + c_n)/(n-j) - c_1\big).$$

All of the above inequalities involve only the *differences* between the characteristics of various individuals. It can therefore be assumed without loss of generality that $\bar{c} = 0$, and hence $c_1 + c_2 + \cdots + c_i \le 0$ for all $1 \le i \le n$. Denoting the last partial sum by Σ_i, and using (3), the above inequalities can be written as

$$(6) \qquad \big[\phi(n+1-i) + \psi(i)\big]\Sigma_i \le \big[\phi(i) + \psi(n+1-i)\big]\Sigma_{i-1},$$

for $2 \le i \le j$, and

$$\big[\phi(j) - \psi(n+1-j) + \phi(n+1-j) - \psi(j)\big]\Sigma_j > \big[\phi(j) - \psi(n+1-j)\big]\Sigma_1.$$

By (4) and $\Sigma_j \le 0$, the last inequality implies $\Sigma_j > \Sigma_1$. On the other hand, (6) implies

$$(7) \qquad \Sigma_j \le \left[\prod_{i=2}^{j} \frac{\Phi(i)}{\Phi(n+1-i)}\right]\Sigma_1 = \left[\prod_{i=2}^{\min\{j,\,n-j\}} \frac{\Phi(i)}{\Phi(n+1-i)}\right]\Sigma_1,$$

where $\Phi(i) = \phi(i) + \psi(n+1-i)$ for all i. However, by (3) and (4), $\Phi(i) - \Phi(n+1-i) = \phi(i)/i - \psi(n+1-i)/(n-i) > 0$ when $i \le n - i$. Therefore, (7) implies $\Sigma_j \le \Sigma_1$, a contradiction. ■

Proof of Theorem 5.1. Suppose, first, that $c_1 + c_n \le 2\bar{c}$. By Lemma A.3, for every nondecreasing partition σ to which some individual u-objects, there is a partition $\tau > \sigma$ such that the lowest-index individual in group 2 does not d-object to τ. By Lemma A.2, the highest-index individual in group 2, namely n, also does not d-object to τ. Therefore, by Lemma A.1, no one d-objects to τ. It follows that the greatest nondecreasing partition to which no one d-objects is stable. A similar argument shows that if

$c_1 + c_n \geq 2\bar{c}$ then the smallest nondecreasing partition to which no one u-objects is stable. ■

It remains to prove the two lemmas in Section 7.

Proof of Lemma 7.2. For $1 < k < m$, say that a partition σ is *k-segregating* if there is a subset K' of K with k elements such that (i) $c_i \leq (\max_{l \in K'} C_l + \min_{l \notin K'} C_l)/2$ for every individual i who is a member of one of the groups in K' and (ii) if all the groups in K' were fused into a single group, then σ would become a segregating partition. Here, as elsewhere, C_l denotes the average characteristic in group l when the partition is σ. Extending the above definition, we will consider *all* partitions to be m-segregating. Given a k-segregating partition $\sigma (1 < k \leq m)$, we will show that there is an improvement path—which is a best-reply improvement path if there is no self-effect—with the initial point σ and a terminal point which is segregating or is k'-segregating for some $1 < k' < k$. Note that if $k < m$ then condition (i) implies $\max_{l \in K'} C_l \leq \min_{l \notin K'} C_l$. We may therefore assume (re-indexing groups, if necessary) that $C_1 \leq C_2 \leq \cdots \leq C_m$ and $K' = \{1, 2, \ldots, k\}$. Condition (i) now reads $c_i \leq (C_k + C_{k+1})/2$ for every individual i such that $\sigma_i \leq k$.

Denote $(C_{k-1} + C_k)/2$ by C. Either (a) $c_i > C$ for some individual i such that $\sigma_i < k$; (b) there are no such individuals, but $c_i > C$ for some individual i such that $\sigma_i = k$; or (c) $c_i \leq C$ for every individual i such that $\sigma_i \leq k$. Note that in case (b) we may assume that $C_{k-1} < C_k$. (If this inequality did not hold then the indices of groups k and $k - 1$ could be interchanged, and case (a) would hold.) Suppose, first, that either (a) holds or (b) holds, but not (c). The construction of the improvement path will proceed in three steps.

Step One. Consider the longest path $\sigma(0), \sigma(1), \ldots, \sigma(T)$ such that $\sigma(0) = \sigma$; all moves are to group k; and, for $1 \leq t \leq T$, the mover in $\sigma(t)$ is the highest-index individual i such that $\sigma_i(t-1) < k$ and $c_i > (\max_{l<k} C_l(t-1) + C_k(t-1))/2$. Here, $C_l(t)$ denotes the average characteristic in group l when the partition is $\sigma(t)$. Note that $\max_{l<k} C_l(t) < C_k(t)$ for all $t \geq 1$. It therefore follows from (4) that this is an improvement path. To prove that this is, in fact, a best-reply improvement path if there is no self-effect, it suffices to show that if $k < m$ then, for all $1 \leq t \leq T$, the mover i in $\sigma(t)$ is such that $c_i \leq (C_k(t-1) + C_{k+1})/2$. If $C_k(t-1) \geq C_k$, this follows from condition (i) in the definition of k-segregation. If $C_k(t-1) < C_k$, then the average characteristic of the movers preceding i must be smaller than C_k. All these individuals have a higher index than i, and their characteristics are therefore greater than or equal to c_i. It follows that $c_i < C_k(t-1) < (C_k(t-1) + C_{k+1})/2$, as had to be shown. Note that if (a) holds, then $T \geq 1$, and therefore $\max_{l<k} C_l(T) <$

$C_k(T)$. If (b) holds, then $T = 0$. However, as pointed out in the previous paragraph, without loss of generality it may be assumed that the last inequality holds in this case, too.

Step Two. Denote $(\max_{l<k} C_l(T) + C_k(T))/2$ by $C(T)$. By the maximality of T, $c_i \le C(T)$ for every individual i such that $\sigma_i(T) < k$. Consider the following path, $\sigma(T), \sigma(T+1), \ldots, \sigma(T')$, in which all the individuals i such that $c_i \le C(T)$ and $\sigma_i(T) = k$ leave group k, one after the other, and join one of the groups $1, 2, \ldots, k-1$. (If there are no such individuals, then $T' = T$.) The first mover is the highest-index individual for whom these inequality and equality hold, then the second-highest, and so no. Each of these individuals joins the lowest-index group l such that there is no alternative group $l' < k$ that individual i would prefer to l. Note that the average characteristic of the (remaining) individuals in group k increases after each move. The average characteristic in each of the groups $1, 2, \ldots, k-1$ may increase (to a value which is less than $C(T)$), decrease, or remain the same. However, since higher-index individuals leave first, if the average characteristic in a group $l < k$ decreases as a result of an individual moving there from k, the new average characteristic in l must be higher than the characteristic of each of the subsequent movers. Therefore, for each mover i, $\min_{l<k}|C_l^* - c_i| \le |C_k(T) - c_i| \le \min_{l \ge k}|C_l^* - c_i|$, where C_l^* is the average characteristic in group l just before i moves. It follows, by (4), that the above path is an improvement path, and is moreover a best-reply improvement path if there is no self-effect. For every individual i, $c_i \le C(T)$ if $\sigma_i(T') < k$, and $c_i > C(T)$ if $\sigma_i(T') = k$. Therefore, if $k = 2$ then $\sigma(T')$ is segregating.

Step Three. If $k > 2$, then repeat Step One, starting with the initial partition $\sigma(T')$. That is, consider the longest path $\sigma(T'), \sigma(T'+1), \ldots, \sigma(T'')$ such that, for every $T' + 1 \le t \le T''$, the mover in $\sigma(t)$ joins group k and is the highest-index individual i such that $\sigma_i(t-1) < k$ and $c_i > (\max_{l<k} C_l(t-1) + C_k(t-1))/2$. (If there are no such individuals for $t = T' + 1$, then $T'' = T'$.) This is clearly an improvement path and is moreover a best-reply improvement path if there is no self-effect. Since each mover is the highest-index individual among the members of groups 1 to $k-1$, the average characteristic in each of these groups either decreases after each move or remains the same. The average characteristic in group k (where initially all individuals have a characteristic greater than $C(T)$) decreases after each move. Therefore, $c_i > (\max_{l<k} C_l(T'') + C_k(T''))/2$ for every individual i who is a mover in the above path. On the other hand, because of the maximality of T'', $c_i \le (\max_{l<k} C_l(T'') + C_k(T''))/2$ for every individual i such that $\sigma_i(T'') < k$. Hence, $c_i < c_j$ for every i and j such that $\sigma_i(T'') < k = \sigma_j(T'')$. It follows that $\sigma(T'')$ is $(k-1)$-segregating.

Suppose now that (c) holds, i.e., $c_i \le C\ (= (C_{k-1} + C_k)/2)$ for every individual i such that $\sigma_i \le k$. Averaging over all individuals in group k, we get $C_k \le C_{k-1}$. Since the reverse inequality is also assumed to hold, c_i must be equal to C for every individual i such that $\sigma_i = k$ or $\sigma_i = k - 1$. Let $k'\ (< k)$ be the lowest index such that $C_{k'} = C$. If $k' = 1$, then the proof is complete: In this case, σ is segregating. Suppose, then, that $k' > 1$. As before, we proceed in (at most) three steps. *Step One*. Consider the best-reply improvement path $\sigma(0)\ (= \sigma), \sigma(1), \ldots, \sigma(T)$ in which all the individuals i such that $\sigma_i < k'$ and $c_i = C$ move to group k'. If $c_i \le (\max_{l<k'} C_l(T) + C)/2$ for every individual i such that $\sigma_i(T) < k'$, then $\sigma(T)$ is segregating if $k' = 2$ and is $(k' - 1)$-segregating if $k' > 2$. Suppose, then, that the reverse inequality, $c_i > (\max_{l<k'} C_l(T) + C)/2$, holds for the highest-index individual i such that $\sigma_i(T) < k'$. *Step Two*. Let $\sigma(T)$, $\sigma(T + 1)$ be the improvement path of length one (best-reply improvement path if there is no self-effect) in which the sole mover is individual i, who joins group k'. *Step Three*. Let $\sigma(T + 1), \ldots, \sigma(T')$ be the best-reply improvement path in which all the individuals in group k' with characteristic C move to group $k' + 1$. Because $\max_{l \le k'} C_l(T') = c_i$ and $\min_{l>k'} C_l(T') = C\ (> c_i)$, the partition $\sigma(T')$ is k'-segregating.

This completes the construction of the improvement path leading from σ to a segregating partition, or to a k'-segregating partition for some $1 < k' < k$. Obviously, in the latter case, the construction may be iterated. After at most $m - 1$ iterations, a segregating partition is reached. To complete the proof of Lemma 7.2 it remains to show that the total number of moves does not exceed $2(m - 1)(n - m)$.

Step Two of the above construction (both if (a) or (b) hold or if (c) holds) describes an improvement path in which certain individuals leave a certain group. There can be no more than $n - m$ such movers, because there is always at least one member in each group. The total number of moves in the (at most $m - 1$) repetitions of Step Two is therefore no more than $(m - 1)(n - m)$. In Step Three, there is an improvement path in which certain individuals *join* a certain group. The same is true in Step One. However, individuals who, at the end of Step Three, are in the destination group for that step do not move in Step One of the following iteration. Therefore, the total number of moves in these two successive steps does not exceed $n - m$. If the initial partition in the last iteration is 2-segregating, then there is no Step Three. We therefore conclude that the total number of moves in all the repetitions of Steps One and Three combined does not exceed $(m - 1)(n - m)$. ■

Proof of Lemma 7.3. Let σ be a segregating partition. Without loss of generality, it may be assumed that σ is nondecreasing.

Case One. The "average distance" model, or the "distance from the average" model without self-effect. Suppose that some individual i d-objects to σ. It follows from Lemma 2.5 that, without loss of generality, i may be assumed to be the lowest-index individual in his group. The first step in our best-reply improvement path will be moving i to group $\sigma_i - 1$. We may then continue in this manner until, after no more than $(m-1)(n-m)$ moves, a nondecreasing partition σ' to which none of the individuals d-objects is reached. (The total number of moves is equal to $\sum_i(\sigma_i - \sigma_i')$. Since, in both σ and σ', each group has at least one member, the difference between $\sum_i \sigma_i$ and $\sum_i \sigma_i'$ cannot exceed $(n-m)(m-1)$.) In the proof of Theorem 2.4 we showed that either no one u-objects to σ'—in which case this partition is stable—or else there is an individual i who u-objects to σ' and, by moving to group $\sigma_i' + 1$, transforms σ' into another nondecreasing partition, $\tau > \sigma'$, to which none of the individuals d-objects. The best-reply improvement path generated by the repeated use of this result must be finite. A stable partition must, in fact, be reached after no more than $(m-1)(n-m)$ moves. Since this partition is nondecreasing, it must be segregating (see the remarks preceding Theorem 2.4).

Case Two. The "distance from the average" model with self-effect and $m = 2$. Suppose that $c_1 + c_n \leq 2\bar{c}$ (the proof for the case in which $c_1 + c_n \geq 2\bar{c}$ is similar). There exists an improvement path comprising only nondecreasing partitions whose initial point is σ, whose length is at most $n-2$, and whose terminal point σ' is such that the highest-index individual in group 1 does not u-object to σ' and the lowest-index individual in group 2 does not d-object to σ'. The length of this improvement path is *exactly* $n-2$ if and only if $\sigma = \sigma^{(1)}$ and $\sigma' = \sigma^{(n-1)}$, or vice versa. In this case, it follows from Lemma A.3 that individual 1 does not u-object to σ'. (If he did so, then $\sigma^{(1)}, \sigma^{(2)}, \ldots, \sigma^{(n-1)}$ would not be an improvement path, because some individuals $1 < i < n$ would have u-objected to the nondecreasing partition $\sigma^{(i)}$ in which he is the highest-index individual in group 1.) This, by Lemmas A.1 and A.2, implies that σ' is stable, and the proof is complete. Assume, then, that individual 1 *does* u-object to σ', and that the length of the above improvement path is therefore no more than $n-3$.

Consider the partition τ resulting from moving individual 1 to group 2. In τ, the average characteristic in group 2 is lower than the average characteristic in group 1. (Otherwise, individual 1 would not have wanted to move there.) But, since σ' was nondecreasing, the characteristic of every individual other than 1 in group 2 is at least as high as the average characteristic in group 1, which implies that all such individuals are better off moving to group 1. Furthermore, if the lowest-index individual moves first, then the second-lowest individual, and so on, then when the time

comes for each of these individuals to leave group 2, he can still make himself better off by moving to group 1. In the end, a partition τ' is reached in which only individual 1 is in group 2. To conform to the notation of previous results, it is convenient at this point to interchange the indices of the two groups, so that τ' becomes a nondecreasing partition.

Consider the longest improvement path with the initial point τ' in which each mover is the current lowest-index individual in group 2. We claim that the terminal point τ'' of this improvement path is stable. By Lemmas A.1 and A.2 and the maximality of the length of the path, no one in group 2 d-objects to τ''. If there were an individual in group 1 who u-objected to τ'' then, by Lemma A.3, there would be an individual $1 < i < n$ in that group that did not d-object to the nondecreasing partition in which he is the lowest-index individual in group 2. But this would contradict the definition of the above improvement path. This contradiction proves our claim. It also proves that not all of the individuals who were in individual 1's group when the partition was σ' are in his group when the partition is τ''. Therefore, the length of the improvement path with the initial point σ' and the terminal point τ'' is less than n. Hence, the total number of moves leading from σ to τ'' is no more than $((n-3)+(n-1)=)$ $2n-4$. ■

REFERENCES

Aumann, R. J. (1959). "Acceptable Points in General Cooperative n-Person Games," in *Contributions to the Theory of Games*, *IV* (A. W. Tucker and R. D. Luce, Eds.), pp. 287–324. Princeton, NJ: Princeton Univ. Press.

Banerjee, S., Konishi, H., and Sönmez, T. (2001). "Core in a Simple Coalition Formation Game," *Soc. Choice Welfare* **18**, 135–153.

Bernheim, B. D., Peleg, B., and Whinston, M. (1987). "Coalition-Proof Nash Equilibria. I. Concepts," *J. Econ. Theory* **42**, 1–12.

Bewley, T. (1981). "A Critique of Tiebout's Theory of Local Public Expenditures," *Econometrica* **49** 713–740.

Bogomolnaia, A., and Jackson, M. O. (1998). "The Stability of Hedonic Coalition Structures," mimeo.

Ellickson, B., Grodal, B., Scotchmer, S., and Zame, W. R. (1999). "Clubs and the Market," *Econometrica* **67**, 1185–1217.

Greenberg, J., and Weber, S. (1993). "Stable Coalition Structures with Unidimensional Set of Alternatives," *J. Econ. Theory* **60**, 62–82.

Jehiel, P., and Scotchmer, S. (1997). "Free Mobility and the Optimal Number of Jurisdictions," *Ann. Econ. Statist.* 219–231.

Kaneko, M., and Kimura, T. (1992). "Conventions, Social Prejudices and Discrimination: A Festival Game with Merrymakers," *Games Econ. Behavior* **4**, 511–527.

Kaneko, M., and Wooders, M. (1982). "Cores of Partitioning Games," *Math. Soc. Sci.* **3**, 313–327.

Karni, E., and Schmeidler, D. (1990). "Fixed Preferences and Changing Tastes," *Amer. Econ. Rev.* **80**, 262–267.

Konishi, H., Le Breton, M., and Weber, S. (1998). "Equilibrium in a Finite Local Public Goods Economy," *J. Econ. Theory* **79**, 224–244.

Le Breton, M., Owen, G., and Weber, S. (1992). "Strongly Balanced Cooperative Games," *Internat. J. Game Theory* **20**, 419–427.

Milchtaich, I. (1996). "Congestion Games with Individual-Specific Payoff Functions," *Games Econ. Behavior* **13**, 111–124.

Milchtaich, I. (1998). "Crowding Games are Sequentially Solvable," *Internat. J. Game Theory* **27**, 501–509.

Monderer, D., and Shapley, L. S. (1996). "Potential Games," *Games Econ. Behavior* **14**, 124–143.

Schelling, T. C. (1978). *Micromotives and Macrobehavior*. New York: Norton.

Tiebout, C. (1956). "A Pure Theory of Local Public Expenditure," *J. Polit. Econ.* **64**, 416–424.

Westhoff, F. (1977). "Existence of Equilibria in Economies with a Local Public Good," *J. Econ. Theory* **14**, 84–112.

Westhoff, F. (1979). "Policy Inferences from Community Choice Model: A Caution," *J. Urban Econ.* **6**, 535–549.

Wooders, M. H. (1980). "The Tiebout Hypothesis: Near Optimality in Local Public Good Economies," *Econometrica* **48**, 1448–1467.

Young, H. P. (1993). "The Evolution of Conventions," *Econometrica* **61**, 57–84.