Journal of Mathematical Sociology, 1993, Vol. 17(4), pp. 227–241
Reprints available directly from the publisher
Photocopying permitted by license only

Printed in the United States of America

FINDING GROUPS WITH A SIMPLE GENETIC ALGORITHM*

LINTON C. FREEMAN

IRU in Mathematical Behavioral Science
Social Science Tower
University of California, Irvine
Irvine, CA 92717

May 13, 1992; revised May 13, 1992

A new solution to the old problem of partitioning a matrix of social proximities into groups is proposed. It draws on a heuristic developed in computer science, the simple genetic algorithm. The algorithm is described and its utility is demonstrated with applications to three standard data sets.

1. INTRODUCTION TO THE PROBLEM

Since the late 1800s, sociologists have been concerned with the human tendency to form groups. Of particular interest are groups composed of relatively small collections of individuals who are linked by recurrent interaction (Tönnies, 1887; Durkheim, 1893/1964; Cooley, 1902). Groups of this sort have been, and continue to be, one of the central interests of the field; they are the focus of the present study.

Homans (1950, p. 84) specified the defining property of such groups:

> If we say that individuals $A,B,C,D,E,\ldots$ form a group, this will mean that at least the following circumstances hold. Within a given period of time, A interacts more often with $B,C,D,E,\ldots$ than he does with $M,N,L,O,P,\ldots$ whom we choose to consider outsiders or members of other groups. B also interacts more with $A,C,D,E,\ldots$ than he does with outsiders, and so on for the other members of the group.

More recently, in a discussion of primate ethology, Sailer and Gaulin (1984) specified the implications of this definition for derermining whether some individual is or is not a member of a group:

> On the basis of co-occurrence, monkey m could be said to belong to group G if it spends more time with monkeys that are in group G than with other monkeys that are not.

This suggests that, given information on people's interaction rates or the time they spend together, we should be able to uncover groups. Consider, for example, a collection containing N individuals. Let us assume that we have data in the form

*This paper has benefitted greatly from suggestions made by several colleagues. In particular, the sage counsel provided by Morry Sunshine has helped to eliminate many of its earlier inadequacies. The program described here, GROUPS, is available from the author on request.

of a symmetric $N \times N$, person by person, matrix P. The entry in any cell of this matrix p_{ij} is some index of the social proximity of a pair of individuals i and j. That proximity is based on those individual's interaction rates, the amount of time they spend in one another's company, their degree of intimacy, or some other record of the strength of their social tie.

Given such a proximity matrix, a natural approach would be to examine all the possible partitions of the N individuals and determine which, if any, of those partitions yield groups in the sense that they were defined by Homans and Sailer and Gaulin. Sailer and Gaulin (1984) showed that it is simple to determine whether any specified partition of a proximity matrix produces acceptable groups. It is required only that we examine whether any individual assigned to a group by the partition in question interacts less with fellow group members than with outsiders. Thus, if we were able to enumerate all the possible partitions among the collection of individuals we could determine which, if any, of those produces groups that meet this condition.

The only trouble with this approach is that, as the number of individuals in the collection is increased, the number of partitions grows at an unacceptably rapid rate; it grows exponentially. This means that with a community of any size at all we simply cannot do the computations in any reasonable amount of time.

Given this restriction on computation, investigators have sought other ways to uncover group-like structures from proximity matrices. For more than 40 years, sociologists have tried to find some alternative computational procedure that would reveal clusters of individuals who were linked by regular interaction. A variety of different procedures have been tried. They include matrix diagonalization (Beum and Brundage, 1950; Coleman and MacRae, 1960), various forms of clustering (Bock and Husain, 1950; Breiger, Boorman and Arabie, 1975), analysis of characteristic roots (MacRae, 1960; Hubbell, 1965; Beaton, 1966; Bonacich, 1972a, 1972b; Richards, 1975; Bonacich and Domhoff, 1981; Weller and Romney, 1990), hierarchical clique analysis (Doreian, 1969; Peay, 1974), multidimensional scaling (Laumann and Pappi, 1973) and Ford-Fulkerson flows (Zachary, 1977).

To a greater or lesser extent, each of these procedures does succeed in uncovering collections of more or less proximate individuals, but since none of them is designed to uncover exactly the kind of patterning specified in the Homans-Sailer-Gaulin definition, we can never be certain that the clusters they reveal reflect the actual group structure in the data.

As an alternative, we might begin by explicitly using the Homans-Sailer-Gaulin definition and trying to uncover groups, not by enumeration, but by using some search strategy. That way we could explore the huge set of possible partitions without looking at every one of them. All we would require is a reasonably good chance of finding those partitions that yield groups. The problem is precisely that of "finding a needle in the haystack." We need to do something short of picking up and examining every single piece of hay, but at the same time, we want to be fairly sure we are going to find the hidden needle.

One established approach to problems of this sort is embodied in a kind of computer program called a genetic algorithm. I will outline this approach in the next section.

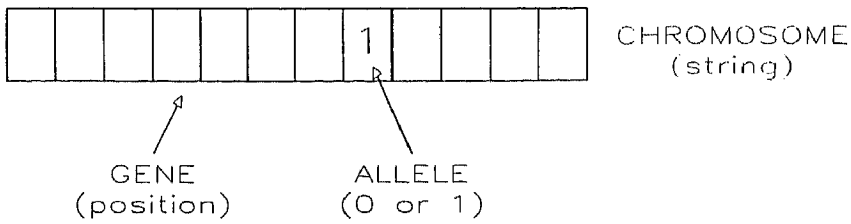


FIGURE 1. Basic structure of a pseudo-organism.

2. SIMPLE GENETIC ALGORITHMS

Holland (1962) developed an imaginative way to search for things that are difficult to find. He reasoned that the natural process of genetic evolution could be viewed as a prototype of this kind of search process. The mechanics of variation and natural selection enable plants and animals to evolve and adapt successfully to a wide variety of ecological niches. In the process of evolution, a population of living beings may be viewed as "searching" for a structural form that will enhance its fitness in a particular context.

Holland proposed, then, that a simulation of the process of genetic evolution might provide a general model of a procedure for such search. Here I will restrict the discussion to the most elementary simple genetic algorithms that stem from this work. My treatment draws heavily on the approach described in a recent book by Goldberg (1989).

A simple genetic algorithm begins with a *population* of pseudo-organisms. At a minimum it includes (1) a process of *reproduction*, (2) a mechanism for genetic *crossover* and (3) some form of *mutation*. These three mechanisms result in the production of new *generations*—new populations of pseudo-organisms.

Each pesudo-organism is represented in the algorithm by a *string* (or chromosome). Each string is a vector containing a number of *positions* (representing genes). Each position in a string is assigned one of a collection of *values* (corresponding to alleles). And each value is capable of being *decoded* into some meaningful *structure* (in effect, a phenotype). These forms are illustrated in Figure 1.

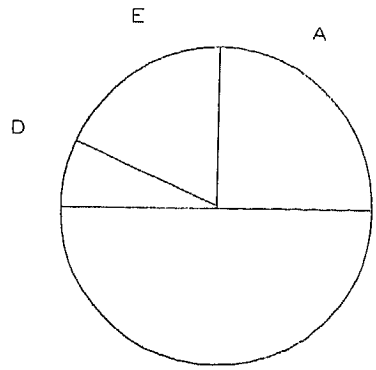
The population evolves through time as a stochastic process. This process is not strictly random; it is directed through the definition of a *fitness function*. The fitness function specifies a numerical value that indicates the extent to which each phenotype approaches some criterion (maximum or minimum) defined by the goals of the search.

The process begins with a random assignment of values to each position in each string in the initial population. At the start, then all the pseudo-organisms are constructed at random.

In the next step, the fitness of each string is calculated. This amounts to evaluating the degree to which each pseudo-organism is adapted to the demands of the ecological niche in which they are located.

Reproduction, then, is biased by these fitness values. In effect, reproduction is governed by spinning a biased roulette wheel as shown in Figure 2. Thus, each pseudo-organism in a given generation is allowed to reproduce proportional to its

Individual	Fitness	Proportion
A	70	25
B	140	.50
C	0	.00
D	35	.125
E	35	.125



Biased Roulette Wheel

FIGURE 2. Reproduction biased according to fitness.

fitness. The reproductive advantages of the individuals of each generation who are most fit raises the average fitness of the population over succeeding generations.

Simple growth of average fitness, however, is not enough. We also need some mechanism that will guarantee that novel forms are introduced. This novelty is introduced by two mechanisms: crossover and mutation.

The pseudo-organisms in the new generation are randomly paired for crossover. For each pair, a crossover site in their strings is chosen, also at random. Each member of the pair retains its own genetic pattern in all positions up to the crossover site, and the two exchange genetic patterns in all positions following the crossover site. This exchange process is illustrated in Figure 3.

And finally, mutation is introduced by changing the value in each position in each string with some small probability. This resists the tendency of the process to get locked into some local minimum or maximum.

The process then returns to the evaluation of fitness—this time of the members of the new generation. It continues to move through these steps—generation after generation—always maintaining a record of the structure of the "best" pseudo-organism—that with the highest fitness value. The process can be ended at any time, and it reveals, as output, the "best" form it was able to locate in its search.

This is a very general search heuristic. It has been used widely in research in artificial intelligence and applied in biology, engineering, psychology, anthropology

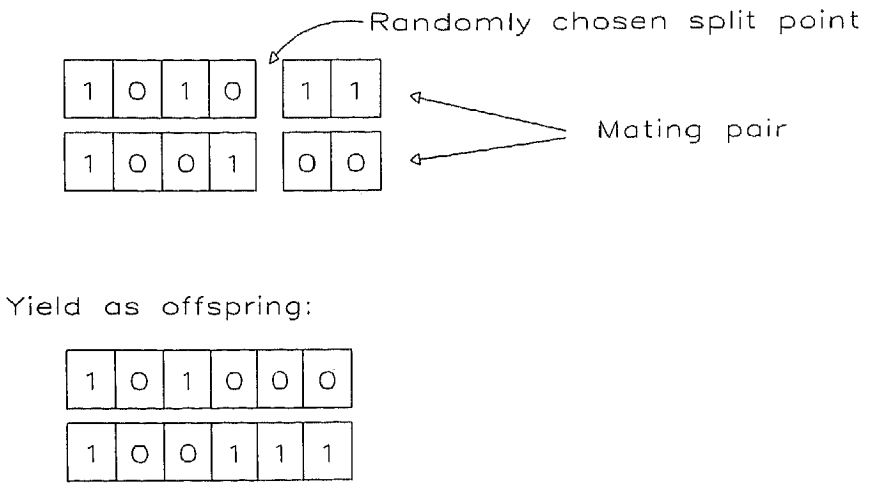


FIGURE 3. Inheritance.

and political science (see Goldberg, 1989 for a recent bibliography). In the next section I will describe how it can be adapted to the sociological problem of the search for human groups.

3. FINDING GROUPS WITH THE SIMPLE GENETIC ALGORITHM

We begin by defining a population containing P pseudo-organisms or strings. Each string contains N positions, where N is the number of individuals under observation. Thus, each of the N positions in each string represents one human individual $\{i, j, \ldots\}$.

Each of these individuals is assigned a binary value (0 or 1). Thus, the individuals in the community are partitioned into two subsets. The patterning of zeros and ones is decoded by pairing each of the N individuals with each of the $N-1$ others. Any given pair may consist of two individuals both of which have the same value (both 0 or both 1) in which case the pair $(i, j) \in S$, or it may consist of two individuals with *different* values (one 0 and one 1) and the pair $(i, j) \in D$. Thus, the individuals in the community are partitioned into two subsets.

The question, then, is to determine the extent to which the pairings—same versus different—correspond to the entries in the social proximity matrix. Each pair of individuals (i, j) in the string is associated with a social proximity p_{ij} in the proximity matrix. Each of these proximities, therefore, can be divided into those in which the i, j pair, is in S and those in which it is in D. Thus, each proximity is classed as $p_{ij}(S)$ if $(i, j) \in S$ and as $p_{ij}(D)$ if $(i, j) \in D$.

From the perspective of the Homans-Sailer-Gaulin definition set down above, any string is fit if and only if it classifies all the individuals correctly. Individuals are classified correctly when the 0's and 1's are assigned in such a way that each individual's interaction with his or her fellow group members is not less than his or her interaction with outsiders. Thus, any given individual i is mis-classified if and

only if

$$\sum_{j \neq i} p_{ij}(S) < \sum_{j \neq i} p_{ij}(D).$$

We could get some sense of the "fitness" or adequacy of a given string simply by counting the number of individuals M it mis-classifies. Such an index, however, turns out to be too crude to be useful here. A more sensitive index would reflect the extent to which pairs of individuals that are in S interact more on average than pairs that are in D.

Each string contains $(N^2 - N)/2$ unordered pairs of individuals. Of these, there are, let us say, $N(S)$ in S and $N(D)$ in D. A string captures the spirit of the notion of group to the degree that

$$\sum_{i \neq j}\sum p_{ij}(S)/N(S) > \sum_{i \neq j}\sum p_{ij}(D)/N(D),$$

that the average proximity of the individuals within each group is large and, at the same time, that the average proximity of the individuals who are in different groups is small.

A fitness measure that is sensitive both to these average proximities and to the misclassification of individuals is

$$F = \frac{\sum_{i \neq j}\sum Pij(S)/N(S)}{\left[\sum_{i \neq j}\sum P_{ij}(D)/N(D)\right](M+1)}.$$

F grows as the average interaction within groups grows, and it shrinks as the average interaction across the group boundaries grows. It shrinks, moreover, as increasing numbers of individuals are mis-classified. When no individuals are mis-classified, it is simply the ratio of in-group interaction to cross-group interaction.

In the present application the whole process was started by defining a population of $P = N$ strings and randomly assigning a 0 or a 1 to each of the N positions in each of those strings. The fitness of each string was determined and each string was permitted to reproduce in proportion to its fitness. Strings were randomly paired and each pair was crossed-over at a randomly selected point. Then mutation occurred. The mutation probability was set at .01 when the algorithm was started.

Fitnesses were calculated at each generation, and if a string with a greater fitness than any previously encountered was produced, its pattern was recorded. If no improvement occurred in 25 generations or any multiple of 25 generations, the mutation probability was momentarily adjusted to .1 for that generation and then returned to .01. If no improvement occurred in 100 generations the program was given a new random start. Finally, when a string was found that produced *no* misclassified individuals, the program stopped and the results were recorded.

In the next section I will show the results of running this version of the simple genetic algorithm to find groups.

4. FINDING GROUPS

Given that this program is based directly on the Homans-Sailer-Gaulin model of group structure and stops only when a group is found, any partitions it produces are *certainly* groups. The problem, then, is to determine whether it can find any acceptable partitions at all and whether those it does find are consistent with the results of ethnographic observation.

Consider the hypothetical two-group data shown in Table 1. These data represent an idealized two-group structure. Individuals A through D are one group with A and B as core members, and C and D as relatively peripheral. Similarly, E through H are another group in which E and F are at the core. There is, moreover, a background level of interaction linking members of the two groups.

When these data were entered into the program, and the program was run repeatedly (500 times), it uniformly produced the A, B, C and D versus E, F, G and H partition. It did so, moreover, quite quickly; it required between 1 and 78 generations and the average was 15.51.

In each program run, the average fitness varied from generation to generation, but it showed a long-range, secular, pattern of growth. A typical run is shown in Figure 4. At generation 11 fitness leapt to a local maximum, then broke out and returned to a lower level. After that it climbed again, and varied up and down, until it found a solution at trial 23.

This result is, of course, not surprising; the algorithm seeks two groups and two groups were present in the data. But what if the data represent three or more groups?

When the three-group data in Table 2 were entered and run 500 times, exactly three partitions were uncovered. On 236 of the trials the program split between A, B, C, D, E, F, G, H and I, J, K, L. On 131 of the trials it split between A, B, C, D and E, F, G, H, I, J, K, L. And on the remaining 133 trials it divided E, F, G, H from A, B, C, D, I, J, K, L. Thus, each of the three groups in the data table was displayed as a minimal, or irreducible group. The frequency with which each group was uncovered was proportional to its ratio of in-group to cross-group proximities.

Suppose, however, there are not only distinct groups in the data, but there are "drifers" or "bridges"—individuals that belong to more than one group and therefore serve as links between groups?

Consider, for example the data of Table 3. There, individuals A, B, C and D are a group, as are I, J, K and L. But individuals E, F, G and H participate equally with members of both groups.

For the data of Table 3, the algorithm produced only two partitions. On 332 of 500 trials the result showed one irreducible group, A, B, C and D, versus all others, and on the other 168 trials the partition was between the second irreducible group, I, J, K and L, and all the others. So the algorithm found the two irreducible groups and it assigned the bridging individuals first to one of them and then to the other. It is capable, then, of distinguishing between the multiple groups shown in Table 2 and the bridging individuals shown in Table 3.

On the face of it, then, the simple genetic algorithm seems to produce results that are consistent with standard sociological intuitions (Freeman, 1992). But the real

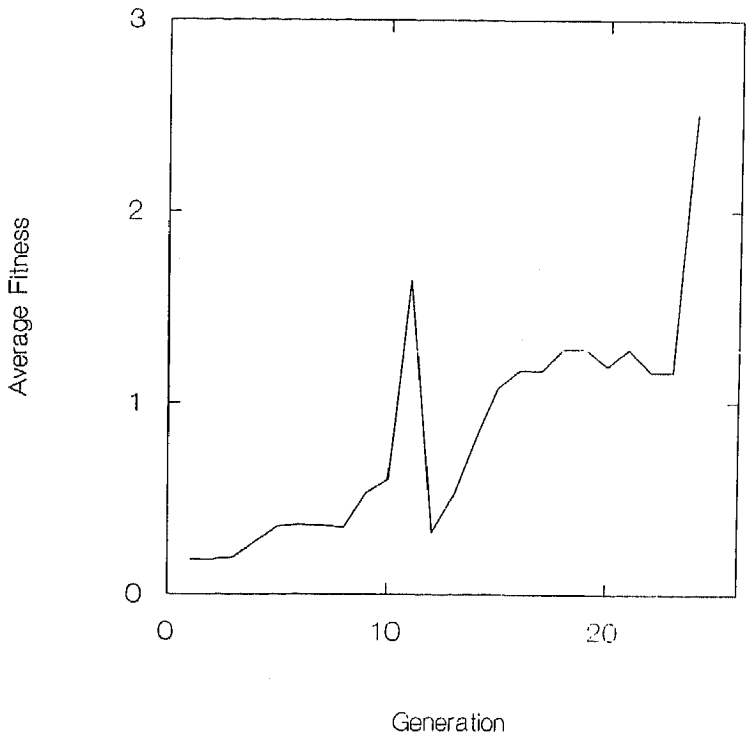


FIGURE 4. Typical growth of average fitness.

TABLE 1
Hypothetical Two-Group Proximities

1	9	8	7	1	1	1	1
9	1	8	7	1	1	1	1
8	8	1	7	1	1	1	1
7	7	7	1	1	1	1	1
1	1	1	1	1	7	6	5
1	1	1	1	7	1	6	5
1	1	1	1	6	6	1	5
1	1	1	1	5	5	5	1

TABLE 2
Hypothetical Three-Group Proximities

1	9	8	7	2	2	2	2	1	1	1	1
9	1	8	7	2	2	2	2	1	1	1	1
8	8	1	7	2	2	2	2	1	1	1	1
7	7	7	1	2	2	2	2	1	1	1	1
2	2	2	2	1	8	7	5	1	1	1	1
2	2	2	2	8	1	7	5	1	1	1	1
2	2	2	2	7	7	1	5	1	1	1	1
2	2	2	2	5	5	5	1	1	1	1	1
1	1	1	1	1	1	1	1	1	7	6	5
1	1	1	1	1	1	1	1	7	1	6	5
1	1	1	1	1	1	1	1	6	6	1	5
1	1	1	1	1	1	1	1	5	5	5	1

TABLE 3
Hypothetical Two-Group Proximities with Co-Memberships

1	9	8	7	2	2	2	2	1	1	1	1
9	1	8	7	2	2	2	2	1	1	1	1
8	8	1	7	2	2	2	2	1	1	1	1
7	7	7	1	2	2	2	2	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	1	2	2	2	2	1	7	6	5
1	1	1	1	2	2	2	2	7	1	6	5
1	1	1	1	2	2	2	2	6	6	1	5
1	1	1	1	2	2	2	2	5	5	5	1

test, of course, is to confront it with actual data. In the next section, I will try it using three standard sets of social proximity data.

5. FINDING GROUPS IN ACTUAL PROXIMITY DATA

To determine the ability of this simple genetic algorithm to find groups, we require data that include precise records of interaction among the members of some community. Three such data sets were used to explore the potential of the genetic algorithm for dealing with real data.

The first of these was collected in the 1930s by Davis, Gardner and Gardner (1941). They drew upon interviews, observations, guest lists, and newspaper accounts to discover the participants in actual social events in a community. In particular, they were concerned with the group structure displayed by a collection of 18 women.

Davis, Gardner and Gardner classified these 18 women into two "cliques." Clique I included women 1 through 8 and clique II included women 10 through 18. Woman 9 they described as a bridge person, belonging to both cliques.

In addition, they reported systematic data on 14 informal social events linking these women. They presented their data in the form of a person by event matrix, where the rows are persons and the columns are social events. But, here, our interest is in *interaction*. If the assumption is made that co-attendance at a small informal social event is a reasonable index of interaction, then it is possible to convert the data into a an 18 by 18 person by person co-attendance matrix (Breiger, 1974).

This interaction matrix was used in the genetic algorithm. The program was run 500 times. In every one of these runs it produced acceptable splits—ones in which no individuals were mis-classified.

All these splits displayed only two patterns. One pattern occurred 327 times. It produced an overall average fitness of 2.199. The other was displayed 173 times. Its average fitness was 2.103. These results are shown in Figure 5.

Figure 5 shows that the partition produced by the split that occurred most often and displayed the greatest ratio of in-group to cross-group interaction, assigned women 1 through 9 to one group, and women 10 through 18 to the other. The second pattern shifted women 8 and 9 from the first group to the second. Thus, in

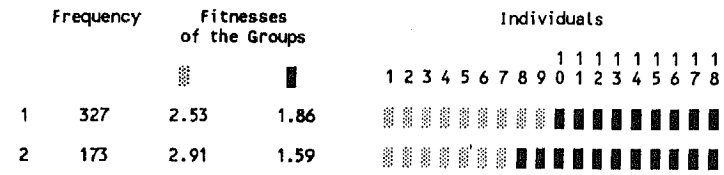


FIGURE 5. Groups in the Davis, Gardner and Gardner data.

that second pattern, the two groups consist of women 1 through 7 versus women 8 through 18. Overall, then, the data display two irreducible groups, women 1 through 7 and women 10 through 18. Women 8 and 9 bridge between these irreducible groups. It should be noted also that whichever of the irreducible groups gets 8 and 9 suffers a loss of fitness by that association.

These results depart only in one detail from those originally reported by Davis, Gardner and Gardner. In the present classification, all the women but 8 are assigned in a way that corresponds to their classification. They assigned woman 8 to the first group, but the data show that in terms of her recorded interaction patterns she, like woman 9, is a member of both groups.

The second data set comes from a study by Freeman, Freeman and Michaelson (1988) who examined a community of 54 recreational users of a southern California beach. They observed a good deal of informal social activity linking these people and cited ethnographic evidence that they were organized mainly into two groups. One group contained 29 individuals (1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 16, 17, 18, 19, 22, 28, 36, 37, 38, 39, 40, 46, 47, 48, 50, 51, 53 and 54). The other main group contained 20 members (9, 10, 14, 15, 20, 21, 24, 25, 26, 27, 29, 31, 32, 33, 41, 42, 43, 44, 49 and 52). Three individuals (23, 30 and 45) appeared to bridge the two groups, and the remaining pair (34 and 35) were seen simply an isolated pair.

Data on interaction were collected by systematic observation at the beach over a period of 31 days. For at least two half-hour periods each day, records of how long each individual interacted with any other individual were made. These data, then, are organized into a 54 by 54 matrix. Cell entries are the numbers of minutes of interaction that was observed between each pair of individuals over the 31 days.

Here again, these data were used to produce 500 runs of the program. The results are a good deal more complicated than those found in the Davis, Gardner and Gardner data; they produced the 25 patterns shown in Figure 6.

As the first pattern in the table shows, individuals 34 and 35 were split off as a distinct group. That agrees with the ethnographic report that suggested that individuals 34 and 35 formed a separate pair.

In the second pattern, the broad outlines of the two main groups that were described in the ethnographic report begin to emerge. The couple (34 and 35) are assigned to the first group and all three bridge persons (23, 30 and 45) are assigned to the second.

The third pattern differs from the second only by shifting one bridge person (30) from the second group to the first. And in the fourth pattern, person 30 is returned and the other bridge persons (23 and 45) are shifted from the second group to the first.

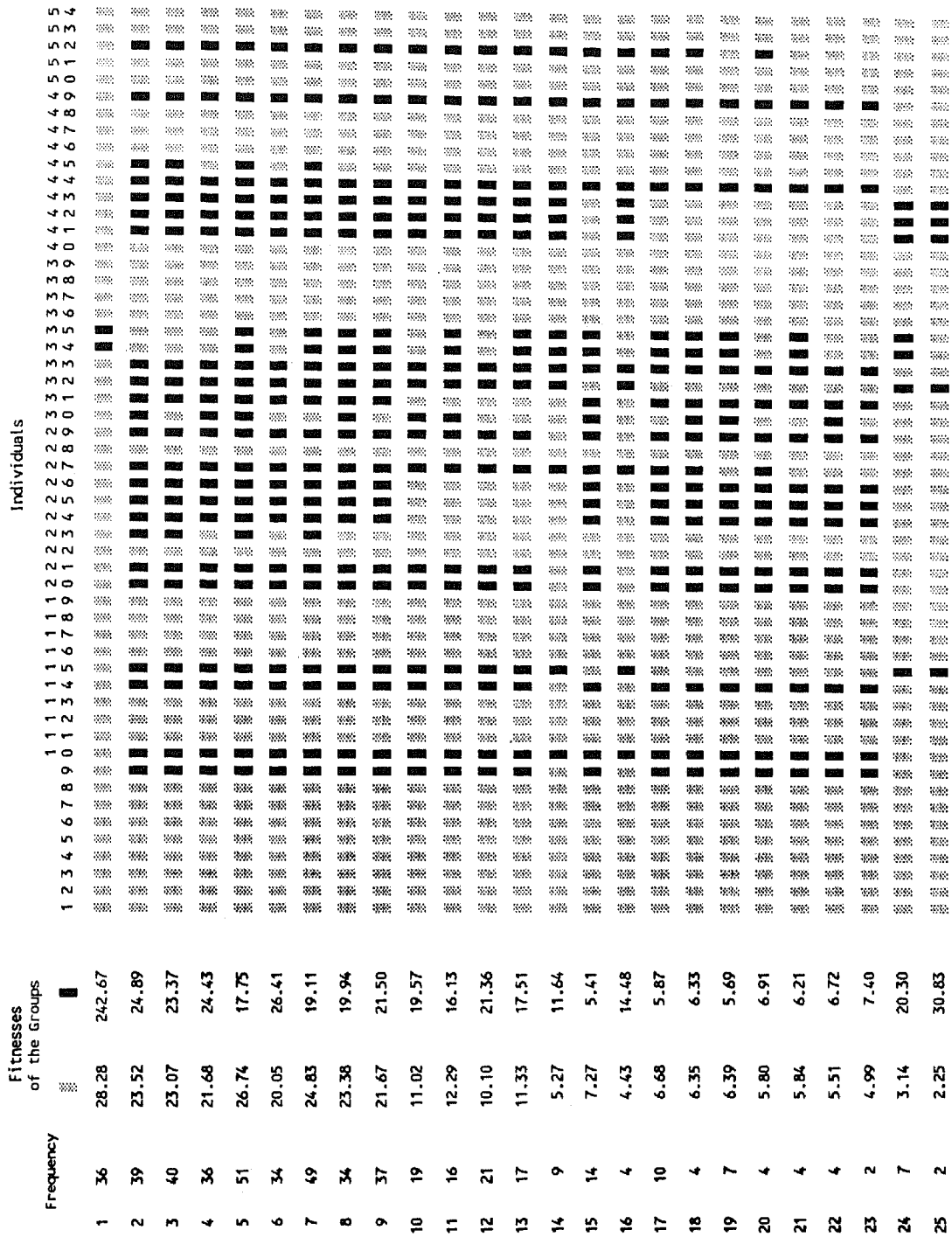


FIGURE 6. Groups in the Freeman, Freeman and Michaelson data.

The fifth pattern specifies one of the main groups as it was described in the ethnography. It places all three bridge persons back in the second group, and at the same time, it assigns the pair (34 and 35 isolated in pattern 1) to that same group.

The sixth pattern specifies the other main group as it was described in the ethnography. It moves the 34, 35 pair and the three bridges to the group uncovered in the fifth pattern.

The seventh pattern shifts individuals 23, 34, 35 and 45. The eighth pattern shifts 23 and 45 again, and the ninth shifts 30 and 45.

Thus, in these nine patterns, we see a set of five individuals who, in various combinations, are being bounced back and forth between the two large groups. We can see the distinction between a shifting an irreducible *group*, like 34 and 35, and *bridges*, like 23, 30 and 45. The group stood alone (in the first pattern) while the bridges drift back and fourth, but never emerge on their own as groups.

As we proceed on down the table, we see a continuation of this kind of switching. In the tenth through the fourteenth and the sixteenth patterns, individuals 24, 25, 26 and 31 are moved as a unit from one group to the other. They are accompanied by various combinations of the points that have been shifting all along (23, 30, 45 and 34 and 35). In addition, in the fourteenth and sixteenth patterns they are accompanied also by individuals 9, 14 and 21.

In the fifteenth and the seventeenth through the twenty-third patterns, individuals 15, 32, 41, 42 and 43 are shifted in this same way from one group to the other. They are accompanied occasionally by individuals 27 and 52 as well as by the 34–35 pair and the three bridging individuals. Finally, in the last two patterns, these individuals, 15, 32, 41, 42 and 43, emerge as a distinct group.

All in all, then, the beach data reveal at least three individuals who are bridges (23, 30 and 45) and three irreducible groups. The strongest group—that with the greatest ratio of in-group to cross-group interaction—is the 34–35 pair. The next strongest is the group consisting of individuals 15, 32, 41, 42 and 43, shown in the twenty-fifth pattern. And the third strongest group was revealed in the fifth pattern. It contains individuals 1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 16, 17, 18, 19, 22, 28, 36, 37, 38, 39, 40, 46, 47, 48, 50, 51, 53 and 54.

This result is, in most details, close to the ethnographic description. There is agreement with respect to the bridging individuals and to the first and the third of the irreducible groups. But the second irreducible group (consisting of individuals 15, 32, 41, 42 and 43) is simply a subset of a larger group described in the ethnography.

As a matter of fact, one of the groups displayed in the sixth pattern in Table 5 agrees exactly with the ethnographic description. It defines a group consisting of individuals 9, 10, 14, 15, 20, 21, 24, 25, 26, 27, 29, 31, 32, 33, 41, 42, 43, 44, 49 and 52. However, this group is not irreducible. Individuals 24, 25, 26 and 31 moved as a unit to the largest group. The same was true of 9, 14 and 21, of 27 and 52, and of 15, 32, 41, 42 and 43. Moreover, these last five individuals are, themselves, a group. Nevertheless, it should be noted that the group defined in the sixth pattern has a greater fitness than any of its subgroups except the one made up of 15, 32, 41, 42 and 43.

All this indicates that this set of individuals—that the ethnographers saw as a group—are organized into a more complex structural form. The data suggest that these individuals form a loose affiliation, made up of various subgroups. Some of these subgroups act as bridges, but they do so as subgroups, not as individuals. Among them, only the one consisting of 15, 32, 41, 42 and 43 is cohesive enough to constitute an irreducible group. Structurally, this whole collection displays greater differentiation than the ethnographic report suggested. Thus, the ethnographic report in this case is not contradicted by the data, but it does seem to oversimplify the structure of the data.

The final data set was collected by Zachary (1977) in a study of 34 members of a university-based karate club. Activities included everything from formal training in karate to social events sponsored by the club. Like other campus clubs, the karate club had officers, but its meetings were informal and decisions were made by consensus. In addition, because of its focus on a martial art, the club employed a part-time instructor.

When the study started, an incipient conflict was brewing between the instructor and the club president. The instructor wanted more pay, and felt he should be able to set fees for lessons. The president felt that his position allowed him to determine the wages of employees, and he wanted to keep costs down.

Over a period of time, club members joined one or the other side of the conflict. Some saw the instructor as their mentor, a person who was only trying to meet his own needs. Others saw him as an employee who was trying to coerce them into paying him more. On the basis of ethnographic evidence, Zachary reported that 16 individuals (1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20 and 22) were in one faction and the remainder were in the other faction.

Zachary also collected systematic data on the interaction among the 34 club members in contexts other than club meetings. Specifically, data were collected in classes, a private non-campus karate club, a university rathskeller, an off-campus bar and at both intercollegiate and open karate tournaments. He reasoned that collections of individuals who interacted in a wide range of contexts would share more information and would be socially closer than those who interacted in a few or none. He constructed a 34 by 34 matrix, therefore, in which the entries were the number of different contexts in which a pair of individuals were observed interacting. This matrix provides a measure of the informal contact linking each pair of individuals.

Again, this matrix was used to produce 500 runs of the program. In this case, the results show 10 different partitionings. Their patterns are shown in Figure 7. They varied in their fitness—in the degree to which in-group interaction predominated over cross-group interaction, but one of them was dramatically better than any of the others. The best pattern, the first, assigned all the individuals exactly as they had been classified by Zachary.

Though it captures Zachary's description, this first pattern does not turn out to be irreducible. The remaining patterns all show the switching associated with bridging, but in these data virtually everyone is switched in one pattern or another. There is no evidence of the kind of highly regular group structure displayed in the other data sets. This may be the result of the fact that the data were collected during a period of conflict during which alliances were made and abandoned as everyone jockeyed

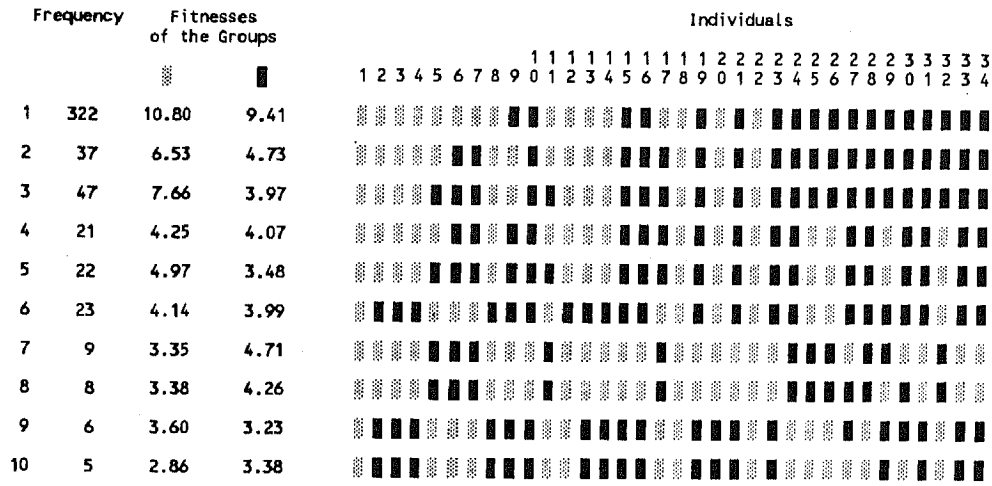


FIGURE 7. Groups in the Zachary data.

for position. In any case, the fact remains that the best pattern displayed by these data corresponds exactly to the ethnographic description.

5. SUMMARY AND CONCLUSIONS

In this paper I have considered a long-term unsolved problem in sociology, how to uncover groups given a matrix of social proximities. I have shown that a simple heuristic based on the genetic algorithm can be used to reveal group structure. As it was used here, the algorithm is not simply a device for uncovering arbitrary collections of individuals who are linked by interaction. Instead, it was adapted to use an explicit theory-based definition of groups introduced by Homans (1950) and again by Sailer and Gaulin (1978).

As programmed here, the algorithm produces bi-partitions. But every bi-partition it produces meets the Homans-Sailer-Gaulin condition; the resulting equivalence sets are groups in exactly the sense they specified. Moreover, it is shown that successive partitionings can uncover multiple groups, subgroups and individuals that bridge across group boundaries.

Overall, the groups uncovered by applying the genetic algorithm correspond in broad outline to those reported in ethnographic accounts. The results produced by the algorithm certainly do not contradict ethnographers' descriptions. Instead, this kind of analysis seems to supplement ethnography; it tends to reveal additional unreported details about group structure.

There is, of course, no guarantee that the repeated use of the heuristic provided by the genetic algorithm will reveal all of the groups contained in a social proximity matrix. It certainly seems to be able to find the groups that are described in ethnographic reports. And, as we discovered here, it can uncover details of group structure that are not apparent from the less systematic procedures of ethnography.

In any case, the results reported here are encouraging enough to warrant further work. Several directions in which such work might proceed are apparent. In the first

place, there is no inherent restriction in the genetic algorithm itself that limits its application to bi-partitions. One potentially important area for future development would involve constructing alternative group models in which the algorithm could directly seek three-group, four-group or general multi-group structures without depending on multiple runs to uncover such structures.

In addition, the genetic algorithm has the potential for a wide range of sociological applications beyond the search for groups. In particular, it could easily be adapted to aid in the search for social actors who occupy equivalent positions, statuses or roles.

REFERENCES

Beaton, A. E. (1966) An inter-battery factor analytic approach to clique analysis. *Sociometry* **29:** 135–145.

Beum, C. O., and Brundage, E. G. (1950) A method for analyzing the sociomatrix. *Sociometry* **13:** 141–145.

Bock, R. D., and Husain, S. D. (1950) An adaptation of Holzinger's *B*-coefficients for the analysis of sociometric data. *Sociometry* **13:** 146–153.

Bonacich, P. (1972a) Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology* **2:** 113–120.

Bonacich, P. (1972b) Technique for analysing overlapping memberships. In *Sociological Methodology 1972*. Herbert L. Costner, ed. Pp. 176–185. San Francisco: Jossey-Bass.

Bonacich, P., and Domhoff, G. W. (1981) Latent classes and group membership. *Social Networks* **3:** 175–196.

Breiger, R. L., Boorman, S. A., and Arabie, P. (1975) An algorithm for clustering relational data, with applications to social network analysis and comparison to multidimensional scaling. *Journal of Mathematical Psychology* **12:** 328–383.

Coleman, J. S., and MacRae, D., Jr., (1960) Electronic processing of sociometric data for groups up to 1, 000 in size. *American Sociological Review* **25:** 722–727.

Cooley, C. H. (1909) *Social Organization*. New York: Charles Scribner.

Davis, A., Gardner, B. B., and Gardner, M. R. (1941) *Deep South*. Chicago: The University of Chicago Press.

Doreian, P. (1969) A note on the detection of cliques in valued graphs. *Sociometry* **32:** 237–242.

Durkheim, E. (1893/1964) *The Division of Labor in Society*. New York: The Free Press.

Freeman, L. C. (1992) The sociological concept of "group": an empirical test of two models. In press, *American Journal of Sociology*.

Freeman, L. C., Freeman, S. C., and Michaelson, A. G. (1988) On human social intelligence. *Journal of Social and Biological Structures* **11:** 415–425.

Goldberg, D. E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison Wesley.

Holland, J. H. (1962) Information processing in adaptive systems. *Information Processing in the Nervous System, Proceedings of the International Union of Physiological Sciences* **3:** 330–339.

Homans, G. C. (1950) *The Human Group*. New York: Harcourt, Brace and Company.

Hubbell, C. H. (1965) An input-output approach to clique identification. *Sociometry* **28:** 377–399.

Laumann, E. O., and Pappi, F. U. (1976) *Networks of Collective Action: A Perspective on Community Influence Systems*. New York: Academic Press.

MacRae, D., Jr. (1960) Direct factor analysis of sociometric data. *Sociometry* **23:** 360–371.

Peay, E. R. (1974) Hierarchical clique structures. *Sociometry* **37:** 54–65.

Richards, W. (1975) *A Manual for Network Analysis (Using the NEGOPY Network Analysis Program). Unpublished ms., Institute for Communication Research, Stanford University.*

Sailer, L. D., and Gaulin, S. J. C. (1984) Proximity, sociality and observation: the definition of social groups. *American Anthropologist* **86:** 91–98.

Tönnies, F. (1887) *Gemeinschaft und Gesellschaft*.

Weller, S. C., and Romney, A. K. (1990) *Metric Scaling: Correspondence Analysis*. Beverly Hills, CA: Sage.

Zachary, W. (1977) An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* **33:** 452–473.