DOCUMENT RESUME

ED 306 004 PS 017 919

AUTHOR Gariepy, Jean-Louis; Kindermann, Thomas
TITLE Quantitative Techniques for the Identification of Social Sub-Groups in Natural Settings.
PUB DATE Apr 89
NOTE 21p.; Paper presented at the Biennial Meeting of the Society for Research in Child Development (Kansas City, MO, April 27-30, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Education; Factor Analysis; Grade 7; *Identification; Junior High School Students; *Network Analysis; *Research Methodology; Research Problems; *Social Networks; *Sociometric Techniques; *Statistical Analysis

ABSTRACT

Addressing a common problem in the analysis of social networks, this study describes quantitative techniques for identifying social subgroups using individual perceptions of social affinities within natural groups. Compared are four analytic methods for abstracting composite representations of sub-structures. These methods, formally evaluated using Confirmatory Factor Analysis (LISREL IX), are illustrated with a class of 20 seventh-graders enrolled in a regular junior high school. The convergence of the different quantitative techniques towards the same structural representation points to the robustness of the social cognition method for the description of social networks. More than robust, the social cognitive procedure is simple and readily applicable in a wide variety of settings. The approach has several additional positive features. It permits analysis to systematically address the questions of cluster sizes and composition, cluster coherence, and dual memberships. Such comparative analyses are especially useful for tracing developmental changes in the nature of prosocial ties within stable peer groups. (RH)

Quantitative Techniques for the Identification of Social Sub-Groups in Natural Settings

Jean-Louis Gariépy

University of North Carolina at Chapel Hill

and

Thomas Kindermann

University of Rochester

Presented at the Biennial Meeting of the Society for Research in Child Development, Kansas City, Missouri, April 1989.

QUANTITATIVE TECHNIQUES FOR THE IDENTIFICATION OF SOCIAL SUBGROUPS IN NATURAL SETTINGS

Jean-Louis Gariépy and Thomas Kindermann, Laboratory of Social Development, Psychology Department, Davie Hall, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27514.

A common problem in the analysis of social networks is to identify naturally occurring affiliative subgroups. This research describes techniques for identifying social subgroups using individual perceptions of social affinities within natural groups. Four analytic methods are compared for abstracting composite representations of sub-structures that constitute common conceptual maps of natural groups. These methods are illustrated with a class of 20 seven-graders enrolled in a regular junior high school. These solutions were formally evaluated using Confirmatory Factor Analysis (LISREL IX). This social cognitive procedure is simple, robust and readily applicable in a wide variety of settings. The quantification methods used for network analysis permitted to systematically address the questions of cluster sizes and composition, cluster coherence as well as dual memberships. Such comparative analyses are especially useful for tracing developmental changes in the nature of prosocial ties within stable peer groups. Finally, comparisons of the present findings with results from more well known psychological applications of sociometric procedures will be discussed.

The role of the social group in shaping individual personality and actions, and as a primary source of influence in development has been repeatedly emphasized in the social sciences. However, as Scott recently observed, in most disciplines, this theoretical emphasis has been confronted with the issue of appropriate methods for social network analysis. Prevailing methods for obtainaing information on social networks have been direct observations or Moreno's "sociometric" procedure. Following this procedure, individuals are requested to nominate their "best friends" and "least liked persons" in their group. In sociology this information was often quantified using graph theory or multidimensional scaling techniques. In psychology and education a "psychometric" solution has evolved which permits the placement of individuals on a standardized dimension of "likeability", "popularity," or "unlikeability."

An alternative procedure for data collection was proposed by Cairns, Perrin, & Cairns in 1985. Their method takes advantage of the finding that children are capable of describing--in free recall--much of the basic information about the social structures of their classrooms. When asked the question: "who hangs around together?", each subject typically generated clusters of persons, and differentiated these clusters from others. When asked the further question, "are there any people who don't have a group?", subjects nominated persons whom they considered to be isolates. The advantage of this procedure is that the information accessed reaches beyond the limited circle of respondent's friendships and makes use of their knowledge of the social network as a whole. Our goal in this presentation is to compare different quantitative techniques for abstracting representations of social networks on the basis of the different cognitive "social maps" generated by this procedure.

We will illustrate our techniques with a data set obtained by Cairns et al. for 20 seven-graders enrolled in a regular junior high-school.

SLIDE 1: RECALL MATRIX

This table summarizes the individual social maps generated by the 17 respondents who returned parental permission. The subjects-to-be-clustered are shown in rows and they are organized in terms of their placement by respondents reported in columns. Here we see that, AMY identified two clusters respectively composed of 10 and 4 individuals. BEA on the other hand, identified three clusters. Note that the 16th subject, PAM, was never assigned to any group. Our analyses focused on the 15 girls but we used the information provided by all female and male respondents. A critical step for the quantification of this information is to summarize it into a second matrix that we call a Co-occurence matrix.

SLIDE 2: THE CO-OCCURENCE MATRIX

The cell entries represent the number of times, across all respondents, that Person *i* was identified to be in the same social group as Person *j*. For example, the first off-diagonal cell indicates that AMY and BEA were assigned to the same group 11 times. Computing these frequencies for every pair of individuals yields a symmetric matrix. The numbers entered on the diagonal represent the total number of times each person was nominated to a group. You can see that as a result of this arrangement 3 fairly separate sub-groups already appear.

Also, if you scan across the columns (or lines) you will notice that certain pairs of individuals like the first two, AMY and BEA, present very similar patterns of co-occurence. By extension, the similarity of person-profiles between every pairs of persons in the group can be determined by computing pearson correlation coefficients. The result is a triangular matrix of correlations that can be used as proximity indices for ordination and clustering techniques.

In practice, the patterns of correlations typically fall into distinct groups, and provide a close initial estimate of the actual clusters. However, two sorts of problems are encountered. First, there is the problem of dual memberships where one person such as Hea in our example may fit into one group or another, or into two groups simultaneously. A second and related problem is the number-of-clusters issue. A quantitative guide to both of these problems is provided in the LISREL VI measurement model. The advantage of this technique over standard factor analysis is that it provides indices of the relative gains in descriptive accuracy between parsimonious models of the network structure and more complex ones. In the present case, a Lambda_X matrix of parameter estimates was constructed with various assumptions about the nature of the network structure.

SLIDE 3: LISREL MODELS

Using as a guide positive correlations that reached the significance level of .05, four different models were generated and compared. In our application, Goodness-of-fit indices and Root Mean Square Residuals suggested that a two cluster model was insufficient to account for the network structure, and that a three cluster model could be significantly improved if individual HEA was assigned dual membership in both the first and the third cluster. This solution is represented by model C. The method proved to be both practical and feasible when we used it for the analysis of a large number of social networks.

Some research applications may require information about the relative proximities that exist among group members both within and between clusters. These proximities can be estimated by means of a Hierarchical Cluster Analysis of the correlation matrix. The next slide presents the results for an average-linkage solution.

SLIDE 4: DENDROGRAM

As you can see, the same three sub-groups were identified with a high degree of separation and high within-group similarities. At a finer level, it is also clear that the groups differed in terms of proximities between their respective members. For instance, compare the cluster of 4 which forms a very tight group to the cluster of 3, at the bottom, where one individual is more loosely attached to the group. Finally, note that individual HEA, who in the previous analysis was assigned dual membership is now included in the largest group. She was, however, the last member to be included.

Unless one uses a Grade Of Membership analysis where "fuzzy sets" can be defined, traditional clustering techniques offer no simple solution to the problem of dual memberships. For this reason, Gower, Rohlf and Legendre, among others, have recommended to conduct Cluster Analysis in conjunction with Principal Coordinate Analysis. This technique, also called Classical Multidimensional Scaling provides a plot in a reduced-dimensional space of the distance relations among the objects to be clustered. When such a plot is obtained it is possible to identify those points that lie between groups of observations and which, in cluster analysis, artificially inflate cluster sizes or precipitate their fusion at higher similarity levels.

SLIDE 5: CLASSICAL MDS

In our application, the first two coordinates explained more than 70% of the variance and justified plotting the association profiles among all subjects in a two-dimensional space. In this solution, some individuals, like the group of four, are represented by a single dot because of highly similar profiles. This analysis reproduced the basic features of the previous technique and further showed that individual Hea actually lies between the largest cluster and the cluster of three. This ordination technique also has its own limitations in that when the variance explained by the first two coordinates tends to be small, a low-dimensional representation may be misleading. In this case, an approximation in two dimensions may be obtained using more general Nonmetric Multidimensional Scaling techniques.

The quantitative techniques presented so far were based on the analysis of similarities between person profiles of relationships. The co-occurrence matrix suggests a second approach which begins with the question: Are certain pairs of individuals more likely to be assigned to clusters together than could be expected by chance?

SLIDE 6: THE CO-OCCURENCE MATRIX AGAIN

Following this approach, a network linkage analysis is essentially a method for determining which ij observed cell frequencies exceed chance expectancy. A simple estimate of the chance expectancy of linkage would be derived from the assumption of equal likelihood of linkage. This may be calculated by dividing the sum of dyadic linkages for each person by the number of persons in the group minus one. But in practice, this estimate must be corrected to account for the differences in nomination frequencies across subjects.

Once the chance and observed probabilities of co-occurrence have been determined, a X^2 is calculated and a figure may be constructed which represents the linkages among persons within the network.

SLIDE 7: NETWORK LINKAGE ANALYSIS

In this figure, two individuals where linked together when the p value for the corresponding X^2 was below .10. Distances between individuals and placement of clusters are arbitrary. One advantage of this method is the information provided on the extent of dyadic ties between group members. In the cluster of four, for example, all possible dyadic linkages were identified, in contrast to the largest cluster where only a subset of such linkages were identified. In this solution, however, the cut-off point for significant linkage did not permit Hea to be linked to any group.

When a dyadic approach is used for network analysis, an appropriate companion solution is Correspondence Analysis which is basically a technique of Factor Analysis for contingency tables. The analysis proceeds in two steps. Like any ordination technique, it is performed on a similarity matrix. Namely, standardized deviations from expected frequencies are calculated using a Chi statistics, which in the present case, provides continuous indices of dyadic distances within the social network. In the second step, a co-variance matrix is computed and its proper values are extracted. A low-dimensional plot is obtained by calculating component scores on the first few coordinates.

SLIDE 8: CORRESPONDENCE ANALYSIS

Again, in this solution, inertia values on each axis permitted an ordination of all subjects on the first two coordinates, and the identification of three spatially separate clusters. The group of four, still represented by a single dot, is located at the origin, while the group of seven and the group of three appear respectively at the two extremes of the second axis. This analysis further indicated that Mia is somewhat peripheral in the group of three, and that Hea, although situated between two groups, is more closely associated to the group of seven.

Discussion

The convergence of these quantitative techniques towards the same structural representation points to the robustness of the social cognitive method for the description of social networks. It must be emphasized that clear solutions do not necessarily require that each group member participates to the interview. Our investigations suggest a lower limit of 8 to 10 respondents for a group of 30 individuals. Perhaps the most surprising feature of the method is that it yields highly replicable structures across independent informants, despite the potential sources of variance for each person.

The validity of the information obtained on social networks through these procedures was confirmed across several domains. For instance, information obtained from direct observation of social interactions, demonstrated that children were more likely to interact with members of their own sub-group than with members of other clusters. Also, using the same techniques for network analysis, Cairns and his co-authors demonstrated that cluster membership was a stronger predictor of early school drop-out than isolated individual characteristics. Along a similar line, Robert Cairns will present tomorrow evidence showing high within-clusters similarities in aggressivity levels as perceived by teachers and peers.

Depending on the goals of the investigator, the quantitative techniques for social network analysis may be used alone or in combination. The several indices provided on group composition, individual proximities and the extent of dyadic connectedness between and within sub-groups, may prove useful for tracking the emergence, organization and stability of friendship in development, and should provide a much needed basis for the evaluation of changes in groups social dynamics.

Table 1: Matrix of Social Clusters Generated by Respondents in 7th Grade

Students Classified		Respondents																
		Girls											Boys					
Gender	Name	Amy	Bea	Cam	Edi	Fay	Gay	Hea	Joy	Lyn	Nia	Ola	Cal	Gig	Hal	Ian	Jan	Ken
Girls	Amy	A_1*	B_2	A_2	A_1	B_2	D_1	A_1	A_3	B_1	—	—	C_2	C_1	D_5	—	—	B_2
	Bea	A_2	B_1*	A_6	A_2	B_5	D_4	A_6	A_4	B_4	A_{11}	—	—	—	D_1	—	—	B_1
	Cam	A_3	B_3	A_1*	A_3	B_4	D_3	A_3	A_2	B_5	A_4	—	C_3	C_4	D_6	—	—	—
	Di	—	B_4	—	A_4	—	D_5	A_5	A_5	B_2	—	—	—	C_3	—	—	A_2	—
	Edi	A_4	B_5	A_5	A_5*	B_3	D_6	A_4	A_1	B_3	A_7	—	C_1	C_2	D_4	—	A_1	—
	Fay	A_7	—	A_4	A_8	B_1*	D_2	A_2	A_6	—	—	—	C_4	C_5	D_3	—	D_1	C_1
	Gay	—	—	A_7	A_6	B_6	D_7*	—	A_7	—	—	—	—	—	—	—	—	C_2
	Hea	A_5	—	A_3	A_7	C_1	A_1	—*	A_8	C_1	—	—	—	—	E_3	—	D_2	—
	Ida	B_1	C_1	B_1	B_2	A_3	—	E_4	B_1	A_1	B_2	C_1	E_1	D_2	C_1	C_3	—	—
	Joy	B_3	C_2	B_3	B_3	A_1	—	E_1	B_4*	A_3	B_3	C_3	E_2	D_3	C_2	C_2	—	—
	Kim	B_2	C_3	B_4	B_1	A_2	—	E_3	B_3	A_2	B_4	C_4	E_4	D_4	C_4	C_4	—	—
	Lyn	B_4	C_4	B_2	B_4	A_4	—	E_2	B_2	A_4*	B_1	C_2	E_3	D_1	C_3	C_1	—	—
	Mia	A_6	—	C_1	—	C_2	A_3	B_2	—	C_2	A_2	—	D_3	—	D_2	—	—	—
	Nia	—	—	C_2	—	C_3	A_4	B_3	—	C_3	A_1*	—*	D_1	—	E_2	—	—	—
	Ola	—	—	C_3	—	C_4	A_2	B_1	—	C_4	A_3	—	D_2	—	E_1	—	—	—
	Pam	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Boys	Arn	A_8	A_1	A_8	C_1	E_2	B_1	C_3	—	—	A_5	A_4	A_3	A_3	A_2	A_1	C_2	A_1
	Bil	A_9	A_2	A_9	C_2	E_5	B_3	C_5	—	—	A_8	A_3	A_4	—	A_3	A_2	—	A_2
	Cal	—	A_3	A_{11}	—	E_1	B_4	C_1	—	—	—	A_1	A_1*	A_2	A_1	A_4	C_1	—
	Dan	A_{10}	A_6	A_{10}	C_3	E_3	B_2	C_4	—	—	A_6	—	A_2	—	—	A_3	—	—
	Edd	—	—	—	—	E_4	—	C_2	—	—	—	A_2	—	—	A_5	—	—	—
	Foz	—	A_4	A_{12}	—	D_5	—	—	—	—	A_{10}	—	A_5	A_1	A_4	—	—	—
	Gig	—	—	D_2	—	D_2	C_1	D_3	—	D_3	—	B_1	B_1	B_1*	B_2	B_2	—	—
	Hal	—	—	D_1	—	D_1	C_4	D_2	—	D_2	—	B_3	B_3	B_3	B_1*	B_3	B_2	—
	Ian	—	—	D_4	—	D_4	C_2	D_4	—	D_1	—	B_4	B_4	B_4	B_4	B_1*	B_1	—
	Jan	—	—	D_3	—	D_3	C_3	D_1	—	—	A_9	B_2	B_2	B_2/A_4	B_3	B_4	—*	—
	Ken	—	A_5	—	—	—	—	—	—	—	—	A_5	—	—	B_5	—	—	—*

Note. A, B, C, D, & E refer to the clusters and to sequence in which the clusters were generated by the respondent, where A was the first cluster, B the second, and on. The subscript numbers refer to the order that the individual name was generated within the cluster, where A_1 refers to first name recalled in first cluster, A_2, the second name in first cluster, and on. The dash (—) indicates that the student's name was not generated (omitted) by the respondent. An asterix (*) refers to self assignment.

Table 2 : Co-Occurrence Matrix

	Amy	Bea	Cam	Di	Edi	Fay	Gay	Hea	Ida	Joy	Kim	Lyn	Mia	Nia	Ola	Row Totals
Amy	13[a]	11[b]	12	7	12	10	5	4	--	--	--	--	2	--	--	63[c]
Bea	11	12	11	6	11	8	5	4	--	--	--	--	3	1	1	61
Cam	12	11	13	7	13	10	5	4	--	--	--	--	3	1	1	67
Di	7	6	7	8	8	5	3	2	--	--	--	--	--	--	--	38
Edi	12	11	13	8	14	10	5	4	--	--	--	--	3	1	1	68
Fay	10	8	10	5	10	12	6	5	--	--	--	--	2	--	--	56
Gay	5	5	5	3	5	6	6	3	--	--	--	--	--	--	--	32
Hea	4	4	4	2	4	5	3	9	--	--	--	--	4	4	4	38
Ida	--	--	--	--	--	--	--	--	14	14	14	14	--	--	--	42
Joy	--	--	--	--	--	--	--	--	14	14	14	14	--	--	--	42
Kim	--	--	--	--	--	--	--	--	14	14	14	14	--	--	--	42
Lyn	--	--	--	--	--	--	--	--	14	14	14	14	--	--	--	42
Mia	2	3	3	--	3	2	--	4	--	--	--	--	9	7	7	31
Nia	--	1	1	--	1	--	--	4	--	--	--	--	7	8	8	22
Ola	--	1	1	--	1	--	--	4	--	--	--	--	7	8	8	22

[a] Diagonals indicate the number of times the individual was named to any group.

[b] Off-diagnonal numbers indicate number of times respondents named two persons to the same cluster.

[c] Row Totals indicate the total number of dyadic linkages for each subject (excluding numbers on the diagonal).

Table 4

Four Alternative Models ($Lambda_X$ Parameter Matrices in LISREL) to Describe Cluster Membership (Ksi Variables)

Persons	Model A I	Model A II	Model A III[a]	Model B I	Model B II	Model B III	Model C I	Model C II	Model C III	Model D I	Model D II
Amy	1	0	0	1	0	0	1	0	0	1	0
Bea	1	0	0	1	0	0	1	0	0	1	0
Cam	1	0	0	1	0	0	1	0	0	1	0
Di	1	0	0	1	0	0	1	0	0	1	0
Edi	1	0	0	1	0	0	1	0	0	1	0
Fay	1	0	0	1	0	0	1	0	0	1	0
Gay	1	0	0	1	0	0	1	0	0	1	0
Hea	1	0	0	0	0	1	1	0	1	1	0
Ida	0	1	0	0	1	0	0	1	0	0	1
Joy	0	1	0	0	1	0	0	1	0	0	1
Kim	0	1	0	0	1	0	0	1	0	0	1
Lyn	0	1	0	0	1	0	0	1	0	0	1
Mia	0	0	1	0	0	1	0	0	1	1	0
Nia	0	0	1	0	0	1	0	0	1	1	0
Ola	0	0	1	0	0	1	0	0	1	1	0
	GFI=.955			GFI=.904			GFI=.969			GFI=.835	
	RMR=.140			RMR=.201			RMR=.115			RMR=.280	

[a] Three Ksi variables (clusters) are specified in Models A, B, and C, and two in Model D. The models differ in the location of the 15 observed variables (persons) and the numbers of clusters.

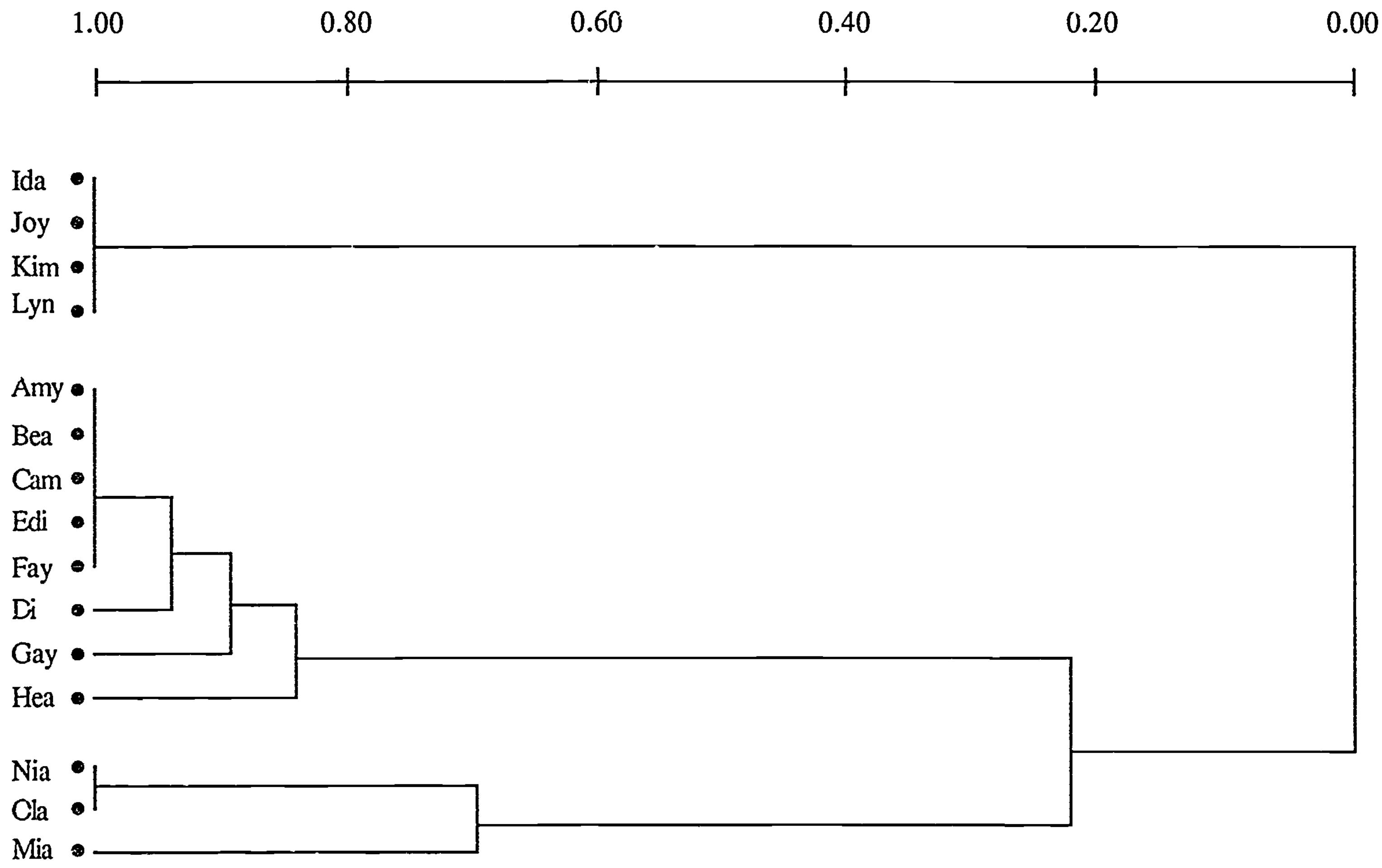

Figure 1. Hierarchical cluster analysis of female social network in Table 1.

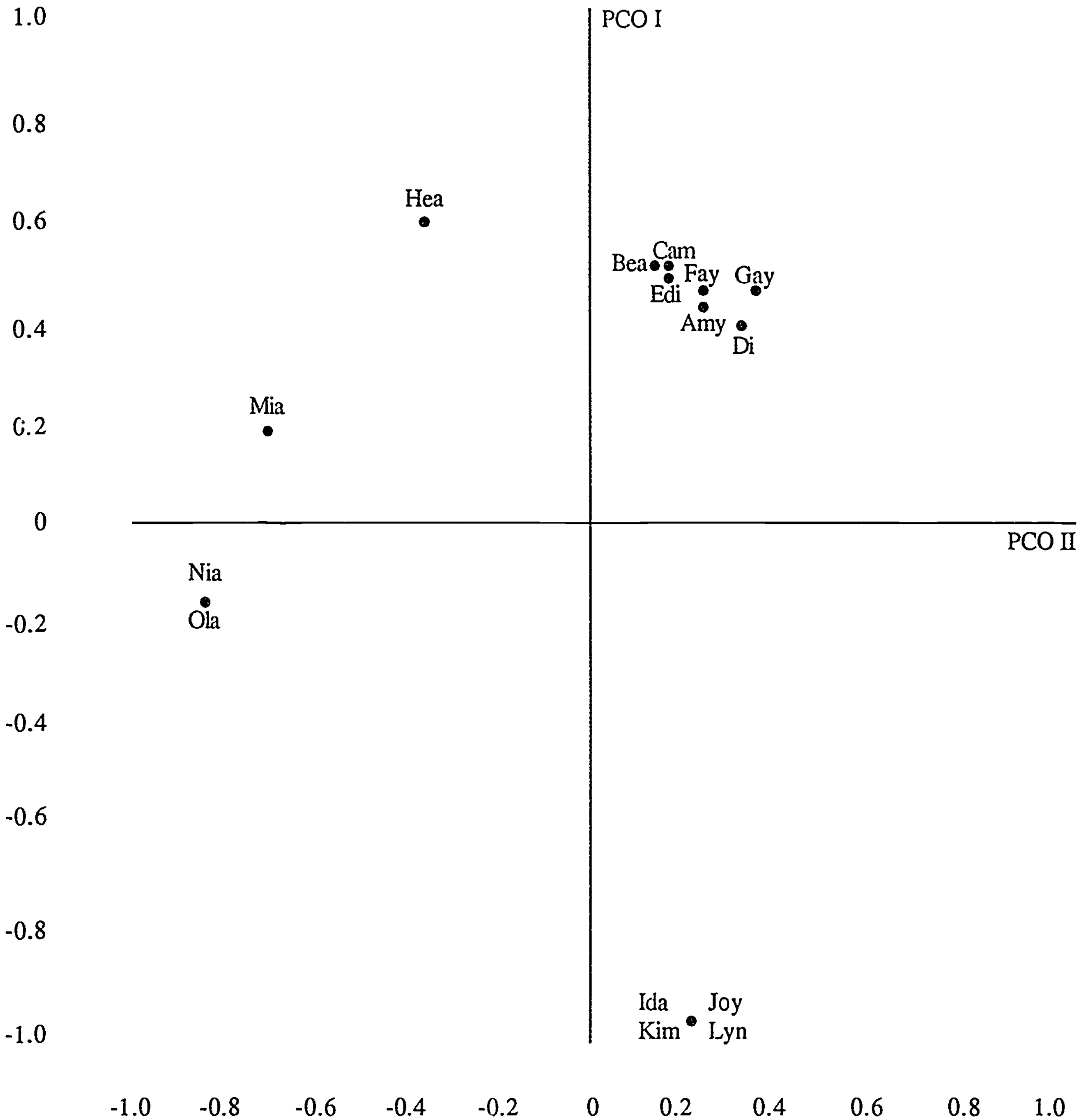

Figure 2. Principal coordinate analysis of female social network in Table 1.

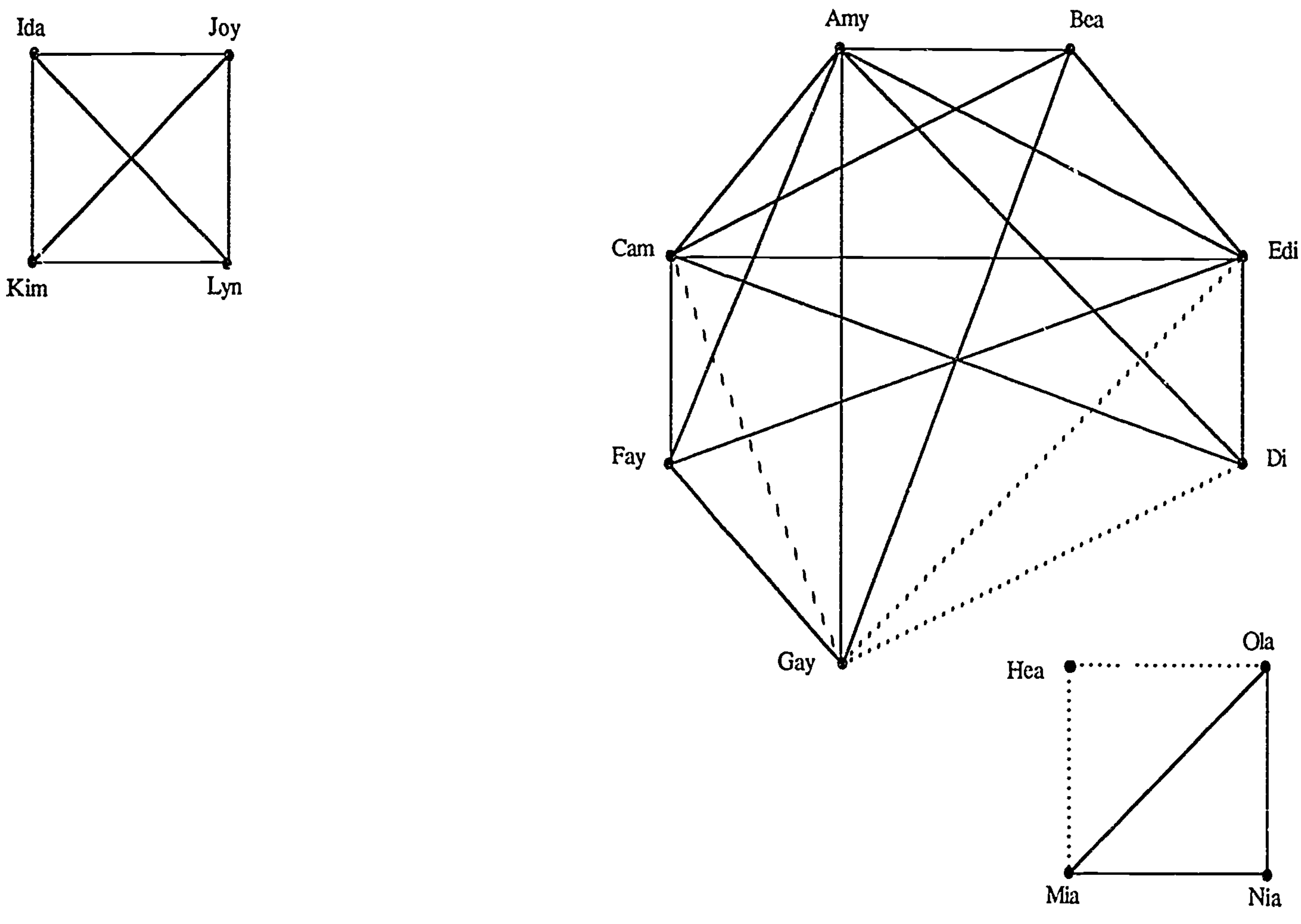

Figure 3. Network linkage analysis of female social network in Table 1.

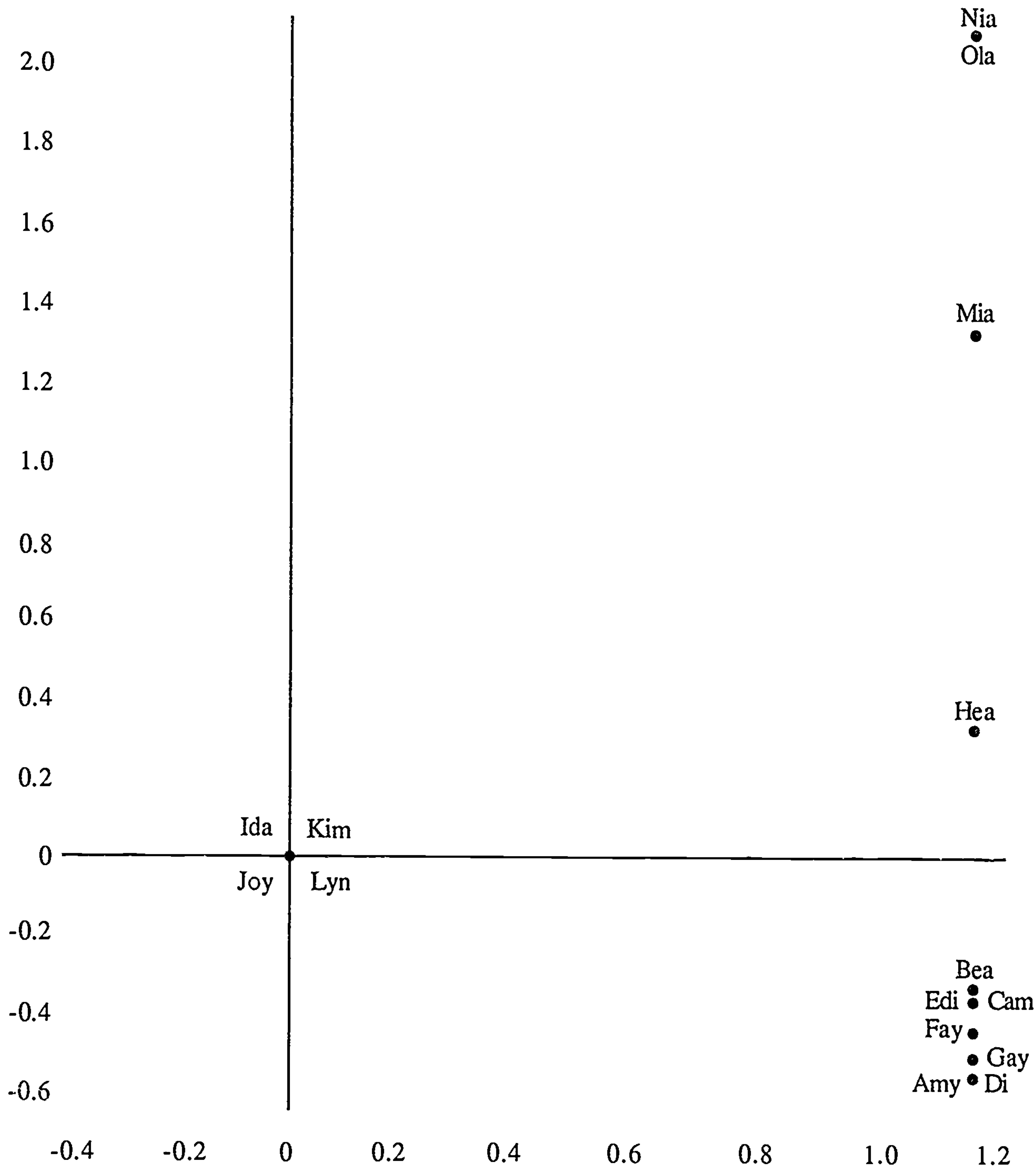

Figure 4. Correspondence analysis of female social network in Table 1.