Available online at www.sciencedirect.com
ScienceDirect
Social Networks 30 (2008) 146–158
SOCIAL NETWORKS
www.elsevier.com/locate/socnet

An algorithm and metric for network decomposition from similarity matrices: Application to positional analysis

Mo-Han Hsieh [a,*], Christopher L. Magee [b]

[a] *Engineering Systems Division, Massachusetts Institute of Technology, 77 Massachusetts Avenue, NE20-392 Cambridge, MA 02139-4307, USA*
[b] *Engineering Systems Division and Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

Abstract

We present an algorithm for decomposing a social network into an optimal number of structurally equivalent classes. The *k*-means method is used to determine the best decomposition of the social network for various numbers of subgroups. The best number of subgroups into which to decompose a network is determined by minimizing the intra-cluster variance of similarity subject to the constraint that the improvement in going to more subgroups is better than a random network would achieve. We also describe a decomposability metric that assesses how closely the derived decomposition approaches an ideal network having only structurally equivalent classes.

Three well-known network data sets were used to demonstrate the algorithm and decomposability metric. These demonstrations indicate the utility of the approach and suggest how it can be used in a complementary way to Generalized Blockmodeling.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Positional analysis; Structural equivalence; Decomposability; *k*-Means method; Generalized Blockmodeling

1. Introduction

In the network analysis literature, two lines of research have been pursued to develop methods of decomposing networks into meaningful subgroups (Wasserman and Faust, 1994). These are (1) research that seeks to identify cohesive subgroups (Frank, 1995); and (2) research that seeks equivalence classes in a network (Breiger et al., 1975; Lorrain and White, 1971). While numerous methods have been proposed to conceptualize the idea of cohesive subgroups (including the algorithm recently proposed by Newman and Girvan (2004)), the recent efforts in social networks research have been on developing methods that identify equivalence classes.

The initial concept of the equivalence class was proposed by Lorrain and White (1971) in the form of structural equivalence. By conceiving nodes in a network as equivalence classes or "positions" that relate in a similar way to other positions, a network can be transformed into a simplified model where nodes are combined into positions and the relations between nodes become relations between positions. For example, if two nodes link to and are linked by exactly the same set of other nodes, they are structurally equivalent to each other. Many definitions of equivalence have been proposed, see (Wasserman and Faust, 1994) for further discussion.

Among the methods that identify structural equivalence classes, Batagelj et al. (1992b) proposed to divide them into direct and indirect methods. While the direct method involves optimizing a pre-specified block model with the network data, an indirect method typically composes two major parts: (1) a definition of dissimilarity for the selected type of equivalence (e.g. the corrected Euclidean-like dissimilarity (Burt and Minor, 1983)) and (2) an algorithm that produces good clustering solutions (e.g. hierarchical clustering). The indirect method is indirect in the sense that the relational information among vertices is first used to create a partition, and the partition is then evaluated with an explicit criterion function (Batagelj et al., 1992b). The evaluation of the partition with a criterion function is not imperative for the indirect method; one example of the criterion function is the specified goodness-of-fit measure proposed by Batagelj et al. (1992a) originally designed for the use of their optimization approach in finding equivalence classes. While most of these indirect methods generate dissimilarity measures

* Corresponding author. Tel.: +1 617 577 5843; fax: +1 617 258 0485.
E-mail address: mohan76@mit.edu (M.-H. Hsieh).

0378-8733/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.socnet.2007.11.002

that are *compatible*[1] with structural equivalence, the decompositions based on these dissimilarity measures are generally not satisfying.

The often used method, CONCOR (Breiger et al., 1975), is considered as having the aspects of both the indirect and direct method (Batagelj et al., 1992b). However, the CONCOR procedure always splits a set of vertices into exactly two subsets. Repeated application of CONCOR results in a series of subdivided bi-partitions of the original network. Thus, the partition outcome is at least partially determined by the procedure, not by the actual structure of the network (Schwartz, 1977).

The most recently developed approach in identifying equivalence classes is Generalized Blockmodeling (GBM) (Doreian et al., 2005). The method considers ideal blockmodels and uses optimization methods to fit them to empirical data. This direct method allows for use of context information in forming hypotheses and gives a criterion function (i.e. inconsistencies) that measures the fit of a specified blockmodel or decomposition structure to the actual data. GBM has been shown to give "better" decompositions of social network data based upon comparing inconsistencies (Batagelj et al., 1992b; Doreian et al., 2005). GBM finds better decompositions by a clear procedure, but as noted in (Doreian et al., 2005) hypotheses with a greater variety of block types can always be found to lower the number of inconsistencies towards zero. In the case of using BGM to find the structural equivalence partition of a network, though it is possible to identify the most appropriate number of subgroups by observing the jump of inconsistencies, to some extent it still involves subjective judgment and thus lacks a fully objective criterion for stopping decomposition.

Another approach to decomposition of networks is based upon network models (Fienberg and Wasserman, 1981; Snijders and Nowicki, 1997; Tallberg, 2005; Wasserman and Anderson, 1987). Recently, the development of stochastic models in the field of cluster analysis has lead to its application to social networks (Handcock et al., 2007; Hoff et al., 2002). The attractive features of using these approaches to find structural equivalence classes include, for example, statistical inferences with full models and statistical criteria for determining the number of classes. However, the potential disadvantages of this approach are the difficulty of model selection and the potentially large number of parameters to be estimated. Both of these disadvantages make theoretical interpretation of positions and blocks for social networks problematic.

The recent survey by Schaeffer (2007) indicates that selecting the best number of clusters (and other "parameter selection" issues) is one of the major open problems of graph clustering. In this paper, we propose a new indirect method for partitioning a network into structural equivalence classes and for this domain develop a method that also addresses the issue of clustering number selection. Overall, the method consists of (1) an unsupervised clustering method, in which vertices are assigned to clusters to minimize the intra-cluster variance of dissimilarity; (2) an approach that takes into consideration not only the dissimilarity between the pair of vertices but also the pair's dissimilarities with all other vertices; (3) a quantitative stopping criteria for determining the number of subgroups that a network should be divided into to better represent its underlying structural equivalence structure. The method is seen as a companion to GBM offering additional insight in certain kinds of studies (where inductive learning is useful) and having a similar limitation.

The paper presents the new method for finding structural equivalence classes and its application to ideal structural equivalence networks in Section 2. In Section 3, we develop a normalized decomposability metric for assessing how close non-ideal networks are to the ideal networks found by our (or any) decomposition methodology. Application of our method including the decomposability metric to three known social networks is presented in Section 4. Brief concluding remarks are given in Section 5.

2. A new method for finding structural equivalence classes

The method starts with any dissimilarity measure of vertices that is *compatible* with structural equivalence. For an n-node network, the dissimilarity measures can be arranged in an n by n matrix, whose entries give the dissimilarity between the row vertices i and the column vertices j. Hierarchical clustering generates the hierarchy of vertices by using these measures and different definitions of dissimilarity between the new clusters. Our method treats the n by n dissimilarity matrix as n data points in the n-dimensional space that we wish to partition. That is, we read row i of the dissimilarity matrix as the n-dimensional coordinates of the ith data point. Since the dissimilarity matrix is symmetric, the coordinates can also be read as the column elements.

With n data points in the n-dimensional space, we then repeatedly apply the k-means method (Hartigan and Wong, 1978; MacQueen, 1967) to partition the n data points into $k = 2$ to $k = n$ clusters. Information about the k-means method and its many variations can be found in (Kaufman and Rousseeuw, 2005). In this study, Lloyd's k-means algorithm (Lloyd, 1982) was used. Lloyd's algorithm begins with a set of k reference points which are randomly selected from the data set. All of the data points are partitioned into k clusters by assigning each point to the cluster of its closest reference point. In each iteration, the centroid for each cluster is calculated. A partition is then made using the newly calculated centroids as reference points for all of the data points. It has been proven (Bottou and Bengio, 1995) that the iterative process will eventually converge to a configuration where each data point is closer to the reference point of its cluster than to any other reference point and each reference point is the centroid of its cluster. Since different initial reference points can generate different partitions, multiple sets of initial points are used to evaluate whether the obtained partition has approached its minimum sum of intra-cluster distances.

For each round of the k-means method that partitions the n data points into k clusters, we have the sum of the within cluster

[1] A dissimilarity measure is compatible to structural equivalence if it satisfies the condition that the dissimilarity of a pair of nodes is zero if and only if the two nodes are structurally equivalent (Doreian et al., 2005, p. 181).

points-to-centroid distances as

$$D_k = \sum_{i=1}^{k} \sum_{j \in S_i} \left\| xj\_ci \right\|^2 \quad (1)$$

where S_i ($i = 1,2,\ldots,k$) is the cluster and c_i is the centroid or mean point of all of the data points x_j in cluster S_i.

In the process of decomposing the network into more subgroups (i.e. as k increases), D_k gradually decreases as more centroids are generated. A smaller D_k is desirable because we want a partition that has a smaller intra-cluster variance. D_k is zero when all of the equivalence classes (including singletons) have been identified by at least one centroid. We define ideal networks as those having only structural equivalence classes (i.e. zero discrepancies for GBM), and for such networks an algorithm that stops further partitioning the network when D_k is zero would be appropriate. However, for most real networks, the monotonically decreasing D_k goes to zero only after numerous singletons have been individually identified as unique equivalence classes. In the case of $k = n$, D_k is always zero because every node is identified as itself an equivalence class. The result of identifying a great number of singletons is relatively meaningless since it does not inform us about the underlying structure of the network. To avoid generating an excessive number of classes for real networks, a quantitative criterion must be designed to appropriately stop further decomposition of the network.

For any assigned number of subgroups, the k-means method seeks to minimize D_k with the same number of centroids. Because nodes of the same equivalence class have the same coordinates, a lower D_k can be obtained by first grouping them with centroids. Therefore, if a network has equivalence classes that have more than one node, D_k decreases significantly with newly added centroids until every such equivalence class has been identified by at least one centroid. The decrease of D_k slows down with larger k when singletons start to appear as classes.

These singletons, with their unique linkage patterns, are similar to randomly wired nodes in a network. We found that the gradual decrease of D_k during the generation of singletons is similar to that of the random networks with the same size and linkage density.[2] We stop further dividing a network into additional subgroups if the decrease of D_k (form k to $k+1$) is smaller or equal to the average decrease of D_k obtained from a sample of these random networks.[3] We thus define a fitness index as simply:

$$F_k = D_k^{\text{random}} - D_k^{\text{real}} \quad (2)$$

where D_k^{random} is the sum of intra-cluster point-to-centroid distances obtained by averaging over the results of a sample of random networks and D_k^{real} is that of the real network. We find the maximum of F_k as a function of k, and the corresponding k represents the most appropriate subdivision of the network because further subdivision is only reducing D_k^{real} at random (or less than random) rates. The nodes belonging to the k different clusters then form the equivalence classes of the network.

To obtain an appropriate estimate of D_k^{random} in Eq. (2), a certain number of random networks have to be sampled. The number of random networks sampled is determined by the standard deviation of D_k^{random} relative to the decrease of D_k^{real}. Since the simulation indicate that D_k^{random} is normally distributed, our procedure is to sample 30 random networks and then determine the appropriate sample size, N, according to

$$N \geq \left[\frac{z_{\alpha/2} s_k}{\Delta F_{k/2}} \right]^2 \quad (3)$$

where z is the ordinate on the normal curve corresponding to the desired probability α (.05 in our case), s_k the sampled standard deviation of D_k^{random}, and ΔF_k is the difference between F_k and either F_{k-1} or F_{k+1}. We iteratively increase sample size until the repeatedly recalculated N satisfies Eq. (3).

In theory, our method should work for ideal networks having only structural equivalence classes because nodes of the same equivalence class cause a larger decrease of the sum of intra-cluster point-to-centroid distances than nodes that belong to no equivalence class (i.e. nodes of random networks). It should be noted that, if an ideal network has a singular node as a structural equivalence class, it is possible that our algorithm will not identify this node as an equivalence class. This is due to the difficulty of differentiating the linkage pattern of a meaningful node from that of a randomly placed node. In this case, our fitness index as shown in Eq. (2) can fail to indicate the most appropriate number of structural equivalence classes and thus the most appropriate decomposition. Nevertheless, our method works for ideal networks with all of their structural equivalence classes having at least two nodes.

Our method works in practice as we have tried the algorithm for a variety of ideal networks, and the algorithm identifies the correct subgroups for all of them. However, there are easy and difficult cases of using the fitness index to identify the right number of classes that the network has.

The difficult cases are the networks whose decrease of the sum of intra-cluster point-to-centroid distances is only slightly higher than that of a random network. Fig. 1 shows two sets of comparison between these difficult and easy cases. Each fitness value in the figure is normalized between zero and one so that we can compare their relative easiness of identifying the peak of fitness index.

Fig. 1(a) shows the fitness index for two ideal networks with the same minimum equivalence class size (i.e. $C = 5$) but different network size (i.e. $n = 25$ and 100). As shown in the figure, it is easier to identify the peak of fitness index for the network with smaller size. Fig. 1(b) shows the fitness index for two ideal networks with the same network size (i.e. $n = 50$) but different minimum equivalence class size (i.e. $C = 2$ and 10). As shown in the figure, identifying the peak of fitness index is now easier for the network with larger minimum equivalence class size.

[2] We use the Erdős-Rényi model to generate random networks. The model considers all pairs of nodes in a graph and puts an edge between the nodes with a fixed probability (which in our case equals to the linkage density). Since a random network with larger size and density has larger step decrease of D_k, the network's decrease of D_k (from k to $k+1$) should be compared with that of the random networks with the same size and density.

[3] These random networks can be viewed as null models for cluster validation in the sense discussed by Gordon (1999).

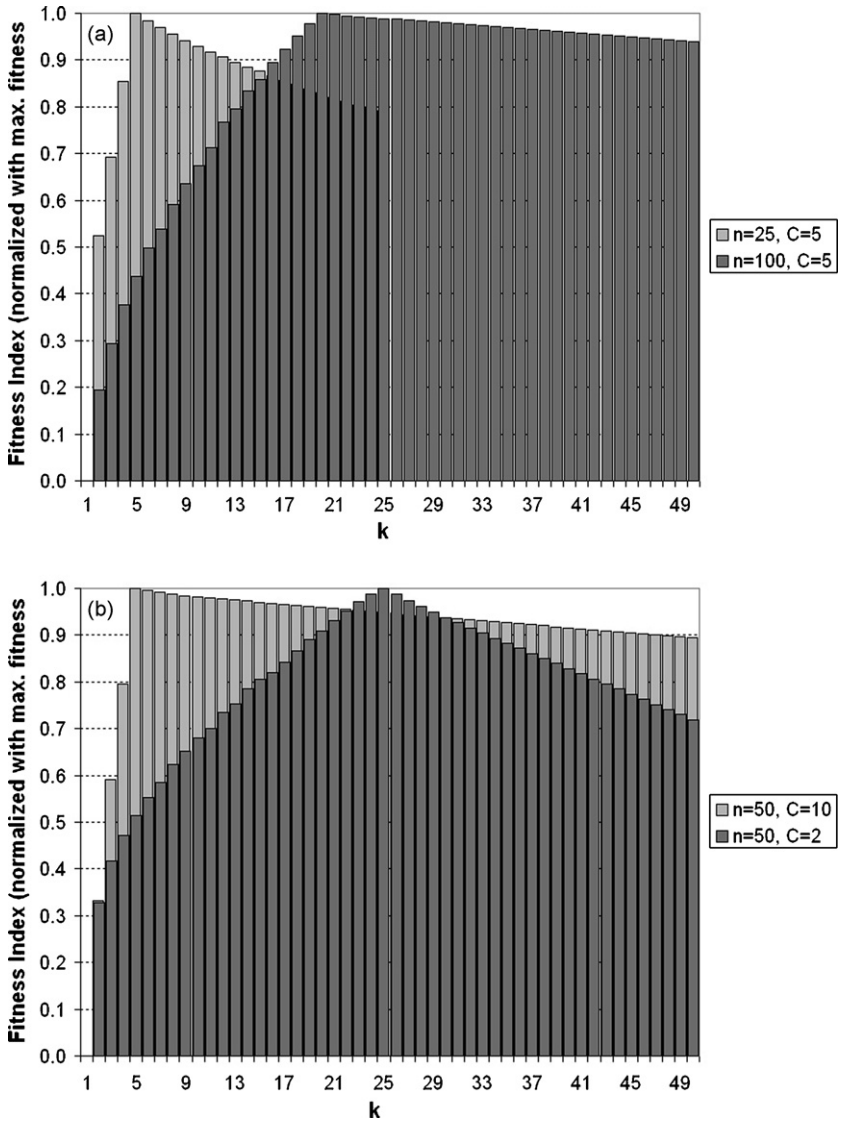

Fig. 1. (a) Fitness index for ideal networks with the same minimum equivalent class size (i.e. $C = 5$) but different network size (i.e. $n = 25$ and 100) and (b) fitness index for two ideal networks with the same network size (i.e. $n = 50$) but different minimum equivalent class size (i.e. $C = 2$ and 10).

In general, the ideal networks having only small equivalence classes and larger network sizes are the difficult cases for use of the fitness index to identify the right number of classes. Nonetheless, the method identifies the correct equivalence classes of these ideal networks. The more important issue is how to assess its value in the less-than-ideal networks that are typically observed for social networks. The next two sections of the paper address this topic.

3. Measuring the decomposability of a network

By applying our class finding algorithm, networks are divided into subgroups that correspond to their underlying structural equivalence structures. However, we want to differentiate among networks whose subgroups are not all ideal equivalence classes. In this case, we define perfect *decomposability* of a network as that achieved when a network is composed of only equivalence classes.

Having a normalized objective measurement of decomposability is useful. For example, we can compare two networks and determine which network is more similar to an ideal network having only equivalence classes. Lower decomposability can be used to infer that the suggested decomposition is more forced and thus should be cautiously utilized in further analysis. Moreover, if other variables (or time series data) are known, the change of the decomposability metric with the variables (or with time) affecting the network can be found. This can allow one to find how various variables influence the structural roles in a given network or a variety of different networks. To be able to compare the decomposability of networks of different size and density, it is necessary to normalize the metric for these effects.

To determine the normalized decomposability of a network, we construct a metric that places networks with only equivalence classes at one end and those without any equivalence class at the other. We use the sum of intra-cluster distance, D_k, of Eq. (1), to quantify the similarity between a real network and an

Table 1
D_{max} for network with different sizes and number of clusters

Number of clusters (k)	Size of network (n)									
	6	7	8	9	10	11	12	13	14	15
2	21.3	32.2	44.6	60.3	76.7	96.4	120	146	168	196
3	14.3	23.2	33.5	46.8	61.7	77.4	97.2	121	144	168
4	8.39	16.4	26.5	36.8	49.5	65.1	82.8	102	125	146
5	4.51	10.5	17.8	27.6	41.4	54.6	68.8	86.9	106	131

ideal network. For an ideal network having only equivalence classes, its sum of intra-cluster distance, D_{ideal}, equals zero. This is because every member of the same equivalence class, when viewed as a node in the multidimensional space, has the same coordinates. Therefore, their intra-cluster distances are zeros and the sum of these distances, D_{ideal}, is zero.

In addition to the value of D_{ideal}, we want the upper bound of the sum of intra-cluster distance, $D_{max(n,k)}$, for networks having n nodes and k clusters. With the lower bound, $D_{ideal} = 0$, and the upper bound, $D_{max(n,k)}$, we can thus obtain the normalized decomposability metric, Q, for the network as

$$Q = 1 - \frac{D_k - D_{ideal}}{D_{max(n,k)} - D_{ideal}} = 1 - \frac{D_k}{D_{max(n,k)}} \quad (4)$$

which defines Q as 1 for perfect decomposability and 0 for $D_k = D_{max(n,k)}$ which is equivalent to no decomposability. To obtain the upper bound, $D_{max(n,k)}$, we are seeking a network that has the maximum possible value of D_k while having the same size and is divided into the same number of clusters as that of the ideal network. To obtain the upper bound of D_k, various Monte Carlo methods can be applied to obtain an approximate solution for the network with size, n, and number of clusters, k. In this study, we used a genetic algorithm (GA) as the optimization method to search for $D_{max(n,k)}$. To apply the GA, the solution domain was represented by rearranging the adjacency matrix of a network into an array of bits, the fitness function was D_k, and the two-point crossover was used to generate a new generation of solutions. For more information and implementation details about GA, see (Mitchell, 1996).

By using the corrected Euclidean-like dissimilarity (Burt and Minor, 1983) as the dissimilarity measure for structural equivalence. Table 1 shows some examples of $D_{max(n,k)}$ (with three significant figures) for network with different sizes and number of classes. In Table 1, the maximum possible D_k for a 9-node network, for example, divided into three classes is $D_{max(n,k)} = D_{max(9,3)} = 46.8$. With this information, we consider three 9-node networks as shown in Fig. 2.

Note that the only difference between network 1 and network 2 is the directed link from node 7 to node 1. network 3 differs from network 2 by its additional links from node 2 to node 7 and from node 4 to node 8. The result of applying our class finding algorithm to network 1 and network 2 shows that the two networks are divided into the same $k = 3$ subgroups (i.e. node 1, 2, and 3, node 4, 5, and 6, and node 7, 8, and 9). Moreover, following Eq. (4), with $k = 3$, we have $D_k = D_3$ for network 1 as 8.71 and for network 2 as 11.2. Therefore, the decomposability metric for network 1 is

$$Q_1 = 1 - \frac{D_3}{D_{max(9,3)}} = \frac{1 - 8.71}{46.8} = 0.81$$

and the decomposability metric for network 2 is

$$Q_2 = \frac{1 - 11.2}{46.8} = 0.76$$

Similarly, our class finding algorithm tells us that network 3 should be divided into still the same $k = 3$ classes. With its sum of intra-cluster distance, D_3, equal to 16.2, we obtain its decomposability metric as

$$Q_3 = \frac{1 - 16.2}{46.8} = 0.65$$

With network 1 having the highest decomposability and network 3 having the lowest, the decomposability metric confirms what visual inspection tells us; network 2 is closer to the ideal network than is network 3 but is further from ideal than is network 1.

It should be noted that, the decomposability metric can be calculated only after we know the number of subgroups that the network should be divided into. Since the decomposability metric monotonically increases as the number of subgroups increases (and equals to unity as every node of the network is itself a subgroup), it cannot be used to determine the appropriate number of subgroups in a network. To do so, we still have to use the fitness index as introduced in Eq. (2). The fitness index compares the decrease of D_k^{real} resulting from increases in the number of subgroups to that of D_k^{randam} and thus can be used for determining the most appropriate number of subgroups.

Since the decomposability can be viewed as a measure of deviation of real networks from ideal networks that contain only equivalence classes, we explored the relationship between a network's decomposability and its deviation from an ideal network. To do this, we examine the decomposability of 10,000 pseudo real networks generated from randomly perturbing[4] all possible linkages of ideal networks (i.e. adding or removing links) with six different percentages. Ideal networks with sizes between 30 and 60 were sampled. Furthermore, we sample ideal networks with the assumption that the number of classes for each network is normally distributed and the size of each class within a network is also normally distributed. Since real networks typi-

[4] The perturbation can be viewed as arising from an error in observation or arising because real social relationships are more complex than the ideal.

Network 1 and its adjacency matrix

	1	2	3	4	5	6	7	8	9
1	0	1	1	0	0	0	0	0	1
2	1	0	1	0	0	0	0	0	0
3	1	1	0	0	0	0	0	0	0
4	0	0	0	0	1	1	0	0	0
5	0	0	0	1	0	1	0	0	0
6	0	0	0	1	1	0	0	0	0
7	0	0	0	0	0	0	0	1	1
8	0	0	0	0	0	0	1	0	1
9	0	0	0	1	0	0	1	1	0

Network 2 and its adjacency matrix

	1	2	3	4	5	6	7	8	9
1	0	1	1	0	0	0	0	0	1
2	1	0	1	0	0	0	0	0	0
3	1	1	0	0	0	0	0	0	0
4	0	0	0	0	1	1	0	0	0
5	0	0	0	1	0	1	0	0	0
6	0	0	0	1	1	0	0	0	0
7	1	0	0	0	0	0	0	1	1
8	0	0	0	0	0	0	1	0	1
9	0	0	0	1	0	0	1	1	0

Network 3 and its adjacency matrix

	1	2	3	4	5	6	7	8	9
1	0	1	1	0	0	0	0	0	1
2	1	0	1	0	0	0	1	0	0
3	1	1	0	0	0	0	0	0	0
4	0	0	0	0	1	1	0	1	0
5	0	0	0	1	0	1	0	0	0
6	0	0	0	1	1	0	0	0	0
7	1	0	0	0	0	0	0	1	1
8	0	0	0	0	0	0	1	0	1
9	0	0	0	1	0	0	1	1	0

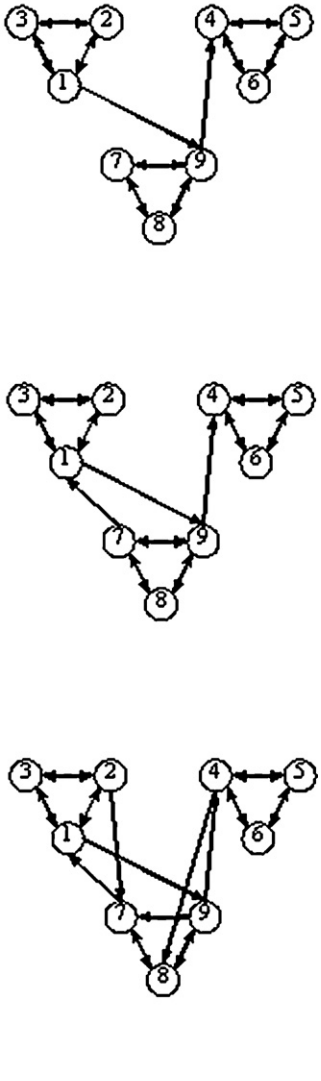

Fig. 2. Networks 1, 2, and 3 and their adjacency matrixes.

cally have very low density, we only sample ideal networks with density lower than 0.2.

Fig. 3 shows the average decomposability metric of the pseudo real networks (with one standard deviation also plotted) versus their percentage of linkage perturbation from ideal networks. Although, the linear relationship between the two has an R-square value of 0.99, the results also show some clear non-linearity. However, if we limit the applicability of the decomposability metric to networks with decomposability of 0.4 and higher, the linear equation gives a reasonable estimate of the linkage perturbation.

With this result, we can calculate the deviation of our previous three networks. Referring to Fig. 3, the decomposability of networks 1, 2, and 3 (calculated above) are equivalent to 4.8, 6.1, and 8.9% linkage perturbation of their underlying ideal network. We note that the lower the decomposability the more questionable the ideal network that we associate with the decomposition is. Other networks with slightly more deviations might also describe the network in these cases.

4. Application of the method

In the previous sections, we propose a method for clustering nodes of networks into structural equivalence classes and a decomposability metric to quantify a network's level of linkage perturbation from a hypothetical underlying ideal network. Because our method can identify the number of classes for any ideal network having only structural equivalence classes (with at least two nodes in each class), in this section we test our method and the decomposability metric with three examples of real networks.

The social network of the 15 office workers reported by Thurman (1979) is used as the first example to evaluate our method. The network is shown in Fig. 4. By applying our method to partition the social network into $k = 2$ to $k = 15$ subgroups, the sum of intra-cluster point-to-centroid distances as a function of k is obtained and is shown in Fig. 5(a) as dark gray bars. To obtain the fitness index, we need the comparable sum of a sample of random networks that have the same size and density as

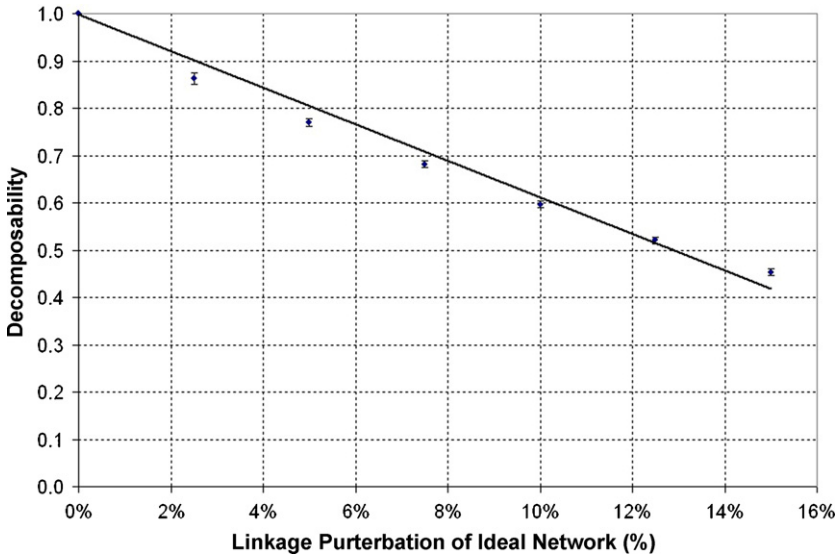

Fig. 3. Decomposability versus linkage perturbation of various ideal networks.

the social network. This sum of intra-cluster point-to-centroid distances as a function of k is shown in Fig. 5(a) as light gray bars. The fitness index generated by subtracting the one of the social network from that of the random networks is shown in Fig. 5(b).

As shown in Fig. 5(b), the fitness index has its maximum at $k = 6$, which by our method indicates that the most appropriate decomposition of the network is into six equivalence classes. Fig. 6 shows these six classes and the block model as revealed by using our method.

As shown in Fig. 6, the first class includes Amy, Katy, and Tina, and the second class includes Ann, Pete, and Lisa. There is strong interaction within and between the two classes. What differentiates them is that the second class has strong interaction with the President. According to Thurman (1979), Pete is characterized as the center of a social circle that included Lisa, Katy and Amy. Ann arrived under the sponsorship of Pete, and Lisa has the ear of the President (Thurman, 1979). It is worth noticing that the fourth class comprises only Emma, who has strong interaction with the President, the members of the second class, and the members of the fifth class. According to Thurman (1979), she plays a special role in the social network and thus identifying her as a unique class seems appropriate considering the context information.

With the network size equal to 15 and the number of subgroups equal to six, we have the upper bound of the sum of intra-cluster distances, $D_{\max(15,6)}$, equal to 113.07. By using Eq. (4), the decomposability metrics for the social network is 0.65.

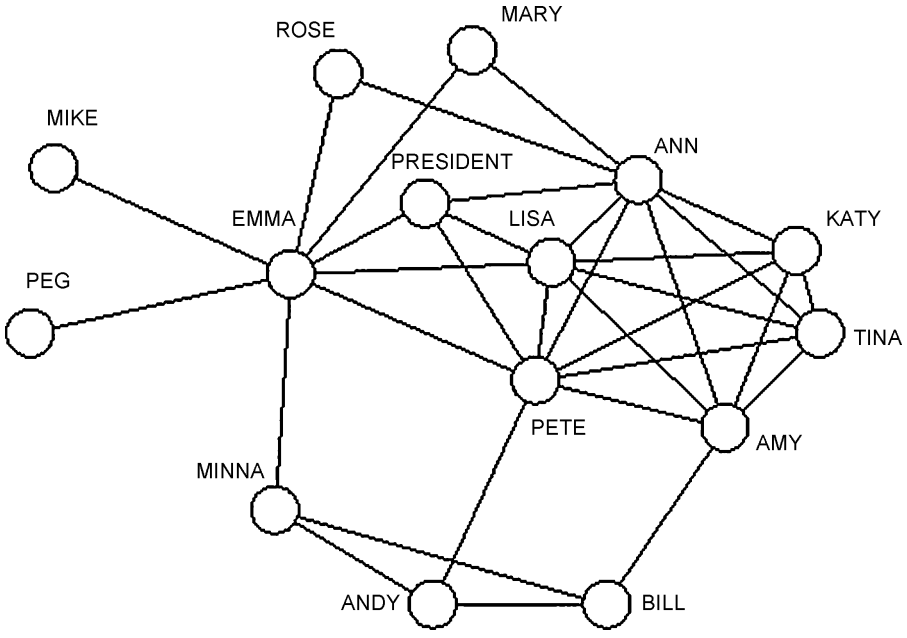

Fig. 4. The social network of the Thurman office data.

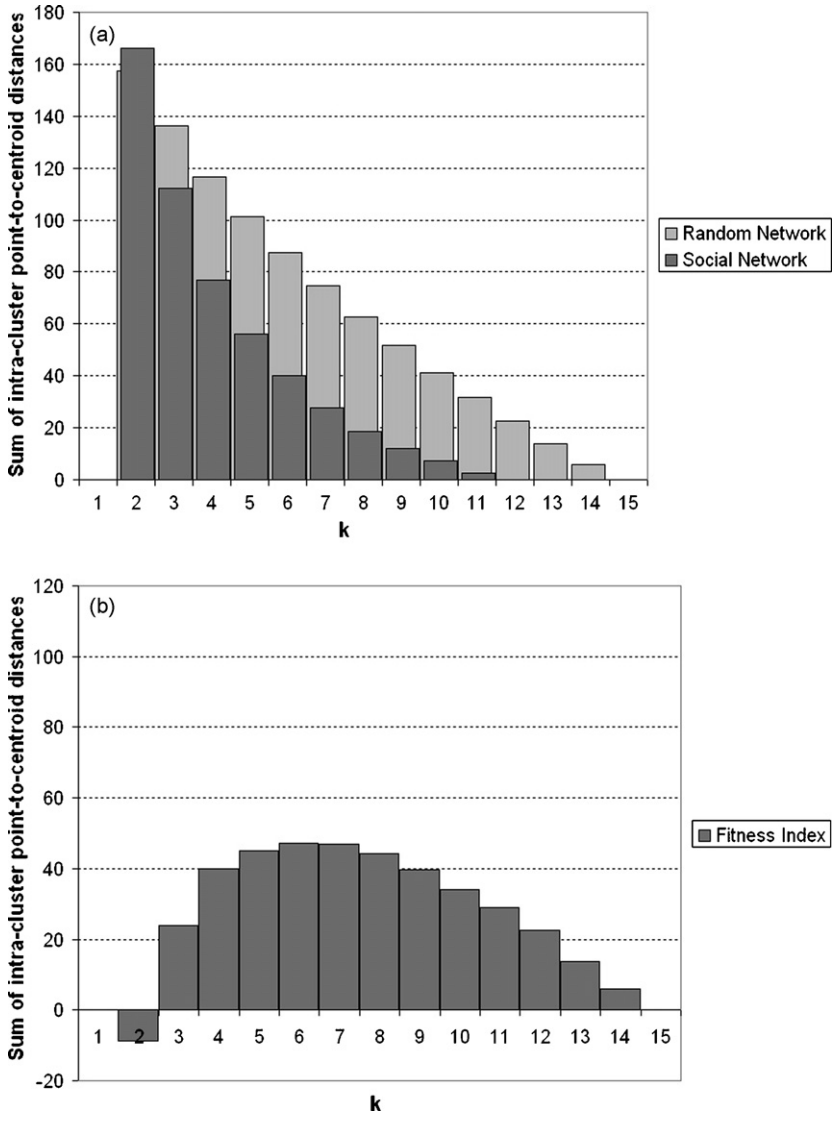

Fig. 5. (a) The sum of intra-cluster point-to-centroid distances of Thurman's office social network (dark gray bars) and that of the random networks with the same size and density (light gray bars) and (b) the fitness index of Thurman's office social network as a function of the number of subgroups.

Class	Members
1	Amy, Katy, Tina
2	Ann, Pete, Lisa
3	President
4	Emma
5	Mary, Rose, Mike, Peg
6	Bill, Andy, Minna

	1	2	3	4	5	6
1	1.00	1.00	.	.	.	0.11
2	1.00	1.00	1.00	0.67	0.17	0.11
3	.	1.00	.	1.00	.	.
4	.	0.67	1.00	.	1.00	0.33
5	.	0.17	.	1.00	.	.
6	0.11	0.11	.	0.33	.	.

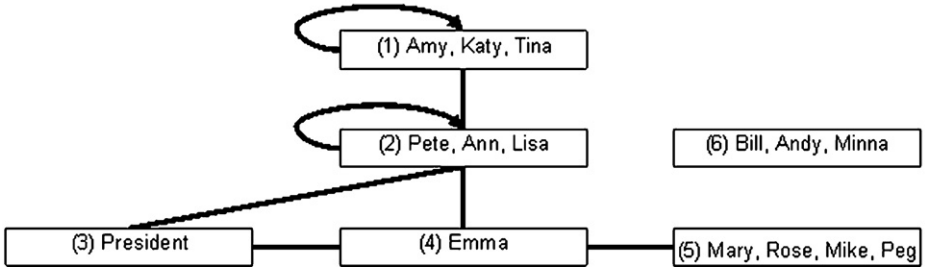

Fig. 6. Class members, block density, and image graph of the Thurman social network found by the algorithm presented in this paper.

Table 2
Kansas SAR inter-organizational network

		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Osage County Sheriff's Department	A	0	1	1	0	1	1	1	0	1	0	1	0	0	0	1	1	1	0	0	1
Osage County Civil Defense Office	B	1	0	1	0	1	1	1	0	1	0	0	0	0	1	0	1	1	0	0	0
Osage County Coroner's Office	C	1	0	0	1	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0
Osage County Attorney's Office	D	1	1	1	0	1	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0
Kansas State Highway Patrol	E	1	0	1	1	0	1	1	0	0	0	1	1	1	1	0	1	1	0	0	1
Kansas State Parks and Resources Authority	F	1	0	1	1	1	0	1	0	1	0	1	1	0	1	0	0	0	0	0	0
Kansas State Game and Fish Commission	G	1	0	1	1	1	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0
Kansas State Department of Transportation	H	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
U.S. Army Corps of Engineers	I	1	0	1	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0	0
U.S. Army Reserve	J	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Crable Ambulance	K	1	0	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
Franklin County Ambulance	L	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
Lee's Summit Underwater Rescue Team	M	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Shawney County Underwater Rescue Team	N	1	0	1	0	1	1	1	0	1	0	0	0	1	0	0	0	1	1	0	1
Burlingame Police Department	O	1	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	0	1
Lyndon Police Department	P	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0
American Red Cross	Q	1	1	1	0	1	1	1	0	1	1	0	0	0	1	0	0	0	0	0	0
Topeka Fire Department Rescue #1	R	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
Carbondale Fire Department	S	1	1	0	0	1	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0
Topeka Radiator and Body Works	T	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0

Moreover, by using the relationship between the decomposability and the linkage perturbation of the ideal network shown in Fig. 3, we can infer that the social network is about 8.9% linkage perturbation from the ideal network. This result indicates that the inferred equivalence structure of the social network might be substituted for easily with more observation or slight changes in interaction patterns.

The inter-organizational Search and Rescue (SAR) network created after a disaster in Kansas (Drabek, 1981) is used as the second example to demonstrate the use of the new method. The SAR network has 20 organizations. The dichotomized communication data among these organizations are shown in Table 2.

To present the basic structure of the network, Drabek used CONCOR to partition the network into five clusters as:

1. Authority position: $\{A, E\}$.
2. Primary support: $\{C, F, G, I, K\}$.
3. Critical resources: $\{D, L, N\}$.
4. Secondary support, 1: $\{M, O, P, Q, R, T\}$.
5. Secondary support, 2: $\{B, H, J, S\}$.

While these five subgroups are potentially useful in understanding this network, Doreian et al. (2005) showed that this partition has 79 inconsistencies when examined with their GBM criterion function for structural equivalence. They found a five-cluster alternative that has only 57 inconsistencies (indicating the weakness of CONCOR discussed in Section 1 to this paper):

1. Authority: $\{A, E\}$.
2. Bodies and survivors: $\{C, F, G, I\}$.
3. Infrastructure: $\{B, D, K, N, P, Q\}$.
4. Primary rescue operators: $\{H, J, L, M, R, S, T\}$.
5. Secondary rescue operators: $\{O\}$.

Applying the method presented in this paper to find the structural equivalence classes of the SAR network, we again partition the network into $k=2$ to $k=20$ subgroups. The sum of intra-cluster point-to-centroid distances of the SAR network and that of the random network with the same size and density is shown in Fig. 7(a). The fitness index is shown in Fig. 7(b).

The fitness index shown in Fig. 7(b) has its maximum at $k=4$, indicating that the most appropriate decomposition is into four equivalence classes. Fig. 8 shows these four classes and the block model as revealed by using our method.

As shown in Fig. 8, our partition differs from that of Doreian et al. (2005) only in that ours combines their two classes, $\{B, D, K, N, P, Q\}$ and $\{O\}$, into one class thus including "secondary rescue operators" with "infrastructure". By using the criterion function for structural equivalence proposed by Doreian et al., our partition has 64 inconsistencies, which is considerably better than the 79 for the five subgroups suggested by CONCOR but seven more than the five subgroup partition proposed by Doreian et al using their direct method. Since more subgroups will decrease the inconsistencies, we examine the five-class decomposition of our method[5]:

1. $\{A, E\}$,
2. $\{C, F, G, I\}$,
3. $\{B, N, O\}$,
4. $\{D, K, P, Q\}$,
5. $\{H, J, L, M, R, S, T\}$.

This partition breaks the third class of our four-class partition into two classes as $\{B, N, O\}$ and $\{D, K, P, Q\}$ thus

[5] Our stopping algorithm indicates that four groups are appropriate but we examine what the k-means method would yield if allowed for five groups only for comparative purposes.

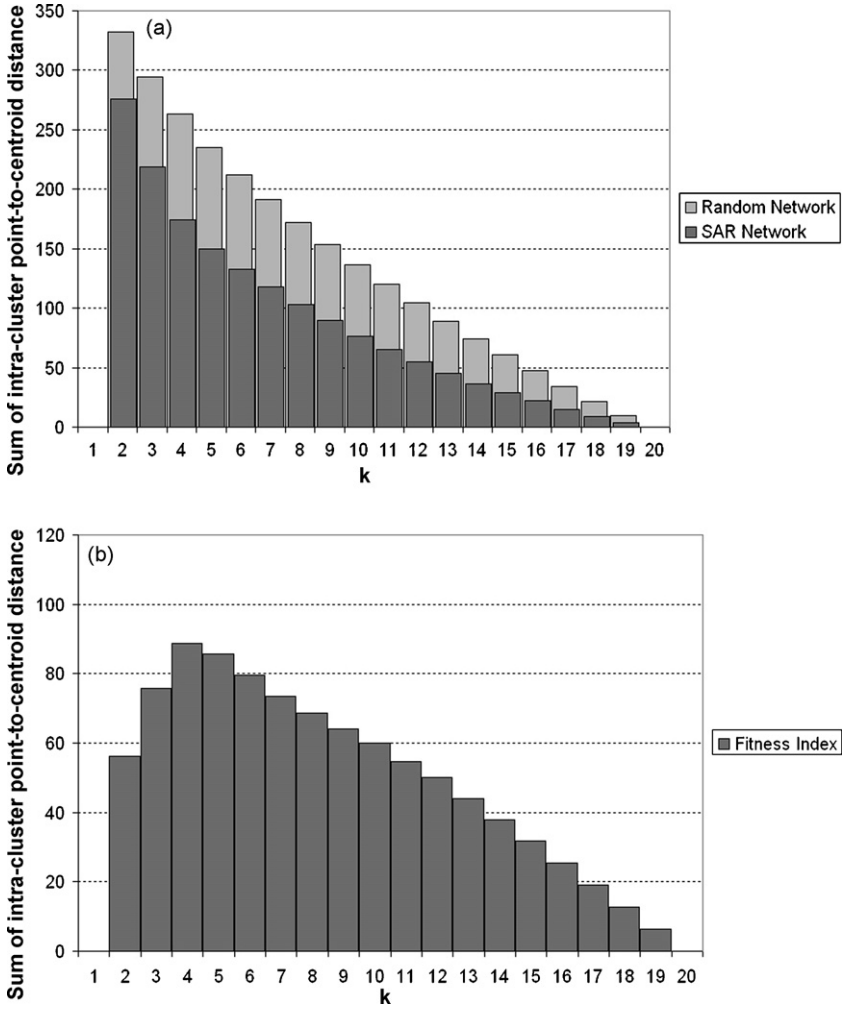

Fig. 7. (a) The sum of intra-cluster point-to-centroid distances of the SAR network and the random networks with the same size and density and (b) the fitness index of the SAR network.

decomposing "infrastructure" but differently than Doreian et al. This decomposition has the same number of inconsistencies (i.e. 57) as that of the different five-class partition of Doreian et al. when examined with their criterion function. Thus, our method appears more effective than CONCOR and relative to GBM is capable of finding interesting decompositions that are worthy of consideration along with various hypotheses arrived at by other information.

With the network size equal to 20 and the number of subgroups equals to four for our first partition and five for the other partition, we have the upper bound of the sum of intra-cluster distance, $D_{\max(20,4)} = 298.51$ and $D_{\max(20,5)} = 271.56$. With these

Class	Members
1	*A, E*
2	*C, F, G, I*
3	*B, D, K, N, O, P, Q*
4	*H, J, L, M, R, S, T*

	1	2	3	4
1	1.0	0.9	0.7	0.3
2	1.0	0.9	0.4	0.1
3	1.0	0.9	0.4	0.2
4	1.0	0.1	0.2	0.0

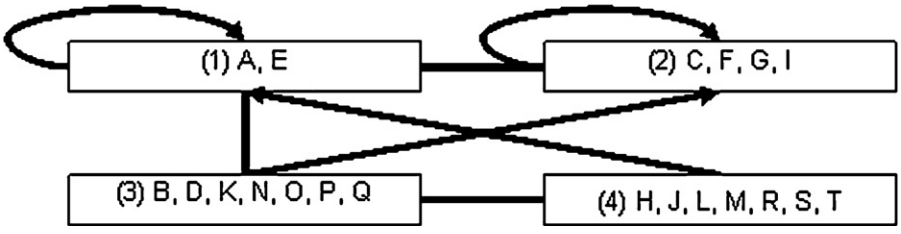

Fig. 8. Class members, block density, and image graph of the SAR network found by the algorithm in this paper.

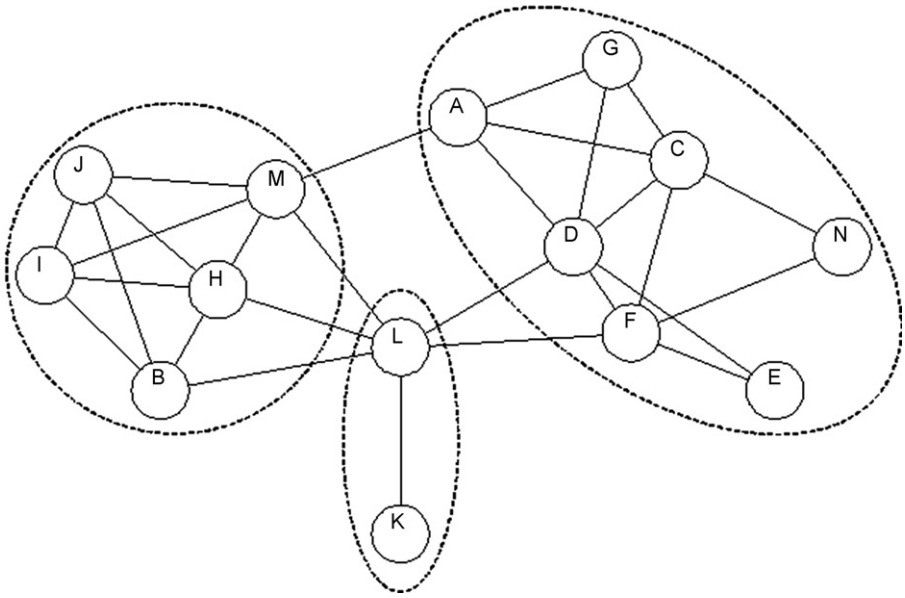

Fig. 9. Political actor network with 32 inconsistencies and decomposability of 0.42.

upper bounds, our first partition has a decomposability metric of 0.42, which is greater than the five subgroup partition of Drabek et al. (i.e. 0.41) and slightly lower than that of the partition of Doreian et al. (i.e. 0.44) and that of our five subgroup partition (i.e. 0.45). We feel it is more important to notice that the decomposability metric of 0.42 is about 15% perturbation from the ideal network. With this high percentage of linkage perturbation, we should be cautious when using *any* of the inferred equivalence structures of the SAR network. Conversely, we can use the low decomposability of the SAR network data and the lack of clarity about structure derived from that data to support the contention that communication structures were weak in this instance (Drabek, 1981).

Our third example is the political actor network reported by Doreian and Albert (1989). In this network, the nodes are the prominent political actors in a local community and the links represent "strong political ally" among the actors. Fig. 9 shows the three-class partition obtained by using CONCOR in the original analysis.

According to Doreian et al. (2005), this partition has 32 inconsistencies when examined with the GBM criterion function for structural equivalence. They proposed a three-cluster alternative shown in Fig. 10 that has only 26 inconsistencies.

By applying our method to find the structural equivalence classes of the network, maximization of the fitness index indicates that the network is best decomposed into four equivalence classes. The four-class partition is shown in Fig. 11. When examined with the criterion function proposed by Doreian et al. (2005), it has 25 inconsistencies, which is one less than that of the partition shown in Fig. 10.

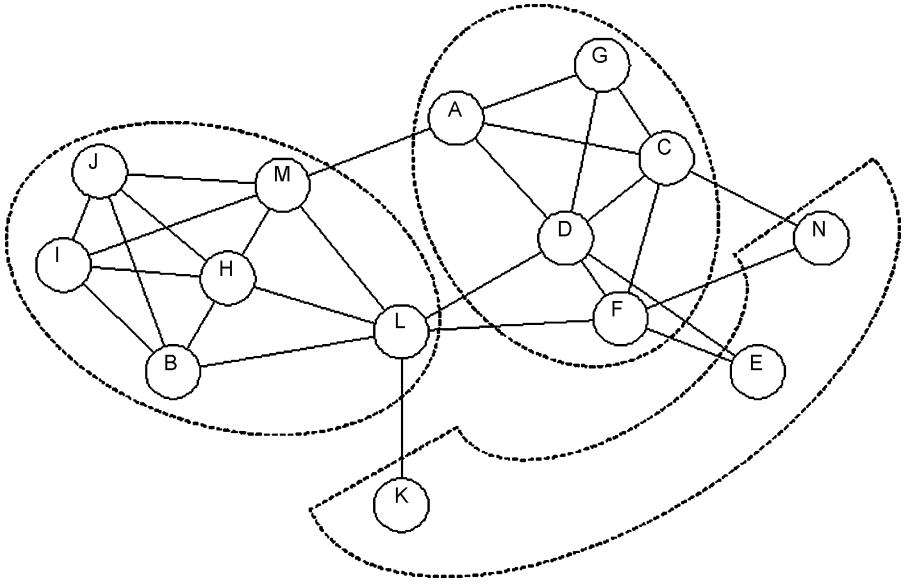

Fig. 10. Political actor network with 26 inconsistencies and decomposability of 0.37.

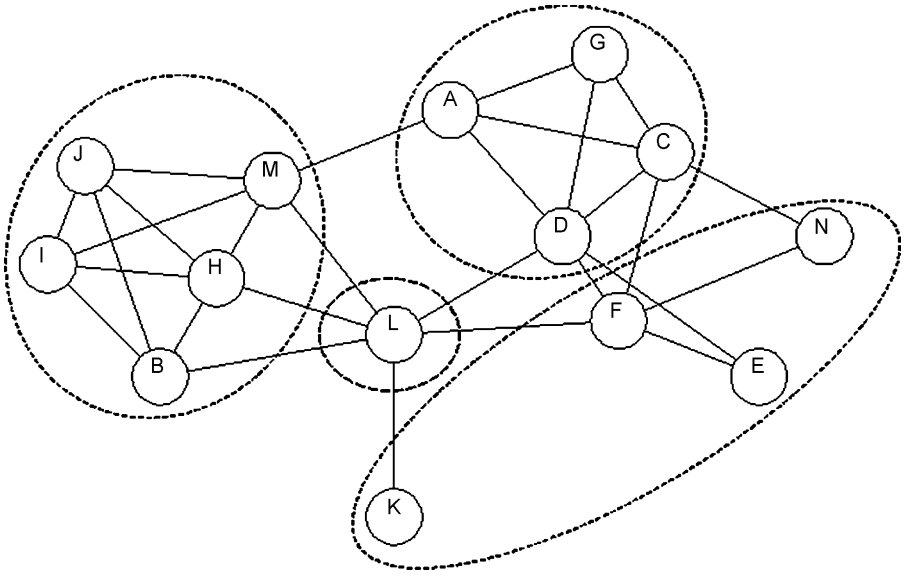

Fig. 11. Political actor network with 25 inconsistencies and decomposability of 0.49.

Since reduced inconsistency is expected with more subgroups, we also explore the three subgroup solution from the inductive method. In this case, our method suggests the *same* partition as derived by CONCOR (i.e. the partition in Fig. 9). Though it has more inconsistencies than that of the partition shown in Fig. 10, the former has a decomposability metric of 0.42 that is higher than 0.37 of the latter. This result clearly shows that our four-class partition, with 25 inconsistencies and a decomposability metric of 0.49, has the best quality in terms of both the criterion function and the decomposability. However, the relatively low decomposability of this network indicates that any of these interpretations is open to change if more or slightly modified data was obtained about these networks. Alternatively, the relatively low decomposability indicates that the structure is significantly deviated from any ideal model and thus the political actor network is relatively weakly structured.

5. Conclusion and discussion

The algorithm described in this paper appears to bring additional theoretical utility to existing methodology for decomposing networks into structural equivalence classes. The theoretical advantage is its ability to find all ideal structural equivalence classes but yet has an objective stopping criterion for continuing decomposition of non-ideal networks. The algorithm also appears to bring additional practical utility to existing tools such as the Generalized Blockmodeling by suggesting different decompositions of clear comparative merit to even well-studied examples as shown in Section 4.

When the algorithm is used in combination with Generalized Blockmodeling, one might obtain the advantages of combining inductive and deductive approaches. For example, with new data sets, one could start with finding the decompositions inductively (best and near best) and by in-context study of these possibly arrive at a new hypothesis to test by various criteria. In general, applying both methods seems to be appropriate in all cases because the results in Section 4 indicate they can deliver slightly different and yet interesting decompositions. In addition, the examples show the potential merit of using our metric for decomposability. The metric provides an objective assessment of the normalized decomposability of various networks (and for various decompositions).

The algorithm can be used in combination with the widely applied hierarchical clustering. For structural equivalence the method described here can quickly suggest a more appropriate decomposition into a specific set of block models. This can be compared with the suggested hierarchy and provide additional structural information of interest. Interesting future research could include (1) application of the algorithm in biological, economic and engineering system classification problems and (2) comparison of the results of this algorithm with the one developed by Newman and Girvan based upon cohesive subgroups in a wide variety of network types.

Acknowledgements

The authors are grateful for the helpful comments by the editor and the reviewers of this paper.

References

Batagelj, V., Doreian, P., Ferligoj, A., 1992a. An optimizational approach to regular equivalence. Social Networks 14, 121–135.

Batagelj, V., Ferligoj, A., Doreian, P., 1992b. Direct and indirect methods for structural equivalence. Social Networks 14, 63–90.

Bottou, L., Bengio, Y., 1995. Convergence properties of the *k*-means algorithm. In: Tesauro, G., Touretzky, D. (Eds.), Advances in Neural Information Processing Systems, vol. 7. MIT Press, Cambridge MA, pp. 585–592.

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

Burt, R.S., Minor, M.J., 1983. Applied Network Analysis: A Methodological Introduction. Sage Publications, Beverly Hills.

Doreian, P., Albert, L.H., 1989. Partitioning political actor networks: some quantitative tools for analyzing qualitative networks. Journal of Quantitative Anthropology, 279–291.

Doreian, P., Batagelj, V., Ferligoj, A., 2005. Generalized Blockmodeling. Cambridge University Press, Cambridge, UK.

Drabek, T.E., 1981. Managing multiorganizational emergency responses: emergent search and rescue networks in natural disaster and remote area settings. Institute of Behavioral Science, University of Colorado, Boulder, Colorado.

Fienberg, S.E., Wasserman, S., 1981. An expotential family of probability-distributions for directed-graphs—comment. Journal of the American Statistical Association 76, 54–57.

Frank, K.A., 1995. Identifying cohesive subgroups. Social Networks 17, 27–56.

Gordon, A.D., 1999. Classification. Chapman & Hall/CRC, Boca Raton.

Handcock, M.S., Raftery, A.E., Tantrum, J.M., 2007. Model-based clustering for social networks. Journal of the Royal Statistical Society Series A-Statistics in Society 170, 301–322.

Hartigan, J.A., Wong, M.A., 1978. A *k*-means clustering algorithm. Applied Statistics 28, 100–108.

Hoff, P.D., Raftery, A.E., Handcock, M.S., 2002. Latent space approaches to social network analysis. Journal of the American Statistical Association 97, 1090–1098.

Kaufman, L., Rousseeuw, P.J., 2005. Finding Groups in Data: An Introduction to Cluster Analysis. N.J. Wiley, Hoboken.

Lloyd, S.P., 1982. Least-squares quantization in Pcm. IEEE Transactions on Information Theory 28, 129–137.

Lorrain, F., White, H.C., 1971. Structural equivalence of individuals in social networks. Journal of Mathematical Sociology, 49–80.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 281–297.

Mitchell, M., 1996. An Introduction to Genetic Algorithms. MIT Press, Cambridge, MA.

Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical Review E 69.

Schaeffer, S.E., 2007. Graph Clustering. Computer Science Review 1, 27–64.

Schwartz, J.E., 1977. An examination of CONCOR and related methods for blocking sociometric data. In: Heise, D.R. (Ed.), Sociological Methodology. Jossey-Bass, San Francisco, pp. 255–282.

Snijders, T.A.B., Nowicki, K., 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. Journal of Classification 14, 75–100.

Tallberg, C., 2005. A Bayesian approach to modeling stochastic blockstructures with covariates. Journal of Mathematical Sociology 29, 1–23.

Thurman, B., 1979. In the office—networks and coalitions. Social Networks 2, 47–63.

Wasserman, S., Anderson, C., 1987. Stochastic a posteriori blockmodels—construction and assessment. Social Networks 9, 1–36.

Wasserman, S., Faust, K., 1994. Social network analysis: methods and applications. Cambridge University Press, Cambridge; New York.