Generalized Blockmodeling with Pajek

Vladimir Batagelj,[1] Andrej Mrvar,[2]
Anuška Ferligoj,[3] and Patrick Doreian[4]

Abstract

One goal of blockmodeling is to reduce a large, potentially incoherent network to a smaller comprehensible structure that can be interpreted more readily. Batagelj, Doreian, and Ferligoj developed a generalized approach to blockmodeling and methods where a set of observed relations are fitted to a pre-specified blockmodel. In the paper this generalized blockmodeling approach as implemented in program Pajek is described. An overview of the blockmodeling procedures in Pajek is given and is illustrated by some examples.

1 Introduction

Blockmodeling has been a main focus of network analysts (Hummon and Carley, 1993) with position as a central concept (Borgatti and Everett, 1992). Blockmodeling seeks to cluster units which have substantially similar patterns of relationships with others, and interpret the pattern of relationships among clusters.

Pajek is a program designed for the analysis of large networks (Batagelj and Mrvar, 1998, 2002, 2003, 2004). Initially generalized blockmodeling, as developed by Batagelj, Doreian, and Ferligoj, was supported by two programs `Model` and `Model2` (Batagelj, 1996). To provide additional support for analysis of smaller parts of large networks the authors of Pajek - Batagelj and Mrvar - decided to include also some procedures for generalized blockmodeling into Pajek. These procedures are time consuming and therefore can be applied only to the networks of moderate size (up to some hundreds).

The aim of the paper is to provide an overview of the blockmodeling procedures implemented in Pajek together with some examples. The basic blockmodeling procedures in Pajek are described in detail in the monograph *Exploratory Network Analysis with Pajek* (de Nooy, Mrvar, and Batagelj, 2004). An extended discussion of generalized blockmodeling can be found in the monograph *Generalized Blockmodeling* (Doreian, Batagelj, and Ferligoj, 2004). Here, only the main ideas and procedures for generalized blockmodeling are given and the basic knowledge of how to use Pajek is assumed.

[1] Faculty of Mathematics and Physics, Univ. of Ljubljana, Slovenia; vladimir.batagelj@uni-lj.si;

[2] Faculty of Social Sciences, University of Ljubljana, Slovenia; andrej.mrvar@uni-lj.si;

[3] Faculty of Social Sciences, University of Ljubljana, Slovenia; anuska.ferligoj@uni-lj.si;

[4] Department of Sociology, University of Pittsburgh, PA 15260, USA; pitpat@pitt.edu

2   Basic definitions

Let us start with some basic definitions.

Let $U = \{X_1, X_2, \ldots, X_n\}$ be a finite set of *units*. The units are related by a binary *relation*

$$R \subseteq U \times U$$

which determine a *network*

$$\mathcal{N} = (U, R)$$

The relation R can be also described by a corresponding binary matrix $\mathbf{R} = [r_{ij}]_{n \times n}$ where

$$r_{ij} = \begin{cases} 1 & X_i R X_j \\ 0 & \text{otherwise} \end{cases}$$

In some applications r_{ij} can be a nonnegative real number expressing the strength of the relation R between units X_i and X_j.

The goal of blockmodeling is to identify, in a given network, *clusters* of units that share structural characteristics defined in terms of R. The units within a cluster have the same or similar connection patterns to other units. They form a *clustering*

$$\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$$

which is a partition of the set of units U:

$$\bigcup_i C_i = U$$

$$i \neq j \Rightarrow C_i \cap C_j = \emptyset$$

As we know from the mathematics: each partition determines an equivalence relation; and conversely: each equivalence relation determines a partition. A clustering $\mathcal{C}$ partitions the relation R into *blocks*

$$R(C_i, C_j) = R \cap C_i \times C_j$$

Each such block consists of units belonging to clusters C_i and C_j and all arcs leading from cluster C_i to cluster C_j. If $i = j$, a block $R(C_i, C_i)$ is called a *diagonal* block.

A *blockmodel* consists of structures obtained by identifying all units from the same cluster of the clustering $\mathcal{C}$. For an exact definition of a blockmodel we have to be precise also about which blocks produce an arc in the *reduced graph* and which do not. The reduced graph can be presented also by its relational matrix, called the *image matrix*.

The goal of blockmodeling is to reduce a large, potentially incoherent network to a smaller comprehensible structure that can be interpreted more readily. Blockmodeling, as an empirical procedure, is based on the idea that units in a network can be grouped according to the extent to which they are equivalent, according to some *meaningful* definition of equivalence.
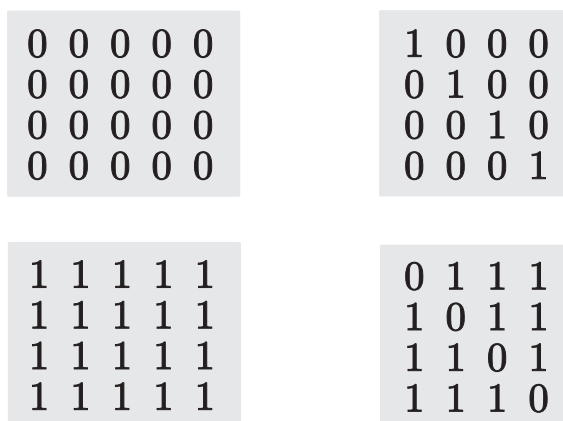
$$
\begin{array}{cccc}
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{array}
\qquad
\begin{array}{cccc}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{array}
$$

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

Figure 1: Ideal blocks for structural equivalence.

3 Equivalences

In general, and without surprise, different definitions of equivalence lead to distinct partitions. Lorrain and White (1971) introduced the first equivalence called structural equivalence. Later White and Reitz (1983) defined regular equivalence.

3.1 Structural equivalence

Units are structurally equivalent if they are connected to the rest of the network in identical ways (Lorrain and White, 1971).

Formally: X and Y are *structurally equivalent* (in the matrix form: $X_i \equiv X_j$) iff

s1.	$XRY \Leftrightarrow YRX$	s1'.	$r_{ij} = r_{ji}$
s2.	$XRX \Leftrightarrow YRY$	s2'.	$r_{ii} = r_{jj}$
s3.	$\forall Z \in U \setminus \{X, Y\} : (XRZ \Leftrightarrow YRZ)$	s3'.	$\forall k \neq i, j : r_{ik} = r_{jk}$
s4.	$\forall Z \in U \setminus \{X, Y\} : (ZRX \Leftrightarrow ZRY)$	s4'.	$\forall k \neq i, j : r_{ki} = r_{kj}$

The matrix form of the definition of structural equivalence can be extended also to the case where r_{ij} are real numbers.

From the definition of structural equivalence it follows that there are four possible ideal blocks (Batagelj, Ferligoj, and Doreian, 1992): *null* (all 0) block and *complete* (all 1) block; and for diagonal blocks also null block with 1's on the diagonal and complete block with 0's on the diagonal. They are presented in Figure 1.

3.2 Regular equivalence

Intuitively, two units are regularly equivalent if they are equally connected to equivalent others (White and Reitz, 1983).

The equivalence relation $\approx$ on U is a *regular equivalence* on network $\mathcal{N} = (U, R)$ iff for all $X, Y, Z \in U$, $X \approx Y$ implies both

R1. $XRZ \Rightarrow \exists W \in U : (YRW \wedge W \approx Z)$
R2. $ZRX \Rightarrow \exists W \in U : (WRY \wedge W \approx Z)$

$$
\begin{array}{ccccc}
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{array}
\qquad
\begin{array}{ccccc}
1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0
\end{array}
$$

Figure 2: Ideal blocks for regular equivalence.

Batagelj, Doreian, and Ferligoj (1992) proved that regular equivalence produces two types of blocks: null blocks and 1-covered blocks (which have in each row and in each column at least one 1). They are presented in Figure 2.

3.3 Generalized concepts of equivalence

An appropriate generalization of the equivalence idea is one where each block, of a particular partition, is free to conform to a different equivalence idea. This led Batagelj (1997) and Doreian, Batagelj, and Ferligoj (1994) to the definition of several types of connection inside and between the clusters, or in another words, different types of ideal blocks. Some of them are presented in Table 1. All these types of blocks are implemented in Pajek. For a given partition they can be used also for large networks (`Operations/ Shrink Network/ Partition`).

4 Establishing blockmodels

The problem of establishing a partition of a network in terms of a considered equivalence is a special case of **clustering problem** that can be formulated as an optimization problem: determine the clustering $\mathcal{C}^*$ for which

$$
P(\mathcal{C}^*) = \min_{\mathcal{C} \in \Phi} P(\mathcal{C})
$$

where $\mathcal{C}$ is a clustering of a given *set of units or units U*, Φ is the set of all possible clusterings and $P : \Phi \to \mathbb{R}$ the *criterion function*. The criterion function must reflect the considered equivalence.

Criterion function can be constructed

- **indirectly** as a function of a compatible (dis)similarity measure between pairs of units, or

- **directly** as a function measuring the fit of real blocks induced by a given clustering to the corresponding ideal blocks with perfect relations within each cluster and between clusters according to the considered types of connections (equivalence).

Table 1: Generalized types of blocks.

0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Null

0	1	0	0	0
1	1	1	1	1
0	0	0	0	0
0	0	0	1	0

Row-Dominant

0	0	1	0	0
0	0	1	1	0
1	1	1	0	0
0	0	1	0	1

Col-Dominant

1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

Complete

0	0	0	1	0
0	0	1	0	0
1	0	0	0	0
0	0	0	1	0

Row-Functional

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	0
0	0	0	1

Col-Functional

0	1	0	0	0
1	0	1	1	0
0	0	1	0	1
1	1	0	0	0

Regular

0	1	0	0	0
0	1	1	0	0
1	0	1	0	0
0	1	0	0	1

Row-Regular

0	1	0	1	0
1	0	1	0	0
1	1	0	1	1
0	0	0	0	0

Col-Regular

4.1 The Indirect approach

In the indirect approach the most important requirement is that the selected dissimilarity measure is *compatible* with the considered equivalence (Batagelj, Ferligoj, and Doreian, 1992). The dissimilarity measure d is compatible with a considered equivalence $\sim$ if for each pair of units holds

$$X_i \sim X_j \Leftrightarrow d(X_i, X_j) = 0$$

Not all dissimilarity measures used in applications are compatible with structural equivalence. For example, the corrected Euclidean-like dissimilarity

$$d(X_i, X_j) = \sqrt{(r_{ii} - r_{jj})^2 + (r_{ij} - r_{ji})^2 + \sum_{\substack{s=1 \\ s \neq i,j}}^{n} ((r_{is} - r_{js})^2 + (r_{si} - r_{sj})^2)}$$

where each unit is described by the vector composed of the corresponding row and column of relational matrix, is compatible with structural equivalence.

The indirect clustering approach does not seem suitable for establishing clusterings in terms of regular equivalence since there is no evident way how to construct a compatible (dis)similarity measure.
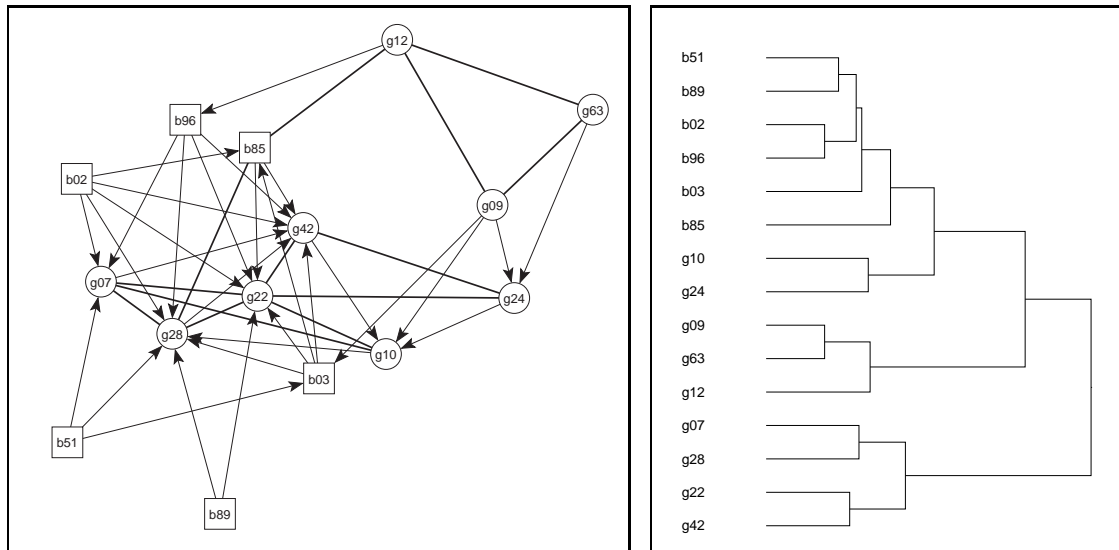
Figure 3: Example network – borrowing learning material (left), dendrogram (right).

In Pajek, the corrected Euclidean-like dissimilarity and several others (Batagelj, Ferligoj, and Doreian, 1992) can be found (`Operations/Dissimilarity`). Also, a hierarchical agglomerative procedure is implemented with several clustering methods (`Net/Hierarchical Decomposition/Clustering/Options`): minimum, maximum, average, general, Ward and Ward[2].

4.2   An Example of the Indirect approach

The blockmodeling procedures implemented in Pajek are demonstrated on a student network. It consists of 15 undergraduate students attending lectures on Social network analysis at Faculty of Social Sciences, University of Ljubljana (2002). Students were asked: *from whom would you borrow learning materials*. The number of choices was not limited. The obtained network is presented on the left side of Figure 3. Circles represent girls with squares representing boys. Bidirectional arcs are replaced by edges. The layout was obtained by the Kamada-Kawai spring embedder procedure implemented in Pajek.

Using the Corrected Euclidean-like dissimilarity and the Ward clustering method we obtain the dendrogram presented on the right side of Figure 3.

In the indirect approach, the dendrogram is used to help identify the number of clusters in a partition. Using the dendrogram to select three clusters we get the partition presented on the left side of Figure 4. (If we had chosen two clusters, then the second pair of clusters would have been merged.) The partition into three clusters is shown using greyscale (white, gray and black vertices).

The obtained partition can also be used to reorder rows and columns of the matrix representing the network. Clusters are delimited using vertical and horizontal lines (see the right side of Figure 4).
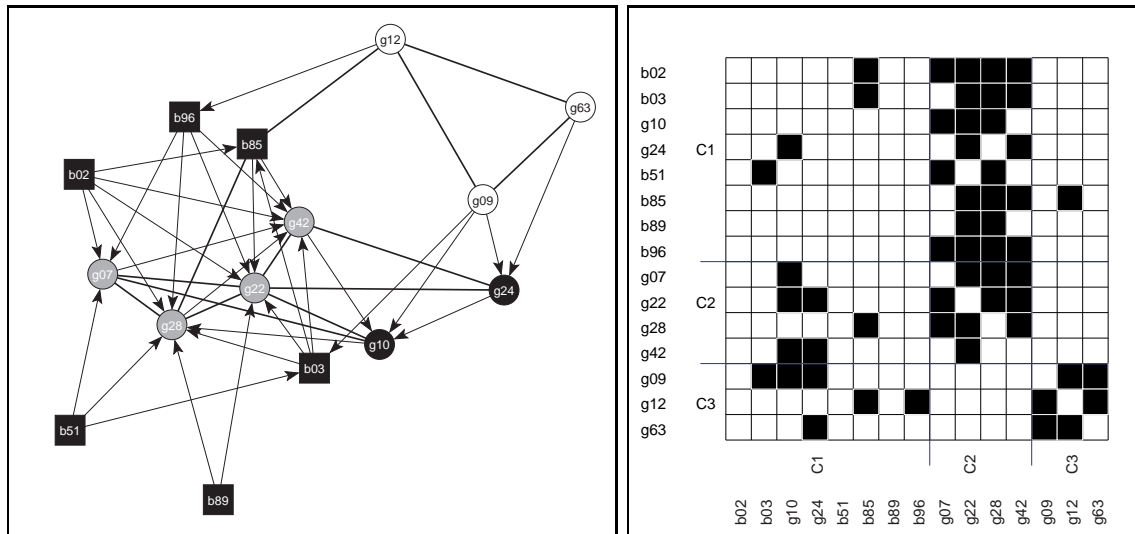
Figure 4: Indirect approach – partition into three clusters (left), reordered matrix (right).

4.3   The Direct approach

The second possibility for establishing blockmodels is to construct an appropriate criterion function directly and then use a optimization algorithm to obtain a clustering solution. Criterion function $P(\mathcal{C})$ has to be **sensitive** to the considered equivalence:

$$P(\mathcal{C}) = 0 \Leftrightarrow \mathcal{C} \text{ defines considered equivalence.}$$

One of the possible ways of constructing a criterion function that directly reflects the considered equivalence is the following one: given a clustering $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$, let $\mathcal{B}(C_u, C_v)$ denote the set of all ideal blocks corresponding to block $R(C_u, C_v)$. Then the global error of clustering $\mathcal{C}$ can be expressed as

$$P(\mathcal{C}) = \sum_{C_u, C_v \in \mathcal{C}} \min_{B \in \mathcal{B}(C_u, C_v)} d(R(C_u, C_v), B)$$

where the term $d(R(C_u, C_v), B)$ measures the difference (error) between the block $R(C_u, C_v)$ and the corresponding ideal block B. The function d has to be compatible with the selected type of equivalence.

It seems that most of the clustering problems obtained using the direct approach are NP-hard. To solve them we can use one of the local optimization procedures (Batagelj, Doreian, and Ferligoj 1992). In Pajek the relocation algorithm is implemented:

Determine the initial clustering $\mathcal{C}$;
   while in the neighborhood of the current clustering $\mathcal{C}$
       there exists a clustering $\mathcal{C}'$ such that $P(\mathcal{C}') < P(\mathcal{C})$
   do move to clustering $\mathcal{C}'$.

The neighborhood in this local optimization procedure is determined by the following two transformations:

- *moving* a unit X_k from cluster C_p to cluster C_q (*transition*);

- *interchanging* units X_u and X_v from different clusters C_p and C_q (*transposition*).

In Pajek the direct approach is available through the menu option `Operations/ Blockmodeling`. Two suboptions are available. In the first one the initial partition is randomly generated by Pajek. In the second one the initial partition has to be given by the user or obtained by some other approach (e.g., by indirect approach). In Pajek structural equivalence and regular equivalence are available as options while a generalized equivalence can be specified by selecting the option `User Defined`. In each cell of the given image matrix the user can specify the allowed types of blocks. The number of repetitions of the local optimization procedure should be also specified by the user.

4.4   Example of the direct approach

The clustering into three clusters of undergraduate students obtained by the direct approach and assuming structural equivalence is exactly the same as the one obtained by indirect approach above. The solution has 28 inconsistencies with a perfect structural equivalence partition. This is minimized value of the magnitude of the criterion function. It is determined as a part of the procedure and is used as a measure of fit. The command sequence is given by (`Operations/Blockmodeling/Random Start`). This leads to a dialogue box in which the equivalence type, the number of clusters, and the number of repetitions are selected.

5   Fitting a pre-specified blockmodel

So far the inductive approaches for establishing blockmodels for a given relation defined over a set of units were considered. Some form of equivalence is specified and clusterings are sought that are consistent with a specified equivalence. Nothing else is specified.

Another view of blockmodeling is deductive in the sense of starting with a blockmodel that is specified in terms of substance prior to an analysis. Batagelj, Ferligoj, and Doreian (1998) presented methods where a set of observed relations are fitted to a pre-specified blockmodel. See also Doreian, Batagelj,and Ferligoj (2004) for an extended discussion of specifying and fitting a variety of pre-specified blockmodels. Given a network, a set of types of ideal blocks, and a reduced model, a solution (partition) can be determined which minimizes the criterion function. It is also possible to fit the network to a partial model and analyze the residual afterward; or to introduce different constraints on the model. For example: units X and Y are of the same type (in the same cluster); or, types of units X and Y are not connected.

In Pajek the pre-specified blockmodel can be described by specifying the cells of image matrix (option `User Defined`). Additional constraints can also be specified but not in a user friendly way using special model files (Batagelj, 1996).

5.1   Example - prespecified block model

In the case of the student network we can consider a core-periphery model: there are diligent students (the core group) with good learning materials and there are less motivated students (the periphery group) who borrow studying material from the students in core group and not from students in their own group. Therefore, our pre-specified blockmodel is:

	C_1	C_2
C_1	com, reg	-
C_2	com, reg	-

In the table `com` stands for complete block, `reg` stands for regular block, and `-` stands for null block.

Using the local optimization algorithm implemented in Pajek we obtained the partition into two clusters presented on the left side of Figure 5. The corresponding reordered matrix is presented on the right side of the figure.
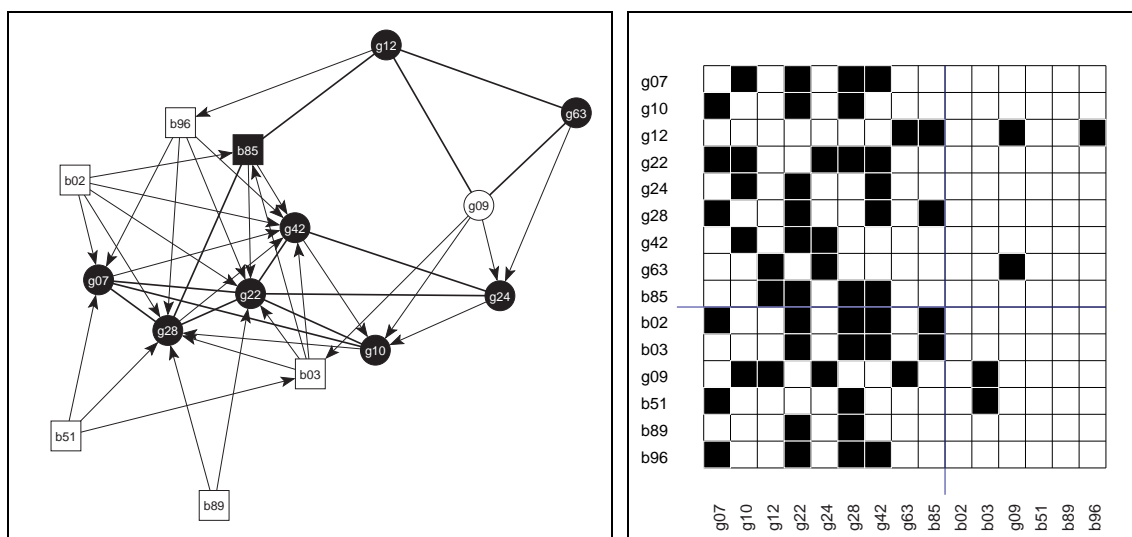

Figure 5: Prespecified blockmodeling – core-periphery (left), reordered matrix (right).

Using Pajek the following image matrix and error matrix with an inconsistency score of 5 were obtained:

	C_1	C_2
C_1	reg	-
C_2	reg	-

	C_1	C_2
C_1	0	3
C_2	0	2

The obtained model fits the network very well: there are 5 inconsistencies in the blocks that were supposed to be null and no inconsistencies in the regular blocks. Although we have used a single network for illustrative purposes, we emphasize that we do not intend a direct comparison of the solutions obtained. We used structural equivalence merely to illustrate the direct and indirect methods. In this section our

focus is on pre-specifying blockmodels. Although we prefer the latter model it is *not* because the number of inconsistencies is smaller. (Indeed, if we use regular equivalence, select two clusters and use the direct approach, we obtain the same partition with the same count of inconsistencies. Also, structural equivalence is stringent and higher inconsistency counts are expected with such models. Comparisons of inconsistency counts across different types of blockmodels are inappropriate.) Our preference is grounded in the idea of specifying something specific about the structure of the blockmodel and testing it. (It is possible to fit a 3-cluster core-periphery model - but as the substantive interpretation is the same, we stay with the simpler 2-cluster model.)

The obtained results are:

- Each student from the core group can find at least one student from the core group from whom (s)he can borrow learning material.

- Each student from the core group is asked from at least one student of the core group for the learning material.

- Each student from the periphery group can borrow learning material from at least one student from the core group.

- Each student from the core group is asked from at least one student from the periphery group for the learning material.
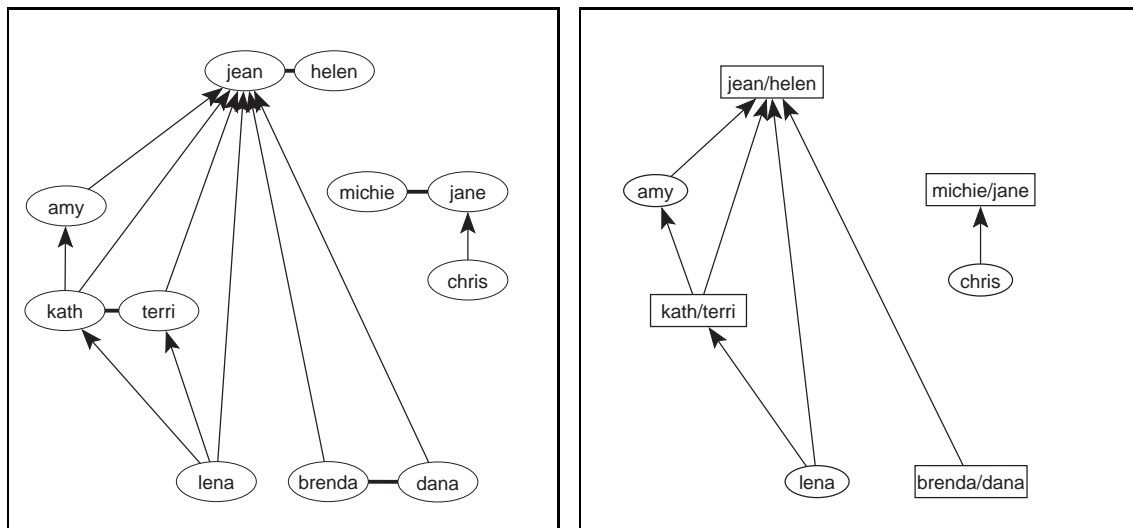

Figure 6: Ranked-clusters blockmodels: Network (left), shrunk network (right).

5.2 Examples of Ranked-Clusters Models

The ideas behind the ranked-clusters model are outlined in Doreian, Batagelj, and Ferligoj (2000), together with the closely related idea of the cyclic-acyclic decomposition of networks. Intuitively put, the structure of these models is one where
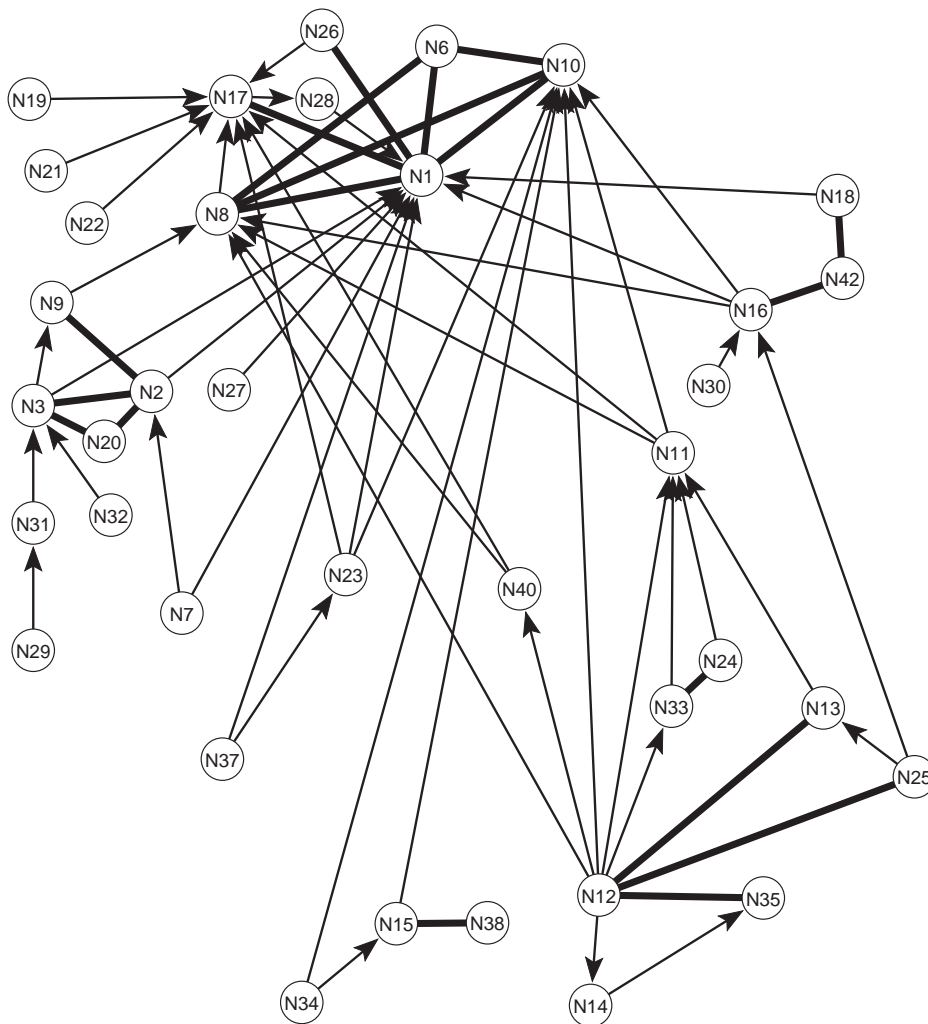
Figure 7: Ranked-clusters model of an interorganizational network.

the diagonal blocks have only symmetric ties and the asymmetric ties goes either up the (hierarchical structure) or down the structure but not both. Figure 6 shows the strong friendship choices of an elite young women's soccer term (with fictitious names). It is drawn with the asymmetric ties going up the structure with the more popular women above the less popular women. On the left is the network of ties and on the right is the shrunken network where the vertices in each diagonal block have been shrunk to a single vertex.

Of course, the structure shown in Figure 6 could be obtained by a visual inspection of the network once it is drawn. A more complex example is given in Figure 7 where the network has environmental social movement organizations in the Milan area (Diani, 1995) and the tie is 'works with'. Surprisingly, it has a ranked-clusters structure with only 8 inconsistencies (all in the form of asymmetric ties being located within diagonal blocks/clusters).

6   Conclusion

The basic blockmodeling procedures are discussed in detail in *Exploratory Network Analysis* with Pajek (de Nooy, Mrvar, and Batagelj, 2004). In this paper we present recent developments of generalized blockmodeling procedures as developed by Batagelj, Doreian and Ferligoj and presented in *Generalized Blockmodeling* (Doreian, Batagelj, and Ferligoj, 2004) and implemented in Pajek.

There are still several algorithms proposed by the group to be implemented in Pajek e.g., blockmodeling of two-mode networks (Doreian, Batagelj, and Ferligoj, 2004) and blockmodeling of multiple networks. Also the user-friendly interface for providing additional constraints in pre-specified blockmodels is planned to be implemented in Pajek.

References

[1] Batagelj, V. (1993): Notes on block modelling. In *Abstracts and Short Versions of Papers*, 3rd European Conference on Social Network Analysis, München: DJI, 1-9.

[2] Batagelj, V. (1996): *Model2 - Program for Generalized Prespecified Blockmodeling*. Manual. Ljubljana.

[3] Batagelj, V. (1997): Notes on blockmodeling. *Social Networks*, **19**, 143-155.

[4] Batagelj, V., Doreian, P., and Ferligoj A. (1992): An optimizational approach to regular equivalence. *Social Networks*, **14**, 121-135.

[5] Batagelj, V., Ferligoj, A., and Doreian, P. (1992): Direct and indirect methods for structural equivalence. *Social Networks*, **14**, 63-90.

[6] Batagelj, V., Ferligoj, A., and Doreian, P. (1998): Fitting pre-specified blockmodels. In C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, and Y. Baba (Eds.): *Data Science, Classification, and Related Methods*. Berlin: Springer, 199-206.

[7] Batagelj, V. and Mrvar, A. (1998): Pajek – A program for large network analysis. *Connections*, **21**, 47-57.

[8] Batagelj, V. and Mrvar, A. (2002): Pajek – Analysis and visualization of large networks. *Lecture Notes in Computer Science*, **2265**. Berlin: Springer, 477-478.

[9] Batagelj, V. and Mrvar, A. (2003): Pajek - analysis and visualization of large networks. In Juenger, M. and Mutzel, P. (Eds.): *Graph Drawing Software*. Berlin: Springer, 77-103.

[10] Batagelj, V. and Mrvar, A. (2004): Pajek home page
`http://vlado.fmf.uni-lj.si/pub/networks/pajek/`

[11] Batagelj, V., Mrvar, A., and Zaveršnik, M. (1999): Partitioning approach to visualization of large graphs. *Lecture Notes in Computer Science*, **1731**, 90-97. Springer-Verlag.

[12] Borgatti, S.P. and Everett, M. G. (1992): Notions of positions in social network analysis. In P. V. Marsden (Ed.): *Sociological Methodology.* San Francisco: Jossey Bass. 1-35

[13] Doreian, P., Batagelj, V. and Ferligoj, A. (1994): Partitioning networks on generalized concepts of equivalence. *Journal of Mathematical Sociology*, **19**, 1-27.

[14] Doreian, P., Batagelj, V. and Ferligoj, A. (2000): Symmetric-Acylic Decompositions of Networks. *Journal of Classification*, **17**, 3-28.

[15] Doreian, P., Batagelj, V. and Ferligoj, A. (2004): Generalized blockmodeling of two-mode network data.. *Social Networks*, **26**, 29-53.

[16] Doreian, P., Batagelj, V., and Ferligoj, A. (2004): *Generalized Blockmodeling.* New York: Cambridge University Press. (Forthcoming)

[17] Hummon, M.P. and Carley, K. (1993): Social networks as normal science. *Social Networks*, **14**, 71-106.

[18] Lorrain, F. and White, H. C. (1971): Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, **1**, 49-80.

[19] de Nooy, W., Mrvar, A., and Batagelj, V. (2004): *Exploratory Network Analysis with Pajek.* New York: Cambridge University Press. (Forthcoming)

[20] Diani, M. (1995): *Green Networks: A Structural Analysis of the Italian Environmental Movement.* Edinburgh: Edinburgh University Press.

[21] White, D. R. and Reitz, K.P. (1983): Graph and semigroup homomorphisms on networks of relations. *Social Networks*, **5**, 193-234.