SAND2011-7065J

Generalized Blockmodeling of Multiple Valued Networks

Dean Jones, Jose Saloio, Nazir Khalil and Nathanael J.K. Brown
Sandia National Laboratories
PO Box 5800
Albuquerque, NM 87185-1138, USA

Linda Nozick*
College Program in Systems Engineering
School of Civil and Environmental Engineering
Cornell University
Ithaca, NY 14853, USA
Phone: (607)255-6496
Fax: (607)255-9004

Abstract

This paper presents an extension to generalized blockmodeling where there are more than two types of objects to be clustered based on valued network data. We use the ideas in homogeneity block modeling to develop an optimization model to perform the clustering of the objects and the resulting partitioning of the ties so as to minimize the inconsistency of an empirical block with an ideal block. The ideal block types used in this modeling are null, complete and a new type that is related to that developed in Žiberna (2007). Three case studies are presented, two based on the Southern Women dataset (Davis et al. 1941) and a third based on passenger air travel in the Continental United States.

Keywords: Blockmodeling; Generalized blockmodeling; Valued network; Regular equivalence

1. Introduction

The goal of blockmodeling is the identification of clusters of objects and the partitioning of the set of ties between those objects into blocks (Lorrain and White, 1971, Breiger et al. 1975, Burt 1976). Much of the early work in blockmodeling focused on structural equivalence as the basis for partitioning. White and Reitz (1983) extended that basis to include regular equivalence. Batageli et al. (1992) suggested using the network data directly to perform this clustering and partitioning rather than summarizing the network information into similarity or dissimilarity matrices and appealing to a generic clustering algorithm. Generalized blockmodeling extends these concepts to include a wide array of block types and the explicit use of optimization to perform the partitioning on the network data directly.

The vast majority of the generalized blockmodeling literature focuses on a single matrix where all the entries are binary. Research including Breiger et al. (1975), Batagelj et al.(1992), Breiger and Mohr (2004), Žiberna (2007) extend these ideas to valued matrices allowing for the representation of the

* Corresponding author

strength of ties. While much of the literature focuses on matrices for which the row and column objects are the same, several authors extend those ideas to matrices for which the row and columns refer to different types of objects, thereby representing two-mode network data.

Our focus is the extension of value based generalized blockmodeling tools (for both one and two-mode network data) to identify the underlying structure when more than two types of objects are under consideration and where information on the strength of the ties can be included. Figure 1 gives an illustration of our focus. In this example there are two types of objects, people (labeled 1, 2, and 3) and locations (labeled A and B). The goal of the tool described in this paper is to support the development of conclusions of the nature "individuals 1 and 2 form a group and that group is associated with location A; whereas, individual 3 is a singleton group and associated with location B". Notice that we could illustrate the first matrix as a one-mode social network and the second network as a two-mode social network and apply existing blockmodeling tools to each matrix separately. Our focus is the simultaneous analysis of multiple networks of both types.

	1	2	3
1		2	0
2	2		1
3	0	1	

	A	B
1	2	0
2	3	1
3	0	2

Figure 1. Example Networks.

The next section develops an explicit optimization model to determine these groups (and blocks). The third section describes a solution procedure for that model. The fourth section applies the model and solution procedure to several illustrative examples. The fifth section gives opportunities for future research.

2. Model Formulation

A key element in the development of the mathematical model to identify these groups is the development of a criterion function. The criterion function provides a mechanism to understand the degree of inconsistency of a block with an ideal block. Remember, when objects are partitioned into groups we can examine the nature of the interaction of one group with another by considering the relevant block formed by the rows of one of the groups and the columns associated with the other. Suppose for each matrix m, an ideal block either has entries which are each equal to or below some value d_m or equal to or above some value e_m and $e_m > d_m$. Based on this definition, we can compute the inconsistency of that block from either ideal and simply assume the ideal that generates the smallest amount of inconsistency is appropriate.

Consider the example in Figure 1. Grouping individuals 1 and 2 together and leaving individual 3, and locations A and B each as singleton sets, provides several interesting blocks to assess each for their level

of inconsistency from an ideal. The intersection of individuals 1 and 2 with location A, based on the right-hand matrix in Figure 1, creates a 2x1 vector that indicates that individuals 1 and 2 visited location A twice and three times, respectively. Suppose d_m is 1 and e_m is 2. Perhaps the two ideals could be characterized as the presence or absence of an association between individuals 1 and 2 and location A. Therefore, the inconsistency from the ideal of no association is then 2-1+3-1=3 whereas the inconsistency from the ideal of an association is 0 (since both entries are equal to or greater than 2).

It is useful to notice that if we have a binary matrix and set d_m equal to 0 and e_m equal to 1, these inconsistency computations match that commonly used for binary blockmodeling. These computations for block inconsistency are similar to those given by Žiberna (2007) except we allow for a different critical value to distinguish "associated" from "not associated". They use restrictions that certain values must be zero, a user-defined value or some function of the entries in the row (or column) must be at least some value.

Suppose there are N types of objects to cluster and m matrices to support that clustering. Also, let $\lambda(m)$ equal 0 if the rows and columns of matrix m correspond to the same object type and one otherwise. We assume that if the interactions, as given by the matrix, are between objects of the same type, the ideal for the level of interaction between objects in the same cluster is defined by e_m whereas the ideal for the level of interaction between objects in different clusters should be defined by d_m. When the rows and columns of the matrix correspond to different types of objects we make no assumption as to which ideal is correct. For the applications to be supported by this formulation, we commonly expect one of the object types to be individuals. Since modeling the interactions among individuals is so important in social network analysis, we provide the capability to assume an ideal for interactions between objects of the same type.

Let $m(r)$ and $m(c)$ be the clusters associated with the row and column objects, respectively. Therefore, the goal of the clustering is to minimize the following objective.

$$\sum_{m|\lambda(m)=0}\left[\sum_{\substack{m(r)\\m(c)|m(r)=m(c)}}\sum_{\substack{i\epsilon m(r)\\j\epsilon m(c)\\i\neq j}} max\left(e_m - m_{ij}, 0\right) + \sum_{\substack{m(r)\\m(c)|m(r)\neq m(c)}}\sum_{\substack{i\epsilon m(r)\\j\epsilon m(c)}} max\left(m_{ij} - d_m, 0\right)\right]$$
$$+\sum_{m|\lambda(m)=1}\sum_{\substack{m(r)\\m(c)}}\min\left[\sum_{\substack{i\epsilon m(r)\\j\epsilon m(c)}} max\left(e_m - m_{ij}, 0\right), \sum_{\substack{i\epsilon m(r)\\j\epsilon m(c)}} max\left(m_{ij} - d_m, 0\right)\right] \quad (1)$$

where m_{ij} is the entry in the (i,j) position in matrix m. The first term in the objective is the penalty associated with matrices that have the same objects for the rows and columns. The second term focuses on matrices for which the object type differs between the rows and the columns. Let's focus on the first component of the first term. There are two situations which are considered and those situations are represented by the first and second components in the brackets, respectively; namely when the cluster for the row objects is the same as the cluster for the column objects and when they are different. When they are the same, we assume that the ideal is the higher level of interaction given by parameter e_m. Therefore, for all entries in the block that represents the cluster with itself, between unique objects, we simply take the maximum of e_m minus the entry and zero. If the interaction is equal to or higher than the minimum

given by the ideal, the penalty is assumed to be zero. If the interaction is lower than this minimum, a positive penalty is assessed. When the cluster for the rows is different than that for the columns, the ideal interaction is at the level of d_m or lower. Hence we take the entry for each pair of objects, one from each cluster and subtract the "allowable" level of interaction. If this interaction is exceeded, a penalty is assessed. It is useful to notice that this penalty structure can be considered to be a generalization of that described by Žiberna (2007).

The second terms focuses on matrices for which the objects that comprise the rows and columns are different. When they are different we must test which ideal is closer to the entries in the block. That is, is the ideal associated with e_m or that associated with d_m a better representation for the interaction between the clusters? Hence, this second term requires the minimum function. The first component within the brackets for this second term computes the penalty if the ideal for the interaction is an association and the second term computes the penalty if the ideal for the interaction is the absence of an association. The minimum function simply selects the penalty to apply for the ideal that is closest to the block values.

The minimum and maximum functions can be replaced by additional variables. However, we do not do this substitution in the interest of clarity and brevity. This conversion is not needed by the solution procedure either; which is a tabu search and described in the next section.

Suppose the set of object types is indexed by n ($n=1,\ldots N$), the set of objects of type n is indexed by t_n ($t_n=1_n,\ldots T_n$) and the set of clusters which contain objects of type n is indexed by k_n ($k_n=1_n,\ldots K_n$). Each object must belong to one and only one cluster where that cluster only contains objects of that type. This restriction is given by the following equation.

$$\sum_{k_n} \gamma_{nk_nt_n} = 1 \qquad \forall n, t_n \tag{2}$$

where $\gamma_{nk_nt_n}$ is a binary variable that is one if object t_n of object type n belongs to cluster k_n and zero otherwise. Notice that this formulation assumes that the number of clusters for each object type is known. It also allows clusters to be empty, if that leads to a better objective value.

To illustrate this formulation, consider the example given in Figure 1. In this example there are two matrices so m ranges from 1 to 2. There are also two types of objects: people and locations, where n is 1 for people and 2 for locations. Suppose we may have up to 2 clusters of people and 2 clusters of locations. Further suppose that e_1 is 2 and d_1 is 1. That is, the ideal for communication between people within the same cluster is 2 or more and the ideal for communication between people in different clusters is 0 or 1. Finally, suppose that e_2 is 2 and d_2 is 1; ideally people that are associated with a location visit that location at least twice and people not associated with a location ideally no more than once.

Now, suppose individuals 1 and 2 are in one cluster and individual 3 is in another cluster. Further, suppose each facility forms a singleton cluster. The computations associated with the objective function for this grouping are as follows. First, consider the first matrix (communication between individuals). There are four blocks for which the penalty stemming from this clustering of individuals is needed. Two blocks are associated with the first component in the first term in the objective (cluster 1 with itself and cluster 2 with itself) and two with the second component in this same term (cluster 1 with cluster 2 and

cluster 2 with cluster 1). The penalty associated with the block formed by cluster 1 with itself is zero because the one pair of individuals has a level of communications which is equal to the minimum allowed as given by e_m. The penalty formed by cluster 2 with itself is zero for the same reason. The penalty associated with the block formed by cluster 1 with cluster 2 is zero because the communication between individual 3 and individuals 1 and 2 does not exceed the maximum allowed as given by d_m. Similarly, the cluster 2 to cluster 1 block also produces no penalty.

3.0 Solution Procedure

This section describes a tabu search algorithm to solve the optimization problems described above. A key element of defining a tabu search algorithm is to identify what constitutes a neighborhood for a solution *S*, $\pi(S)$, where a solution is a mapping of each object to a single cluster for that type of object. Our assumption is that a neighboring solution is exactly the same except a single object has moved from one cluster to another cluster. Rather than investigating all solutions in the current neighborhood, we compute the contribution to the objective for each object based on its current cluster assignment. The objects of each type are then rank ordered within there respective types. The user then provides the percent of swaps to consider for each type of object to evaluate the neighborhood. For example, the user might say for objects of type 1 explore swaps for the top 10% of objects as ranked by their contribution to the objective. For simplicity let *PerExp* be the percentage of objects of each type to conduct swaps to explore a neighborhood with the understanding that this percentage could vary by object type.

In order to reduce the likelihood of cycling we maintain a tabu list of moves that have occurred over the last tabu tenure (TL) iterations. The entries on this list are simply a list of the objects and the clusters they have moved from and into. This allows us to create rules based on this information that minimize the chance of cycling.

This algorithm is a random, multi-start algorithm. That is, we execute the algorithm with a solution picked at random a fixed number of times (MaxIter). For each iteration, we keep the best solution found. The solution reported is then the best identified over all iterations. The algorithm is then as given below.

Algorithm (minimizing)

1. Specify the maximum number of clusters for each type of object. Also set the number of trial initial solutions, *MaxIter*. Let *I = 1* and be the counter for the number of times the tabu search algorithm has been invoked. Set the tabu tenure *TL* (we have assumed 7 but this is a parameter that might need to be tuned). Set the maximum number of tabu search steps TabuMaxIter.

2. Randomly assign objects to clusters for that object to obtain an initial solution $S = S_I^*$ Evaluate the objective value for the solution, let that objective value be *BestObj*, the best objective value found as of this point in this trial and S_I^* be the solution associated with that objective function value (S_I^* will always be the solution associated with the best objective function value found to date in trial *I*).

3. Start the tabu algorithm to improve the quality of the solution identified in Step 2.

3a. Initialize the tabu list to null. Let *TabuIter = 1.*

3b. Evaluate the objective value of each solution in the *PerExp* portion of the neighborhood to explore and find the best $S' \in \pi(S)$. Go to 3c.

3c. Update the best solution found for this trial based on which case this new solution *S'* falls into.

Case 1. If objective value of solution (S') >= BestObj for this trial and the move from S to S' is not in the tabu list, then let $S' = S$, update tabu list and let *TabuIter = TabuIter + 1*. Go to 3b. Notice that this implies that all neighboring solutions are worse than the best solution found to date for this random start for trial *I*. We accept this inferior solution by making $S' = S$, but we do not accept this solution as the best found to date in this trial, S_I^*,with the hope that it will lead to better solutions through future moves.. Notice that we do not update the value *BestObj* or S_I^* but we do move to this inferior solution.

Case 2. If objective value of solution (S') < BestObj, let $S' = S_I^*$(update the best solution), $S' = S$, update the tabu list, set BestObj equal to the objective value of the solution (S'), and let *TabuIter = TabuIter + 1*. Go to 3b. Notice that this case refers to the identification of a neighboring solution that is better, so we accept this superior solution, even if the move is on the tabu list. If the move is on the tabu list, we increase the risk of cycling by accepting this solution, but likely that is worth the improvement in solution quality.

Case 3. If objective value of solution (S') >= BestObj and the move from S to S' is in the tabu list, then record the best tabu objective (BestObj) and solution (S_I^*) for this initial solution and go to step 4. This implies that we terminate this trial. If we were to except this move to S' we would be accepting an inferior solution which is "undoing" a previous move and could lead to cycling.

3d . If TabuIter < TabuMaxIter, go to 3b. Otherwise, go to step 4.

4. Update the number of iterations and check the stopping rule: *I = I + 1*. If *I ≤ MaxIter*, go to step 2. Otherwise, compare the best objective values from the different initial solutions, S_I^* , and report the one with the best objective value.

4.0 Illustrative Examples

In this section we focus on several examples, each with different characteristics. First, we focus on a single matrix for which the columns and rows represent different types of objects, and the relationship between each pair of objects (where each object in the pair is of a different type) is binary. This example demonstrates how the above modeling approach can be used to perform generalized block modeling of

two-mode network data. Next we focus on a matrix of one-mode network data but for which the relationships are valued. This example illustrates the use of the modeling approach to perform valued generalized block modeling. Finally, we turn to an example which involves three types of objects for which one type of object forms the rows of one matrix and the columns of another. This example illustrates how blockmodeling can be used to simultaneously address matrices of different structure.

Our first and second examples are based on the dataset described in Davis et al. (1941) focused on Southern Women and their participation in social events. The mapping of individuals to the events they attended is given in Table 1, where a one indicates attendance at the event and a zero indicates that the person did not attend the event. For an interesting and detailed discussion of this data set see Freeman (2003).

Table 1. Matrix of the Participation of Women in Social Events (Davis et al. 1941).

Individual	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
Evelyn	1	1	1	1	1	1	0	1	1	0	0	0	0	0
Laura	1	1	1	0	1	1	1	1	0	0	0	0	0	0
Theresa	0	1	1	1	1	1	1	1	1	0	0	0	0	0
Brenda	0	0	1	1	1	1	1	1	0	0	0	0	0	0
Charlotte	0	0	1	1	1	0	1	0	0	0	0	0	0	0
Frances	0	0	1	0	1	1	0	1	0	0	0	0	0	0
Eleanor	0	0	0	0	1	1	1	1	0	0	0	0	0	0
Ruth	0	0	0	0	1	0	1	1	1	0	0	0	0	0
Verne	0	0	0	0	0	0	1	1	1	0	0	1	0	0
Myra	0	0	0	0	0	0	0	1	1	1	0	1	0	0
Katherine	0	0	0	0	0	0	0	1	1	1	0	1	1	1
Sylvia	0	0	0	0	0	0	1	1	1	1	0	1	1	1
Nora	0	0	0	0	0	1	1	0	1	1	1	1	1	1
Helen	0	0	0	0	0	0	1	1	0	1	1	1	0	0
Olivia	0	0	0	0	0	0	0	0	1	0	1	0	0	0
Flora	0	0	0	0	0	0	0	0	1	0	1	0	0	0
Pearl	0	0	0	0	0	1	0	1	1	0	0	0	0	0
Dorothy	0	0	0	0	0	0	0	1	1	0	0	0	0	0

Freeman (2003), after employing a range of analytical procedures, concluded that the data suggest that the women should be partitioned into two groups where membership in the first group is Evelyn, Laura, Theresa, Brenda, Charlotte, Frances, Eleanor, Pearl and Ruth with the remainder in the second group. Freeman (2003) also presents a consensus analysis using 21 procedures. All 21 procedures suggested that all pairs of women from the set Evelyn, Laura, Theresa, Brenda, Charlotte and Frances belonged together. Further, they also suggested that all pairs of women from the set Myra, Katherine, Sylvia, Nora and Helen also belonged together. Freeman (2003) does not suggest a partitioning of the events.

We apply the formulation and solution procedure described in the previous two sections to this dataset when the maximum number of person clusters is two and the maximum number of event clusters is three. Table 2 illustrates the suggested clustering. The number of inconsistencies associated with this solution is 51. That is, of the 252 entries in the matrix in Figure 1, 51 are not consistent with the clustering suggested by the model. The key difference in the assignment of women to clusters suggested by this model and that discussed in Freeman (2003) is that Pearl and Ruth are part of the second cluster based on this model. This assignment stems from the fact that women in cluster one are associated with events E3-E7 but Ruth and Pearl each, only attended one of those events. The second cluster of women are

associated with the second cluster of events (E8 and E9), which both Pearl and Ruth also attend. In addition, the first cluster of women is also associated with these events.

Table 2. Matrix of Southern Women Data with 2 Clusters for Women and 3 for Events.

Individual	E3	E4	E5	E6	E7	E8	E9	E1	E2	E10	E11	E12	E13	E14
Evelyn	1	1	1	1	0	1	1	1	1	0	0	0	0	0
Laura	1	0	1	1	1	1	0	1	1	0	0	0	0	0
Theresa	1	1	1	1	1	1	1	0	1	0	0	0	0	0
Brenda	1	1	1	1	1	1	0	0	0	0	0	0	0	0
Charlotte	1	1	1	0	1	0	0	0	0	0	0	0	0	0
Frances	1	0	1	1	0	1	0	0	0	0	0	0	0	0
Eleanor	0	0	1	1	1	1	0	0	0	0	0	0	0	0
Pearl	0	0	0	1	0	1	1	0	0	0	0	0	0	0
Ruth	0	0	1	0	1	1	1	0	0	0	0	0	0	0
Verne	0	0	0	0	1	1	1	0	0	0	0	1	0	0
Myra	0	0	0	0	0	1	1	0	0	1	0	1	0	0
Katherine	0	0	0	0	0	1	1	0	0	1	0	1	1	1
Sylvia	0	0	0	0	1	1	1	0	0	1	0	1	1	1
Nora	0	0	0	1	1	0	1	0	0	1	1	1	1	1
Helen	0	0	0	0	1	1	0	0	0	1	1	1	0	0
Olivia	0	0	0	0	0	0	1	0	0	0	1	0	0	0
Flora	0	0	0	0	0	0	1	0	0	0	1	0	0	0
Dorothy	0	0	0	0	0	1	1	0	0	0	0	0	0	0

In Table 2, the third cluster of events (E1, E2 and E10-E14) does not appear to tell much of a story with respect to either group of women. This suggests that looking for a solution that has three clusters of women might be useful. That solution is illustrated in Table 3. The number of inconsistencies associated with this solution is 41 (a reduction of 10 over the previous solution) which translates into about 16% of the entries in the matrix in Table 1. This solution removes Frances and Eleanor from the first cluster of women in the previous solution and groups them with Ruth, Verne, Myra, Olivia, Flora, Pearl and Dorothy. The third group of women is composed of Katherine, Sylvia, Nora, and Helen. Under this clustering, the first group of women is associated with events E1-E7. All three clusters of women are associated with E8 and E9 and the third cluster is associated with events E10-E14.

Table 3. Matrix of Southern Women Data with 3 Clusters for Women and 3 for Events.

Individual	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
Evelyn	1	1	1	1	1	1	0	1	1	0	0	0	0	0
Laura	1	1	1	0	1	1	1	1	0	0	0	0	0	0
Theresa	0	1	1	1	1	1	1	1	1	0	0	0	0	0
Brenda	0	0	1	1	1	1	1	1	0	0	0	0	0	0
Charlotte	0	0	1	1	1	0	1	0	0	0	0	0	0	0
Frances	0	0	1	0	1	1	0	1	0	0	0	0	0	0
Eleanor	0	0	0	0	1	1	1	1	0	0	0	0	0	0
Ruth	0	0	0	0	1	0	1	1	1	0	0	0	0	0
Verne	0	0	0	0	0	0	1	1	1	0	0	1	0	0
Myra	0	0	0	0	0	0	0	1	1	1	0	1	0	0
Olivia	0	0	0	0	0	0	0	0	1	0	1	0	0	0
Flora	0	0	0	0	0	0	0	0	1	0	1	0	0	0
Pearl	0	0	0	0	0	1	0	1	1	0	0	0	0	0
Dorothy	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Katherine	0	0	0	0	0	0	0	1	1	1	0	1	1	1
Sylvia	0	0	0	0	0	0	1	1	1	1	0	1	1	1
Nora	0	0	0	0	0	1	1	0	1	1	1	1	1	1
Helen	0	0	0	0	0	0	1	1	0	1	1	1	0	0

It is useful to notice that this solution differs from the consensus analysis given in Freeman (2003) in that all pairs of women from the set Evelyn, Laura, Theresa, Brenda, Charlotte and Frances are concluded to belong together and that all pairs of women from the set Myra, Katherine, Sylvia, Nora and Helen also are concluded to belong together. The motivation from this model to omit Frances from the first cluster (which contains the other five women) is that she only attends 3 of the events E1 through E7 so the penalty is lower by one if she is placed in the second cluster (rather than the first). As for Myra, she only attended 2 of the 5 events in the third event cluster (E10-E14) so the penalty is one less to place her in the second person cluster rather than the third (person cluster).

It is useful to notice that we did not have to pre-specify any block types to produce solutions that are consistent with the literature (Doreian, 2004). The model concluded whether the block ideal should be a complete block (all ones) or a null block (all zeros). Also, it is very easy to see the motivation behind the groupings the model has suggested.

Next, to explore the application of this formulation to valued generalized block modeling, we convert the two-mode network data associated with the Southern Women dataset into a one-mode representation where the objects are the women and the relationships are the number of events pairs of women attended. That data is given in Table 4 formatted to illustrate the clusters (max of two allowed) and the blocks (within cluster minimum value of 3 and between cluster maximum value of 2). This is the same clustering suggested by Freeman (2003). This solution is rather insensitive to the within cluster minimum value and the between cluster maximum value, so it is a very stable solution using this formulation.

Table 4. One-Mode Analysis of Southern Women Data.

	Evelyn	Laura	Theresa	Brenda	Charlotte	Frances	Eleanor	Ruth	Pearl	Verne	Myra	Katherine	Sylvia	Nora	Helen	Olivia	Flora	Dorothy
Evelyn	8	6	7	5	3	4	3	3	3	2	2	2	2	2	1	1	1	2
Laura	6	7	6	5	3	4	4	3	2	2	1	1	2	2	2	0	0	1
Theresa	7	6	8	6	4	4	4	4	3	3	2	2	3	3	2	1	1	2
Brenda	5	5	6	6	4	4	4	3	2	2	1	1	2	2	2	0	0	1
Charlotte	3	3	4	4	4	2	2	2	0	1	0	0	1	1	1	0	0	0
Frances	4	4	4	4	2	4	3	2	2	1	1	1	1	1	1	0	0	1
Eleanor	3	4	4	4	2	3	4	3	2	2	1	1	2	2	2	0	0	1
Ruth	3	3	4	3	2	2	3	4	2	3	2	2	3	2	2	1	1	2
Pearl	3	2	3	2	0	2	2	2	3	2	2	2	2	2	1	1	1	2
Verne	2	2	3	2	1	1	2	3	2	4	3	3	4	3	3	1	1	2
Myra	2	1	2	1	0	1	1	2	2	3	4	4	4	3	3	1	1	2
Katherine	2	1	2	1	0	1	1	2	2	3	4	6	6	5	3	1	1	2
Sylvia	2	2	3	2	1	1	2	3	2	4	4	6	7	6	4	1	1	2
Nora	2	2	3	2	1	1	2	2	2	3	3	5	6	8	4	2	2	1
Helen	1	2	2	2	1	1	2	2	1	3	3	3	4	4	5	1	1	1
Olivia	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	2	2	1
Flora	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	2	2	1
Dorothy	2	1	2	1	0	1	1	2	2	2	2	2	2	1	1	1	1	2

Table 5 gives the suggested clustering when 3 clusters are allowed. Notice that the third cluster results from combining Pearl from the previous first cluster and Olivia, Flora and Dorothy from the previous second cluster. This third cluster is made up of individuals that do not attend very many events (in comparison to the other women) and what events they do attend tend to be somewhat common among them. For example, all four women attended event E9. Two of the four women attended E8 and E11. Of the events for which at least one of the four attended, E6 had the minimum attendance from the group with only Pearl attending. Table 6 illustrates the more clearly. Table 6 is the same one-mode transformation of the data but this time the entries have been normalized based on the total number of events each pair of individuals attended. This table has been formatted based on the results of applying this model to the normalized data, a maximum of three clusters and 0.5 for within cluster minimum value and 0.49 as the between cluster maximum value. Notice that the clusters identified are the same. The average value for the entries associated with blocks along the diagonal (excluding the diagonal entries of

1.0) is about twice as large as for blocks off the diagonal. The normalized data with these same parameters but with a maximum number of clusters of two yields the same solution as given in Table 4.

Table 5. One-Mode Analysis of Southern Women Data with Three Clusters.

	Evelyn	Laura	Theresa	Brenda	Charlotte	Frances	Eleanor	Ruth	Verne	Myra	Katherine	Sylvia	Nora	Helen	Olivia	Flora	Pearl	Dorothy
Evelyn	8	6	7	5	3	4	3	3	2	2	2	2	2	1	1	1	3	2
Laura	6	7	6	5	3	4	4	3	2	1	1	2	2	2	0	0	2	1
Theresa	7	6	8	6	4	4	4	4	3	2	2	3	3	2	1	1	3	2
Brenda	5	5	6	6	4	4	4	3	2	1	1	2	2	2	0	0	2	1
Charlotte	3	3	4	4	4	2	2	2	1	0	0	1	1	1	0	0	0	0
Frances	4	4	4	4	2	4	3	2	1	1	1	1	1	1	0	0	2	1
Eleanor	3	4	4	4	2	3	4	3	2	1	1	2	2	2	0	0	2	1
Ruth	3	3	4	3	2	2	3	4	3	2	2	3	2	2	1	1	2	2
Verne	2	2	3	2	1	1	2	3	4	3	3	4	3	3	1	1	2	2
Myra	2	1	2	1	0	1	1	2	3	4	4	4	3	3	1	1	2	2
Katherine	2	1	2	1	0	1	1	2	3	4	6	6	5	3	1	1	2	2
Sylvia	2	2	3	2	1	1	2	3	4	4	6	7	6	4	1	1	2	2
Nora	2	2	3	2	1	1	2	2	3	3	5	6	8	4	2	2	2	1
Helen	1	2	2	2	1	1	2	2	3	3	3	4	4	5	1	1	1	1
Olivia	1	0	1	0	0	0	0	1	1	1	1	1	2	1	2	2	1	1
Flora	1	0	1	0	0	0	0	1	1	1	1	1	2	1	2	2	1	1
Pearl	3	2	3	2	0	2	2	2	2	2	2	2	2	1	1	1	3	2
Dorothy	2	1	2	1	0	1	1	2	2	2	2	2	1	1	1	1	2	2

Table 6. One-Mode Analysis of Southern Women Data Using Normalized Relationships with Three Clusters.

	Evelyn	Laura	Theresa	Brenda	Charlotte	Frances	Eleanor	Ruth	Verne	Myra	Katherine	Sylvia	Nora	Helen	Olivia	Flora	Pearl	Dorothy
Evelyn	1.00	0.80	0.88	0.71	0.50	0.67	0.50	0.50	0.33	0.33	0.29	0.27	0.25	0.15	0.20	0.20	0.55	0.40
Laura	0.80	1.00	0.80	0.77	0.55	0.73	0.73	0.55	0.36	0.18	0.15	0.29	0.27	0.33	0.00	0.00	0.40	0.22
Theresa	0.88	0.80	1.00	0.86	0.67	0.67	0.67	0.67	0.50	0.33	0.29	0.40	0.38	0.31	0.20	0.20	0.55	0.40
Brenda	0.71	0.77	0.86	1.00	0.80	0.80	0.80	0.60	0.40	0.20	0.17	0.31	0.29	0.36	0.00	0.00	0.44	0.25
Charlotte	0.50	0.55	0.67	0.80	1.00	0.50	0.50	0.50	0.25	0.00	0.00	0.18	0.17	0.22	0.00	0.00	0.00	0.00
Frances	0.67	0.73	0.67	0.80	0.50	1.00	0.75	0.50	0.25	0.25	0.20	0.18	0.17	0.22	0.00	0.00	0.57	0.33
Eleanor	0.50	0.73	0.67	0.80	0.50	0.75	1.00	0.75	0.50	0.25	0.20	0.36	0.33	0.44	0.00	0.00	0.57	0.33
Ruth	0.50	0.55	0.67	0.60	0.50	0.50	0.75	1.00	0.75	0.50	0.40	0.55	0.33	0.44	0.33	0.33	0.57	0.67
Verne	0.33	0.36	0.50	0.40	0.25	0.25	0.50	0.75	1.00	0.75	0.60	0.73	0.50	0.67	0.33	0.33	0.57	0.67
Myra	0.33	0.18	0.33	0.20	0.00	0.25	0.25	0.50	0.75	1.00	0.80	0.73	0.50	0.67	0.33	0.33	0.57	0.67
Katherine	0.29	0.15	0.29	0.17	0.00	0.20	0.20	0.40	0.60	0.80	1.00	0.92	0.71	0.55	0.25	0.25	0.44	0.50
Sylvia	0.27	0.29	0.40	0.31	0.18	0.18	0.36	0.55	0.73	0.73	0.92	1.00	0.80	0.67	0.22	0.22	0.40	0.44
Nora	0.25	0.27	0.38	0.29	0.17	0.17	0.33	0.33	0.50	0.50	0.71	0.80	1.00	0.62	0.40	0.40	0.36	0.20
Helen	0.15	0.33	0.31	0.36	0.22	0.22	0.44	0.44	0.67	0.67	0.55	0.67	0.62	1.00	0.29	0.29	0.25	0.29
Olivia	0.20	0.00	0.20	0.00	0.00	0.00	0.00	0.33	0.33	0.33	0.25	0.22	0.40	0.29	1.00	1.00	0.40	0.50
Flora	0.20	0.00	0.20	0.00	0.00	0.00	0.00	0.33	0.33	0.33	0.25	0.22	0.40	0.29	1.00	1.00	0.40	0.50
Pearl	0.55	0.40	0.55	0.44	0.00	0.57	0.57	0.57	0.57	0.57	0.44	0.40	0.36	0.25	0.40	0.40	1.00	0.80
Dorothy	0.40	0.22	0.40	0.25	0.00	0.33	0.33	0.67	0.67	0.67	0.50	0.44	0.20	0.29	0.50	0.50	0.80	1.00

The final example is focused at the core of the contribution of this paper, the simultaneous analysis of multiple matrices where the matrices are single-mode, two-mode or some combination of both and multiple matrices include the same objects. For this example we use the average airfare between the top 48 airports (as measured by enplanements) in the continental United States in 2010 and the percentage of air service provided by carrier (based on passengers carried in 2010) at each of these airports, available from the Bureau of Transportation Statistics (2011). The top 50 airports in the United States includes Honolulu (HNL) and San Juan, Puerto Rico (SJU), which are outside the continental United States, hence we focus on 48 airports. The airfare data is available from the airline origin and destination survey dataset, which is effectively a 10% sample of all tickets sold in the U.S., and the air service provided by airport and carrier, is available in the Air Carrier Statistics (Form 41 Traffic).

The average airfare data across all pairs of these 48 airports is about $390 with about 10% of fares exceeding $630. To reduce the impact of these large fairs, which often reflect very little volume, we censor large values by assuming that these tickets were $630. Since the formulation described above assumes that larger values for the relationships in the matrices are indicative of closer ties, we simply subtract the fare from $630; thereby mapping larger fares to smaller values in the airfare table. As for the percentage of service at each of the 48 airports by carrier, we represent 10 of the largest carriers explicitly

and combine the remainder into an umbrella carrier, "Other". Regional carriers that are closely aligned with specific large carriers are included with that carrier. For example, American Eagle customers are grouped with American Airlines' customers. Where the association is less clear, regional carriers are included in the category "Other".

The airfare matrix has 48*47 entries which are in the hundreds of dollars. The carrier and airport matrix has 11*48 entries which range from zero to one. We multiply the entries in the carrier and airport matrix by 4,000 so that the tables have equal importance in the modeling. Tables 7 and 8 give the suggested clustering of airports and carriers based on the following parameters.

- Maximum of 10 clusters for airports.
- Maximum of 7 clusters of carriers.
- Desirable fare between airports in the same cluster of below $260 (bottom 25th percentile of fares).
- Desirable fare between airports in different clusters of above $340 (top 50th percentile of fares).
- We assume 40% or more service provided at an airport by a single carrier is indicative of a close relationship.
- We assume that 20% or less service provided at an airport by as single carrier is indicative of a "minimal" relationship.

Table 7. Suggested Clustering of Airports.

Cluster ID	Airports
1	DAL, LAS, MDW
2	CLT, PHX
3	ATL, CVG, DTW, MEM, MSP, SLC
4	BNA, OAK, SFO, SJC, SMF, STL
5	IAH
6	BWI, DCA, EWR, IAD, JFK, LGA, PHL
7	SAT, AUS, BOS, CLE, FLL, IND, MKE, MSY, PIT, RDU, SEA
8	DFW, MIA, ORD
9	DEN, LAX, MCO, PDX, TPA
10	HOU, MCI, SAN

Table 8. Suggested Clustering of Carriers.

Cluster ID	Carriers
1	Alaska Airlines (AL), Jet Blue (JB), Sky West (SKY), Spirit (SP)
2	Continental (CO)
3	Southwest (SW)
4	Delta (DL)
5	US Airways (US)
6	United (UN), Other Carriers (O)
7	American (AA)

Table 9 gives the average fare between pairs of airport clusters and Table 10 gives key information about each airport cluster including the average, minimum and maximum value for the Herfindahl-Hirschman Index (HHI) index (Herfindahl, 1950, Hirschman, 1964), and the average percent of air service provided by each carrier cluster. The HHI index is a measure of the market concentration where larger values indicate higher concentration of service by fewer carriers. It is useful to notice that the model has

suggested forming singleton clusters for several of the largest airlines.

Table 9. Average Fares Between Pairs of Airport Clusters (without re-scaling)

Airport Cluster	1	2	3	4	5	6	7	8	9	10
1	$631	$523	$390	$523	$414	$399	$392	$373	$467	$627
2	$535	$317	$388	$422	$446	$446	$373	$452	$423	$537
3	$391	$391	$361	$318	$383	$408	$330	$344	$397	$319
4	$498	$424	$317	$268	$470	$405	$333	$399	$467	$586
5	$413	$443	$382	$458		$578	$435	$401	$514	$519
6	$406	$445	$411	$407	$581	$251	$432	$470	$498	$374
7	$466	$370	$331	$333	$443	$434	$288	$355	$393	$350
8	$367	$447	$343	$397	$406	$469	$349	$472	$427	$357
9	$455	$417	$399	$468	$534	$501	$391	$430	$357	$379
10	$616	$522	$317	$591	$531	$372	$340	$356	$363	$312

Based on Table 9, the cheapest average fares are between cities in the same cluster with the exception of airport clusters 1, 3 and 8. The average airfares between airports in clusters 1, 3 and 8 are $631, $361 and $472, respectively. However, the average airfare between cities in cluster 1 with cities in cluster 8 is only $367. Travel between LAS and MDW is about $1,150 whereas travel between DAL, MDW and LAS is rather inexpensive. With the rescaling, the average fare between cities in cluster 1 is closer to $450, still higher than cluster 8, but closer. Given the assumptions as to "acceptable" within cluster fares, and between cluster fares, the only penalty assessed is on travel between LAS and MDW. As for cluster 3, the average airfare between airports in this cluster is $361 whereas the average airfare between cities in cluster 3 with cities in clusters 10 and 4 is about $317. Finally, the average travel cost between cities in cluster 8 is about $472 where as the average cost for travel from airports in cluster 8 with airports in cluster 3 is about $344.

Table 10. Key Characteristics of Airport Clusters.

Airport Cluster ID	Average HHI	Min HHI	Max HHI	Percent of Service						
				CC 1	CC 2	CC 3	CC 4	CC 5	CC 6	CC 7
1	0.62	0.19	0.91	1%	2%	73%	7%	2%	5%	3%
2	0.40	0.27	0.52	1%	2%	16%	7%	56%	6%	5%
3	0.46	0.40	0.51	1%	1%	4%	67%	2%	9%	4%
4	0.31	0.17	0.58	2%	2%	47%	8%	4%	10%	11%
5	0.55	0.55	0.55	0%	74%	0%	3%	4%	8%	3%
6	0.23	0.15	0.34	2%	10%	10%	16%	14%	14%	13%
7	0.14	0.04	0.21	3%	8%	20%	16%	6%	13%	11%
8	0.51	0.27	0.66	1%	3%	0%	7%	5%	9%	65%
9	0.13	0.11	0.16	3%	6%	22%	15%	7%	15%	10%
10	0.34	0.20	0.74	2%	3%	51%	10%	4%	8%	9%

CC-Carrier Cluster

Clusters 1, 4 and 10 are all cities with extensive service by Southwest Airlines (Southwest serves less than 40% of passengers at only 2 of the 13 airports). Both cities in cluster 2 are US Airways hubs (US Airways service in Charlotte is above 70% and in Philadelphia it is over 40%). Delta Airlines provides

over 60% of the air service at all six airports in cluster 3. The three airports in cluster 8 are American Airlines hubs. Clusters 7 and 9 are comprised of airports with service from a variety of carriers. These two clusters have the lowest average and maximum HHI index of all 10 clusters. Airport cluster 6 is comprised of all the major airports in the Northeast (except Boston) and the average and maximum HHI is also lower for that cluster than all other clusters except for clusters 7 and 9. Finally, IAH, which is a hub for Continental airlines, forms a singleton cluster. IAH fares, are on average, in the top 20% of airport fares and is the only airport tightly associated with Continental Airlines, which are likely contributing the desirability of treating it as a singleton cluster.

5.0 Conclusions

This paper described an extension to generalized blockmodeling where there are more than two types of objects to be clustered based on valued network data. The ideas in homogeneity block modeling are used to develop an optimization model to perform the clustering of the objects and the resulting partitioning of the ties so as to minimize the inconsistency of an empirical block with an ideal block. The ideal block types used in this modeling were null, complete and a new type that is related to that used in Žiberna (2007). A tabu search solution procedure was developed to solve the resultant optimization.

This modeling approach for valued network data is dependent on two parameters e_m and d_m where the ideal for the level of interaction between objects in the same cluster is at least e_m and the ideal for the level of interaction between objects in different clusters is assumed to be no more than d_m. These two parameters provide more flexibility to tailor the analysis to application then that given in Žiberna (2007), which relies on a single parameter for this purpose.

Three case studies using the formulation and solution procedure were described, two based on the Southern Women dataset (Davis et al. 1941) and a third based on passenger air travel in the Continental United States. The clustering suggested by this formulation for the Southern Women dataset is consistent with that given in Freeman (2003). As for the air transportation analysis, the formulation identified clusters of airports where they serve as hubs for the same carrier, are airports that are served by specific carriers, are located close together or where no carrier has significant market dominance. As for carriers, large national carriers were mostly assigned to singleton clusters

There are opportunities for future work in at least two complementary directions. One opportunity focuses on the explicit introduction of uncertainly. For example, as these ideas are used in practice, some of the information available on the ties between objects could be subject to some uncertainty. As an illustration, if one were to apply these tools to attempt to understand the activities of a market competitor (industrial competition), and there were observations as to who is frequenting different locations as an indicator of the character of the activities undertaken at that location, that data might be limited by the ability to collect this information. One mechanism to include this in the analysis is to simply ignore relationships in the computation of the objective function value that are subject to these issues. Alternatively, weights could be associated with each tie to indicate the quality of the information that lead to that estimate of the relationship. Further research to explore what mechanisms to use in order to incorporate this information into the analysis is very important.

A second opportunity focuses on the opportunity to extend this tool to address data collected over time. For example, suppose we had the same types of information collected a multiple points in time; it would be useful to identify a clustering and partitioning of ties that departs as little of possible from block ideals over all time periods. We might require that the solution include membership in clusters that is invariant over time or we might allow the membership to change, but with a penalty. Allowing the membership to change over time is useful in that organizational structures are often fluid and understanding the nature of the fluidity is very useful.

References

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

Breiger, R.L., Boorman, S., Arabie, P., 1975. An algorithm for clustering relational data with applications to social network analysis. Journal of Mathematical Psychology 12, 329–383.

Breiger, R.L., Mohr, J.W., 2004. Institutional logics from the aggregation of organizational networks: operational procedures for the analysis of counted data. Computational and Mathematical Organization Theory 10, 17–43.

Bureau of Transportation Statistics, 2011. Aviation Data Library, accessed July 18, 2011, on-line at http://www.transtats.bts.gov/databases.asp?Mode_ID=1&Mode_Desc=Aviation&Subject_ID2=0.

Burt, R.S., 1976. Positions in networks. Social Forces 55, 93–122.

Davis, A., Gardner, B., Gardner, M.R., 1941. Deep South. University of Chicago Press, Chicago.

Doreian, P., Batagelj, V., and Ferligoj, A., 2004. Generalized blockmodeling of two-mode network data. Social Networks, 26, 29-53.

Doreian, P., Batagelj, V., and Ferligoj, A., 2005. Generalized Blockmodeling. Cambridge University Press, New York, New York.

Freeman, L.C., 2003. Finding social groups: a meta-analysis of the Southern Women data. In: Breiger, R., Carley, C., Pattison, P. (Eds.), Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers. National Research Council, The National Academies Press, Washington, DC, pp. 39–97.

Lorrain, F., White, H.C., 1971. Structural equivalence of individuals in social networks. Journal of Mathematical Sociology 1, 49–80.

O.C. Herfindahl, 1950. Concentration in the U.S. Steel Industry, Ph.D. Thesis, Columbia University, New York.

O.A. Hirschman, 1964. The paternity of an index. The American Economic Review 54, 761–762.

Hlebec, V., 1996. Metodoloˇske znaˇcilnosti anketnega zbiranja podatkov v analizi omreˇzji: Magistersko delo. FDV, Ljubljana.

Žiberna, A., 2007. Generalized block modeling of valued networks. Social Networks, 29, 105-126.