The Impact of Non-Response Treatments on the Stability of Blockmodels

Žnidaršič, Anja
University of Maribor, Faculty of Organizational Sciences Kidričeva 55a
SI-4000 Kranj, Slovenia
E-mail: anja.znidarsic@fov.uni-mb.si

Ferligoj, Anuška
University of Ljubljana, Faculty of Social Sciences
Kardeljeva ploščad 5
SI-1000 Ljubljana, Slovenia
E-mail: anuska.ferligoj@fdv.uni-lj.si

Doreian, Patrick
University of Pittsburgh, Department of Sociology
Pittsburgh, PA, United States
University of Ljubljana, Faculty of Social Sciences
Kardeljeva ploščad 5
E-mail: pitpat@pitt.edu

1 Introduction

Social network data are collected in a variety of ways and these data are analyzed in terms of structural properties and network processes. Yet collected social network data, most likely, contain different types of errors. One widely used technique for delineating structural patterns of relationships is blockmodeling. We do not know how vulnerable these methods are to missing data problems. However, given that they are *positional* - focusing on actor locations defined for the *whole* network - it is reasonable to expect that blockmodeling is highly vulnerable. We focus on actor non-response, as one form of missing data, and data processing strategies designed to treat such missing data. We examine the impact of these data processing strategies on the results of blockmodeling by using real networks and simulations based on them. A set of 'known' networks are used, errors due to actor non-response are introduced and the resulting networks are then treated in different ways. Blockmodels are fitted to these networks and compared. The outcome indicator is the level of correspondence of known block structures and the corresponding block structures of the treated networks. We use the Adjusted Rand Index and the proportion of incorrect blocks in a blockmodel to describe the level of (dis)similarity. The amount and type of non-response, as well the treatments of this form of missing data, all have an impact on the resulting blockmodel structures.

2 Actor non-response and treatments

Non-response in social networks can appear in two forms; *item* non-response (or missing tie) (Rumsey, 1993; Borgatti et al., 2006; Huisman and Steglich, 2008; Huisman, 2009) and *actor* non-response (Stork and Richards, 1992; Costenbader and Valente, 2003; Kossinets, 2006; and Huisman, 2009). We consider only (the more consequential) actor non-response together and five possible missing (or non-reported) data treatments.

Each non-respondent in a network with n actors causes $n-1$ missing ties. The actor response

rate is the same as the relational response rate and is equal to $1 - m/n$, where m is the number of non-respondents (Knoke and Yang, 2008). For example, in a network with 13 actors where two of them refuse to respond, the actor response rate is 0.84. There are 24 ties missing ties, the proportion of partially non described ties between non-respondents and respondents is equal to 0.14.

Missing data treatments due to actor non-response have been studied by several authors (Stork and Richards, 1992; Robins et al., 2004; Huisman and Steglich, 2008; Huisman, 2009). The first treatment is the *available case approach* where only completely described ties are take into account. In practice, this means that not only the rows of nonrespondents in a sociomatrix deleted, the corresponding columns are deleted also. Robins et al. (2004, pg. 260) argued that this approach in fact leads to the re-specification of network boundaries.

The second treatment is called *reconstruction* (Stork and Richards, 1992; Huisman, 2009). The main advantage of this treatment, compared to the complete case treatment, is the use of all partially described ties between respondents and non-respondents. The row of missing ties is replaced with corresponding column for each missing respondent. If the network has more than one non-respondent, the ties between two actors who refused to respond cannot be obtained without additional imputation. In the simplest solution, the zeros are imputed instead of the missing ties between two non-respondents. Stork and Richards (1992, pg. 198) argued that the two criteria should be satisfied when reconstruction is used: (i) the non-respondents and the respondents should not differ systematically from each other, and (ii) and the available data from the respondents are useful and reliable description of ties between two actors. The second criterion can be more easily satisfied in undirected networks (e.g., conversation) than in directed ones (e.g., giving advice).

In general, imputation methods replace a missing data with an appropriate estimate. In the simplest case, one used too often in different social network studies and analyses, the unreported ties are replaced by zeros. (Therefore, this treatment is called *null tie imputation*). In this treatment, non-respondents are retained in the data set and their outgoing ties are set to 0. Such partially described ties between respondents and non-respondents are used for analyses.

Another imputation procedure is to use the mean of incoming ties. In the case of binary networks, this means that a tie (one) is imputed if a missing actor is popular. Operationally, if at least half of the responders report a tie to this actor, the presence of a tie is imputed. For actors receiving less than half of the potential incoming ties, zero is imputed. This procedure is called *imputation based on modes*, because the modal incoming values of ties are imputed for missing rows.

A fifth treatment is the combination of the reconstruction procedure and imputations based on modes for ties between two non-respondents (*reconstruction and modes*).

Robins et al. (2004, pg. 258) emphasized that "non-respondents create significant and potentially insidious problems for network analysis". This can be an especially serious problem in analyses where arrangements of ties in larger structures or sub-structures are studied. One such widely use procedure is blockmodeling. Ferligoj et al., (2011) emphasized that "obtaining a better understanding of the vulnerability of establishing blockmodels to errors of measurement is an important open problem", and actor non-response is one part of a wider conception of errors in social network research design.

3 Blockmodeling and indices for comparison of two blockmodels

The goal of blockmodeling is to reduce a large difficult to understand network to a smaller comprehensible and more simply interpretable structure (Batagelj et al., 2004). The practical core of the blockmodeling procedure is to partition the network actors into clusters (discrete subgroups called *positions*), and, at the same time, to partition the set of ties into *blocks* which determine the ties between positions (Faust and Wasserman, 1992; Doreian et al., 2005). A block is defined as the relation

between two clusters of actors. The actors within a cluster and between the clusters should have the same (or similar) pattern of ties based on a selected equivalence. The result of blockmodeling is a compact representation of a network, a blockmodel, which represents the essential structure of a network. It can be represented by a reduced graph or by an image matrix. The units in this reduced network are clusters of equivalent actors from the original network representing positions. Arcs in a reduced graph represent relations between positions (Doreian et al., 2005). The blockmodeling concepts of partitions and blocks can be viewed also in terms of positions and roles (Faust and Wasserman, 1992; Ferligoj et al., 2011).

Here, we focus on generalized blockmodeling based on structural equivalence (Batagelj et al., 1992; Doreian et al., 2005) and use a direct approach where the best partition is identified based on minimal values of the criterion function. Actors are structurally equivalent if they are connected in exactly the same way to same neighbors (a formal definition is presented in Doreian et. al., 2005, pg. 172). Batagelj et al. (1992) proved that there are just two possible (ideal) blocks consistent with structural equivalence: null and complete. The criterion function is defined by the difference between empirical blocks and corresponding ideal blocks.

The blockmodeling procedures have been implemented in the program Pajek (Batagelj and Mrvar, 2010a,b), and in the R-package called Blockmodeling (Žiberna, 2008). Both were used in this (current) study.

Blockmodeling results in a partition of actors into positions where relationships within and between positions determining the image matrix. The stability of a blockmodel to non-response (and treatments of these non-responses) are measured by two indices where the original ('known') blockmodel and the 'treated' blockmodel of a network (where there are errors that have been treated) are compared. One index, the Adjusted Rand Index (ARI), measures the differences between the two partitions in terms of their composition (Hubert and Arabie, 1985; Saporta and Youness, 2002). The lower the ARI, the worse is the correspondence of the position memberships for two partitions. According to the extended simulation study of Steinley (2004) we will say that a blockmodel is stable, in terms of agreement between partitions or that correspondence of the position memberships is acceptable, if the mean of the Adjusted Rand index ($mARI$) satisfies $mARI \geq 0.8$. (A more stringent requirement is that $mARI \geq 0.9$.) Perhaps more important, due to the nature and the purpose of blockmodeling, is whether the identified blocks, given the positions for the treated network, correspond (or not) to the block types in the known blockmodel. Therefore, our second index for comparing two blockmodels is the proportion of incorrect block types in treated blockmodel compared to a known blockmodel. This index will be denoted by $ErrB$. Higher values mean lower concordance between two blockmodels. Somewhat arbitrarily, we consider results that have where $mErrB \geq 0.2$ to be unacceptable.

4 Two real networks and simulations based on them

The results of an extensive study of the impact of non-response treatments on the stability of blockmodels on real and simulated network data were presented in Žnidaršič et al. (2010). The main finding was that the selection of the best or most appropriate non-response data treatment depends on the symmetry of a network. Based on simulations, two recommendations were made: (i) for symmetric (or largely symmetric) networks, the best treatments are reconstruction and combination of reconstruction with imputations based on modes, and (ii) for non-symmetric (or largely non-symmetric) networks, the best approaches are complete case and imputations based on modes. The extent of symmetry of

a network is measured directly with reciprocity[1]. Another very important recommendation was to *never* use null tie imputation even though it is the easiest and is most tempting approach for treating non-response (Žnidaršič et al., 2010).

The goal of this paper is to examine these findings by considering a pair of directly comparable networks (in terms of size and substantive type). One has far greater symmetry than the other and we will label them as 'symmetric' and 'non-symmetric' even though they are not pure cases regarding the presence or absence of symmetry. The procedure of our simulation study is as follows: take a real network, establish the blockmodel of this real 'known' network, delete selected numbers of actors according to a mechanism for generating non-respondents[2], treat the missing data with each of five treatments presented in Section 2, establish the blockmodel of the 'treated' network, and compare the 'known' and 'treated' blockmodels using the stability indices ARI and $ErrB$. The factorial design (for each starting real network) in our study has 105 cells (for the combination of three missing mechanisms, five treatments of non-response, and seven numbers of actors with non-response). Within each cell, the generation of incomplete data was repeated 20 times for networks with one non-respondent actor, 50 times for combinations of two non-respondents and 100 times for combinations of three or more non-respondents.

In this paper the networks of two successful Little League baseball teams of boys reported by Fine (1987) and extensively studied in terms of generalized blockmodeling by Doreian et al. (2005) are used in the simulation study. Boys were asked to name their three best friend in a team[3].

The first network, the Transatlantic Industries (TI) team, is an example of a more symmetric network, with a reciprocity value of 0.54 and the second one the Sharpstone Auso (SA) team with a smaller reciprocity equal of 0.26.

Doreian et al. (2005, pg. 199) reported for the Transatlantic Industries friendship network a blockmodel with two clusters based on structural equivalence. Figure 1 presents the TI network with two positions denoted with white and gray vertices (left panel) and an image matrix with a complete block on the diagonal between boys of the first cluster and three null blocks. The value of the criterion function is equal to 29 (the number of ties in null blocks and two non existing ties in the complete block).

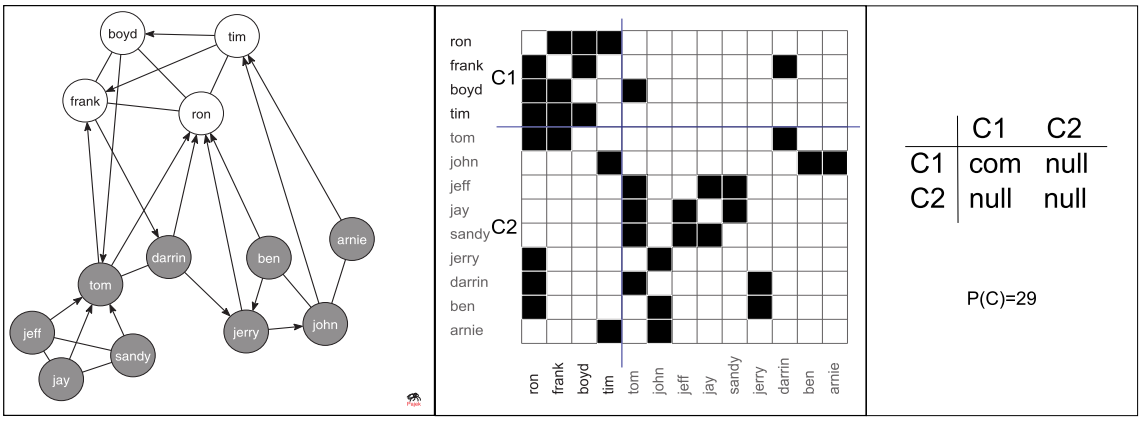


Figure 1: The Transatlantic Industries network (left), a two position solution based on structural equivalence (middle), and the corresponding image matrix (right)

[1]Reciprocity (Huisman, 2009) measures how symmetric is a network and it is defined by $reciprocity = \frac{2 \cdot M}{M+A}$, where M indicates the number of mutual dyads and A the number of asymmetric dyads.

[2]Actors were deleted in three ways: at random, based on their outdegree, and based on their indegree.

[3]The fixed choice design can also be a source of errors (e.g., Holland and Leinhardt, 1973; Kossinets, 2006).

The second example is a non-symmetric network of Sharpstone Auto (SA) team. The block-modeling based on structural equivalence into two clusters has one well fitting partition with 17 inconsistencies. The partition is presented on the left panel of Figure 2, and the image matrix with two complete blocks and two null blocks showing a core-periphery structure with ties to the core position (the right panel).

Both networks, TI and SA, can be partitioned also with a finer-grained partition (Doreian et al. 2005, pg. 196-201). In these blockmodels clusters with only one actor in a cluster were obtained. For our study such partitions are not appropriate as the sensitivity of (especially) ARI index would be high. Therefore, we decided to study the stability of the blockmodels with two clusters of both networks presented in the middle panel on Figures 1 and 2.

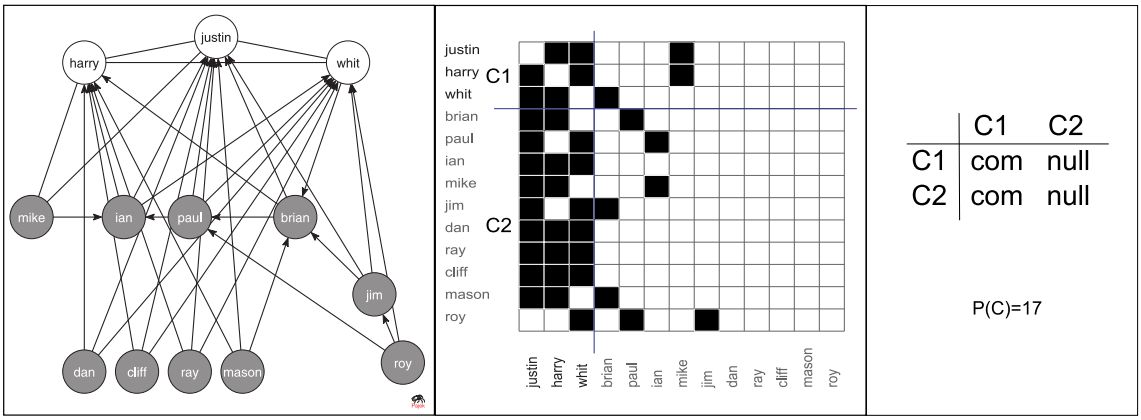
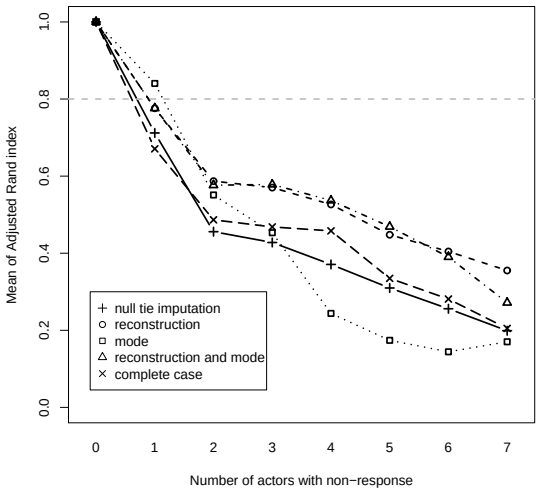


Figure 2: The Sharpstone Auto network (SA) (left), two clusters solution based on structural equivalence (middle), and image matrix (right)

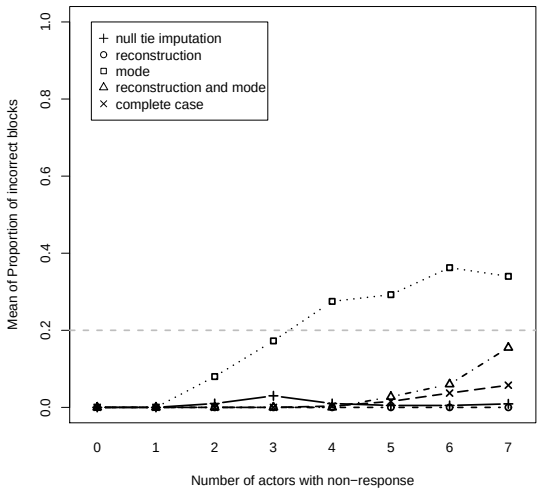
The results of the simulation study for the TI network are presented in Figure 3. In the left panel the results of mean values of the Adjusted Rand Index are plotted against the number of non-response actors which ranges from 1 to 7 (with the proportion of non-response taking the values 0.08, 0.15, 0.23, 0.31, 0.38, 0.46, and 0.54). When one actor is randomly selected as a non-respondent, the $mARI$ values are above 0.8, indicating an acceptable agreement between the 'known' and 'treated' partitions, appears only for the imputations based on modes. For three or more missing actors, the imputations based on modes performs the worst. A similar pattern is shown for the null tie imputation treatment. The best treatments are reconstruction and combination of reconstruction plus modes imputations. The three null blocks in the real blockmodel of TI network (Figure 1) are obviously not the ideal ones. The mean density of the null blocks is relatively high and is equal to 0.17. This could be one reason for high vulnerability of stability of blockmodels in terms of the agreement between partitions. The right panel of Figure 3 shows results for proportion of correctly identified block types in a blockmodel. For the whole range of missing actors the values of $mErrB$ are below 0.2 for all treatments except for the imputations based on modes. The best treatments for preserving the blockmodel structure are reconstruction, complete case, and null tie imputations.

Figure 4 shows the results for the non-symmetric SA network. The stability of partition is higher than for the TI network. The best treatment in terms of the agreement of partitions ($mARI$) and the proportion of correctly identified block types is imputations based on modes. For more than half of non-respondents in a network (7 non-respondents out of 13 actors in the network) the modes imputation can obtain the correct blockmodel for each combination of nonrespondents ($mErrB$=0). The second best treatment is the complete case approach where the values of ARI are above 0.8 in

the presence of six non-respondents or less and the values of $mErrB$ are below 0.2 for the whole range of introduced missing actors. The reconstruction plus modes treatment can return the known partition when three or fewer non-respondents are present. The reconstruction procedure and null ties imputations are the worst treatments according to both indices.

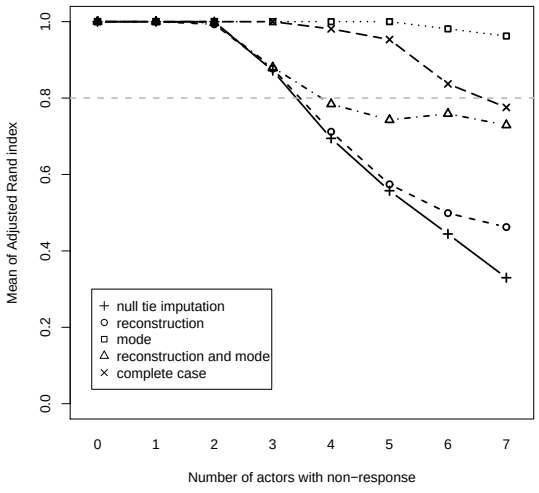


(a) Mean of the Adjusted Rand Index, $mARI$

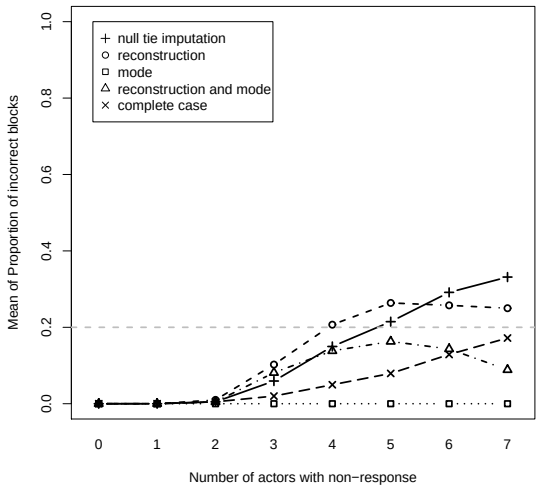


(b) Mean of Incorrect block types, $mErrB$

Figure 3: Results of the simulation study based on the Transatlantic Industries network for data missing completely at random



(c) Mean of the Adjusted Rand Index, $mARI$



(d) Mean of Incorrect block types, $mErrB$

Figure 4: Results of the simulation study based on the Sharpstone Auto network for data missing completely at random

5 Conclusions

The results reported here for these two networks fully confirm the results obtained previously by Žnidaršič et al. (2010). The performance of the missing data treatments for nonresponse in social networks depends on the symmetry of the networks. Treatments that are the best for symmetric networks are reconstruction and combination of reconstruction and modes imputations. For the non-symmetric network the best treatments are the modes imputations and the complete approach. The treatments that are the best for symmetric networks perform the worse in the case of non-symmetric networks and vice versa. We did a similar simulation study also for not at random deletion of actors becoming non-respondents based on indegree or outdegree. The results show that the performance of missing data treatment does not depend on the selected mechanism of non-respondents.

Table 1: Results of non-response treatments on the stability of blockmodels

Blockmodel	Symmetric *Reciprocity* = 0.54		Non-symmetric *Reciprocity* = 0.26	
Treatment	ARI	ErrB	ARI	ErrB
Complete case	∘	+	+	+
Reconstruction	+	+	-	-
Mode imputations	-	-	+	+
Null tie imputations	-	∘	-	-
Reconstruction + mode	+	+	∘	∘

The null tie imputation and the complete case approach have different performances, but we do not advise using either of them. The null tie imputation performs always the worst. In the complete case approach we lose information about the location of actor(s) in a position, because non-respondents are deleted from the network. Further work will be extended to larger networks and also to other types of equivalences.

REFERENCES

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

Batagelj, V., Mrvar, A., Ferligoj A. and Doreian, P., 2004. Generalized Blockmodeling with Pajek. *Metodološki zvezki 1 (2)*, 455-467.

Batagelj, V. and Mrvar, A., 2010a. *Pajek 1.28 – program for large network analysis.* Available at: http://pajek.imfm.si/doku.php?id=download .

Batagelj, V. and Mrvar, A., 2010b. *Pajek, Program for Analysis and Visualization of Large Networks, Reference Manual, List of commands with short explanation, version 1.28.* Available at: http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.pdf .

Borgatti, S. P., Carley, K. M. and Krackhardt, D., 2006. On the robustness of centrality measures under conditions of imperfect data. *Social Networks 28*, 124-136.

Costenbader, E. and Valente T. W., 2003. The stability of centrality measures when networks are sampled. *Social Networks 25*, 283-307.

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

Faust, K. and Wasserman S., 1992. Blockmodels: Interpretation and evaluation. *Social Networks 14*, 5-61.

Ferligoj, A., Doreian, P. and Batagelj, V., 2011. Positions and Roles. In: Scott J. and Carrington P. J. (Eds.), *The SAGE Handbook of Social Network Analysis.* Cambridge Sage Publications, Thousand Oaks, pp. 434-446.

Fine, G. A., 1987. *With the Boys: Little League Baseball and Preadolescent Culture.* University of Chicago Press, Chicago.

Holland, P. W. and Leinhardt, S., 1973. The structural implications of measurement error in sociometry. *The Journal of Mathematical Sociology 3*, 85-111.

Hubert, L. and Arabie, P., 1985. Comparing partitions. *Journal of Classification 2*, 193-218.

Huisman, M., 2009. Effects of missing data in social networks. *Journal of Social Structure 10.* Available at: http://www.cmu.edu/joss/content/articles/volume10/huisman.pdf.

Huisman, M. and Steglich, C., 2008. Treatment of non-response in longitudinal network studies. *Social networks 30*, 297-308.

Knoke, D. and Yang, S., 2008. *Social networks analysis.* Sage Publications, Los Angeles. 2nd edition.

Kossinets, G., 2006. Effects of missing data in social networks. *Social networks 28*, 247-268.

Robins, G., Pattison, P. and Woolcock, J., 2004. Missing data in networks: exponential random graph (p*) models for networks with non-respondents. *Social networks 26*, 257-283.

Rumsey, D. J., 1993. *Nonresponse models for social network stochastic processes.* Ph.D. thesis. The Ohio State University.

Saporta, G. and Youness, G., 2002. Comparing two partitions: Some proposals and experiments. In: Hardle, W., Ronz, B. (Eds.), *Proceedings in Computational Statistics.* Physica Verlag, Berlin, pp. 243-248.

Steinley, D., 2004. Properties of the Hubert-Arabie Adjusted Rand index. *Psychological Methods 9*, 386-396.

Stork, D. and Richards, W. D., 1992. Nonrespondents in communication network studies: problems and possibilities. *Group and Organization Management 17*, 193-209.

Žiberna, A., 2008. Blockmodeling 0.1.7: An R package for Generalized and classical blockmodeling of valued networks. Available at: http://www2.arnes.si/∼aziber4/ .

Žnidaršič, A., Ferligoj, A. and Doreian, P., 2010. Non-response in social networks: The impact of different non-response treatments on the stability of blockmodels. Submitted.