Missing network data and their impact on clustering results

Anja Žnidaršič[1]
Patrick Doreian[2,3]
Anuška Ferligoj[2]

[1]University of Maribor, Faculty of Organizational

[2]University of Ljubljana, Faculty of Social Sciences

[3]University of Pittsburgh, Department of SociologySciences

ARS
Napoli, May 16, 2017

Content

Missing network data in (valued) networks...

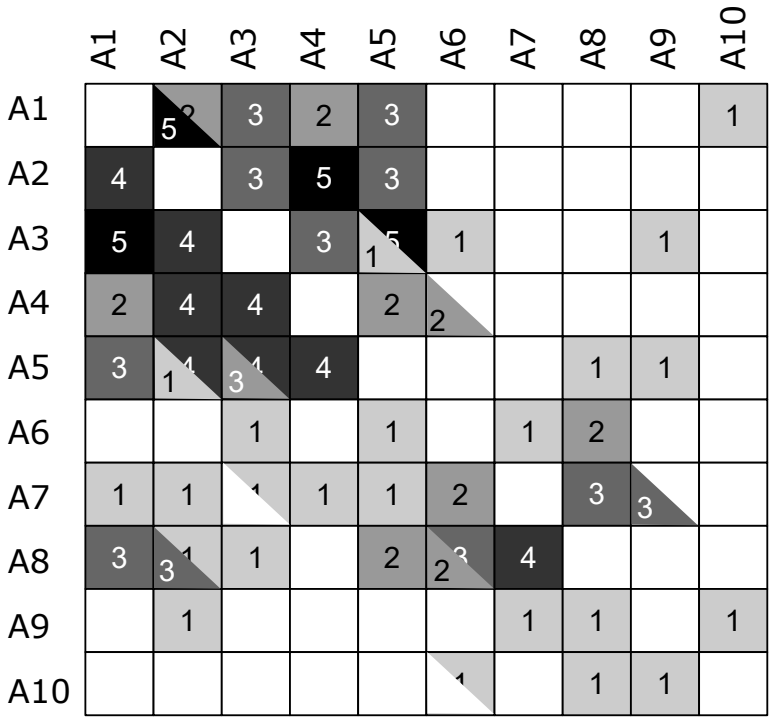
... could be classified into three broad categories:

- boundary specification problems,
- questionnaire design,
- errors due to network members (actors).

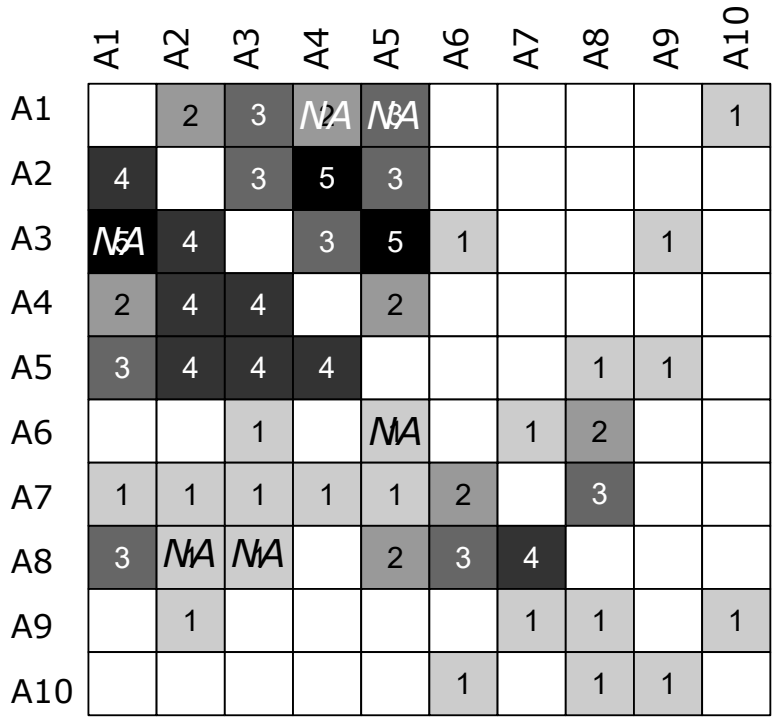
Errors due to actors in (valued) networks...

... could be divided into three subcategories:

- complete actor non-response,
- item non-response (regarding specific ties), and
- reporting errors in the recorded ties.



(a) Erroneously reported tie values



(b) Item non-response

Actor non-response in networks

Each non-respondent leads to $n - 1$ missing ties, where n is a number of actors in a network.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
A1		2	3	2	3					1
A2	4		3	5	3					
A3	5	4		3	5	1			1	
A4	2	4	4		2					
A5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
A6			1		1		1	2		
A7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
A8	3	1	1		2	3	4			
A9	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
A10						1		1	1	

Network consist of 10 actors and 3 refused to respond. The actor response rate (and relational response rate) is 0.7.

Effects of actor non-response in social networks

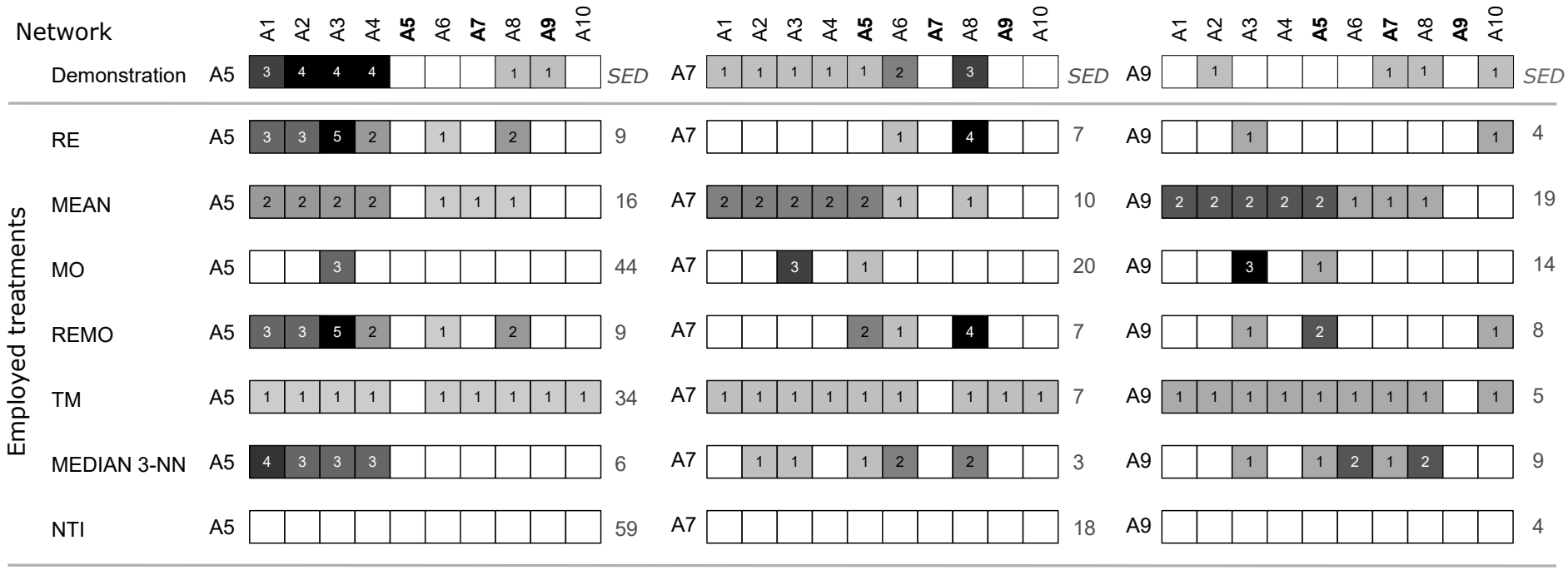
Effects of actor non-response on different network properties in binary networks such as network density, average vertex degree, out-degree or in-degree, clustering coefficient or transitivity, assortativity, mean inverse geodesic distance etc. have been examined previously by several authors (Stork and Richards, 1992; Costenbader and Valente, 2003; Kossinets, 2006; Huisman, 2009; Žnidaršič et al., 2012).

We decided to extend those studies to investigate the non-response effects in valued networks, more precisely:

- impact on indirect blockmodeling in simulated networks (Žnidaršič et al., 2017a),
- impact on several centrality measures (weighted betweenness, weighted closeness, weighted reciprocity,...) in simulated networks (Žnidaršič et al., 2017b).

Here, impact of actor non-response in case of real valued network data on three types of clustering will be presented.

Actor non-response treatments on demonstration network



A5, A7, A9 are non-respondents in the demonstration network.

RE - Reconstruction, MEAN -Imputations of mean values of incoming ties, MO - Imputations of modal values of incoming ties, REMO - Reconstruction and modal values, TM - Imputations of total mean, MEDIAN 3-NN - Median of 3-nearest neighbour of incoming ties, NTI - Nul tie imputations

SED - Squared Euclidean distance between vector of original tie values and corresponding vector of treated tie values.

Figure: Results of seven actor non-response treatments for the demonstration network with three non-respondents

Actor non-response treatments

Table: Characteristics of the whole demonstration network and the seven treated networks

Network		Magnitude of changed ties in treated network according to the whole one without diagonal								Average (imputed) tie values of outgoing ties of non-respondents			Network characteristics				QAP corr.
		-4	-3	-2	-1	0	1	2	3	A5	A7	A9	Arcs	recW	densW	Mean tie value	
Demonstration network										1.89	1.11	0.44	49	0.691	1.222	2.245	
Treated networks	RE			1	11	73	5			1.78	0.56	0.22	42	0.843	1.133	2.429	0.953
	MEAN			4	4	69	9	4		1.22	1.33	1.44	54	0.591	1.278	2.130	0.882
	MO	2	2	1	10	72	1	1	1	0.33	0.44	0.44	37	0.533	1	2.432	0.812
	REMO			1	10	73	6			1.78	0.67	0.33	44	0.827	1.156	2.364	0.952
	TM		3	2	1	74	10			1	1	1	59	0.623	1.178	1.797	0.878
	kNNMedian				10	75	4	1		1.44	0.78	0.78	46	0.660	1.178	2.304	0.955
	NTI	3	2	1	11	73				0	0	0	32	0.506	0.878	2.469	0.817

RE - reconstruction, MEAN - imputations of the mean values of incoming ties, MO - imputations of the modal values of incoming ties, REMO - reconstruction and imputations based on modal values of incoming ties, TM - imputations of the total mean, kNNMedian - imputations of median of 3-nearest neighbours based on incoming ties, NTI - and null tie imputations
recW - weighted reciprocity, densW - weighted density, Mean tie value - mean of tie values (without zeros)

Impact of actor non-response treatments on demonstration network on indirect blockmodeling

Table: Cluster membership of the actors the whole demonstration network and the seven treated networks

		Actor's membership in clusters based on indirect blocmodeling			
Network		Cluster 1	Cluster 2	Custer 3	ARI
Demonstration network		*A*1, *A*2, *A*3, *A*4, *A*5	*A*6, *A*7, *A*8	*A*9, *A*10	
Treated networks	RE	*A*1, *A*2, *A*3, *A*4, *A*5	*A*6, *A*7, *A*8	*A*9, *A*10	1
	MEAN	*A*2, *A*3, *A*4	*A*1, *A*5	*A*6, *A*7, *A*8, *A*9, *A*10	0.378
	MO	*A*1, *A*2, *A*4, *A*5	*A*3	*A*6, *A*7, *A*8, *A*9, *A*10	0.501
	REMO	*A*1, *A*2, *A*3, *A*4, *A*5	*A*6, *A*7, *A*8	*A*9, *A*10	1
	TM	*A*2, *A*3, *A*4	*A*1, *A*5	*A*6, *A*7, *A*8, *A*9, *A*10	0.378
	kNNMedian	*A*1, *A*2, *A*3, *A*4, *A*5	*A*6, *A*7, *A*8	*A*9, *A*10	1
	NTI	*A*1, *A*2, *A*3, *A*4	*A*5	*A*6, *A*7, *A*8, *A*9, *A*10	0.501

RE - reconstruction, MEAN - imputations of the mean values of incoming ties, MO - imputations of the modal values of incoming ties, REMO - reconstruction and imputations based on modal values of incoming ties, TM - imputations of the total mean, kNNMedian - imputations of median of 3-nearest neighbours based on incoming ties, NTI - and null tie imputations
ARI - Adjusted Rand Index between whole partition and corresponding treated partition

Basic scheme of simulations procedure for directed valued networks

1. For each real network, do the following:
   - 1.1 Establish the partition of the whole network using indirect blockmodeling employing corrected Euclidean distance and Ward's clustering method.
   - 1.2 Construct the 'observed' data for a wide range for the number of non-respondents to create the measured networks. This was done by randomly selecting actors to become non-respondents and deleting all of their outgoing ties.
   - 1.3 Employ each of the seven non-response data treatments separately to impute values replacing missing data to create the treated networks.
   - 1.4 Establish a partition of each treated network using indirect blockmodeling with the Corrected Euclidean distance and Ward's clustering method.
   - 1.5 Compare the partitions of the whole original and treated networks by the Adjusted Rand Index (*ARI*)
2. Binarize the network and:
   - 2.1 repeat the strategy from step 1.1 to 1.5..
   - 2.2 adopt the above strategy by using direct blockmodeling under structural equivalence.

   (Comparison of the partitions of the whole binary network and the treated networks was done by *ARI* and the proportion of incorrect block types (*mErrB*).)

Simulations were prepared using  package `blockmodeling` (Žiberna, 2008) and Program 

(Batagelj and Mrvar, 1996-2017a,-).

Real network of providing help in problem solving (*PHPSc*)

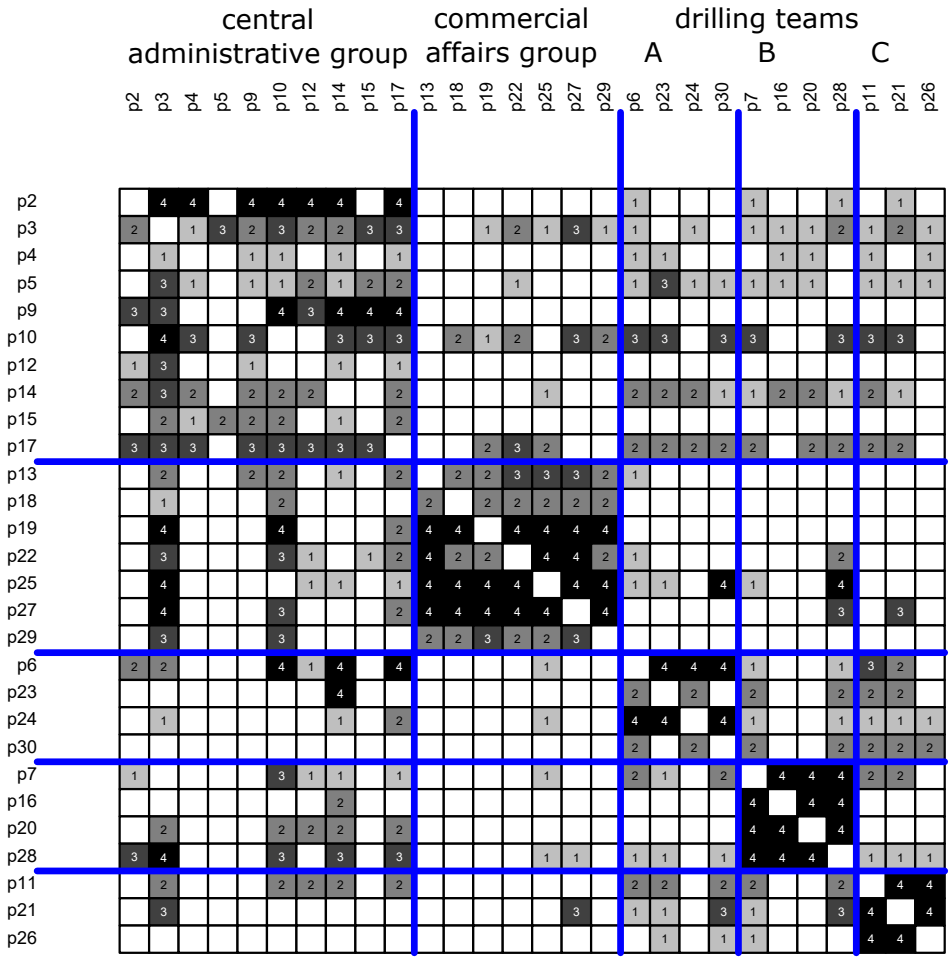


Figure: The confirmed network for providing help in problem solving (*PHPSc*) partitioned by work units

Valued *PHPSc* network and indirect blockmodeling

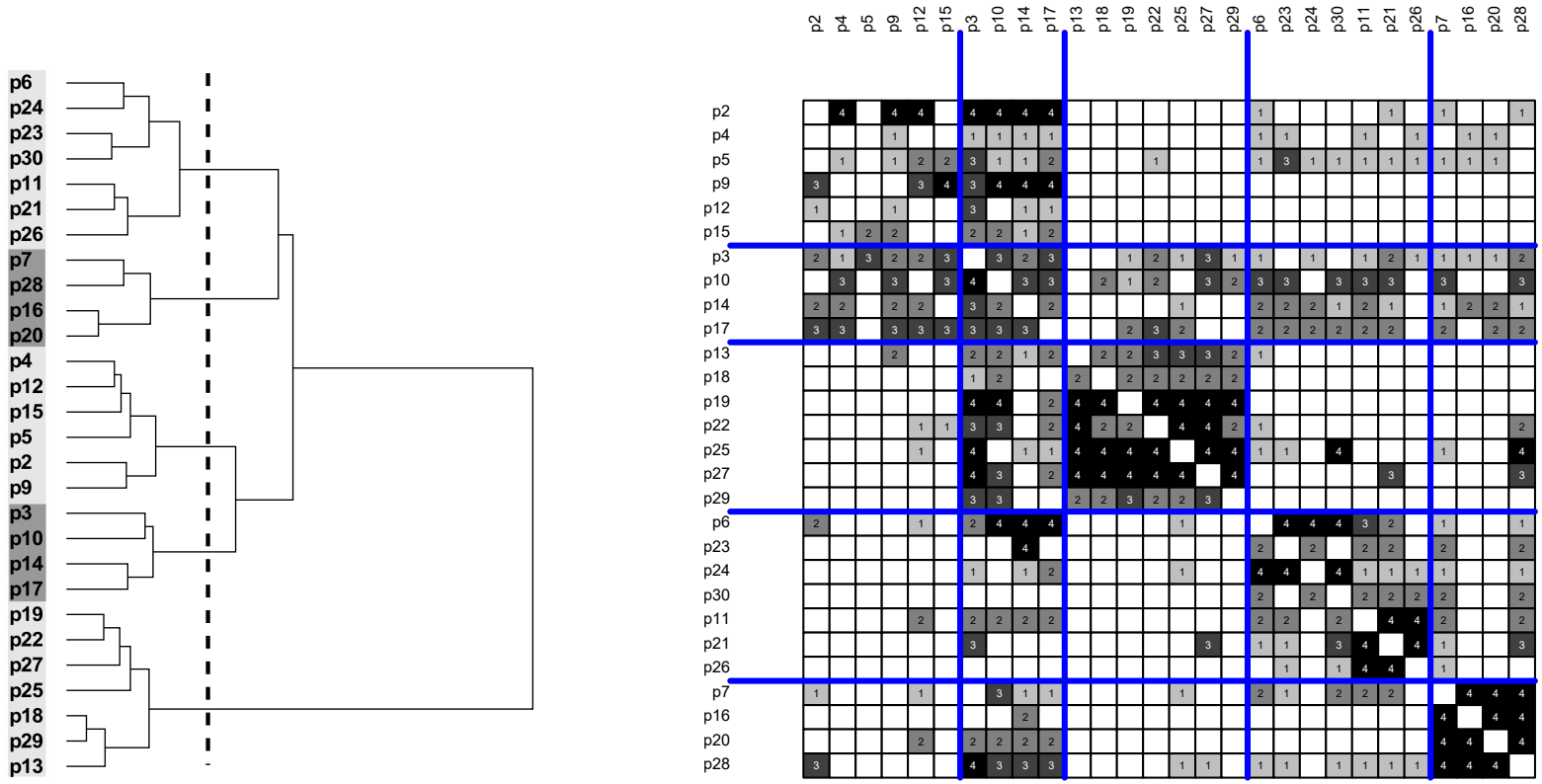


Figure: Dendrogram for the indirect blockmodeling of the *PHPSc* network and matrix representation with partitions into 5 clusters based on indirect blockmodeling

Results of simulation study for indirect blockmodeling of valued *PHPSc* network

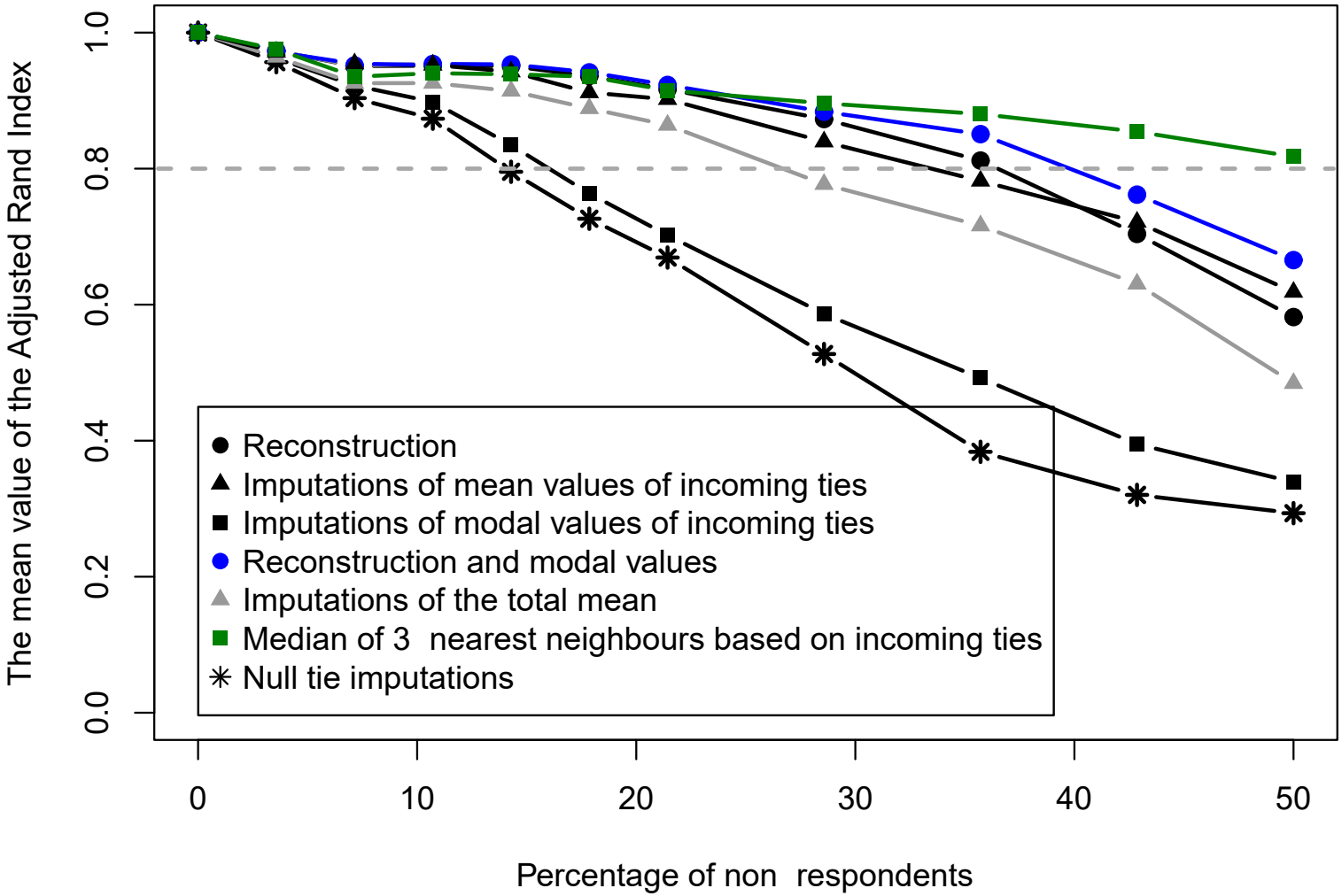


Figure: Results of simulation study for indirect blockmodeling of valued *PHPSc* network

Binarized *PHPSc* network and indirect blockmodeling

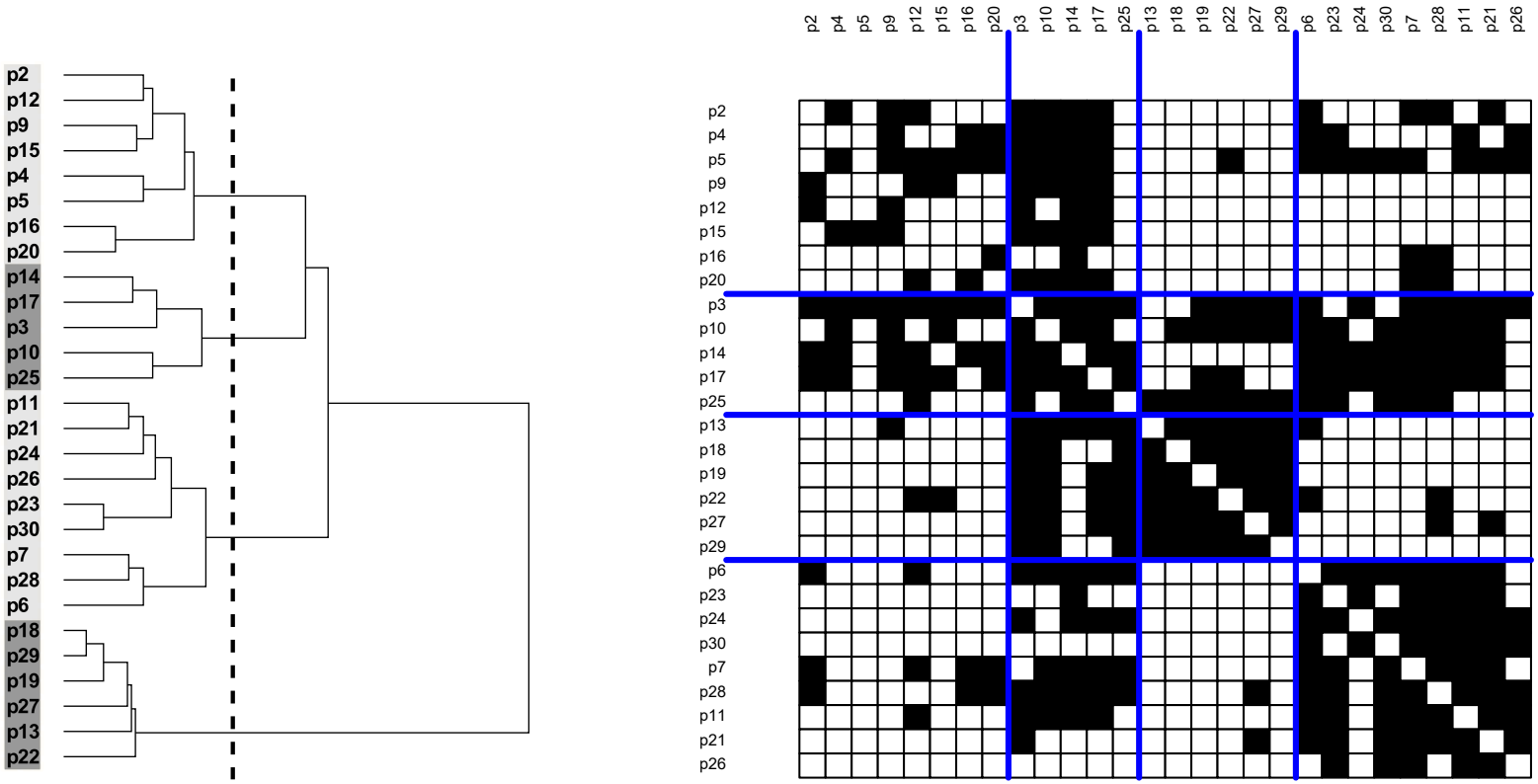


Figure: Dendrogram for the indirect blockmodeling of the *PHPSc* network and matrix representation with partitions into 4 clusters based on indirect blockmodeling

Results of simulation study for indirect blockmodeling of binarized *PHPSc* network

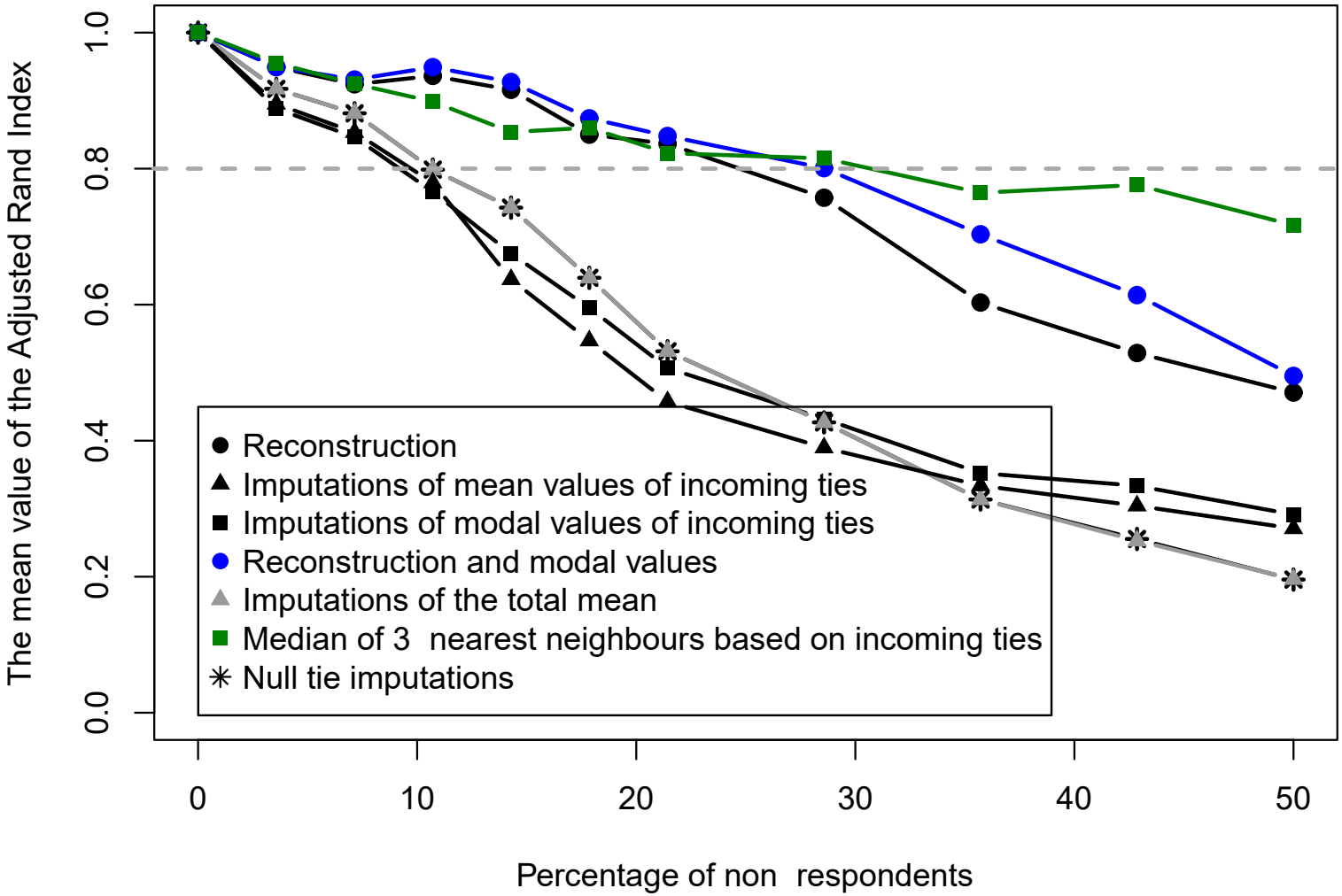


Figure: Results of simulation study for indirect blockmodeling of binarized *PHPSc* network

Binarized *PHPSc* network and direct blockmodeling based on structural equivalence

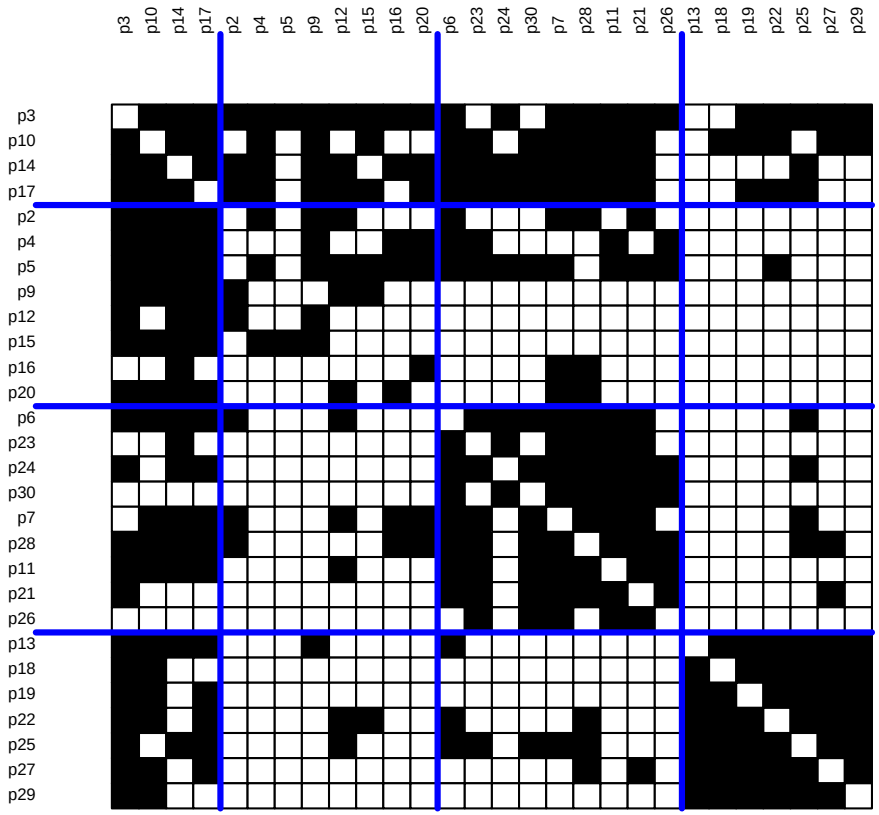


Figure: Matrix representation of binarized *PHPSc* network with partitions into 4 clusters based on direct blockmodeling

Results of simulation study for direct blockmodeling based on structural equivalence of binarized *PHPSc* network

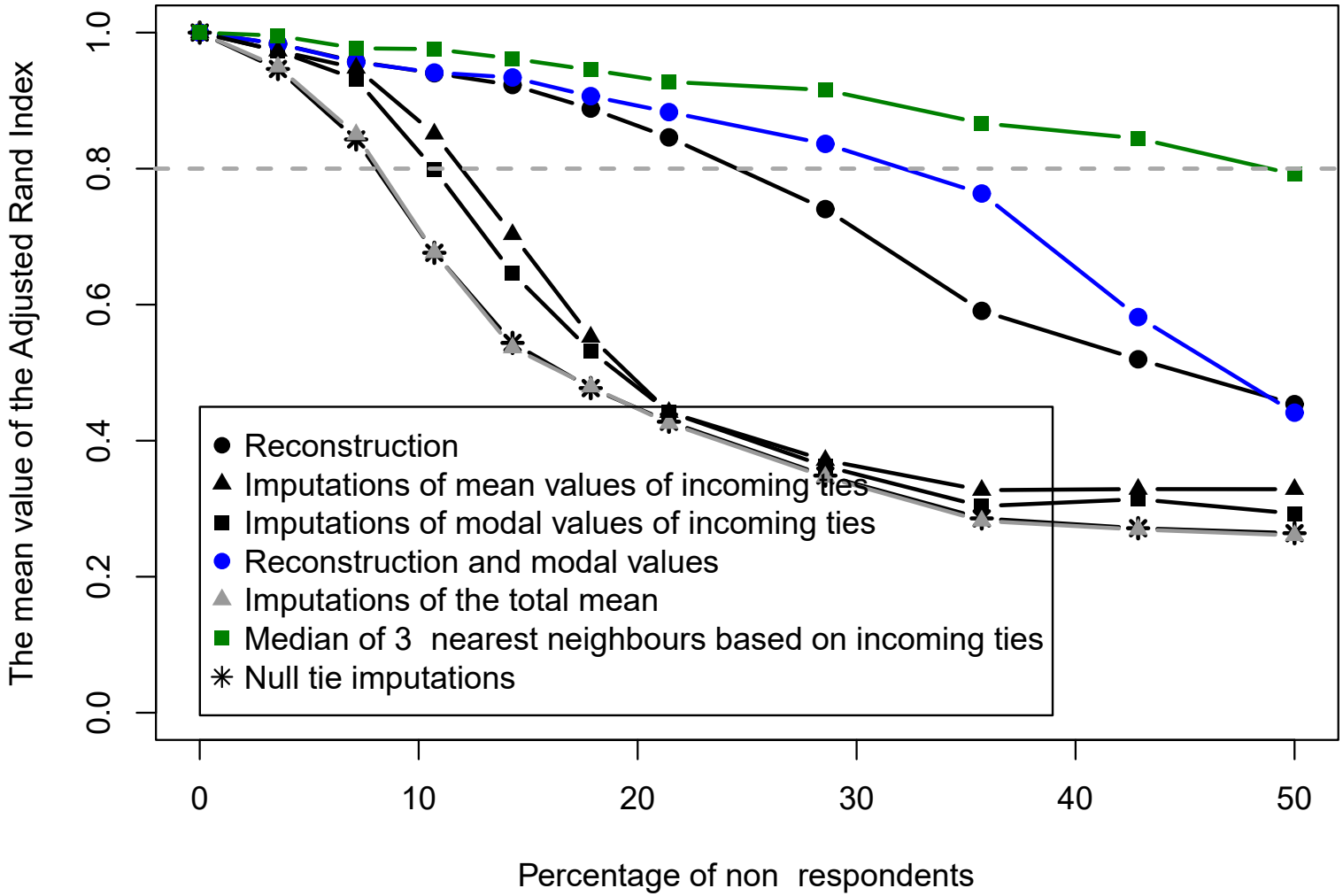


Figure: Results of simulation study for direct blockmodeling of binarized *PHPSc* network - the Adjusted Rand Index

Results of simulation study for direct blockmodeling based on structural equivalenc of binarized *PHPSc* network

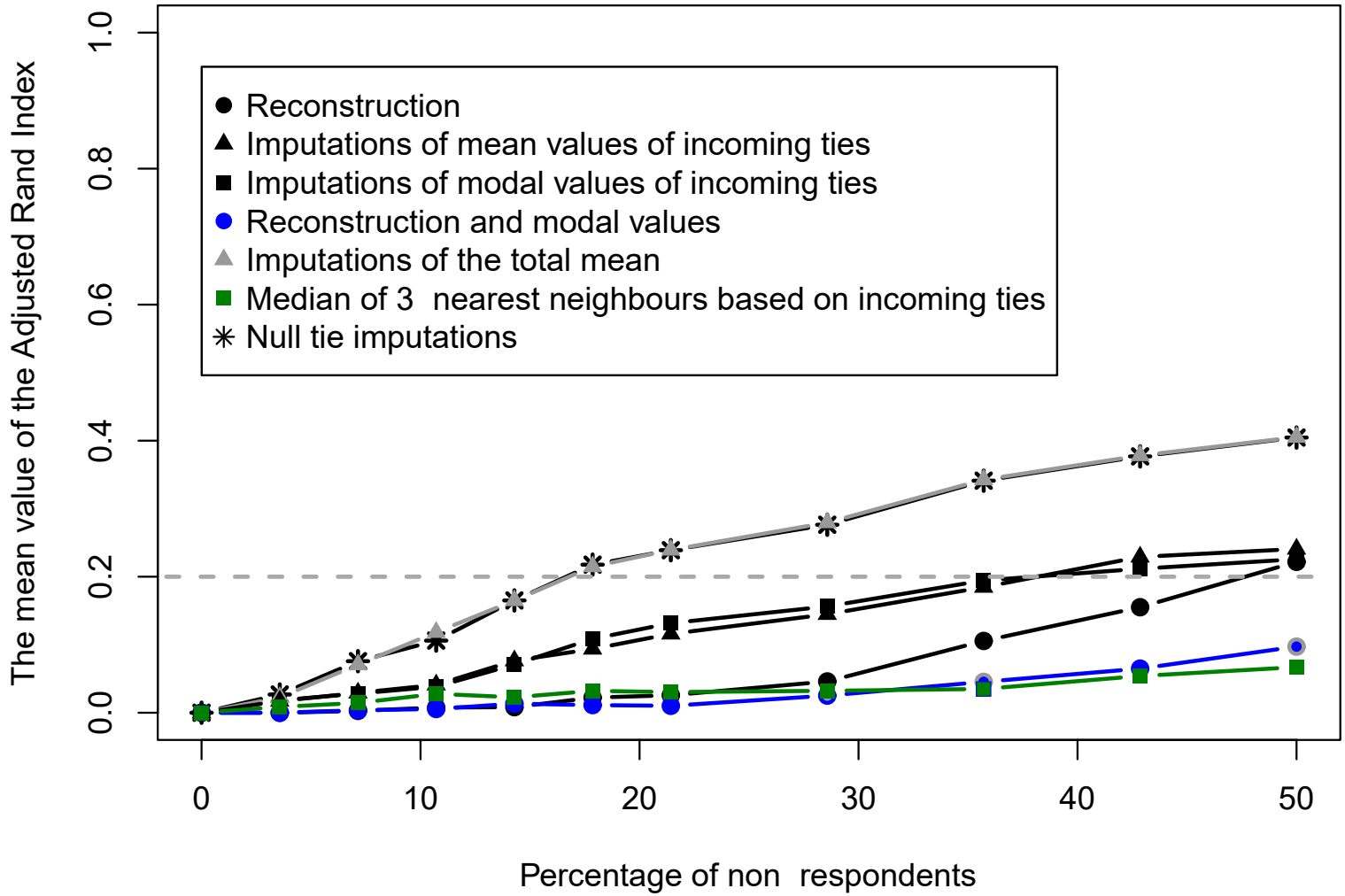


Figure: Results of simulation study for direct blockmodeling of binarized *PHPSc* network - the Proportion of incorrect block types

Conclusions

General guidelines for researchers

- Never disregard non-respondents.
- Recode absent ties with NA and report actor response rate.
- Do not impute 0's instead of absent ties, since there exist better solutions.

Guidelines for the researchers about non-response treatments

- Regardless the hypothesized blockmodel structures, the clustering procedure employed, and the level of weighted reciprocity, the most preferable actor non-response treatment is using 3-nearest neighbours based on incoming ties.
- The second best overall treatment is reconstruction combined with modal values for ties between non-respondents, especially if a network is highly symmetrical.

References

Batagelj, Vladimir, and Andrej Mrvar. 1996-2017a. *Pajek 5.01*. (May 5, 2017).

———. 1996-2017b. *Pajek and pajek-xxl, program for analysis and visualization of large networks, reference manual, list of commands with short explanation, version 5.01*.

Costenbader, Elizabeth, and Thomas W. Valente. 2003. The stability of centrality measures when networks are sampled. *Social Networks* 25(4):283 – 307.

Huisman, Mark. 2009. Effects of missing data in social networks. *Journal of Social Structure* 10(1).

Kossinets, Gueorgi. 2006. Effects of missing data in social networks. *Social Networks* 28(3):247 – 268.

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

Žiberna, Ales. 2008. *Blockmodeling 0.1.7: An R package for generalized and classical blockmodeling of valued networks*.

Žnidaršič, Anja, Patrick Doreian, and Anuška Ferligoj. 2017a. Actor non-response in valued social networks: The impact of different non-response treatments on the stability of blockmodels. *Social Networks* 48:46–56.

———. 2017b. Stability of centrality measures in valued networks regarding different actor non-response treatments and macro-network structures. *Network Science* Accepted.

Žnidaršič, Anja, Anuška Ferligoj, and Patrick Doreian. 2012. Non-response in social networks : the impact of different non-response treatments on the stability of blockmodels. *Social Networks* 34:438–450.

Thank you for your attention.