UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Anja Žnidaršič

Stability of blockmodeling

(Stabilnost bločnega modeliranja)

Doktorska disertacija

Ljubljana, 2012

UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Anja Žnidaršič

Mentorica: red. prof. dr. Anuška Ferligoj

Somentor: prof. dr. Patrick Doreian

Stability of blockmodeling

(Stabilnost bločnega modeliranja)

Doktorska disertacija

Ljubljana, 2012

Acknowledgements

This thesis would not have been possible without the help, guidance and support of my supervisor Professor Anuška Ferligoj. I would also like to thank my cosupervisor Professor Patrick Doreian for his enthusiasm and never-ending ideas.

I am also indebted to my committee members, Professor Valentina Hlebec and Professor Andrej Blejec, whose remarks helped me to improve the thesis.

Most of all, I would like to express my deepest gratitude to my family and friends for their support and all my classmates who provided me with great information resources and were always prepared for constructive debate.

Univerza
v Ljubljani
Fakulteta
za družbene vede

Kardeljeva ploščad 5
1000 Ljubljana, Slovenija
telefon 01 58 05 120

I Z J A V A O A V T O R S T V U
doktorske disertacije

Podpisani/-a Anja Žnidaršič, z vpisno številko 74060825, sem avtor/-ica doktorske disertacije z naslovom: Stability of blockmodeling (Stabilnost bločnega modeliranja).

S svojim podpisom zagotavljam, da:

- je predložena doktorska disertacija izključno rezultat mojega lastnega raziskovalnega dela;
- sem poskrbel/-a, da so dela in mnenja drugih avtorjev oz. avtoric, ki jih uporabljam v predloženem delu, navedena oz. citirana v skladu s fakultetnimi navodili;
- sem poskrbel/-a, da so vsa dela in mnenja drugih avtorjev oz. avtoric navedena v seznamu virov, ki je sestavni element predloženega dela in je zapisan v skladu s fakultetnimi navodili;
- sem pridobil/-a vsa dovoljenja za uporabo avtorskih del, ki so v celoti prenesena v predloženo delo in sem to tudi jasno zapisal/-a v predloženem delu;
- se zavedam, da je plagiatorstvo – predstavljanje tujih del, bodisi v obliki citata bodisi v obliki skoraj dobesednega parafraziranja bodisi v grafični obliki, s katerim so tuje misli oz. ideje predstavljene kot moje lastne – kaznivo po zakonu (Zakon o avtorski in sorodnih pravicah (UL RS, št. 16/07-UPB3, 68/08, 85/10 Skl.US: U-I-191/09-7, Up-916/09-16), prekršek pa podleže tudi ukrepom Fakultete za družbene vede v skladu z njenimi pravili;
- se zavedam posledic, ki jih dokazano plagiatorstvo lahko predstavlja za predloženo delo in za moj status na Fakulteti za družbene vede;
- je elektronska oblika identična s tiskano obliko doktorske disertacije ter soglašam z objavo doktorske disertacije v zbirki »Dela FDV«.

V Ljubljani, dne 8. 5. 2012 Podpis avtorja/-ice: _____

Abstract

Stability of blockmodeling

Social networks consist of actors and relations among them. An obvious graphical representation is a graph with vertices for actors and arcs for ties between them. Such a raw presentation usually does not provide a satisfying representation. The purpose of social network analyses is to detect simple and useful descriptions of the fundamental structures of relationships from large and seemingly incoherent networks. A widely used technique for finding such structural patterns is generalized blockmodeling.

The fact is that the social network data (usually gathered with surveys) are measured with errors. An error in social network analysis occur when there is an extra tie or a missing tie according to the true underlying and unobservable structure. Types of errors found in literature are classified into three categories: the boundary specification problem, errors caused by questionnaire format, and errors caused by actors. The boundary specification problem concerns rules of inclusion for actors in a network. In the realist approach actors in the network determine the boundaries of the network themselves, while in the nominalist approach the boundaries are determined by the researcher. Errors caused by questionnaire can be divided into three subcategories: errors caused by free or fixed choice design, using recall or recognition method, and direction of question. Errors caused by actors can be divided into three subgroups: actor non-response, non-response on tie, and measurement errors.

Combining both facts, wide usefulness of generalized blockmodeling and that social networks are measured with errors, results in the decision to investigate the impact of errors to the results of blockmodeling.

A result of a blockmodeling is a partition of actors determining positions of actors and an image matrix with determined block types. Actors are partitioned into clusters based on selected type of equivalence (the most known is the structural equivalence) or simply with selection of allowed block types according to philosophy of generalized equivalence. According to the result of blockmodeling two indices for comparison of blockmodels are presented. The first one is the Adjusted Rand Index which measures the agreement between pairs of partitions, and the second index compares block types and their position in the image matrices to compute the proportion of incorrect blocks.

In extensive simulation studies the blockmodels established from the whole starting network and blockmodels obtained from the measured network with introduced errors are compared with indices of blockmodeling stability described above. In the simulations studies two types of networks are used: real networks known from the literature and simulated network based on a desired structure where the size of the network, the number of clusters, types of blocks and their positions in image matrix, and probabilities of ties in blocks are taken into account.

First, the blockmodels of real networks gathered without limitation of number of choices (free choice design) are compared to the measured blockmodels established from networks with limited number of choices. In addition, the impact of direction of question (asking about giving or receiving social support) on blockmodeling is presented.

An extensive part of dissertation is dedicated to the actor (and tie) non-response. We investigated the impact of different non-response treatments on the identified blockmodels. Three different regimes of actor non-response are used (random selection of nonrespondents and based on indegree or outdegree) and then with those measured networks different non-response treatments are used. The simplest nonresponse treatment is the complete-case approach where nonrespondents are deleted from the network and therefore a smaller network is obtained. In the reconstruction procedure the unobserved outgoing ties are replaced with corresponding incoming ties for that actor. The deficiency of the reconstruction procedure is that for two nonrespondents the reconstruction of ties between them is not possible. Another treatment uses the modal value of incoming ties and is termed imputation based on mode. For binary networks this implies imputing ones if actors are popular given their received ties. The reconstruction procedure can be combined with imputations based on mode for ties between nonrespondents. Based on the results of the simulations we established that selection of the best nonresponse treatment depends on the level of the symmetry of the network.

The tie non-response occurs if an actor participates in the research, but does not provide the response on all network members. Recommendations about the best missing data treatment are almost the same as in the case of actor non-response.

The stability of blockmodeling to randomly changed ties is performed with different types of both, networks and equivalences. The structural equivalence turns out to be highly stable, which is not true in the case of the regular and generalized equivalence. The detailed insight into the performance of regular equivalence in network with randomly introduced small amount of errors is provided.

In addition, we try to answer the question if the relative changes in network characteristics and actors properties are able to predict the stability of blockmodeling and to what extent. One of the main conclusions is that all indices, which were used as predictors, have more power to predict the position membership of the actors than the percent of incorrectly identified block types.

In real studies the real underlying structure (presented with whole networks) is un-

known, which makes the comparison between whole and measured network impossible. Therefore, the impact of properties of measured network (alone) to the blockmodeling stability was investigated.

At the end limitations of the simulation studies are pointed out together with ideas for further research. We also provide short instructions for the researchers as the summary of our findings.

Keywords: social network, blockmodeling, stability of blockmodeling, error, non-response treatment

Povzetek
Stabilnost bločnega modeliranja

Socialna omrežja so sestavljena iz akterjev in relacij med njimi. Grafično jih lahko predstavimo z grafi, kjer točke predstavljajo akterje, usmerjene povezave pa relacije oziroma povezave med njimi. Takšna groba predstavitev navadno ne zagotavlja zadovoljivega prikaza. Namen analize socialnih omrežij je poiskati iz velikih, navadno nepovezanih omrežij, preprost in uporaben opis temeljnih struktur. Pogosto uporabljana tehnika za iskanje takih strukturnih vzorcev je posplošeno bločno modeliranje.

Dejstvo je, da so podatki socialnih omrežij (zbrani navadno z anketami) merjeni z napakami. Napaka se v socialnem omrežju pojavi, ko je v omrežju dodatna povezava ali ko povezava manjka glede na pravo prikrito strukturo. Tipe napak, ki smo jih našli v literaturi, smo razvrstili v tri skupine: problem določitve mej omrežja, napake, povzročene z zasnovo vprašalnika, ter napake, povzročene s strani akterjev. Problem določitve mej omrežja se nanaša na pravila za vključevanje akterjev v omrežje. V realističnem pristopu akterji sami določijo meje omrežja, medtem ko pri nominalističnem pristopu meje omrežja določi raziskovalec. Napake, povzročene z zasnovo vprašalnika, se nadalje delijo v tri podskupine: napake zaradi omejevanja oziroma neomejevanja števila izbir, uporaba metode prepoznavanja oziroma spominske metode in napake zaradi smeri zastavljenih vprašanj. Napake, povzročene s strani akterjev, so sestavljene iz treh podskupin: neodgovori akterjev, neodgovori na povezavi ter merske napake.

Kombiniranje obeh predstavljenih dejstev, torej pogostosti uporabe bločnega modeliranja in dejstva, da so socialna omrežja merjena z napakami, je pripeljalo do odločitve, da preučimo vpliv napak na rezultate bločnega modeliranja.

Rezultat bločnega modeliranja je razvrstitev akterjev, ki določa položaj akterjev, ter bločna matrika z določenimi tipi blokov. Akterji so razvrščeni v skupine na podlagi izbrane enakovrednosti (najbolj znana je strukturna enakovrednost) ali preprosto z izborom dovoljenih tipov blokov glede na koncept posplošene enakovrednosti. Skladno z rezultati bločnega modeliranja smo izbrali dva kazalnika za primerjavo bločnih modelov. Prvi kazalnik je prilagojeni Randov kazalnik, ki meri ujemanje med dvema razvrstitvama. Drugi kazalnik primerja tipe blokov in njihov položaj v bločni matriki ter se izračuna kot delež napačno razvrščenih blokov.

V obsežnih simulacijskih študijah z obema predstavljenima kazalnikoma primerjamo

bločne modele, dobljene iz popolnih omrežij, z bločnimi modeli izmerjenih omrežij. Pri tem uporabljamo dva tipa omrežij: iz literature znana realna omrežja in simulirana omrežja na podlagi želene strukture, kjer smo upoštevali velikost omrežja, število skupin, tipe blokov in njihov položaj v bločni matriki ter verjetnosti povezav v posameznih blokih.

Najprej tako primerjamo bločna modela, dobljena iz omrežij brez omejitev, z bločnim modelom, dobljenim iz omrežja z omejenim številom izbir. V nadaljevanju predstavimo vpliv smeri zastavljenega vprašanja (npr. dajanje oziroma sprejemanje socialne opore) na bločni model.

Obsežen del disertacije je posvečen neodgovorom akterjev ter neodgovorom na povezavah. Raziskovali smo vpliv različnih tretmajev za manjkajoče podatke na postavljeni bločni model. Nerespondente smo generirali na tri različne načine (naključno in na podlagi vhodne oziroma izhodne stopnje), nato pa smo na teh izmerjenih omrežjih uporabili različne tretmaje. Najpreprostejši tretma je pristop popolnih podatkov, kjer so nerespondenti odstranjeni iz omrežja, tako da dobimo v bistvu manjše omrežje. Pri rekonstrukciji so nezabeležene izhodne povezave zamenjane z ustreznimi vhodnimi povezavami. Pomanjkljivost rekonstrukcije je, da manjkajočih povezav med dvema nerespondentoma ne moremo nadomestiti brez dodatnih imputacij. Tretji tretma uporablja modus vhodnih povezav za nadomeščanje manjkajočih vrednosti, zato ga imenujemo imputacije na podlagi modusa. Rekonstrukcijo lahko kombiniramo z imputacijami na podlagi modusa za povezave med nerespondenti. Na podlagi rezultatov simulacij smo ugotovili, da je izbira najboljšega tretmaja za manjkajoče podatke zaradi neodgovorov odvisna od stopnje simetrije omrežja.

Neodgovor na povezavi povzročijo akterji, ki sicer sodelujejo v raziskavi, vendar ne zagotovijo odgovorov o vseh povezavah. Priporočila glede najboljšega tretmaja so zelo podobna kot v primeru neodgovorov akterja.

Stabilnost bločnega modeliranja na naključno spremenjene povezave smo izvedli z različnimi tipi omrežij in enakovrednosti. Strukturna enakovrednost se je izkazala kot izjemno stabilna, medtem ko velja za regularno in posplošeno enakovrednost ravno obratno. Prikazali smo tudi natančnejši vpogled v obnašanje regularne enakovrednosti v primeru majhnega odstotka naključnih napak.

V nadaljevanju smo poskušali odgovoriti še na vprašanje, ali lahko relativne spremembe v karakteristikah omrežja in značilnostih akterjev napovedo stabilnost bločnega modeliranja in v kakšnem obsegu. Ena izmed glavnih ugotovitev je, da so vse v modelih uporabljene spremenljivke uspešnejše pri napovedovanju razvrstitve akterjev kot pri napovedovanju odstotka napačno razvrščenih blokov.

V realnih raziskavah ne poznamo resnične prikrite strukture (predstavljene s popolnim omrežjem), zato je tudi primerjava popolnega in izmerjenega omrežja nemogoča. Zato smo namesto razlik v karakteristikah omrežja preučili tudi vpliv lastnosti izmerjenega omrežja na stabilnost bločnega modeliranja.

Na koncu so predstavljene omejitve raziskave oziroma simulacij skupaj z idejami za nadaljnje raziskovalno delo. Na podlagi dobljeni rezultatov smo podali tudi nekaj kratkih napotkov oziroma priporočil za raziskovalce socialnih omrežij, ki se nanašajo tako na samo zasnovo raziskave kot tudi na analizo omrežij z bločnim modeliranjem.

Ključne besede: socialno omrežje, bločno modeliranje, stabilnost bločnega modeliranja, napaka, tretma neodgovorov

Contents

List of Figures

List of Tables

1 Introduction

The aim of this dissertation is to investigate how stable are established blockmodels, and hence blockmodeling, to different types of errors in the research design. The impact of different types of errors to different network characteristics has been examined by several authors but never with generalized blockmodeling (Doreian et al., 2005).

1.1 A short overview

Generalized blockmodeling is an useful, increasingly and widely used technique for finding structural patterns in social networks. Network consists of a set of actors with relation(s) defined on them. A relation can be any type of contact, connection, or a tie between a pair of actors (Knoke and Yang, 2008). The goal of the blockmodeling is to reduce a large incoherent network to a smaller comprehensible and simply interpretable structure (Batagelj et al., 2004). In more detail, the purpose 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 are determined by the positions (Faust and Wasserman, 1992; Doreian et al., 2005).

Actors are partitioned into clusters based on some type of equivalence. The best known and widely used types are structural and regular equivalence. The extension is generalized equivalence which can be defined by a set of allowed block types.

The appropriate ways of comparing two blockmodels are established. The result of using a blockmodeling procedure is a partition (of actors) determining positions and image matrix with selected block types. The stability of a blockmodel to an error can

be defined or measured with two indices. The whole starting blockmodel and the measured blockmodel from network with introduced errors have to be be compared. The first index, the Adjusted Rand Index, measures the agreement between both partitions and the second index compares block types in image matrices and their positions and is calculated as the percent of incorrectly identified block types. The described indices agree with two central ideas of social network analysis pointed out by Doreian (2008).”The first is that the structure of a social network, as a whole, is important to collective outcomes at the level of the network. The second is that the location occupied in a network is important for outcomes at the actor level” (Doreian, 2008, pg. 3).

The errors in the research design can be classified into three categories: the boundary specification problem, errors caused by design, and errors caused by actors. A questionnaire can be a large source of errors, especially with specification of number of choices and recall method. The impact on the established blockmodel also has the direction of question where the perceptions of giving or receiving of social support can be gathered. An important source of errors could also be actors themselves. They could refuse to respond to the entire questionnaire or only to a particular tie. For actor (and tie) non-response different possible treatments are examined, such as the complete-case approach, reconstruction procedure and imputations. The measurement errors where there is a discrepancy between the true value of a concept and the observed (or measured) value of that concept. The definition of measurement error in the social network analysis is presented together with its main sources.

In this dissertation an evaluation of an impact of different errors to the blockmodeling results is investigated. First, the scheme of simulation studies is presented together with whole networks and their blockmodels used in the simulations. Beside the real networks known from the literature simulated networks with desired parameters are used as well. In the studies the amount and types of errors are controlled. The most stable type of equivalence and the best treatments in case of non-response are determined.

The predictive power of differences in network characteristic between whole starting

network and measured network with introduced errors to the stability of blockmodeling is examined. In addition, the characteristics of measured network and their impact of blockmodeling stability is presented.

Simulations are performed in an R environment with a package called Blockmodeling (Žiberna, 2008) and the visualization of networks is made in Pajek (Batagelj and Mrvar, 2010a,b).

1.2 Structure of the dissertation

The first part includes Chapters 2 to 4 where the main topics are introduced with the overview of the relevant literature. In Chapter 2 basic definitions of a network and relations with aim of the social network analysis are presented. The network characteristics and measures of centrality and prestige that are analyzed in the simulations are presented. In Chapter 3 the purpose and concepts of generalized blockmodeling are presented together with definitions of structural, regular, and generalized equivalence. An overview of different approaches to blockmodeling is presented with an emphasis on generalized blockmodeling. In Chapter 4 the review and classification of errors in research design are presented. The main categories of errors are: the boundary specification problem, errors introduced by design, and errors caused by actors. The comparison of errors from the social network data collection process and from the ordinary surveys is provided.

In Chapter 5 two indices for measuring the blockmodeling stability are presented: the Adjusted Rand Index and the proportion of incorrect block types.

Chapter 6 presents the basic scheme of simulations and two broad types of networks which are used in the evaluation of blockmodeling stability. First, the networks are classified into real and simulated networks and later also according to the type of equivalence used in blockmodeling procedure.

Chapter 7 is the core chapter of the thesis where the evaluation of stability of blockmodeling on design errors is presented. Section 7.1 evaluate the blockmodeling results if the number of choices is limited to the fixed number of actors instead of the free choice design. The impact of the direction of a question on the established blockmodeling is presented in Section 7.2. Extensive studies of actor non-response are presented in Section 7.3. Section 7.4 presents an overview of results of a tie non-response in real networks presents. Random measurement errors and their impact on stability of blockmodeling are presented in Section 7.5 where we also try to answer the question of which type of equivalence produces the most stable results.

In Chapter 8 the impact of differences in network characteristics on the stability of blockmodeling and the impact of characteristics of measured network on the blockmodeling results are examined.

In the last chapter, Chapter 9, the obtained results are evaluated. In addition, an overview of the thesis is provided with the scientific contributions. At the end, the extended summary in Slovene is provided.

2 Networks and basic definitions

In this chapter networks and network characteristics are introduced. The analysis of social networks examines the relationships of social actors, rather than only the characteristics of individuals as general social research. The main network characteristics, which are used for the estimation of blockmodeling stability in Chapter 8, are presented at the end of the chapter.

2.1 Networks

Social networks are fundamental to social life. In an intuitive way, a social network can be defined as follows: "Social network consist of a finite set or sets of actors and the relation or relations defined on them" (Wasserman and Faust, 1998, 21).

Vertex is the smallest unit in a network (de Nooy et al., 2005). In case of social networks the term vertex is replaced by the term **actor** (Wasserman and Faust, 1998; Knoke and Yang, 2008), where it can represent individuals or collective social units, such as formal or informal organizations. Our main focus in the dissertation are social networks from their research design to the established blockmodels, therefore mainly the term actor will be used. There are some exceptions, where synonyms are used, especially in general definitions or in direct citations.

The number of distinct sets of units in a network determines the mode of a network. If the actors in a network are from one set, then we have one-mode network. If the actors in the network are classified into two sets, then we have two-mode network. Usually, ties in the two-mode network lead from actors from the first set to actors in the other set

(Wasserman and Faust, 1998). In this dissertation only one-mode networks are used.

"**A relation** is generally defined as a specific kind of contact, connection, or tie between a pair of actors" (Knoke and Yang, 2008, 7). Knoke and Kuklinski (1982) distinguish two parts of the relation; content and form. The content refers to the type of connection (e.g. be a friend, helping, gossip...). Two basic aspects of form are the intensity or strength of the link between two actors and level or frequency of contacts.

In mathematical notation, the network can be written as $N = \{A, R_1, R_2, \ldots, R_r\}$, where $A = \{a_1, a_2, ..., a_n\}$ is a finite set of actors (or units). Connections among actors are described using one or more binary relations $R_i \subseteq A \times A, i = 1, \ldots, r$. If a network has only one relation (as all networks in this dissertation), it can be represented by a graph, in which actors are presented by vertices (or nodes), directed relations by arcs, and mutual relations by edges (de Nooy et al., 2005).

Another representation of a network is a sociomatrix R with n rows and n columns, where entry r_{ij} indicates the strength and/or sign of a relation between actors i and j (Wasserman and Faust, 1998, 80). The relation in the valued network can take value from categorical or interval scale, with all possible values or just nonnegative ones. If the relation between actors can be positive or negative, then we have signed network. In the case of binary networks, which are used in this dissertation, only the presence or absence of a tie is important. The values of r_{ij} are 0, if actor i is not in relation with actor j ($a_i R a_j$), and 1 otherwise. The diagonal elements of sociomatrix R are so called 'self-choices', which are usually (in social network analysis) undefined and set to 0.

As written above, networks can be classified into several types based on different criteria; number of sets of actors, number of relations and possible values of relations (the measurement scale of relations). In this dissertation, according to the above classification, the one-mode binary networks with one relation will be used. Another distinction can be made based on the number of actors; networks with some 10 actors are called small, and networks with some 1000 actors are called large. Because the blockmodel-

ing method used in the dissertation is computationally intensive, only small networks (mainly up to 20 actors) are used.

2.1.1 Social network analysis

"The goal of the network analysis is to create, from raw relational data, a useful description of a system of relationships" (Stork and Richards, 1992, 194). Wasserman and Faust (1998, 4) stated that "social network analysis is based on an assumption of the importance of relationships among interacting units". Probably the most important relational concepts are that actors and their actions should be viewed as interdependent rather than independent units, and the relational ties between actors are channels to transfer the (material or nonmaterial) resources.

Knoke and Yang (2008, pg. 4-6) state three underlying assumptions about pattern relations and their effects:

(i) Structural relations are more important for understanding observed behavior than general attributes as age, gender, and ideology.

(ii) Social networks affect perceptions, beliefs, and actions through variety of structural mechanisms that are socially constructed by relations among actors.

(iii) Social relations should be viewed as a dynamic process and not as static structure.

Social network analysis can be also viewed as a collection of methods and models (Wasserman and Faust, 1998; Scott, 2000). Scott (2000, 38) stated that "social network analysis emerged as a set of methods for the analysis of social structures, methods that specifically allow an investigation of the relational aspects of these structures".

One useful, increasingly and widely used technique for finding structural patterns in networks is generalized blockmodeling, which is the main focus of this dissertation. It is presented in the next chapter.

2.2 Network characteristics

The numerical characteristics of a network can be a single number as in case of network density, network reciprocity or number of different types of dyads (presented in Section 2.2.1). Another set of measures is calculated for each vertex in a network, e.g. measures of centrality and/or prestige. The most important measures, which are also used for analysis of stability of blockmodeling in Chapter 8, are presented in Sections 2.2.2 and 2.2.3.

2.2.1 Characteristics of a network as a whole

We will denote the number of vertices (or actors) in a graph with n, and number of directed lines or arcs with m. The density of the network describes general level of linkage among the actors in a network (Scott, 2000, pg. 93). In the directed network it is calculated as number of arcs in a network, divided by the number of all possible arcs in a network (Wasserman and Faust, 1998):

$$\Delta = \frac{m}{n(n-1)} \,. \tag{2.1}$$

The equation 2.1 presumes that loops are not allowed in a network, otherwise the denominator in the above equation will be n^2. The density is a real number between 0 and 1, and is equal to 1 if the network is complete (each vertex has an arc to all other vertices).

Another concept is to look at pairs of actors to investigate the relationship between them. A dyad is a subgraph, a subset of two nodes from the network and all arcs between them. There are $\binom{n}{2} = \frac{n(n-1)}{2}$ dyads in a network with possible states (Holland and Leinhardt, 1970; Wasserman and Faust, 1998). A mutual relationship between actor i and j exist when there is an arc from actor i to actor j ($i \rightarrow j$) and an opposite arc from actor j to actor i ($j \rightarrow i$). We can also say that there is an edge from actors i and j. There is an asymmetric relationship between actors if there is either an arc $i \rightarrow j$ or $j \rightarrow i$, but not both. The null dyad occurs, if neither actor from the pair has a tie to the other actor. The dyad census is a triple (M, A, N), where M is a number of mutual

dyads, A is a number of asymmetric dyads and N is a number of null dyads in a network.

Reciprocity (Huisman, 2009) measures how symmetric is a network and it is defined, for directed networks, as

$$reciprocity = \frac{2 \cdot M}{2 \cdot M + A}, \tag{2.2}$$

where M indicates the number of mutual dyads and A the number of asymmetric dyads.

Beside the dyad structure properties, also the triad structure of the network can be investigated. A triad is defined as a subgraph of three nodes with arcs between them (Faust, 2007). In the extensive study she showed that the majority of variance in triadic census can be explained by "properties that are more local than triadic - network density, the indegree and outdegree, and the distribution of mutual, asymmetric, and null dyads" (Faust, 2007, 242-243). This is the main reason why triadic census was not included in our study.

2.2.2 Measures of centrality and prestige

Variety of measures can be calculated from the structure of a network to determine the most important or the most central actors in a network. Measures of centrality and prestige can be defined in two different ways according to 'objects' of interest. If we investigate the position of individual actors within the network we talk about actor centrality (where the result is the calculated number for each actor). On the other hand, the term centralization is used when we want to characterize the whole network with a single numeric value (de Nooy et al., 2005).

The most important distinction among the centrality measures is based on the type of relations in a network (Batagelj, 1993). If the relations in a network are considered to be directed, then measures of importance are calculated. Two subgroups of measures of importance (or prestige) can be calculated; the measures of influence take into account the number of outgoing ties and the measures of support are calculated based

on the number of incoming ties. The term centrality measures is used when we talk about undirected networks. It is important to emphasize that the measures described below only present the subset of the most known and widely accepted measures in the literature.

Measures of centrality and prestige based on degree

The degree centrality is the simplest measure, where actor is the most central in a network, if it has the most ties to other actors (Wasserman and Faust, 1998). Absolute measure of degree centrality of actor a_i is defined as the number of ties with other actors in a network (2.3):

$$c_D(a_i) = deg(a_i) = \sum_{j=1}^{n} r_{ij} \,, \tag{2.3}$$

where r_{ij} is in binary network equal to 1 if there is a tie between actors i and j, and 0 otherwise (in valued networks r_{ij} represents the weight of a tie). The degree of an actor depends on the size of a network, which means that $c_D(a_i)$ can not be used for comparison of networks with different sizes. The normalized or relative degree centrality is defined as

$$C_D(a_i) = \frac{c_D(a_i)}{n-1} \,, \tag{2.4}$$

where the maximal absolute degree of a unit (in a network without loops) is $n-1$. $C_D(a_i)$ can take the values from 0 to 1, where 1 indicates that unit i is connected to all other units and 0 indicates that unit is isolated.

In directed networks the concept of degree centrality can be extended to indegree centrality, where we take into account just incoming ties of an actor and we measure the support. The influence of an actor can be measured by outdegree centrality, where just outgoing ties are counted. The relative measures of indegree and outdegree centrality are calculated in the same way as relative all-degree centrality in Equation 2.4. All three types of centrality are examples of local measures, where only immediate neighbours of a unit are take into account. This seems to be the main deficiency of those measures (Scott, 2000).

Measures of closeness centrality and prestige

Global measures consider all units connected by paths with a given unit (Batagelj, 1993). The measure of global centrality is closeness centrality first suggested by Sabidussi in 1966 (in Freeman 1978-1979). The Sabidussi's index of absolute actor closeness (Freeman, 1978-1979; Wasserman and Faust, 1998) is defined as

$$c_C(a_i) = \sum_{j=1, i \neq j}^{n} \frac{1}{d(a_i, a_j)} , \tag{2.5}$$

where $d(a_i, a_j)$ is the geodesic distance between actors a_i and a_j. The relative closeness centrality (Equation 2.6) is obtained from Equation 2.5 with multiplication by $(n-1)$[1]:

$$C_C(a_i) = (n-1) \cdot c_C(a_i) = \sum_{j=1, i \neq j}^{n} \frac{n-1}{d(a_i, a_j)} , \tag{2.6}$$

If network is not strongly connected, only reachable actors are taken into account and the result is weighted with number of reachable actors. The relative closeness centrality ranges between 0 an 1 (when actor is adjacent to all other actors) and can be interpreted as "the inverse average distance between actor i and all other actors" (Wasserman and Faust, 1998, 185). "The closer a vertex is to all other vertices, the easier information may reach it, the higher its centrality" (de Nooy et al., 2005, 127).

In the directed networks the prestige can be computed according to outgoing arcs (out-closeness or closeness based on outdegree), which can be interpreted in how many steps we can reach all other actors from the selected one or according to incoming arcs (in-closeness or closeness based on indegree), which can be interpreted as how close is the selected actor to all others. The relative out-closeness and in-closeness can be computed with formula in Equation 2.6, where just outgoing and incoming ties, respectively, are taken into account.

Measure of betweenness centrality

The betweenness centrality similarly as closeness centrality takes into account the geodesic

[1]When the actor is adjacent to all other actors, the smallest possible distance of selected actor to all other actors is obtained. In this case the maximum of absolute closeness centrality $c_C(a_i)$ is obtained and it is equal to $\frac{1}{n-1}$. In normalization of $c_C(a_i)$, the index is divided by its maximal value, which is equal to multiplication with $n-1$.

distance. It reveals how important an actor is due to his position in a network to control the flow of information. The main idea of betweenness centrality is that "an actor is central if it lies between other actors on their geodesics" (Wasserman and Faust, 1998, 189). The betweenness centrality of an actor a_i is defined as

$$c_B(a_i) = \sum_{j<k}^{n} \frac{g_{jk}(a_i)}{g_{jk}} , \tag{2.7}$$

where $i \neq j \neq k$ and g_{jk} is the number of all geodesics between two actors j and k and $g_{jk}(a_i)$ is the number of geodesics between actors j and k that contain actor i.

The relative betweenness centrality is defined separately for directed and undirected networks (Wasserman and Faust, 1998, 189). In case of undirected networks, the maximal value of $c_B(a_i)$ is $\binom{n-1}{2} = \frac{(n-1)(n-2)}{2}$, which is the number of different pairs of actors not including actor a_i (the maximal index is obtained when actor a_i falls on all geodesics among pairs of other actors). The maximal value of absolute betweenness is, in the case of directed network, equal to $(n-1)(n-2)$, because the order of choosing actors in pair is important. The relative betweenness can be calculated as:

$$C_B(a_i) = \begin{cases} \frac{c_B(a_i)}{\frac{(n-1)(n-2)}{2}}; & \text{for undirected networks} \\ \frac{c_B(a_i)}{(n-1)(n-2)}; & \text{for directed networks} \end{cases} . \tag{2.8}$$

The minimal value of relative betweenness (Equation 2.8) is 0, when actor is not located on any geodesics among pairs of other actors and maximal value 1, when all geodesics among pair of other actors include this actor.

2.2.3 Measures of prestige

Prestige measures are computed for directed networks only, since for this measures the direction is an important property of the relation.

Proximity prestige

For calculation of proximity prestige, the idea of an influence domain is used. The influence domain (or input domain) of an actor in a directed network "is the number or proportion of all other vertices (actors) which are connected by a path to this vertex

(actor)" (de Nooy et al., 2005, 193). If the network is strongly connected, there are all other actors in influence domain of every actor, so the distinction between actors is poor (de Nooy et al., 2005). The I_i is defined as the number of actors in the influence domain of actor i and equals the number of actors which can reach the actor i.

Proximity prestige was suggested by Lin in 1976 (in Wasserman and Faust 1998, 203) and is defined as

$$P_P(a_i) = \frac{I_i/(n-1)}{\sum_{j=1}^{n} d(a_j, a_i)/I_i}\,, \tag{2.9}$$

where the sum is taken over all actors j in the influence domain of actor i. The proximity prestige $P_P(a_i)$ of actor a_i is the proportion of all actors (except itself) in its influence domain ($I_i/(n-1)$) divided by the mean distance from all actors in its influence domain ($\sum_{j=1}^{n} d(a_j, a_i)/I_i$) (Wasserman and Faust, 1998; de Nooy et al., 2005). The proximity prestige index $P_P(a_i)$ has maximal value 1, when all actors in a network are adjacent to actor a_i, and minimal value 0, if actor a_i is unreachable. The proximity prestige for strongly connected networks is equal to input closeness centrality, therefore its computation is reasonable only for weakly connected networks.

Hubs and authorities

Two measures of prestige, which are especially useful in case of directed networks of web pages, are hubs end authorities. Kleinberg (1998, 8) stated, in context of web pages, that "Hubs and authorities exhibit what could be called a *mutually reinforcing relationship*: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs." In terminology of social network, we can say that an actor is a good hub, if it points to many good authorities, and an actor is a good authority, if it is pointed to by many good hubs.

For each actor a two weights are calculated; weights of good hubs x_a, and weights of good authorities y_a. Weights are computed according to network A by solving the eigenvector problem of matrices AA^T, where the first eigenvector represent weights for good hubs, and the first eigenvector of a matrix A^TA represents weights of good authorities.

3 Blockmodeling

In this section the purpose of blockmodeling is presented (Section 3.1) together with different types of equivalence (Section 3.2) and the main concepts of generalized blockmodeling (Section 3.4).

3.1 Description and purpose of blockmodeling

One goal of blockmodeling is to reduce a large incoherent network to a smaller comprehensible and simply interpretable structure (Batagelj et al., 2004). In more detail, the purpose of blockmodeling procedures is to partition the network actors into clusters (subgroups called *positions*), and, at the same time, partition the set of ties into *blocks* which are determined by the ties between actors in positions (Wasserman and Faust, 1998; Doreian et al., 2005). Similarly, (Batagelj, 1997, 143) stated that blockmodeling has two basic subproblems: (i) partitioning of units or determining the classes (clusters) that form the vertices in a model, (ii) determining the links in a model (and their values). In the second definition Batagelj emphasizes the importance of determining the ties in a model, which essentially consist of the ties within and between clusters.

The actors within a cluster should have the same (or a very similar) pattern of ties based on a selected equivalence. The resulting blockmodel is a compact representation of a network, a model, which represents essential structure of a network, which is easier to interpret. The blockmodel can be presented by relational matrix called 'image' matrix or by a 'reduced' graph. The units in this image representation of a blockmodel are positions made up of equivalent actors and the arcs (summarizing blocks) represent ties between positions (Doreian et al., 2005).

Beside the notation used in Section 2.1, some additional notation will be needed. We denoted the network of actors and relations between them as:

- $N = \{A, R_1, R_2, \ldots, R_r\}$, where $A = \{a_1, a_2, ..., a_n\}$ is a finite set of actors. Connections among actors are described using one or more binary relations $R_i \subseteq A \times A, i = 1, \ldots, r$.

- Relation R can be presented with the sociomatrix $R = \left[r_{ij}\right] \in \mathbb{R}^{n \times n}$, where r_{ij} indicates the strength and/or sign of a relation between actors i and j.

In addition to this general social network notation, the special notation for blockmodel procedure is:

- $C = \{C_1, C_2, \ldots, C_k\}$ is a partition of actors A into k clusters. Beside the clustering of actors, the clustering C partitions also the relation R into blocks

$$R\left(C_i, C_j\right) = R \cap C_i \times C_j \, .$$

  "Each block is defined in terms of units (actors) belonging to clusters C_i and C_j and consist of all arcs from units (actors) in cluster C_i to units (actors) in cluster C_j. If $i = j$ the block $R\left(C_i, C_i\right)$ is called diagonal block" (Doreian et al., 2005, 169).

Doreian et al. (2005, 169) emphasized that the term 'block' has two meanings in the literature. In the first usage a block is a set of actors grouped together into a blockmodel and the second usage, which is preferred by authors, is that "a block is a relation between two clusters of units (actors)".

3.2 Types of equivalence

Blockmodel partitioning is based on "the idea that units (actors) in a network can be grouped according to the extent to which they are equivalent in terms of some meaningful definition of equivalence" (Doreian et al., 2005, 170). To summarize Faust's (1988) discussion, there are two basic approaches to the equivalence, regardless to the definition used:

(i) the equivalent actors have the same connection pattern to the **same** neighbors;

(ii) the equivalent actors have the same or similar connection pattern to (possibly) **different** neighbors.

The first type of equivalence is the structural equivalence and the second one is generalized type of the structural equivalence called regular equivalence. Both formal definitions are presented below.

3.2.1 Structural equivalence

The structural equivalence is probably the most commonly used type of equivalence. Lorrain and White (1971) give the definition as: Actors are structurally equivalent if they are connected to the rest of the network in identical ways. The definition can be written in mathematical notation as follows (Batagelj et al., 1992b; Doreian et al., 2005):

Actors (or units) x and y are structurally equivalent ($x \equiv y$) if and only if

(i) $xRy \Leftrightarrow yRx$,

(ii) $xRx \Leftrightarrow yRy$,

(iii) $\forall z \in A \setminus \{x, y\} : (xRz \Leftrightarrow yRz)$,

(iv) $\forall z \in A \setminus \{x, y\} : (zRx \Leftrightarrow zRy)$, where A is set of actors[2].

In the matrix notation, the above definition can be written as follows: Actors a_i and a_j[3] are structurally equivalent if and only if

(i) $r_{ij} = r_{ji}$,

(ii) $r_{ii} = r_{jj}$,

(iii) $\forall k \neq i,\ j : r_{ik} = r_{jk}$,

(iv) $\forall k \neq i,\ j : r_{ki} = r_{kj}$.

[2] In the cited literature notation E and U is used for the set of units.

[3] x_i and x_j is used in the cited definition instead of a_i and a_j

Batagelj et al. (1992b) prove that for structural equivalence there are only four possible ideal blocks (Table 3.1): null, complete, null diagonal (zeros on the diagonal, one out of diagonal) and diagonal complete (one on the diagonal and zeros otherwise). Null and complete blocks are the only possibilities for nondiagonal blocks.

Table 3.1: Four possible ideal blocks for structural equivalence

	C_j			
C_i	0	0	0	0
	0	0	0	0
	0	0	0	0
	0	0	0	0

null

	C_j			
C_i	1	0	0	0
	0	1	0	0
	0	0	1	0
	0	0	0	1

diagonal complete

	C_j			
C_i	1	1	1	1
	1	1	1	1
	1	1	1	1
	1	1	1	1

complete

	C_j			
C_i	0	1	1	1
	1	0	1	1
	1	1	0	1
	1	1	1	0

null diagonal

Batagelj et al. (1992b, 67) emphasized that "although the definition of structural equivalence is 'local' it has 'global' implications - structurally equivalent units behave in the same way also to all other units. A position is defined in terms of all other units in a network".

3.2.2 Regular equivalence

The formal definition of regular equivalence was written by White and Reitz (1983). It is generalization of structural equivalence and the idea is that actors "are equivalent if they link in equivalent ways to other units that are also equivalent" (Doreian et al., 2005, 173).

The equivalence relation $\approx$ on A is a regular equivalence on network $N = \{A, R\}$ if and only if for all $x, y, z \in A$, $x \approx y$ implies both

(i) $xRz \Rightarrow \exists w \in A : (yRw \wedge w \approx z)$

(ii) $zRx \Rightarrow \exists w \in A : (wRy \wedge w \approx z)$

Batagelj et al. (1992a, 67) prove that for regular equivalence only null and regular blocks are possible. The term regular block is used for one-covered blocks, which have at least one 1 in each row and column. An example of regular block is presented in Table 3.2.

3.2.3 Generalized equivalence

The concept of generalized equivalence was first introduced by (Doreian et al., 1994). Batagelj et al. (1992a,b) showed that the structural equivalence is consistent only with complete and null blocks and in regular equivalence only null and regular blocks are allowed. In generalized concept of equivalence these requirements about connection patterns are relaxed to allow broader types of connections. Two generalizations are proposed by (Doreian et al., 1994, pg. 2):

(i) blocks, obtained by clustering based on generalized equivalence, can conform to different types of equivalence,

(ii) each block in the image matrix can have a particular pattern where specific types of equivalence are special cases of more general patterns.

First, we will present the weakening of the regular equivalence property (Doreian et al., 1994, 2005). The one-covered or regular block has at least one 1 in each block and in each column. This property can be viewed separately for rows and columns. A block is *row-regular* if all of it rows are one covered, and a block is *column-regular* if all of its columns are 1 covered. Blocks are presented in Table 3.2. The regular block is both row-regular and column-regular.

The weakening of structural equivalence leads to *row-dominant* and *column-dominant* blocks. The *row-dominant* block has at least one row one-covered, which means that there is at least one actor from the first cluster with ties to all actors from the second cluster. The *column-dominant* block has at least one actor from the second cluster who is receiving ties from all actors in the first cluster. That means that *column-dominant* block has at least one column one-covered. It is obvious, that complete block is both

row-regular and column-regular.

In a *row-functional* block each actor from the first cluster has a tie to exactly one actor from the second cluster. A *column-functional* block is one where each actor from the second cluster has exactly one actor from the first cluster linked to it. If the block is squared, the row-functional and column-functional blocks are the sparest possible regular blocks.

Table 3.2: Examples of ideal blocks (ties between actors of cluster C_i and C_j) for generalized type of equivalence

	C_j				
C_i	1	1	1	1	1
	1	1	1	1	1
	1	1	1	1	1
	1	1	1	1	1

complete

	C_j				
C_i	0	1	0	0	0
	1	1	1	1	1
	0	0	0	0	0
	0	0	0	1	0

row-dominant

	C_j				
C_i	0	0	1	0	0
	0	0	1	1	0
	1	1	1	0	0
	0	0	1	0	1

col-dominant

	C_j				
C_i	0	1	0	0	0
	1	0	1	1	0
	0	0	1	0	1
	1	1	0	0	0

regular

	C_j				
X	0	1	0	0	0
	0	1	1	0	0
	1	0	1	0	0
	0	1	0	0	1

row-regular

	C_j				
C_i	0	1	0	1	0
	1	0	1	0	0
	1	1	0	1	1
	0	0	0	0	0

col-regular

	C_j				
C_i	0	0	0	0	0
	0	0	0	0	0
	0	0	0	0	0
	0	0	0	0	0

null

	C_j				
C_i	0	0	0	1	0
	0	0	1	0	0
	1	0	0	0	0
	0	0	0	1	0

row-functional

	C_j				
C_i	1	0	0	0	
	0	1	0	0	
	0	0	1	0	
	0	0	0	0	
	0	0	0	1	

col-functional

3.3 Different approaches to blockmodeling

Batagelj et al. (1992b, 66) distinguish two main approaches to blockmodeling problems

(i) The *indirect approach* reduced the blockmodeling problem to standard data anal-

ysis problem (cluster analysis, multidimensional scaling) where the dissimilarity matrix between actors based on selected type of equivalence is computed.

(ii) In the *direct approach* criterion function based on selected type of equivalence is constructed and the best partition that best fit the selected criterion function is directly searched with optimization algorithm.

Both direct and indirect approaches have been implemented in Pajek (Batagelj and Mrvar, 2010a,b), in an R-package called Blockmodeling (Žiberna, 2008) in UCINET (Borgatti et al., 2002), and in some other programs.

The dissertation focuses on generalized blockmodeling which is an example of direct approach and is in detailed presented in the next section.

3.4 Generalized blockmodeling

The concept of generalized blockmodeling was proposed by (Doreian et al., 2005). Doreian et al. (2005, 25-26) stated three main characteristics of generalized blockmodeling compared to conventional blockmodeling:

(i) In the direct approach only the network data are used (without transformation to dissimilarity measures first).

(ii) A broader set of block types is introduced, which can enable the analysts a better capture of the network structure.

(iii) The blockmodel is specified beyond just permitting block types and allow specification of location of block types and membership of the clusters. This means that analyst's knowledge can be incorporated in prespecifed blockmodels prior to blockmodeling analysis.

In his dissertation (Žiberna, 2007) distinguished four types of generalized blockmodeling: binary blockmodeling[4], valued blockmodeling, homogeneity blockmodeling and

[4]The term is used for the generalized blockmodeling of binary networks as introduced in Doreian et al. (2005).

implicit blockmodeling. He stated that the most important difference between the above types of generalized blockmodeling is "an appropriate definition of the inconsistencies of the empirical blocks with the ideal ones" (Žiberna, 2007, 46). The last three types of blockmodeling analyze the valued networks and are in detailed presented and evaluated in his dissertation.

In this dissertation the main focus will be on generalized blockmodeling with binary networks. The term 'generalized' is mainly omitted in the dissertation and just the term 'blockmodeling' is used.

3.4.1 Criterion function

As mentioned before, the criterion function evaluates the partition and based on the minimal value the best partition is selected. If the value of criterion function is zero, that means that obtained partition perfectly matches the selected equivalence.

The criterion function was first presented in (Batagelj et al., 1992b,a) and is extendedly presented also in (Doreian et al., 2005, 185-187, 223-226).
First, some additional notation will be presented:

- $C = \{C_1, C_2, \ldots, C_k\}$ is a partition of actors A into k clusters.
- $R(C_i, C_j)$ denotes an empirical block and $T(C_i, C_j)$ is an allowed block or ideal block corresponding to block $R(C_i, C_j)$. The set of all feasible block types is denoted by $\mathcal{T}$.

The *deviation* $\delta\left(C_i, C_j; T\right)$ measures the deviation or inconsistency of an empirical block $R(C_i, C_j)$ from the nearest ideal block $T \in T(C_i, C_j)$. The deviations measures for different block types are presented in Table 3.3.

Based on the deviation $\delta\left(C_i, C_j; T\right)$ the *block-error* or *block inconsistency* of block $R(C_i, C_j)$ for type T can be defined as

$$\varepsilon\left(C_i, C_j; T\right) = w(T)\delta\left(C_i, C_j; T\right) \tag{3.1}$$

Table 3.3: Deviation measures for different types of blocks

Connection or block type	block type inconsistencies $\delta\left(C_i, C_j; T\right)$	position of the block
Null	s_t	nondiagonal
	$s_t + min(0, n_r - 2s_d)$	diagonal
Complete	$n_r n_c - s_t$	nondiagonal
	$n_r n_c - s_t + min(2s_d - n_r, 0)$	diagonal
Row-dominant	$(n_c - m_r - 1)n_r$	diagonal, $s_d = 0$
	$(n_c - m_r)n_r$	otherwise
Column-dominant	$(n_r - m_c - 1)n_c$	diagonal, $s_d = 0$
	$(n_r - m_c)n_c$	otherwise
Row-regular	$(n_r - p_r)n_c$	
Column-regular	$(n_c - p_c)n_r$	
Regular	$(n_r - p_r)n_c + (n_c - p_c)n_r$	
Row-functional	$s_t - p_r + (n_r - p_r)p_c$	
Column-functional	$s_t - p_c + (n_c - p_c)p_r$	

Legend:

s_t - total block sum = number of 1s in a block

s_d - diagonal block sum = number of 1s in a diagonal

n_r - number of rows in a block

n_c - number of columns in a block

s_t - total block sum = number of 1s in a block

p_r - number of non-null rows in a block

p_c - number of non-null columns in a block

m_r - maximal row sum

m_c - maximal column sum

where $w(T) > 0$ is a weight of type T. Usually, the weights are set to 1, although different block types can contribute differently and departures from one type of blocks can be seen more important than from the others.

The block inconsistency (Equation 3.1) can be extended to the set of feasible block types

$\mathcal{T}$ as

$$\varepsilon\left(C_i, C_j; T\right) = \min_{T \in \mathcal{T}} \left(C_i, C_j; T\right). \tag{3.2}$$

Block inconsistencies are combined in *total inconsistency or blockmodeling criterion function* $P(C)$ as a sum of inconsistencies within each block (Equation 3.2) across all blocks as

$$P(C; \mathcal{T}) = \sum_{C_i, C_j \in C} \varepsilon\left(C_i, C_j; \mathcal{T}\right). \tag{3.3}$$

The criterion function (Equation 3.3) has two properties

(i) $P(C) \geq 0$ and

(ii) $P(C) = 0$ if and only if we obtain the exact blockmodeling, that is if the partition C perfectly matches the selected equivalence defined with allowed block types.

Doreian et al. (1994, 2005) emphasized that different definitions of equivalence lead to distinct partition. Therefore each selection of block types from general set of blocks presented in Table 3.2 will usually not lead lead to the same partition as that obtained with structural or regular equivalence.

Another remark made by Doreian et al. (1994, 25) is, that generalized blockmodeling with the full set of allowed block types will "permit in most case, the establishment of blockmodel that fit very well". The blockmodels with zero inconsistencies can be fitted to most empirical networks. This should not be seen as a problem, but as an opportunity for close examination of the blocks and therefore richer characterization of the fundamental empirical network.

3.4.2 A clustering algorithm

As mentioned before, generalized blockmodeling directly searches for the best fitting partition according to the selected criterion function.

Batagelj et al. (1992b, 80) proposed the use of a local optimization relocation algorithm:

Determine the initial clustering C;

repeat:

If in the neighborhood of the current clustering C there exists a clustering C' such that $P(C') < P(C)$ **then** move to clustering C'.

Usually, the neighborhood of the current clustering C is determined by two following transformations:

(i) **moving** an actor from one cluster to another cluster,

(ii) **interchanging** of two actors from different clusters.

In order to obtain a good solution, the above procedure should be repeated with different random initial partitions C. Because this is a local optimization algorithm, there is no guarantee that the partition (or partitions) with minimal value of criterion function $P(C)$ will be found. In our simulations, presented later in the dissertation, we run each individual blockmodeling algorithm with 100 randomly selected partitions. In general, the risk of missing some optimal best fitting partitions increases with increasing size of a network. Therefore for larger network the number of starting partitions should be larger, which means that the optimization algorithm could be very time consuming.

According to presented definitions of different types of equivalences (Section 3.2), where main advantage of generalized equivalence is its adaptability to the problem of interest or to data, and remark made by Batagelj et al. (1992b) that definition of structural equivalence is 'local' and it has 'global' implications, the following thesis is set up:

Thesis 1. *Structural equivalence gives more stable results than regular (or other generalized types) equivalence.*

The stability of blockmodeling in the above thesis will be measured with two indices of network stability: the Adjusted Rand Index and the proportion of incorrect block types presented in Chapter 5.

Beside the *Thesis 1* the research question about predictive power of different network characteristics to the stability of blockmodeling was established as well. The more detailed notation is as follows:

Research question 1. *To what extent are the (relative) differences in network characteristics (e.g. network density, reciprocity, number of different types of dyads) and correlation and/or Euclidean distance between vectors of vertex properties (e.g. centrality measures) able to predict the results of blockmodeling (stability of partition and type of blockmodel)?*

The network characteristics and vertex properties from the research question are presented in Section 2.2, while the results of simulation studies established based on that question are presented in Chapter 8.

4 Design errors

In this chapter errors in research process introduced through questionnaire design, selection of actors and/or misreporting or unreporting of actors are presented. At the end of chapter brief comparison of errors in social network analysis and errors in usual social science research is presented.

Surveys and questionnaires[5] are the most used techniques (in the absence of archival data also the most practical techniques) for gathering social network data (Marsden, 2005; Wasserman and Faust, 1998). "Surveys allow investigators to decide on relationships to measure and on actors/objects to be approached for data"(Marsden, 2005, 10). All methods have the potential for introducing different types of errors, including measurement errors, and it is necessary to consider the implications of these errors in two ways. One is to consider how certain types of errors can be reduced (a very good thing in its own right) and the other is to assess the impact of errors on the results obtained from using network analytic tools. Of course, the two implications of errors are not unrelated, even though in the dissertation we mainly focus on the second assessment.

Errors have a variety of sources including boundary specification problems, , actor non-response, and censoring vertex degree through a fixed choice design in network surveys (Kossinets, 2006). These errors are often caused by the study design. Another source of errors can be actors themselves through non-response and/or respondent inaccuracy. Marsden (2005, 10) emphasized that surveys introduce artificiality and findings or collected data "rest heavily on the presumed validity of self-reports". All

[5]Wasserman and Faust (1998, 45) allege three main different questions formats: roster vs. free recall, free vs. fixed choice and ratings vs. complete rankings.

types of design errors found in literature are composed and presented in broader set on Figure 4.1. The first cut is to distinguish boundary specification problems, questionnaire design, and errors due to respondents. In the next sections causes and researched consequences of different error types are presented in detail.

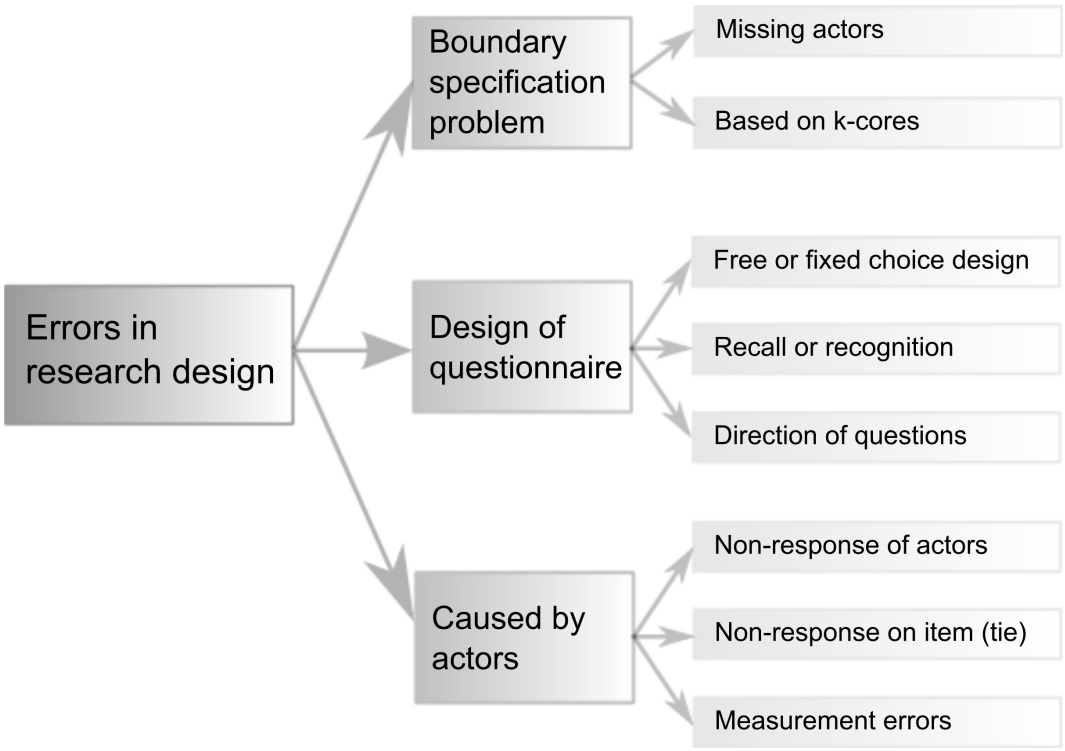


Figure 4.1: Scheme of errors in research design

The main research question in the dissertation is to estimate the sensitivity of blockmodeling to the errors in the research design. More precisely, the following research question was established:

Research question 2. *How stable is blockmodeling procedure to different types and amounts of errors?*

4.1 Boundary specification problem

The boundary specification problem concerns rules of inclusion for actors in a studied network (Laumann et al., 1989). The distinction between two approaches can be made,

the *realist* and the *nominalist* approach. The first approach for defining network boundaries is *realist* in the sense that actors in the network determine the boundaries of their network alone and identify their common membership in a network. In a *nominalist* approach, network boundaries are defined by researchers, most often using some membership criterion. The criterion for inclusion of actor can be defined based on actor (his/her status or influence), relations or activity and it can also be a combination of all three factors. Neither approach is guaranteed to identify network boundaries correctly and, while they can be combined, the resulting data are still likely to have errors regarding boundaries.

Doreian and Woodard (1994, 268) observed that the risk of incorrectly specified boundaries "is particularly acute for analyses resting on a positional conceptualization of social structure where a position is defined in terms of the pattern of ties to all other actors of the network". They proposed using k-cores[6] for determining network boundaries. However, this is predicated on the collection of some network data and may be best accompanied by an expanding network selection based on the in-degree (of yet to be included actors). The incorrect location of a boundary can take three forms:

(i) including actors who do not belong,

(ii) excluding actors who do belong, and

(iii) both incorrect inclusions of some actors and exclusions of others.

Exclusion of actors through an incorrect boundary location implies that all of their potential data are lost regardless of the instrument used.

Kossinets (2006) showed that deletion of actors from a network, which illustrates incorrectly specified network boundaries, can significantly change network characteristics. E.g., the results of simulations (on a collaboration network) revealed that mean vertex degree decreases when the fraction of missing vertices increases.

[6] A k-core is a (cohesive) subset of network actors where subset actors are connected to at least k actors from the same subset. This can be computed based on in-degree, out-degree or all-degree of vertices.

Laumann et al. (1989, 63) emphasized that specification of rules of inclusion "pertain both to the selection of actors or nodes for the network and to the choice of types of relationships among those actors to be studied". If the relations are used to indicate network boundaries this strategy is termed as *relational* strategy. Marsden (2011)[371] emphasized that "positional, event-based, reputational, and relational criteria can be used together to identify population for network studies, a study might begin with a list of actors included on a positional basis, and supplement it using event-based or reputational criteria before fieldwork begins".

The boundary specification problem needs to be distinguished from the network sampling. In the whole network studies when boundaries are set, all actors are included in the research without additional sampling and it is assumed that all dyadic relationships are observed or measured[7]. Marsden (2011, 371) emphasized that the boundary specification problem "often results in a complete listing of actors or rooster of the study population".

The assumption about completely observed presence or absence of ties between all actors is clearly not true in practice when networks are collected through sample surveys (Handcock and Gile, 2007). Compared to the social science surveys, network surveys rarely draw samples (Marsden, 2011). The sampling procedure is needed in the large-scale social systems where the complete enumeration of the members is not possible (Granovetter, 1976; Scott, 2000). The sampling in social networks should lead general principles from survey research: "a representative sample of cases is drawn from the sample population in question, their relations are investigated, and sampled networks are constructed that will be homologous to the partial system that occur in the population as a whole" (Scott, 2000, 59). The reality is not so promising and simple, because in social network area there are no rules for "judging the quality of relational data derived from a sample" (Scott, 2000, 59) and "we are left guessing about the representativeness of the patterns in the social relations found" (Granovetter, 1976, 1288).

[7]Of course, other errors from research design (e.g. non-response, errors due to fix choice design....) can be present.

The simplest sampling method[8] for social network requests a full membership list, then the random samples with replacement are drawn and each sample member is asked a sociometric question, e.g. whom he or she knows from the list (Granovetter, 1976; Erickson et al., 1981; Erickson and Nosanchuk, 1983). The described procedure[9] was successfully used for estimation of network density and average outdegree of an actor, which are the global properties of the network. On the other hand, "it is almost impossible to go beyond such basic parameters to measure the more qualitative aspects of network structure"(Scott, 2000, 59). In the sampling procedure local properties are more difficult to estimate and rare events are poorly represented (Granovetter, 1976, 1301).

4.2 Introduced by design

Network instruments are another source for introducing errors. Three different question formats are often considered when designing instruments for collecting social network data:

(i) free or fixed choice designs,

(ii) using recall or recognition of actors, and

(iii) giving or receiving social support with different direction of question.

[8]E.g. Lee et al. (2006) distinguished three kinds of sampling methods; node sampling, link sampling and snowball sampling. In his definition of node sampling network consists of randomly chosen nodes and ties among them. In link sampling procedure firstly ties are randomly selected and secondly all nodes attached to those ties are kept in the network.

[9]It should be noted that the sampling procedure has also its own problems where "the network sample or samples may be based on a list imperfectly reflecting the target population, the samples may be drawn non-randomly from the list, and response may be non-random", e.g actors with smaller degree may be less active and therefore less willing to participate in the study. (Erickson and Nosanchuk, 1983, 367).

4.2.1 Free or fixed choice design

The questionnaire or name generator in social network collection process could have instructions about predetermined number of actors (or choices) which each network member has to select or not. The first question format is known as *fixed choice design* and the second as *free choice design*.

The potential problems of *fixed choice design* were pointed out by (Holland and Leinhardt, 1973, 90). They distinguished three possibilities which may occur in a fixed-choice design where l choices are allowed:

(i) true structure is exactly the same as the observed structure in a sociogram (e.g. an actor has exactly l friends),

(ii) a subset of choices is presented in a sociogram (e.g. an actor has more than l friends), and

(iii) the true structure is a subset of ties in a sociogram (e.g. an actor has less than l friends and he has to choose more persons to satisfy the requirements of design).

Therefore, when a fixed number (l) of choices is specified in an instrument, actors with more ties than the threshold l are forced to leave out alters (case (ii)) and actors with fewer ties can add nonexistent ties to reach the threshold and to satisfy the requirements of design (case (iii)). While it is possible that the true structure is exactly the same as the observed structure with a fixed choice design, it is likely that either the true structure is contained within the observed structure or, even more likely, the observed structure is contained within the real structure[10]. The second case implies the presence of missing data for specific ties and in the third case the ties are misreported.

"Fixed choice nominations can easily lead to a non-random missing data pattern" (Kossinets, 2006, 253). Popular individuals with many contacts are more likely to be chosen by their friends and friends of your friends are very likely to be your friends too - transitivity patterns in a network (Feld, 1991; Newman, 2003). Marsden (2011, 373)

[10]Of course, combination of both problems can be present in the collected network data.

emphasized the practical advantages of fixed choice nominations in survey administration where "they simplify and specify a sociometric task for respondents, thereby reducing burden".

A free-choice design has the potential to allow the collection of richer network data, but it does not automatically eliminate errors. These can also arise from respondents having different interpretations of the terms in a question. For example, the term 'friend' can have different components ranging from 'acquaintance' to best friend (Holland and Leinhardt, 1973; Hlebec and Kogovšek, 2006). Additionally, the graphical appearance of the name generator can affect responses more than the wording of questions or specific instructions provided by researchers (Manfreda et al., 2004; Vehovar et al., 2008). In a web survey for collecting ego-centered networks, it was found out that respondents were more influenced by graphical design than by wording of question and instructions. From among 30 provided spaces for alter naming, 15% of respondents in shorter and 12% in longer version of name generator filled in all possible spaces.

4.2.2 Recall or recognition

Another instrument design choice concerns the elicitation of recalled or recognized actors. The *recognition method* provides all actors a complete list, *or rooster*, of other members in a network (Wasserman and Faust, 1998). On the other hand, if the respondents have to nominate the actors without the help of rooster, the questionnaire format is called *(free recall)*.

Brewer and Webster (2000, 362) based on literature review argued that "no prior research has examined the effects of forgetting on the measurement of structural properties of personal and social network". In their study of friendship ties collected among students on the campus[11], they found out that on average actors recalled 80% of their friends. The forgotten (or recognized) friends seem to be more peripheral than recalled ones. Beside expected differences in network density of recalled and recognized

[11]Considering an e-mail correspondence with Cynthia Webster in March 2010 the dataset is not available anymore, because the authorization period to use the data has expired.

networks, there were also significant differences in number of cliques, degree centralization and closeness. Brewer (2000) found out that people forget a substantial proportion of interactions with others and that the number of recalled ties correlates with the number of forgotten ties. Unfortunately, there seems to be no good predictors of actors' level of forgetting.

The delineated structure of networks based on a recall or a recognition design option can differ and the notion of 'forgotten' alters (if they do belong to the network) can imply serious missing data errors. Brewer (2000, 29) emphasized that "forgotten network ties would make recalled network data incomplete and possibly distort measurement of various characteristics and structural features of personal and social networks". Therefore, collection methods based on recognition or objective records should be used to assure complete or at least more accurate network data (Stork and Richards, 1992; Brewer, 2000). Use of rooster also simplifies the reporting task for actors by reminding them of eligible network members (Marsden, 2011). He emphasized that both methods require careful manipulation with members' names. The recall method has to ensure that actors named by different names (e.g. nicknames, spelling variations..) are correctly matched and names on a list in recognition method should be known to (or recognized by) all network members.

The average number of recognized ties tends to be higher than the number of recalled ties (Hlebec, 1992, 1993). Hlebec and Ferligoj (2001) also found out that free recall procedure produces strong ties and it is robust to the effects of measurement procedure and respondents mood.

Although ego-centered networks are not subject of our research, the results of Bell et al. (2007, 288) are interesting. They studied patterns of forgetting in ego-centered networks with three different relations. The main finding was that more intimate relations are less likely to be forgotten. For example, people forget 6% of 30-days sex partners, 18% of drug use partners and more than quarter (26%) of close friends. His findings that more specific name generators produce less errors was confirmed also by

Wright and Pescosolido (2002) who established that forgetfulness is largely random, because he did not find any systematic predictors (with regard to actor's or network's characteristics.)

4.2.3 Direction of questions

Some social relations have an intrinsic direction. An obvious example is the provision of social support. Support relationships can be gathered in ways that are attentive to directional flow of support (Stork and Richards, 1992; Ferligoj and Hlebec, 1999). The original question asks about alters from whom respondents request support and a reversed question asking about alters asking them for support. The perceptions of given and received social support (in small networks up to 30 actors) are equally stable (Ferligoj and Hlebec, 1999; Hlebec and Ferligoj, 2002) but they also differ. Asking about support flow in only one direction runs the risk of missing some or all social support ties for an actor.

4.3 Caused by actors

Errors due to actors (beyond those introduced by the instrument design) can be divided into three categories:

(i) actor non-response,

(ii) non-response regarding an item or a tie, and

(iii) measurement errors in recorded ties.

4.3.1 Actor non-response

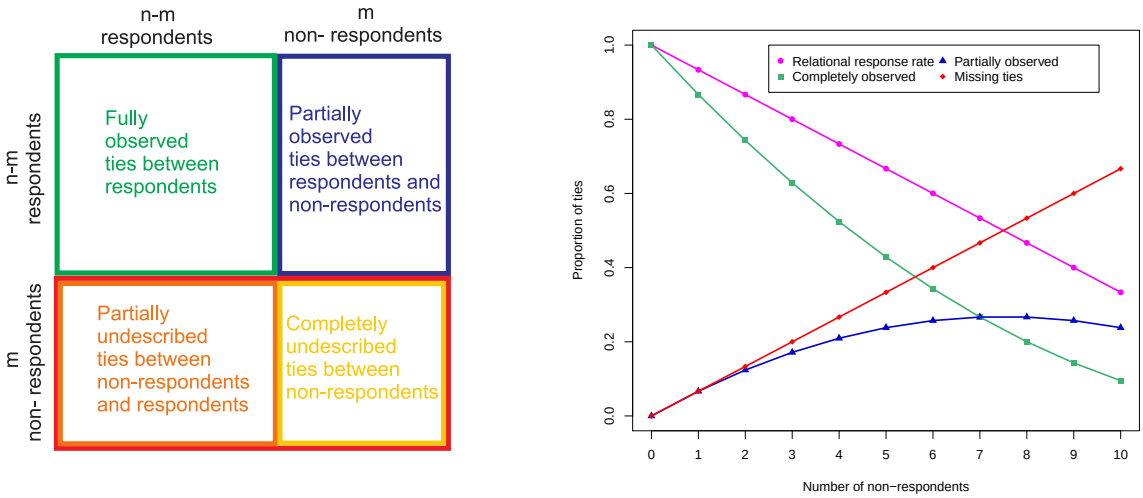
Let n denote the number of vertices in a network and m the number of actors providing no responses (for whatever reason). Each non-respondent implies $(n-1)$ missing ties. The actor response rate is $(1 - m/n)$. It is straightforward to show that the *relational response rate*, which is defined as the proportion of potentially observed ties that are

measured, is also $(1 - m/n)$ (Knoke and Yang, 2008).

For the $(n - m)$ respondents, there are $(n - m)(n - m - 1)$ ties for which all data can be collected, these ties are denoted as *fully observed ties* between respondents (Figure 4.2(a)). Assuming that data are obtained for all of them, the proportion of these fully observed network ties is $(n - m)(n - m - 1)/n(n - 1)$. There are $(n - m)m$ partial descriptions of ties between respondents and non-respondents. The proportion of these partially observed ties is $(n - m)m/n(n - 1)$. The number of *missed or undescribed ties* is $m(n - 1)$ and their proportion is $m(n - 1)/n(n - 1) = m/n$. They consist of:

(i) *partially undescribed* ties between non-respondents and respondents, the proportion of these ties is $m(n - m)/n(n - 1)$,

(ii) *completely undescribed* ties between non-respondents, whose proportion is $m(m - 1)/n(n - 1)$.



(a) Scheme

(b) Proportions of types of ties

Figure 4.2: Types of ties in network with non-respondents

Figure 4.2(b) shows these proportions for a network with actors ($n = 15$) and number of non-respondents between 0 and 10 ($0 \leq m \leq 10$). As shown in Figure 4.2(b), the relational response rate declines linearly with the number of actors not responding, consistent with Knoke and Yang (2008). The proportions of the fully observed part of the network ties decline in a more extreme (curvilinear) way. The proportion of

unobserved (or missed) ties increases in a linear fashion. The curve for the partially observed part of the network first increases and then decreases as the level of non-response increases. As a specific example, for network with 15 actors ($n = 15$) and three non-respondents ($m = 3$), the actor response rate (and the relational response rate) is 0.8, the proportion of fully described ties is 0.63, the proportion of partially described ties is 0.17 and the proportion of missing ties is 0.20. Among the missing ties the proportion of completely undescribed ties is equal to 0.03. Figure 4.3.1 suggests that non-response can be a major problem, one that gets worse as it increases.

Stork and Richards (1992) reviewed the network literature and reported that response rates varied from 65% to 90%. Costenbader and Valente (2003), based on sample of 59 networks, a wider range between 51% and 100%. They excluded four networks from their analysis because their response rates were lower than 50%. The extreme kinds of examples shown in the right side of Figure 4.2(b) are possible in empirical research.

The effects of actor non-response on network properties such as network density, average vertex degree, out-degree or in-degree, clustering coefficients, transitivity, assortivity, and geodesic distances have been examined (Stork and Richards, 1992; Costenbader and Valente, 2003; Kossinets, 2006; Huisman, 2009). The effects of four types of random errors (edge deletion, node deletion, edge addition, and node addition) on centrality measures were studied by Borgatti et al. (2006). "Both the node-removal and edge-removal cases can be thought of as forms of sampling from the network, since what remains after removal is a random sample of the network" Borgatti et al. (2006, 126). According to our definition (discussed in detail in Section 4.3.1.1), the node removal is an example of actor non-response treatment known as *complete-case approach* and the edge-removal case is in fact item non-response (Section 4.3.2). They used random networks of different sizes and densities and established that the effects of different kinds of errors are similar to each other. The accuracy of centrality measures (degree, closeness, betweenness, and eigenvector centrality) predictably declines with higher amount of introduced errors.

Robins et al. (2004, 258) point out that "many network studies are based on the premise that in order to understand some social phenomenon of interest, it is necessary to understand the arrangement of network ties into *larger network structures and sub-structures*". In any analysis where this premise is correct, a missing tie causes limited possibilities to describe the context or position of actors with missing ties in a whole network. Blockmodeling (Doreian et al., 2005), presented in detailed in Section 3, is one way of delineating the wider structures and substructures of a network and the results obtained may be vulnerable to the impact of non-response. The results of blockmodeling stability to actor non-response is presented in Section 7.3.

4.3.1.1 Non-response (or missing) data treatments

Stork and Richards (1992) suggest that the presence of non-respondents for collected network data can be treated in three different ways: (i) using a complete-case analysis, (ii) using an available-case analysis, and (iii) imputing data values as replacements of the missing data. We expand this list to consider five different missing data treatments. The impacts of these procedures on delineated blockmodels are discussed in Section 7.3.

The complete-case approach

If we have non-respondents in a network, the outgoing ties for each non-respondent are not reported. The results are rows of missing ties in the matrix representation of the observed network. Consider the example shown in Figure 4.3 (left panel) having three non-respondents (B2, B6, and G1). Note that some of the respondents (e.g. B1 and B5) report ties to the non-respondents (both B1 and B5 report ties to B2 and B6).

The complete-case approach, also known as 'listwise' deletion of actors (Huisman and Steglich, 2008), removes not only the rows for the non-respondents but also their columns. Removing columns means that all partially described ties (as reported by one actor) are removed from the analysis. These incoming ties to non-respondent actors, while recorded, are deleted. The result is the smaller network as shown in Figure 4.3 (right panel).

	B1	**B2**	B3	B4	B5	**B6**	**G1**	G2	G3	G4	G5
B1		1	1	0	1	1	0	0	0	0	0
B2	NA		NA	NA	0	NA	NA	NA	NA	NA	NA
B3	1	1		1	1	0	0	0	0	0	0
B4	0	1	1		1	0	0	0	0	0	0
B5	1	1	1	0		1	0	0	0	0	0
B6	NA	NA	NA	NA	NA		NA	NA	NA	NA	NA
G1	NA	NA	NA	NA	NA	NA		NA	NA	NA	NA
G2	0	0	0	0	0	0	1		1	1	1
G3	0	0	0	0	0	0	0	1		1	0
G4	0	0	0	0	0	0	0	1	1		1
G5	0	0	0	0	0	0	0	1	1	0	

	B1	B3	B4	B5	G2	G3	G4	G5
B1		1	0	1	0	0	0	0
B3	1		1	1	0	0	0	0
B4	0	1		1	0	0	0	0
B5	1	1	0		0	0	0	0
G2	0	0	0	0		1	1	1
G3	0	0	0	0	1		1	0
G4	0	0	0	0	1	1		1
G5	0	0	0	0	1	1	0	

Figure 4.3: Network where three non-respondents (B2, B6 and G1) provide no outgoing ties (on the left) and the smaller network obtained with complete-case approach (on the right)

Robins et al. (2004) argue that this approach amounts to respecifying the network boundary nonrespondents are removed to create a smaller network. The complete case analysis might be valid when non-respondents are missing completely at random. They emphasized that especially for large proportion of non-respondents the simple exclusion (the complete-case approach) of actors is not appropriate method, because there is no evidence whether the non-respondents are missing completely at random, which means that the connection patterns between respondents and non-respondents are the same. However, if this does not hold then the results may be biased because the sample of remaining actors may be unrepresentative (Schafer and Graham, 2002). Stork and Richards (1992, 197) argue that the complete approach "seriously weakens any analysis at the system level". On the other hand, Schafer and Graham (2002) emphasized that the main advantage of complete-case approach is simplicity and that method can be effected if only a small part of the network is discarded, i. e. if the proportion of non-respondents is small.

Reconstruction

An available case approach uses both the completely described ties between respondents and the partially described ties between respondents and non-respondents. In doing

so, some data transformations are used.

Reconstruction of the missing outgoing ties of non-respondents occurs when they are replaced by the observed *incoming* ties to those actors (Stork and Richards, 1992; Huisman, 2009). Therefore 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 or unmeasured ties is replaced with corresponding column for each missing respondent. The result is that ties involving non-respondents and respondents become symmetric. Stork and Richards (1992) emphasized that reconstruction is not the same as imputation, because in the reconstruction procedure no new ties are added. The reconstruction simply allows that the relationship between two persons, in essence, can be measured by using one report of the tie. However, for two non-respondents the reconstruction of ties between them is not possible. Some additional imputations are required to record data for them. In the simplest case, those unavailable ties (marked as NA in Figure 4.4) have zeros imputed. More satisfactory imputations are presented later in this section on page 54.

Figure 4.4 present network with 8 respondents and three non-respondents B2, B6 and G1. Each row of a non-respondent is in reconstruction procedure replaced by corresponding column, e.g a second row is replaced by a second column. Because we have three non-respondents in the network, six ties ($3 \cdot (3-1) = 6$) between them can not be obtained without additional imputations. In the simplest solution, the zeros are imputed instead of the missing ties between two non-respondents (right panel on Figure 4.4).

Stork and Richards (1992, 198) argued that the two criteria should be met when reconstruction is used:

(i) the non-respondents and the respondents should not differ systematically from each other, and

(ii) the available data from the respondents are useful and reliable description of ties between two actors.

	B1	**B2**	B3	B4	B5	**B6**	**G1**	G2	G3	G4	G5
B1		1	1	0	1	1	0	0	0	0	0
B2	1		1	1	1	NA	NA	0	0	0	0
B3	1	1		1	1	0	0	0	0	0	0
B4	0	1	1		1	0	0	0	0	0	0
B5	1	1	1	0		1	0	0	0	0	0
B6	1	NA	0	0	1		NA	0	0	0	0
G1	0	NA	0	0	0	NA		1	0	0	0
G2	0	0	0	0	0	0	1		1	1	1
G3	0	0	0	0	0	0	0	1		1	0
G4	0	0	0	0	0	0	0	1	1		1
G5	0	0	0	0	0	0	0	1	1	0	

	B1	**B2**	B3	B4	B5	**B6**	**G1**	G2	G3	G4	G5
B1		1	1	0	1	1	0	0	0	0	0
B2	1		1	1	1	0	0	0	0	0	0
B3	1	1		1	1	0	0	0	0	0	0
B4	0	1	1		1	0	0	0	0	0	0
B5	1	1	1	0		1	0	0	0	0	0
B6	1	0	0	0	1		0	0	0	0	0
G1	0	0	0	0	0	0		1	0	0	0
G2	0	0	0	0	0	0	1		1	1	1
G3	0	0	0	0	0	0	0	1		1	0
G4	0	0	0	0	0	0	0	1	1		1
G5	0	0	0	0	0	0	0	1	1	0	

Figure 4.4: A network with three non-respondents (B2, B6 and G1) obtained by reconstruction with unavailable ties between non-respondents (left) and with imputed zeroes for ties between non-respondents (right)

The second criterion can be more easily met in undirected networks (e.g. conversation) than in directed ones (e.g. giving advice).

The respondents and non-respondents should be compared (if possible) in two ways: "using individual-level data and 'using data that described their pattern of communication[12]" (Stork and Richards, 1992, 198). In the comparison on individual actor level variables such as gender, age, education, professional training, level in the organization and other variables reasonable in the research study could be used. Patterns between respondents and non-respondents could be investigated in terms of received ties (incoming ties) by studying frequency and strength of ties and characteristics of the nominators (from whom ties are received).

Imputations

Imputations of ties in social networks replace missing ties by estimates to create an apparently full data set. There are four types of simple imputation procedures where each missing value is imputed only once (Schafer and Graham, 2002; Huisman, 2009):

(i) imputation of unconditional means,

(ii) imputations from unconditional distributions,

[12]In more general networks, patterns of relation under investigation.

(iii) imputations of conditional means, and

(iv) imputations from conditional distributions.

Here we focus on the first group of imputations. Huisman (2009) outlines three possible methods for imputing unconditional means in social networks. Only two of those methods are relevant here, *using the total mean* and *using means of incoming ties*. The third possibility of imputing unconditional means is the *average number of outgoing relations* of an actor or 'person mean'. For complete actor non-response, where all outgoing ties are missing, this method is inapplicable.

Using the total mean

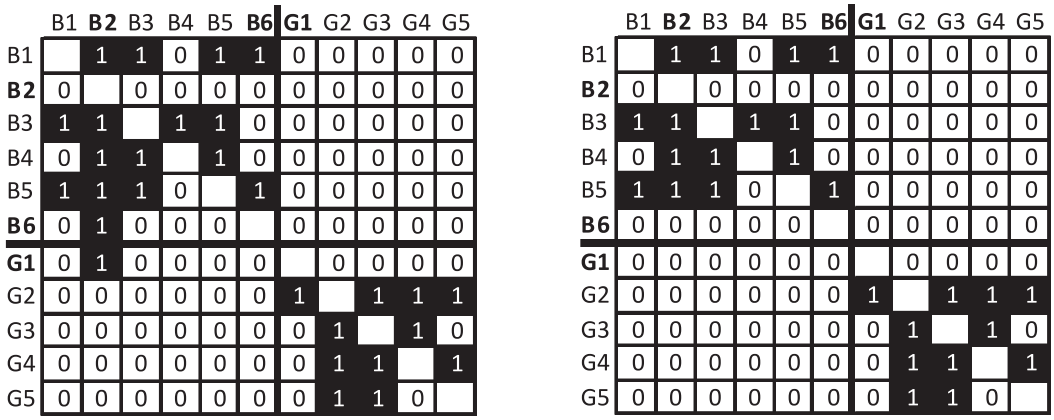
The first method uses the *average number of ties in the network*. This is the 'total mean' of the observed ties which is also the density of a network. For binary networks this means imputing zeros instead of missing ties in sparse networks and ones in dense networks. Some threshold is required for this imputation. Huisman (2009) used 0.5 as the threshold in his simulation study. We note that (Costenbader and Valente, 2003) reported network densities between 0.01 and 0.49 for a sample of 59 networks.

Using means of incoming ties

The second option imputes the *average value of incoming ties of an actor* which is known also as the *item mean*. For binary networks this implies imputing ones if actors are popular given their received ties. Operationally, this also requires a threshold. When this is set at 0.5, a tie is imputed if the actor is chosen by at least half of respondent actors (left panel of Figure 4.5). More precisely, for each missing outgoing tie x_{ij} $(i \neq j)$ of the non-respondent i, the mean value of all available incoming ties of actor j is imputed. For binary networks, this implies the imputation of the modal value of the incoming ties and this procedure is termed *imputations based on the mode*.

Left panel in Figure 4.5 shows imputed values based on mode value (of incoming ties). There are three non-respondents, denoted with bold letters, B2, B6 and F1 (missing ties presented on left panel in Figure 4.3). Actor B2 has four incoming ties out of eight possible ties from respondents which meet the criteria that is chosen by at least half of the

respondent actors. Therefore 1 (a tie) is imputed for a tie between non-respondents B5 and B6 to actor B2. On the other hand, other actors have less than four incoming ties, therefore no ties (zeros) are imputed.



	B1	**B2**	B3	B4	B5	**B6**	**G1**	G2	G3	G4	G5
B1		1	1	0	1	1	0	0	0	0	0
B2	0		0	0	0	0	0	0	0	0	0
B3	1	1		1	1	0	0	0	0	0	0
B4	0	1	1		1	0	0	0	0	0	0
B5	1	1	1	0		1	0	0	0	0	0
B6	0	1	0	0	0		0	0	0	0	0
G1	0	1	0	0	0	0		0	0	0	0
G2	0	0	0	0	0	0	1		1	1	1
G3	0	0	0	0	0	0	0	1		1	0
G4	0	0	0	0	0	0	0	1	1		1
G5	0	0	0	0	0	0	0	1	1	0	

	B1	**B2**	B3	B4	B5	**B6**	**G1**	G2	G3	G4	G5
B1		1	1	0	1	1	0	0	0	0	0
B2	0		0	0	0	0	0	0	0	0	0
B3	1	1		1	1	0	0	0	0	0	0
B4	0	1	1		1	0	0	0	0	0	0
B5	1	1	1	0		1	0	0	0	0	0
B6	0	0	0	0	0		0	0	0	0	0
G1	0	0	0	0	0	0		0	0	0	0
G2	0	0	0	0	0	0	1		1	1	1
G3	0	0	0	0	0	0	0	1		1	0
G4	0	0	0	0	0	0	0	1	1		1
G5	0	0	0	0	0	0	0	1	1	0	

Figure 4.5: Network with three non-respondents (B2, B6 and G1) obtained by imputations based on mode (left) and by null tie imputation (right)

Null tie imputations

Robins et al. (2004, 261) note that "even when reconstruction is not appropriate, it may still be useful to retain non-respondents in the data set, but only to analyze those network constructs that can be defined in terms of incoming ties". In such a case the matrix has rows of 0s for each non-respondent and this treatment will be marked as *null tie imputations* (Figure 4.5). Although, the use of null tie imputation is not consistent with the spirit of their proposal, we include it for the sake of completeness. If the network is sparse (the network density is below 0.50), the null imputation treatment is equal to the total mean treatment.

Reconstruction plus mode-based imputations

In the reconstruction procedure, ties between non-respondents cannot be replaced without additional imputations. It is possible to combine the reconstruction procedure with imputations based on the mode for those ties (Figure 4.6). As presented on left panel in Figure 4.4, we have six ties between non-respondents. Unobserved ties from non-respondents B6 and G1 ti actor B2 are replaced by a tie with imputations based on

mode (because half of the respondents (four actors of eight respondents) nominate the non-respondent B2), for other unobserved ties zeros are imputed.

	B1	**B2**	B3	B4	B5	**B6**	**G1**	G2	G3	G4	G5
B1		1	1	0	1	1	0	0	0	0	0
B2	1		1	1	1	0	0	0	0	0	0
B3	1	1		1	1	0	0	0	0	0	0
B4	0	1	1		1	0	0	0	0	0	0
B5	1	1	1	0		1	0	0	0	0	0
B6	1	1	0	0	1		0	0	0	0	0
G1	0	1	0	0	0	0		1	0	0	0
G2	0	0	0	0	0	0	1		1	1	1
G3	0	0	0	0	0	0	0	1		1	0
G4	0	0	0	0	0	0	0	1	1		1
G5	0	0	0	0	0	0	0	1	1	0	

Figure 4.6: Network with three non-respondents obtained by reconstruction plus imputations based on the mode.

There are also other possibilities for imputing ties between non-respondents. For example, Huisman (2009) suggested random imputations proportional to the observed density where the probability of a tie is proportional to the observed network density. We do not consider these other alternatives.

Other possible non-response data treatments include: a reconstruction procedure where ties between non-respondents are imputed randomly with a probability proportional to the network density (Huisman, 2009); imputation by preferential attachment where the probability of a tie from actor i to actor j depends on the indegree of actor j (Huisman and Steglich, 2008) and 'hot deck' imputations where actor attributes are used. Huisman (2009) used both categorical data (about actors) and structural properties (e.g. indegree) to locate a completely observed donor actor as a source to substitute ties for a non-responding actor. We do not consider actor attributes here.

Robins et al. (2004, 261) argued that none of the strategies for missing data problem is universally successful and that "judgments about the appropriateness of any strategy will almost certain depend both on the researchers' beliefs about the underlying

process by which the network data are generated and on the kind of network characteristic that the researcher intends to measure".

The actor non-response is an important source of errors in research design which is too often overlooked by the researchers. Instead of clear notation of absent ties, they are coded with zeros as non-existing ties. The problem of non-response can be solved (or at least reduced) with selection of the appropriate treatment. Therefore we will try to answer the following question:

Thesis 2. *The stability of blockmodeling with non-respondents (compared to the whole network without non-respondents) is higher when reconstruction is used than imputations of unconditional means (based on the number on incoming ties).*

4.3.2 Non-response on item or tie

Item or tie non-response occurs, when an actor participates in the study, but data on particular items or ties are missing, because a respondent does not indicate presence or absence of particular tie or ties (Rumsey, 1993; Borgatti et al., 2006; Huisman and Steglich, 2008; Huisman, 2009). "While we may have an explicitly defined network it is always possible that we are missing certain important edges" (Adar and Ré, 2007, 27).

The partial informations for the incompletely observed actors are available and should be used to "obtain (better) estimates of the structural properties of the actors and the network, and may give information on the nature of the missing data mechanism" (Huisman, 2009, 3).

The missing data on relations (or non-response on tie) were studied in personal networks by Burt in 1987. He found out that missing relations are the weak relations and that complete network data are collected among close discussion partners. The respondents with different characteristics tend to have different impact on incomplete data. For example, discussion relations reported by women are less likely to be missing than those reported by men, relations are more likely to be incomplete if reported

by blacks, and higher education have negative effect on completely reported relational data. He emphasized that "Missing data are doubly a course to a survey network analysis. First, network items are more complex than the usual opinion survey item and so might seem more likely to generate missing data. Second, network analysis is especially sensitive to missing data" (Burt, 1987, 63).

Treatments of actor non-response (presented in 4.3.1.1) can be used also to treat missing data due to item non-response. Instead of a whole row of unavailable ties, in item-nonresponse only a few ties are missing. In *the complete-case approach* the respondents with incompletely reported ties will be removed from the analysis (deletion of rows with missing data and the corresponding columns)[13]. The use of *reconstruction* procedure in item non-response case means that unreported tie r_{ij} is replaced with an observed tie r_{ji}. In case when both ties r_{ij} and r_{ji} are unobserved, reconstruction procedure is not possible. In the simplest case a zero is imputed (treatment called reconstruction) and in second case imputations based on mode are used (treatment called reconstruction plus mode). In the *null tie imputation* procedure a zero is imputed instead of the unobserved tie. If a tie r_{ij} is missing, the mode value of incoming ties of actor j can be used in *imputations based on mode.*

4.3.3 Measurement errors

Measurement error occurs when there is a discrepancy between the true value of a concept and the observed (measured) value of that concept. The common notation is that observations or measurements of a concept are an additive combination of true score plus error (or noise) and this error is known or referred to as measurement error (Wasserman and Faust, 1998, 59). The first introduction of measurement error in social network analysis (in accordance with standard definition of measurement error) was made by Holland and Leinhardt (1973) in 1973. They assumed (87) that "all groups possess an underlying pattern of generalized affect that is not directly observable but which generates the responses of group members in sociometric tests. We call this un-

[13]If there is a large proportion of actors with non-reported tie(s), the use of complete case approach is meaningless or even impossible (each actor has at least one unreported tie)

derlying pattern the true structure of the group, and distinguish it from the observed structure or sociogram." They emphasize that the true structure is a hypothetical construct that is necessarily unobservable. This means that "there is no single, best or obvious mathematical representation for it" (Holland and Leinhardt, 1973, 87). They suggested also one possible representation of networks, a sociogram, which is a directed graph model where nodes represent actors and arcs represent relations between pairs of actors.

Holland and Leinhardt (1973, 87) defined the measurement error as missing or extra tie in a network as:

> "In sociometry measurement error occurs when, regardless of the cause, the response made by a subject in a sociometric test fails to agree with the underlying true structure. Measurements error in sociograms occurs for two reasons: (1) no choice is recorded in the sociogram for a sentiment relation that exist in the true structure or (2) a choice is recorded in the sociograms for which there is no corresponding sentiment relation in the true structure."

The effects and representations of measurement errors in binary networks can be extended to valued networks where measurement error occurs when a wrong value (of strength tie) is recorded.

Measurement errors can be random or systematic (Ferligoj et al., 1995; Viswanathan, 2005). "Random error is any type of error that is inconsistent or does not repeat in the same magnitude or direction except by chance" (Viswanathan, 2005, 98). The extent of random error is assessed with reliability indexes of measurement. Systematic error has consistent effects. This means that the differences between measured and true scores tend to be consistently positive or negative. The presence (or absence) of systematic errors is examined through the validity of measurement.

In the case of social networks, the definition of random measurement errors can be written as follows: "If misreporting is random, then even as each individual may re-

port some interactions that do not exist, and omit others that do, the likelihood of any particular misrepresentation would be unrelated with any other misrepresentations and also unrelated to any characteristics of the individuals involved" (Feld and Carter, 2002, 367).

Feld and Carter (2002) also presented two types of systematic error in the measurement of network ties:

(i) individuals over/underreporting others, which is called an *expansiveness bias*,

(ii) and individuals being over/underreported by others, which is called an *attractiveness bias*.

These error types are likely to arise from self-reported ties regarding the presence or absence of social ties, the most common method for gathering social network data (Marsden, 1990). There has been an extensive discussions regarding the differences between ties reported by respondents (cognitive ties) and social ties recorded by researchers observing interactions of the respondents (observed ties) (Bernard and Killworth, 1977; Killworth and Bernard, 1979 - 1980; Bernard et al., 1982; Freeman et al., 1987; Hammer, 1985; Krackhardt, 1987). The findings about accuracy of social network data were rather negative. Bernard and Killworth (1977, 17) argued that "people do not know, with any accuracy, those with whom they communicate". In the summary of the reviewed literature Bernard et al. (1984, 503) emphasized that "half of what informants report is probably incorrect in some way." The inaccuracy of reports refer to a list of nominations of an actor (e.g. with whom they communicate or have a relation) and frequency of relations (Bernard et al., 1982).

Knoke and Yang (2008) define the discrepancy between self-reported ties and actual behavioral ties as informant bias. We do not enter this debate as to which form is 'accurate' or 'inaccurate', but note only that real ties may be omitted from collected network data and ties may be recorded when they do not exist.

The impact of different measurement characteristics on the reliability and validity of whole networks was extensively studied by Ferligoj and Hlebec (Ferligoj and Hlebec,

1999; Hlebec, 1999, 2001; Hlebec and Ferligoj, 2002). They found out that different dimensions of social support are not equally reliable. The informational and the emotional support are the most reliable and the material support is the least reliable. The perceptions of received support and given support (direction of a question) are equally stable. Similarly, the data collection method (recall or recognition) does not have a large effect on the quality of measures. The least reliable is the binary response scale and the most reliable is five point category scale (regardless of whether the labels are used or not). Zemljič and Hlebec (2001, 2005) investigated the stability of measures of centrality and prominence in whole networks. They found out that global measures (e.g. flow betweenness) are more sensitive to measurement errors than local measures (e.g. in-degree), that in-measures are more stable than out-measures and the reliability of mentioned measures is higher when network is denser.

Adar and Ré (2007, 23) argued that translating the research techniques of social network analysis to large scale network leads to new set of measurement instruments for data collection where researchers "can no longer be completely confident that data about individuals, or the connections between them, are accurate". An example of large scale data are internet communities (e.g. Facebook, MySpace), where biases (or errors) can appear due to application design where friends are added by default or connections are added trough spamming process in an automated way.

4.4 Comparison of errors in social network data collection process and in ordinary surveys

In previous sections the design errors in social network collection process were presented. In this section we will compare those errors to errors arising in (ordinary) social science surveys.

Groves (2004) distinguished four main types of errors:

i) Coverage error arising from failure to give any chance of selection into the sample to some people from the population.

ii) Nonresponse error where data on all persons in the sample are not collected.

iii) Sampling error arising from non-observation, because not all members from the population are measured.

iv) Measurement errors due to inaccuracies in recorded responses on the survey instruments. These can be divided into errors due to effects of an interviewer on the respondent's answers, errors due to respondents (e.g. ranging from inability to answer the question, lack of an effort or information to obtain the correct answer...), errors due to the weaknesses in wording of the survey questionnaires, and errors due to mode of data collection.

Biemer and Lyberg (2003, 35-37) defined the total survey error as "the difference between a population mean, total, or other population parameter and the estimate of the parameter based on the sample survey". The total survey error includes "all potential sources of error that can arise between planing the survey and reporting the final results" (Biemer, 2010, 27) and can be divided into sampling error due to selecting a sample instead of the entire population, and nonsampling error due to mistakes or system deficiencies. The nonsampling errors can be made on any stage of survey process and can be viewed as mistakes or unintentional errors in contrast to sampling errors, which are "intentional" and can be controlled through adjustment of sample size. The survey design should be optimized, which means "finding a balance between sampling errors and nonsampling errors so that the overall total survey error is as small as possible for the budget available for the survey" (Biemer and Lyberg, 2003, 38) .

In comparison to Groves (2004) classification presented above, where three categories of nonsampling errors are presented (coverage error, nonresponse error, and measurement error). Biemer and Lyberg (2003, 39) decomposed the nonsampling errors to five parts presented in Table 4.1. The specification error and processing error are added to the coverage error, nonresponse error and measuremet error.

Specification errors occur when the concept which should be measured and the concept implied in question differ and could be caused by poor communication between per-

Table 4.1: Five major sources of nonsampling error and their potential causes (Biemer and Lyberg, 2003; Biemer, 2010)

Sources of errors	Types of errors
Specification error	Concepts
	Objectives
	Data elements do not align with objectives
	Questions lack relevance for the research purposes
Frame error	Omissions
	Erroneous inclusion
	Duplications
	Faulty information
Nonresponse error	Whole unit
	Within unit
	Item
	Incomplete information
Measurement error	Information system
	Setting
	Mode of data collection
	Respondent
	Interview
	Instrument
Processing error	Editing
	Data entry
	Coding
	Weighting
	Tabulation

sons involved in the survey process (e.g. between the sponsor, the researcher, the questionnaire designer and the data analyst). Specification errors should be distinguished from the measurement errors, because "they pertain specifically to the problem of measuring the wrong concept in a survey, rather than measuring the right concept poorly" (Biemer, 2010, 31) .

The second source of nonsampling errors is construction of the sampling frame, which is "usually a list of the target population members that will be used to draw the sam-

ple" (Biemer and Lyberg, 2003, 40) . The frame errors can arise "in the process for constructing, maintaining, and using the sampling frame(s) for selecting the survey sample" (Biemer, 2010, 33). In the 'frame list' (or more generally the data base) the population members can be omitted (also noncoverage error), duplicated (especially if the frame list is compiled from two or more different sources (Piazza, 2010)) or erroneously included. For example, Piazza (2010, 141) emphasized that if the noncoverage error is small, up to 5%, the "sampling from such a list could bias results only slightly".

The nonresponse category includes unit nonresponse, where a sampling unit refuse to respond to any part of the survey or the (mail, post) survey is never returned to the researcher. The partially completed questionnaires caused the item nonresponse. Biemer (2010, 10) distinguished also the incomplete response which occurs especially in the open-ended questions where some information is provided, but the response is very short and inadequate. The decision about nonresponding to an item seems to be related to the item's topic, therefore the item nonresponse bias seems to greater than the unit nonresponse bias (Dixon and Tucker, 2010). The item nonresponse is also easier to estimate compared to the unit nonresponse, because more information about actors who refused to respond is available. Krosnick and Presser (2010, 263) emphasized that the key factor in minimizing the response errors in social surveys are questionnaires which should be "crafted in accordance with best practices". Recommendations about best practices should arrise from both experience and methodological research.

According to Biemer and Lyberg (2003) the key sources of measurement errors are respondents (providing incorrect informations), interviewer (influence to the respondent or incorrectly recorded responses), the questionnaire which is poorly designed (e.g. misunderstanding of terms used), and also the mode of the survey (e.g. telephone survey, face-to-face interview..). Biemer (2010) emphasized that measurement errors are often the most damaging source of errors in surveys.

The last, fifth source of nonsampling errors are errors due to data editing, data entry, data coding (especially the for open-ended questions) and assignment of survey

weights.

In our scheme of errors in the social network research design on page 39 in Section 4.1 we have three main categories of errors: boundary specification problem, design of questionnaire and caused by actors. The boundary specification problem can be compared with frame error from classification made by (Biemer and Lyberg, 2003; Biemer, 2010) , especially with omission of units or erroneous inclusion of actors when setting network boundaries.

Our second category of errors caused by design of questionnaire is partially related with specification errors and relevance of the questions to the research purposes. All three subcategories of those errors arise in the phase of designing the questionnaire (instruction about free or fixed number of nominations, list of actors or recall procedure, and wording of questions) which is not specifically highlighted in the Biemer's classification of nonsampling errors. On the other hand all those question formats can be affected with so called reporting error from the cognitive survey response process which has four components: comprehension of the question, retrieval of the relevant information, integration of this information via judgment or estimation process, and the reporting of the resulting judgment or estimate (Tourangeau and Bradburn, 2010). The errors in this process can occur due to misunderstanding of the question, having trouble of remembering the information and formulation of the answer.

The third category of errors caused by actors covers the sources of errors from Biemer's classification; the nonresponse error with both unit or actor nonresponse and item or tie nonresponse, and measurement errors.

5 Stability of blockmodeling

In this chapter two indices of blockmodeling stability are presented: the Adjusted Rand Index (Section 5.1.1) and the portion of incorrect block types 5.1.2.

5.1 Comparison of two blockmodels

As noted in Section 3.4, the result of using a blockmodeling procedure is a partition (of actors) determining positions of actors and image matrix with selected block types. **The stability of a blockmodel to an error** can be defined, or measured, with two indices where the original blockmodel and the obtained blockmodel from network with introduced errors are compared. The first index, the Adjusted Rand Index, measures the differences between the two partitions in terms of their composition. The lower the index is, the worse is the correspondence of the position membership. Equally important - perhaps more important - is whether the identified blocks, given the positions for the treated network, correspond (or not) to the block types in the true blockmodel. Further, the correct block types need to be in their correct blockmodel locations. This is measured by a second index calculated as the percent (or proportion) of the incorrectly located blocks. The higher the rate is, the worse is the identification of block types in the blockmodel.

5.1.1 The Adjusted Rand Index

When we identify two different blockmodels from the same network we would like to answer the natural question: how strong is the agreement between two obtained partitions? One of the most widely used and popular indices for comparing partitions or, more precisely, measuring concordance between them is the Rand index (Hubert

and Arabie, 1985; Saporta and Youness, 2002). It is computed based on how pairs of actors are disposed in two partitions U and V of the same data set of size n. The total number of possible combinations of pairs is $\binom{n}{2}$ and they can be classified into four groups:

(i) actors in a pair are in the same clusters in partition U and in the same clusters in partition V,

(ii) actors in a pair are in the same cluster in partition U and in different clusters in partition V,

(iii) actors in a pair are in different clusters in partition U and in the same cluster in partition V,

(iv) actors in a pair are in the same clusters in partitions U and V.

The frequencies of those types of pairs are presented in Table 5.1.

Table 5.1: Contingency table for classification of pairs from two partitions

	Partition V	
Partition U	Pair in same group	Pair in different groups
Pair in same group	a	b
Pairs in different groups	c	d

The Rand Index is the fraction of agreement and can be computed as

$$RI = \frac{a+d}{a+b+c+d} = \frac{a+d}{\binom{n}{2}} \,. \tag{5.1}$$

Formula (5.1) can be written more exactly with notation presented in Table 5.2. Let's say that given set of n actors is partitioned into two partitions U and V. Partition U has R clusters and partition V has (in general) C clusters, n_{ij} denotes the number of actors that belong to cluster U_i in partition U and to cluster V_j in partition V. The frequencies from Table 5.1 can be computed as

$$a \;=\; \sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - n}{2} \,,$$

$$
\begin{aligned}
b &= \sum_{i=1}^{R} \binom{n_{i\cdot}^2}{2} - a = \frac{\sum_{i=1}^{R} n_{i\cdot}^2 - \sum_{i=1}^{R} \sum_{C=1}^{j} n_{ij}^2}{2}, \\
c &= \sum_{j=1}^{C} \binom{n_{\cdot j}^2}{2} - a = \frac{\sum_{j=1}^{C} n_{\cdot j}^2 - \sum_{i=1}^{R} \sum_{C=1}^{j} n_{ij}^2}{2}, \\
d &= \binom{n}{2} - a - b - c = \frac{\sum_{i=1}^{R} \sum_{c=1}^{C} n_{ij}^2 + n^2 - \sum_{i=1}^{R} n_{i\cdot}^2 - \sum_{j=1}^{C} n_{\cdot j}^2}{2}. \qquad (5.2)
\end{aligned}
$$

Table 5.2: Notation used to compute the Rand Index and the Adjusted Rand Index

		Partition V				
	Cluster	V_1	V_2	$\dots$	V_C	Sums
Partition U	U_1	n_{11}	n_{12}	$\dots$	n_{1C}	$n_{1\cdot}$
	U_2	n_{21}	n_{22}	$\dots$	n_{2C}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	U_R	n_{R2}	n_{R2}	$\dots$	n_{RC}	$n_{R\cdot}$
	Sums	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot C}$	$n_{\cdot\cdot} = n$

The Rand Index from Equation (5.1) can be in above notation (see Equations (5.2)) computed as

$$
RI = \frac{\binom{n}{2} + \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}\left(\sum_{i=1}^{R} n_{i\cdot}^2 + \sum_{j=1}^{C} n_{\cdot j}^2\right)}{\binom{n}{2}}. \qquad (5.3)
$$

The distribution of the Rand index is far from normal and depends on "the number of clusters, on their proportions and separability" (Saporta and Youness, 2002, 247). The Rand index has some imperfections so that the expected value of the Rand index of two random partitions does not take a constant value Santos and Embrechts (2009); Vinh et al. (2009). The values of Rand Index also approache to its upper limit of unity as the number of clusters increases (Rand in Santos and Embrechts, 2009, 387). There is agreement in the literature that a correction (or normalization) for chance is necessary and that Adjusted Rand index should be used (Yeung and Ruzzo, 2001; Steinley, 2004; Warrens, 2008; Santos and Embrechts, 2009; Vinh et al., 2009). The Adjusted Rand Index (ARI) is computed as

$$
ARI = \frac{Rand\ Index - Expected\ Index}{Maximum\ Index - Expected\ Index} =
$$

$$= \frac{\binom{n}{2}(a+d) - ((a+b)(a+c)+(c+d)(b+d))}{\binom{n}{2}^2 - ((a+b)(a+c)+(c+d)(b+d))}. \quad (5.4)$$

In notation from Table 5.2 the Adjusted Rand Index (Equation 5.4) can be calculated as

$$ARI = \frac{\binom{n}{2}\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \sum_{i=1}^{R}\binom{n_{i\cdot}}{2}\sum_{i=1}^{C}\binom{n_{\cdot j}}{2}}{\frac{1}{2}\binom{n}{2}\left(\sum_{i=1}^{R}\binom{n_{i\cdot}}{2} + \sum_{i=1}^{C}\binom{n_{\cdot j}}{2}\right) - \sum_{i=1}^{R}\binom{n_{i\cdot}}{2}\sum_{i=1}^{C}\binom{n_{\cdot j}}{2}}. \quad (5.5)$$

The Adjusted Rand Index has expected value zero and maximal value one. Besides, "there is a wider range of values that the Adjusted Rand Index can take on, thus increasing the sensitivity of the index" (Yeung and Ruzzo, 2001, 764). Steinley (2004) showed that Adjusted Rand Index is invariant to changes in the number of clusters, number of objects to be classified and relative cluster size. He presented (based on 168.000 simulations) some general guidelines for interpreting of values of the Adjusted Rand Index. Proposed values for determining the cluster recovery or agreement between two partitions of the Adjusted Rand Index (ARI) are:

(i) $ARI > 0.9$ indicates excellent accordance,

(ii) $ARI > 0.8$ indicates good accordance,

(iii) $ARI > 0.65$ can be viewed as moderate accordance, and

(iv) $ARI < 0.65$ indicates poor accordance.

In our evaluation of stability of blockmodeling in Chapter 7 we would say that blockmodel is stable in terms of agreement between partitions or that correspondence of the position memberships is acceptable if the mean of Adjusted Rand index is above 0.8.

5.1.2 The proportion of incorrect block types

The second index for comparison of blockmodels is the proportion of incorrect block types in one blockmodel compared to a reference blockmodel. Let I_1 be the image of original (true) blockmodel and I_2 the image of a blockmodel obtained from a network with introduced error, also named 'measured' blockmodel (see (5.6)). Consider the following two blockmodels, where I_1 e.g. presents the original blockmodel structure.

$$I_1 = \begin{bmatrix} \text{com} & \text{null} \\ \text{null} & \text{com} \end{bmatrix} \qquad I_2 = \begin{bmatrix} \text{com} & \text{null} \\ \text{null} & \text{null} \end{bmatrix} \quad (5.6)$$

The proportion of incorrect blocks ($ErrB$) measures the blockmodels disagreement and is defined as the number of block disagreement in blockmodels divided by the number of blocks in a blockmodel.

$$ErrBlock = \frac{number\ of\ block\ disagreements}{number\ of\ blocks\ in\ a\ blockmodel} = \frac{1}{4} = 0.25 \tag{5.7}$$

If the two blockmodels agree perfectly then $ErrB = 0$. However, when the two image matrices disagree regarding the location of blocks then $ErrB > 0$. In our example one block in I_2 differs from an image matrix of original blockmodel I_1 (see (5.6)). For two presented image matrices the proportion of incorrect blocks is 0.25 (see Equation (5.7)).

The boundary of ARI index is established based on extended simulations of Steinley (2004), while the boundary for $ErrB$ values is drawn somewhat arbitrary. We took results where the mean of the proportion of incorrectly identified blocks ($mErrB$) exceeds 0.2 to be unacceptable.

These two indices provide a clear and straightforward way of measuring the correspondence between two blockmodels. Their relevance is suggested by the importance of two central ideas of social network analysis pointed out with Doreian (2008). "The first is that the structure of a social network, as a whole, is important to collective outcomes at the level of the network. The second is that the location occupied in a network is important for outcomes at the actor level". In terms of the blockmodeling procedure, the whole blockmodel (or image matrix) is important at the network level and position of actor in model is important at individual level. Both have to be depicted accurately to examine these two basic network ideas.

6 The design of simulation studies for evaluation of stability of blockmodeling

This chapter represents the outline of the simulation studies for evaluation of stability of blockmodeling to different types of design errors presented in Chapter 4.

The basic scheme of simulations is presented in Section 6.1 and presents the framework of all simulation studies in the dissertation. The additional specific steps in simulations (e.g. use of different non-response treatments in actor non-response) and more detailed characteristics of simulations (e.g. number of simulated starting networks, percent of introduced errors..) are presented in subsections of Chapter 7 before the results of simulations.

This is followed by the presentation of networks used in simulation studies (Section 6.2). Two main types of networks are used; real networks from the literature (Section 6.2.1) and simulated networks (Section 6.2.3 to Section 6.2.4) with known structure based on some type of equivalence presented in Section 3.2.

6.1 A basic scheme of simulations

The basic scheme of our simulation study is straightforward:

1. **Select** a network from the literature or
   generate a whole network under a known starting model.

2. **Establishing a blockmodel of the whole (real) network** that has two parts:
   (i) the known (real) partition of the actors of the whole network into positions; and
   (ii) the image matrix with the known distribution of block types by location.

3. Let *nGen* denote the number of simulations for a given combination of network type, type of design error, amount of error, and (in some cases) treatment regime. **For i=1:nGen**, do the following:

   (a) **Construct the network with introduced design errors (the measured networks)**

   (b) **Establish a blockmodel of the measured network** that also has two parts:
   (i) the partition of the actors of the measured network into positions; and
   (ii) the blockmodel image of the measured network.

   (c) **Compare the resulting blockmodels of the whole and the measured networks** using:
   (i) the Adjusted Rand Index to compare the two sets of positions (as described in Section 5.1.1); and
   (ii) the proportion of incorrect block types (as described in Section 5.1.2).

4. **Investigate the impact of design errors** in terms of the mean of the values of ARI - denoted as $mARI$ - and the mean of the proportion of incorrect blocks - denoted by $mErrB$.

In the dissertation we use the following terms:

- a *whole network* that is known network (or starting network in the simulations),

- a *measured network* which is obtained from the whole network by introducing some kind of design error, and
- a *measured and treated* (or *treated*) network that is obtained by treating a measured network to deal with the introduced errors[14].

6.2 Networks used in evaluation of stability

In this section networks used in evaluation of stability of blockmodeling are presented. First, networks are divided to real and simulated networks. Additional classification of networks is made based on selected type of equivalence in the blockmodeling procedure.

6.2.1 Real whole networks partitioned based on structural equivalence

6.2.1.1 A boy-girl liking ties network

The first real network presents a liking relationship between boys and girls in a classroom (used by Doreian et al. (2005, 237) and is presented in Figure 6.1 (left). There are clearly two subgroups, based on gender, each with many internal ties. The best fitting model based on structural equivalence with two clusters is the one shown in Figure 6.1 (right). There are 12 inconsistencies and they are all null ties within the two diagonal blocks. This served as a prototype for the symmetric blockmodel structure in Section 6.2.3.1.

6.2.1.2 The student note borrowing network

Data for a note borrowing network for 15 undergraduate students attending lectures of a course were collected by Hlebec (1993) and used by Batagelj et al. (2004). The students were asked: "From whom would you borrow learning materials?" The number of choices was not fixed. This network is presented in Figure 6.2 (left) together with

[14]Missing data treatments are used in case of actor non-response in Section 7.3

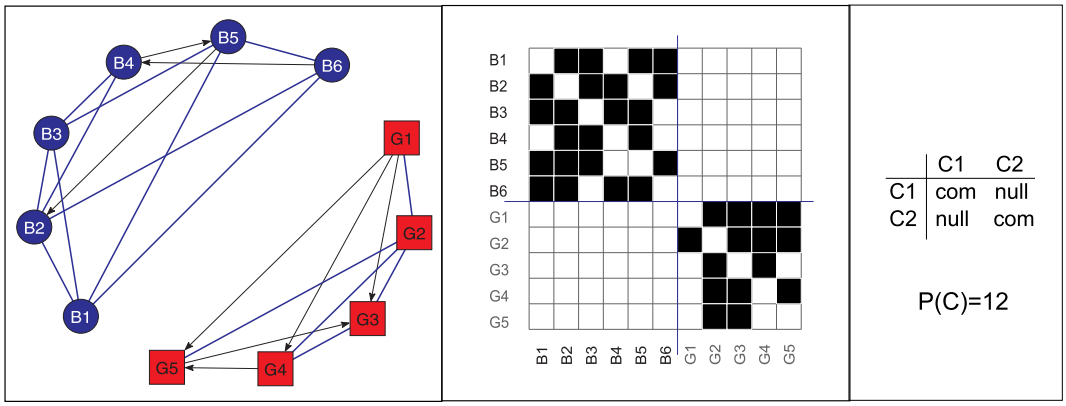


Figure 6.1: A boy-girl network of liking ties (left), two partitions based on structural equivalence (middle) and image matrix (right)

a fitted blockmodel using structural equivalence (shown in the middle panel of Figure 6.2). There are three clusters (positions) that are labeled C1, C2, and C3. Boys are represented by squares and the girls by circles. Position memberships in the network diagram on the left of Figure 6.2 are indicated by the colors of the vertices. Blue indicates membership in cluster C1, red shows membership in C2 and green indicates membership in C3. The fitted blockmodel with 28 inconsistencies is on the right in Figure 6.2 (and is the prototype for the (second) non-symmetric blockmodel structure (in Section 7.3.3.3)).

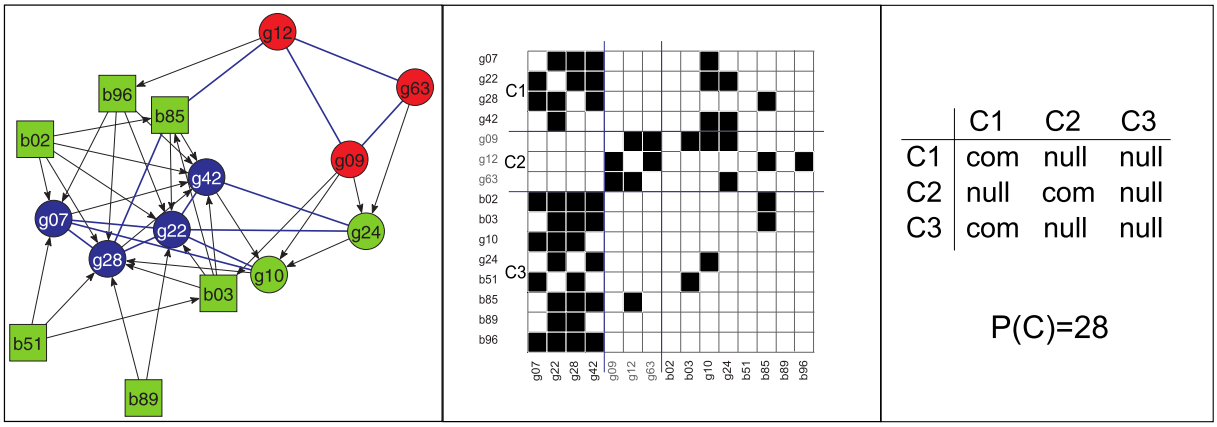


Figure 6.2: The note borrowing network (left), three partitions based on structural equivalence (middle) and image matrix (right)

6.2.1.3 The networks of emotional support

The networks of emotional support are part of a large study of quality of measurement instrument for social network data (Ferligoj and Hlebec, 1998, 1999; Hlebec, 1999, 2001; Hlebec and Ferligoj, 2001; Zemljič and Hlebec, 2005). Networks measure different types of social support (instrumental, informational, emotional support and social companionship) with different scales (e.g. binary, five-point ordinal scale with or without labels...). Both giving and receiving of social support were measured with two data collection techniques, recognition and free recall of actors.

In our studies two networks measuring giving and receiving (original and reversed question) of emotional support were used. Networks with major characteristics are presented in Section 7.2.2.

6.2.2 Real whole networks partitioned based on generalized types of equivalence

6.2.2.1 A Student Government data

Student Government data were collected by Hlebec in 1992 during an experiment on different methods for collecting social network data. The networks originally consist of twelve members and advisors of the Student Government of the University of Ljubljana and their cognition about communication interactions, but one respondent refused to cooperate in the experiment. Therefore, several authors in later papers (Hlebec, 1993, 1999; Doreian et al., 2005; de Nooy et al., 2005) take into account just eleven actors without non-respondent. As described in Section 4.3.1 they use the complete approach to treat the non-response. The incoming ties about non-respondent are available in degree of Hlebec (1992) and we used them to examine their impact on obtained blockmodeling. First, the complete data sets with summary of main results on blockmodeling are presented and at the end the incoming ties for non-respondent are described.

Two methods or designs of questions were used in experiment, recall and recognition,

and there was no limitation on number of listed persons (Hlebec, 1992, 1993). The data about (cognitive) communication flow, limited to matters of the Student Government during the last six months, among actors were elicited through the following three questions:

1. Who of the members and advisors of the Student government do you (most often) informally discuss with?

2. Which members and advisors of the Student Government do you (most often) ask for an opinion?

3. Which of the members and advisors of the Student Government (most often) ask you for an opinion?

Six networks (obtained with two methods (recall and recognition) and three questions) are presented in Table 6.1. Networks consist of seven ministers (labeled from m1 to m7), one prime minister (pm) and three advisors (a1, a2, and a3).

Hlebec (1993) reported that networks obtained with recognition have richer structure than those obtained by the recall method. The average size of recognized egocentric network (the average number of persons named by each actor) was higher than recalled one. Similarly, the minimum and maximum number of persons named was on average higher for recognized networks for all three relations. The hypotheses that the size of recalled network and the recognized network is due to different types of measurement (that means that we expect that the respondent with larger recalled network would have larger recognized network) were confirmed for discussion and asked for an opinion relations.

The recall discussion network was extensively explored in terms of generalized blockmodeling by Doreian et al. (2005). They started the investigation for the best or the most suitable blockmodel by restricting the block types to null, complete, regular, row-dominant and column-dominant. The blocmodeling procedure was applied for partitions into two five clusters and the results (obtained partitions and minimum number

Table 6.1: Student Government data

(a) Discussion ties - recall

	M1	PM	M2	M3	M4	M5	M6	M7	A1	A2	A3
M1	0	1	1	0	0	1	0	0	0	0	0
PM	0	0	0	0	0	0	0	1	0	0	0
M2	1	1	0	1	0	1	1	1	0	0	0
M3	0	0	0	0	0	0	1	1	0	0	0
M4	0	1	0	1	0	1	1	1	0	0	0
M5	0	1	0	1	1	0	1	1	0	0	0
M6	0	0	0	1	0	0	0	1	1	0	1
M7	0	1	0	1	0	0	1	0	0	0	1
A1	0	0	0	1	0	0	1	1	0	0	1
A2	1	0	1	1	1	0	0	0	0	0	0
A3	0	0	0	0	0	1	0	1	1	0	0

(b) Discussion ties - recognition

	M1	PM	M2	M3	M4	M5	M6	M7	A1	A2	A3
M1	0	1	1	0	0	1	0	0	0	1	0
PM	0	0	0	1	0	1	0	1	0	0	0
M2	1	1	0	1	1	1	1	1	0	1	0
M3	0	0	0	0	0	0	1	1	0	0	0
M4	0	1	0	1	0	1	1	1	0	1	0
M5	0	1	0	1	1	0	1	1	0	0	0
M6	0	0	0	1	0	0	0	1	1	0	1
M7	0	1	0	1	0	0	1	0	0	0	1
A1	0	1	1	0	0	0	1	1	0	0	1
A2	1	1	1	0	1	0	0	0	0	0	0
A3	0	0	0	1	0	1	0	1	0	0	0

(c) Asking for an opinion ties - recall

	M1	PM	M2	M3	M4	M5	M6	M7	A1	A2	A3
M1	0	1	0	0	0	1	0	0	0	0	0
PM	0	0	0	1	0	0	0	1	0	0	0
M2	1	1	0	1	0	1	0	1	0	0	0
M3	0	1	0	0	0	0	0	0	0	0	0
M4	0	1	0	0	0	1	0	0	0	0	0
M5	0	1	0	0	0	0	0	0	0	0	0
M6	0	0	0	1	0	0	0	1	1	0	0
M7	0	1	0	1	0	0	0	0	0	0	1
A1	0	0	0	0	0	0	1	1	0	0	1
A2	1	1	1	0	1	0	0	0	0	0	0
A3	0	0	1	0	0	1	0	1	0	0	0

(d) Asking for an opinion ties - recognition

	M1	PM	M2	M3	M4	M5	M6	M7	A1	A2	A3
M1	0	1	1	0	0	0	0	0	0	0	0
PM	0	0	0	1	0	0	0	1	0	0	0
M2	1	1	0	0	0	0	0	0	0	1	0
M3	0	0	0	0	0	0	1	1	0	0	0
M4	0	1	0	0	0	1	0	1	0	1	0
M5	1	1	1	1	1	0	1	1	0	0	0
M6	0	0	0	1	0	0	0	1	1	0	0
M7	0	1	0	1	0	0	0	0	0	0	1
A1	0	1	0	0	0	0	1	1	0	0	0
A2	1	1	1	0	1	0	0	0	0	0	0
A3	0	0	0	1	0	1	0	1	0	0	0

(e) Asked for an opinion ties ties - recall

	M1	PM	M2	M3	M4	M5	M6	M7	A1	A2	A3
M1	0	1	1	0	0	0	0	0	0	1	1
PM	1	0	1	1	1	1	1	1	0	1	0
M2	0	1	0	0	1	0	0	0	0	0	0
M3	0	1	1	0	0	1	1	1	0	0	1
M4	0	1	0	0	0	1	1	0	0	1	0
M5	0	1	0	0	0	0	0	0	0	0	1
M6	0	0	0	1	0	0	0	1	1	0	1
M7	0	1	1	1	0	1	1	0	1	0	1
A1	0	0	0	0	0	0	0	0	0	0	0
A2	0	1	0	0	1	0	0	0	0	0	0
A3	0	1	0	0	0	0	0	0	1	0	0

(f) Asked for an opinion ties - recognition

	M1	PM	M2	M3	M4	M5	M6	M7	A1	A2	A3
M1	0	1	1	1	0	0	0	1	0	1	1
PM	1	0	1	1	0	1	0	1	0	1	0
M2	1	0	0	0	0	0	0	0	0	0	0
M3	0	0	1	0	0	1	1	1	1	0	0
M4	1	1	0	0	0	1	0	0	0	0	0
M5	1	1	0	0	1	0	0	0	0	0	1
M6	1	0	0	1	0	1	0	1	1	0	0
M7	0	1	0	1	0	1	1	0	1	0	1
A1	0	0	0	0	0	0	1	0	0	0	0
A2	1	0	1	0	1	0	0	0	0	0	0
A3	0	0	0	0	0	0	1	0	1	0	0

of inconsistencies for those partitions) are presented in Table 6.2[15]. Label C_r^s denotes partitions into s clusters and r is a counting index. For example there are two partitions into two clusters C_1^2 and C_2^2 with zero inconsistency (labeled I_{min}).

[15] A constraint was used to have at least two vertices in each cluster. Otherwise, even more equally well fitting partitions exist

Table 6.2: Optimal partitions for Student Government recall discussion network and allowed block types {null, com, rdo, cdo, reg}

	Partition	I_{min}
C_1^2	$\{m_1, pm, m_2, m_3, m_5, m_6, m_7, a_1, a_2\}\{m_4, a_2\}$	1
C_2^2	$\{m_1, a_2\}\{pm, m_2, m_3, m_4, m_5, m_6, m_7, a_1, a_3\}$	1
C_1^3	$\{m_1, pm, m_2, m_3, m_4, m_5, m_7\}\{m_6, a_3\}\{a_1, a_2\}$	0
C_2^3	$\{m_1, m_2, a_2\}\{pm, m_3, m_4, m_5, m_6, m_7\}(a_1, a_3\}$	0
C_3^3	$\{m_1, m_2\}\{pm, a_3\}\{m_3, m_4, m_5, m_6, m_7, a_1, a_2\}$	0
C_4^3	$\{m_1, m_4\}\{pm, a_3\}\{m_2, m_3, m_5, m_6, m_7, a_1, a_2\}$	0
C_1^4	$\{m_1, m_2\}\{pm, m_4\}\{m_3, m_5, m_6, m_7, a_2\}\{a_1, a_3\}$	0
C_2^4	$\{m_1, m_2, a_2\}\{pm, m_4\}\{m_3, m_5, m_6, m_7\}(a_1, a_3\}$	0
C_3^4	$\{m_1, m_2, a_2\}\{pm, m_4, m_6, m_7\}\{m_3, m_5\}\{a_1, a_3\}$	0
C_1^5	$\{m_1, m_2\}\{pm, m_3\}\{m_4, a_3\}\{m_5, a_1, a_2\}\{m_6, m_7\}$	1
C_2^5	$\{m_1, m_2, a_2\}\{pm, m_4\}\{m_3, m_5\}\{m_6, m_7\}(a_1, a_3\}$	1
C_3^5	$\{m_1, m_2, a_2\}\{pm, m_4\}\{m_3, m_6\}\{m_5, m_7\}\{a_1, a_3\}$	1
C_4^5	$\{m_1, a_2\}\{pm, m_3\}\{m_2, a_3\}\{m_4, m_5\}(m_6, m_7, a_1\}$	1
C_5^5	$\{m_1, a_3\}\{pm, m_5\}\{m_2, m_7, a_1\}\{m_3, m_4\}\{m_6, a_2\}$	1

Different allowed block types contribute to a new set of well-fitting partitions. Doreian et al. (2005, 228-233) represented detailed results of generalized blockmodeling also for the following combination of allowed block types: {null, rdo, cdo} (Table 6.3) and {null, cdo}. Summarized results about value of criterion function for different number of clusters are presented also for structural and regular equivalence.

According to Hlebec (1992) the refusal actor (denoted by R) for the Student Government recall discussion network has two incoming ties, from prime minister (pm) and minister 2 (m_2). The network is presented on Figure 7.46 in Section 7.3 (the last panel titled '*Null tie imputations*'), where the refusal actor is drawn with square and other 11 actors are drawn with circle.

Table 6.3: Optimal partitions for Student Government recall discussion network and allowed block types {null, rdo, cdo}

	Partition	I_{min}
C_1^2	$\{m_1, a_2\}\{pm, m_2, m_3, m_4, m_5, m_6, m_7, a_1, a_3\}$	1
C_1^3	$\{m_1, m_4\}\{pm, a_3\}\{m_2, m_3, m_5, m_6, m_7, a_1, a_2\}$	0
C_2^3	$\{m_1, m_2\}\{pm, a_3\}(m_3, m_4, m_5, m_6, m_7, a_1, a_2\}$	0
C_1^4	$\{m_1, m_2\}\{pm, m_4\}\{m_3, m_5, m_6, m_7, a_2\}\{a_1, a_3\}$	0

6.2.3 Simulated whole networks based on structural equivalence

The real whole networks partitioned based on structural equivalence (presented in Section 6.2.1) were used to construct the simulated networks. The starting whole networks were constructed based on specified image matrix, starting partition and probability of a tie in complete and in null block.

6.2.3.1 A completely symmetric blockmodel structure

The prototype for the completely symmetric blockmodel was the boy-girl liking ties network presented in Section 6.2.1.1. A two-cluster partition for a network with 10 actors with five actors in both the first cluster and second cluster was used. The cluster membership is denoted by $(1, 1, 1, 1, 1, 2, 2, 2, 2, 2)$. The image matrix has two complete blocks on the diagonal and null blocks out of diagonal and it is shown in Equation (6.1).

$$IM = \begin{bmatrix} \text{com} & \text{null} \\ \text{null} & \text{com} \end{bmatrix} \tag{6.1}$$

Ties were constructed to be consistent with this image matrix. Ties were added with probability $pTie_{null}$ where the image matrix has null blocks and for the complete blocks ties were added with probability $pTie_{com}$. The values used for $pTie_{null}$ and $pTie_{com}$ are shown in Table 6.4.

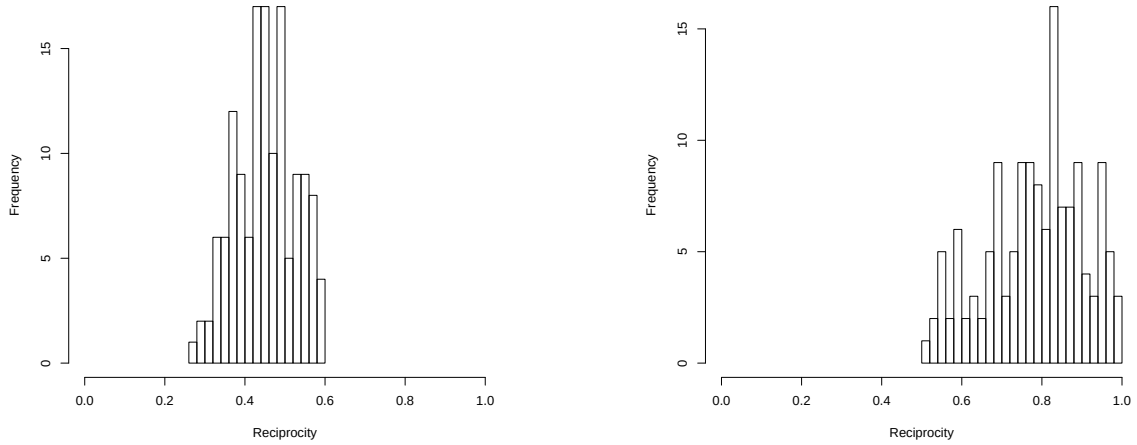
Ten networks were generated for each combination of parameters ($pTie_{com}$ and $pTie_{null}$) shown in Table 6.4. This results in 140 created different whole networks. Every constructed network was checked to see if the structure obtained with blockmodeling pro-

Table 6.4: Selected combinations of probabilities for a symmetric blockmodel structure

$pTie_{com}$	$pTie_{null}$
1	0.1, 0.2
0.95	0.0, 0.1, 0.2
0.9	0.0, 0.1, 0.2
0.8	0.0, 0.1, 0.2
0.7	0.0, 0.1, 0.2

cedure was consistent with the structure shown in Equation (6.1).

The basic network properties were examined. The histogram for density of simulated networks is presented in Figure 6.3(a). The minimal density of simulated network is 0.278 and the maximal density is equal to 0.600 ($Q_1 = 0.400,\ Me = 0.456,\ Q_3 = 0.503$). The mean density of simulated networks is 0.452 with standard deviation 0.075.



(a) Density (b) Reciprocity

Figure 6.3: Histograms of density and reciprocity for the completely symmetric blockmodel structure

The extent to which a network is symmetric was measured by reciprocity (Huisman, 2009) and was calculated for each whole network. It is defined for directed networks

as

$$reciprocity = \frac{2 \cdot M}{2 \cdot M + A}\ , \tag{6.2}$$

where M indicates the number of mutual dyads and A the number of asymmetric dyads. The descriptive statistics for this measure over the 140 whole networks are ($Min = 0.50, Q_1 = 0.70, Me = 0.79, Q_3 = 0.88, Max = 1.00$) and confirm that these networks were highly symmetric (Figure 6.3(b)).

6.2.3.2 A first non-symmetric blockmodel structure

The second structure for a simulated whole network is based on the image matrix in Equation (6.3) with a partition having three positions.

$$IM = \begin{bmatrix} \text{com} & \text{null} & \text{null} \\ \text{null} & \text{com} & \text{null} \\ \text{com} & \text{null} & \text{com} \end{bmatrix} \tag{6.3}$$

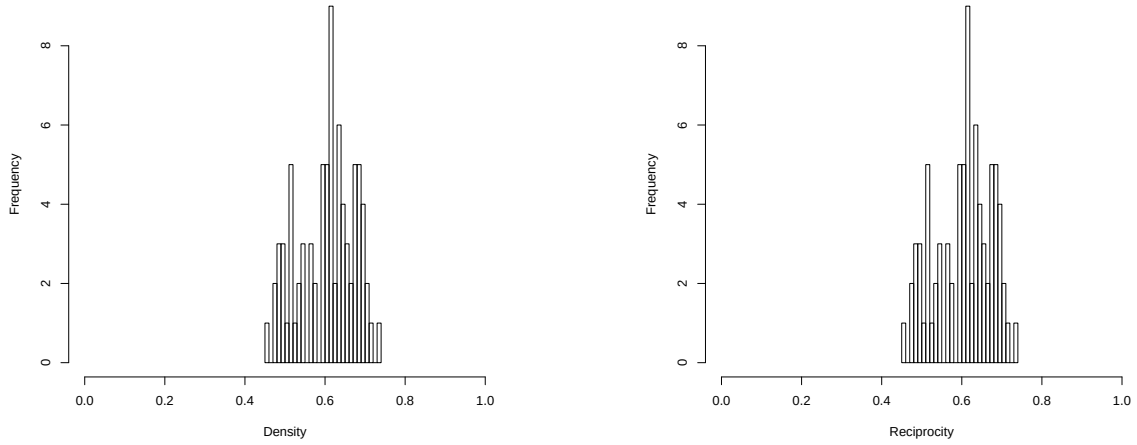
Note that the lower left block and all three diagonal blocks in Equation (6.3) are complete. The membership of the three-cluster partition for a network with 15 vertices is denoted by $(1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3)$. The construction of the whole networks was done in the same manner as for the completely symmetric structure with regard to null and complete blocks. The probabilities used for these constructions are shown in Table 6.5.

Table 6.5: Selected combinations of probabilities for whole networks with both non-symmetric blockmodel structures

$pTie_{com}$	$pTie_{null}$
1	0.1, 0.2
0.9	0.0, 0.1, 0.2
0.8	0.0, 0.1, 0.2

Again, 10 networks were constructed for each combination of probabilities ($pTie_{com}$ and $pTie_{null}$) in Table 6.5. The descriptive statistics for reciprocity (Figure 6.4(b)) across the 80 whole networks ranges from $Min = 0.46$ to $Max = 0.73$ ($Q_1 = 0.55, Me =$

$0.61, Q_3 = 0.66$). The mean density for whole starting networks is 0.61 with standard deviation 0.07 ($Min = 0.46, Q_1 = 0.55, Me = 0.61, Q_3 = 0.66, Max = 0.74$).



(a) Density (b) Reciprocity

Figure 6.4: Histograms of density and reciprocity for the first non-symmetric blockmodel structure

6.2.3.3 A second non-symmetric blockmodel structure

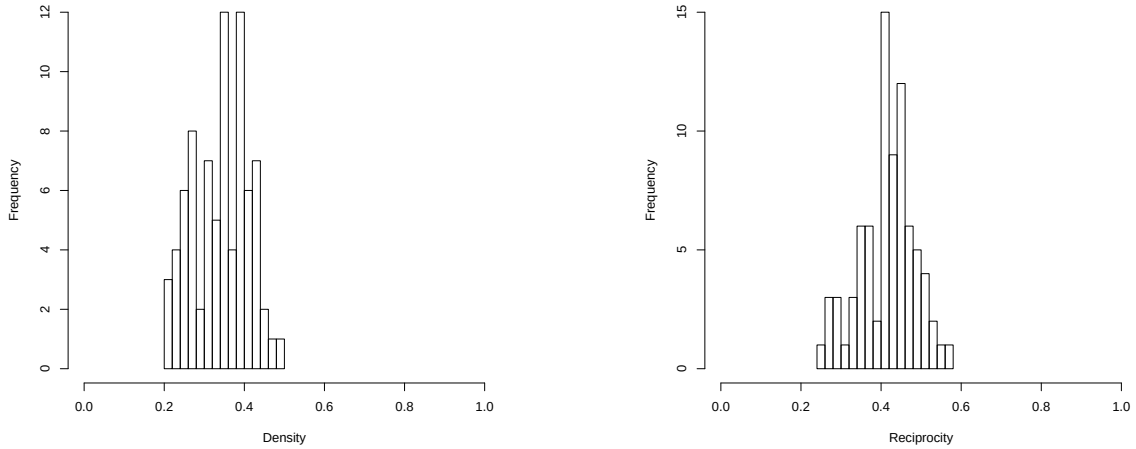
The prototype for the third structure for whole networks based on structural equivalences was the note borrowing network (presented in Section 6.2.1.2). The starting networks were constructed based on the image matrix shown in Equation (6.4).

$$IM = \begin{bmatrix} \text{com} & \text{null} & \text{null} \\ \text{null} & \text{com} & \text{null} \\ \text{com} & \text{null} & \text{null} \end{bmatrix} \tag{6.4}$$

Again, there are three clusters with the same cluster membership as the networks with the first non-symmetric blockmodel structure. The only difference is the presence of a null block on the diagonal.

Ten networks were generated for each combination of probabilities ($pTie_{com}$, $pTie_{null}$) that were presented in Table 6.5. The summary description of the reciprocity measures ranges from 0.26 to 0.57 with a median of 0.42 ($Q_1 = 0.37,\ Q_3 = 0.46$). Replacing a diagonal complete block with a null block created networks with slightly lower

reciprocity measures than for the networks from the first example of non-symmetric blockmodel structure.



(a) Density (b) Reciprocity

Figure 6.5: Histograms of density and reciprocity for the second non-symmetric blockmodel structure

6.2.4 Simulated whole networks based on regular equivalence

Two well known structures (Doreian et al., 2005, 235-236) were selected for the starting structure for simulated networks based on regular equivalence: cohesive subgroups model and core-periphery model.

6.2.4.1 The cohesive subgroup model

The cohesive subgroup model, with intraposition ties and with no ties between positions, is usually schematically presented as:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where 0*s* indicates the null blocks and 1*s* can indicates also additional blocks beside the complete blocks (Doreian et al., 2005). When the regular blocks will be used in places

with 1s, the term *regular cohesive subgroups* model will be used.

The more detailed procedures for generating networks with regular blocks is as follows (Žiberna, 2007, 171):

(i) The starting partition is used to split empty matrix of 0s into blocks and to determine the size of a network.

(ii) With blocks where the blockmodel indicated null blocks, nothing is changed.

(iii) In regular block each cell had a probability to become 1 equal to $pTie_{reg}$:

$$pTie_{reg} = \frac{1}{min(n_r, n_c) - 1}, \tag{6.5}$$

where n_r and n_c are the number of rows in a block and number of columns in a block, respectively.

(iv) Each block is checked for regularity, that is, each row and column are checked if they had at least one 1. If not, the regularity is enforced by adding 1 to a randomly chosen cell from that row or column.

Two different network sizes were used in simulations; the smaller network with 10 actors and a network with 15 actors. Additional parameters in simulations were the number of clusters and the number of actors in each cluster.

Two-clusters partitions

In the smaller network with 10 actors two-cluster partitions were used. The starting image matrix IM_1 is presented in Equation (6.6). The membership of the two-cluster partition for a network with 10 actors is determined with the following starting partitions:

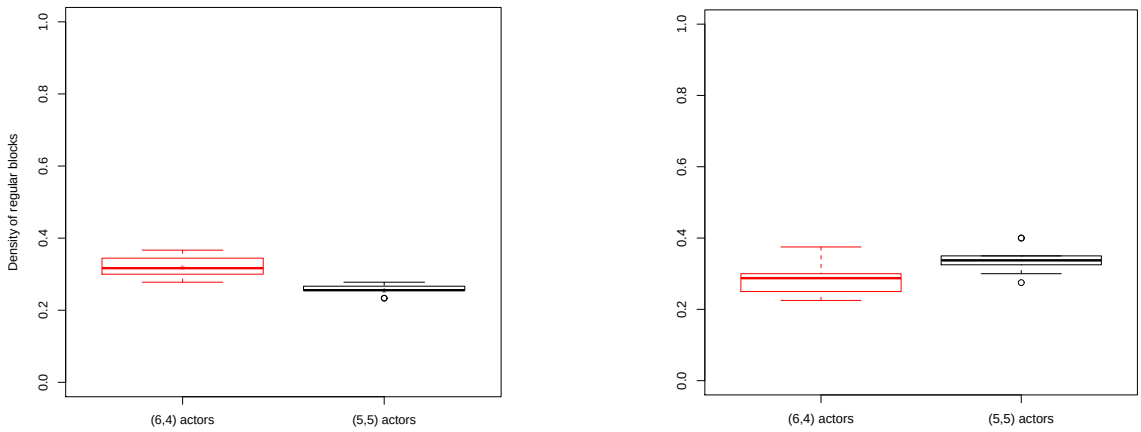
- (5, 5) actors: $C_{55} = (1, 1, 1, 1, 1, 2, 2, 2, 2, 2)$,
- (6, 4) actors: $C_{64} = (1, 1, 1, 1, 1, 1, 2, 2, 2, 2)$.

In the first example (C_{55}) both clusters have 5 actors, and in the second example (C_{64}) the first cluster has 6 actors and the second one has 4 actors.

$$IM_1 = \begin{bmatrix} \text{reg} & \text{null} \\ \text{null} & \text{reg} \end{bmatrix} \qquad IM_2 = \begin{bmatrix} \text{reg} & \text{null} & \text{null} \\ \text{null} & \text{reg} & \text{null} \\ \text{null} & \text{null} & \text{reg} \end{bmatrix} \tag{6.6}$$

For each starting two-cluster partition, 10 networks were constructed for calculated probability $pTie_{reg}$ in regular blocks. For the C_{55} partition in regular core-periphery model, the probability of a tie in regular blocks according to Equation 6.5 is equal to $\frac{1}{4}$. The probability of tie in the biggest regular block in partition is $\frac{1}{5}$ and in the smallest block with four actors that probability is equal to $\frac{1}{3}$.

The density (Figure 6.6(a)) across the 10 whole networks for C_{55} partition ranges from $Min = 0.14$ to $Max = 0.19$ ($Q_1 = 0.16, Me = 0.17, Q_3 = 0.18$). The mean density for whole starting networks is 0.166 with standard deviation 0.015. The density across the 10 whole networks with (6, 4) actors partition ranges from $Min = 0.14$ to $Max = 0.20$ ($Q_1 = 0.16, Me = 0.17, Q_3 = 0.19$, mean=0.169, sd=0.019).



(a) Network density

(b) Density of regular blocks

Figure 6.6: Boxplots of density and density of regular blocks for the regular cohesive subgroup two-cluster models

The density of regular blocks is presented in Figure 6.6(b). The mean value of density of regular blocks across 10 whole starting networks with (6, 4) actors partition is 0.375

(sd=0.034). The minimal regular block density is 0.32, and the maximal value is equal to 0.43 ($Q_1 = 0.35, Me = 0.38, Q_3 = 0.40$). The density of regular blocks across the 10 whole networks with (5, 5) actors partition ranges from $Min = 0.22$ to $Max = 0.38$ ($Q_1 = 0.28, Me = 0.30, Q_3 = 0.30$, mean=0.293, sd=0.041). The presented densities are higher than $pTie_{reg}$ from Equation 6.5, because after generation of ties blocks are checked for regularity and additional ties are enforced if necessary.

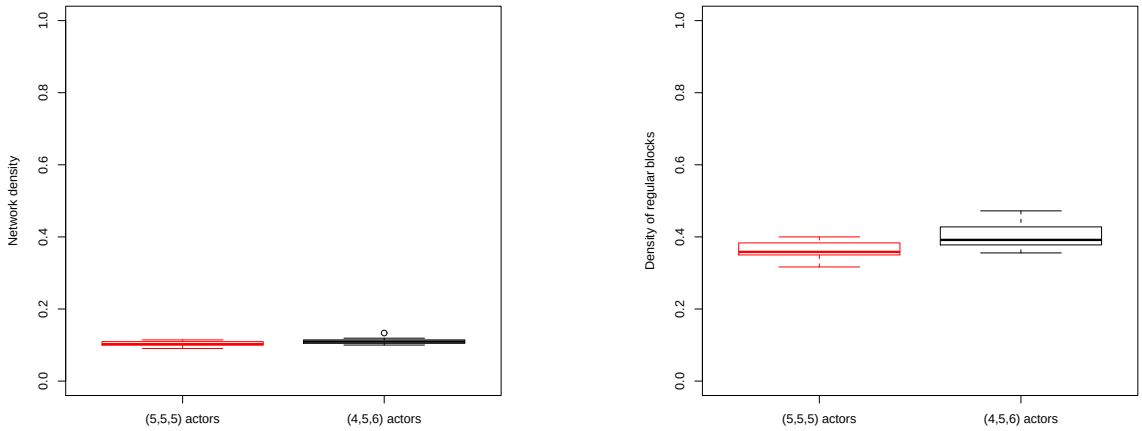
Three-clusters partitions

The membership of the three-cluster partition for a network with 15 actors and block-model structure presented with IM_2 in Equation 6.6 is determined with the following starting partitions:

- (5, 5, 5) actors: $C_{555} = (1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3)$,
- (4, 5, 6) actors: $C_{456} = (1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3)$.

In the first example (C_{555}) all three clusters have 5 actors, and in the second example (C_{456}) the first cluster has 4 actors, the second cluster has 5 actors, and the third cluster has 6 actors.

Densities of whole starting networks are lower compared to two-cluster partition, because in two-cluster models we have half regular and half null blocks, and in three-cluster models there are six null blocks out of nine. The mean density of 10 starting whole networks for C_{555} patition is 0.102 with standard deviation 0.007 ($Min = 0.09, Me = 0.10, Max = 0.11$). The mean density for C_{456} partition is a little bit higher and is equal to 0.112 (sd=0.009), and values are in range from 0.10 to 0.13 (Figure 6.7(a)). On the other hand, densities of regular blocks are more similar to two-cluster starting networks. The mean density of regular blocks for C_{555} partition is 0.362 with standard deviation 0.026 ($Min = 0.32, Me = 0.36, Max = 0.40$). The densities of regular blocks for (4,5,6) actors partition are in range from 0.36 to 0.47 ($Q_1 = 0.38, Me = 0.39, Q_3 = 0.47$) with mean value 0.401 and standard deviation 0.036 (Figure 6.7(b)).



(a) Network density

(b) Density of regular blocks

Figure 6.7: Boxplots of density and density of regular blocks for the regular cohesive subgroup three-cluster models

6.2.4.2 The core-pheriphery model

Our core-pheriphery model has one central position or core position, which is internally cohesive and connected with all other positions. The other positions from the periphery are connected to this core position and are not connected to other periphery positions neither are internally cohesive. The model is schematically presented as:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

where 0*s* and 1*s* have the same meaning as explained in Section 6.2.4. When instead of 1*s* regular blocks are used, the term *regular core-periphery model* will be used.

Two-clusters partitions

Two different network sizes (as in previous section) will be used. In the smaller network with 10 actors two-cluster partitions will be used. The starting image matrix IM_3 is presented in Equation (6.7). The membership of the two-cluster partition for a network with 10 actors is determined with the following starting partitions:

- (6, 4) actors: $C_{64} = (1, 1, 1, 1, 1, 1, 2, 2, 2, 2)$,

- (5, 5) actors: $C_{55} = (1, 1, 1, 1, 1, 2, 2, 2, 2, 2)$,
- (4, 6) actors: $C_{46} = (1, 1, 1, 1, 2, 2, 2, 2, 2, 2)$.

In the first example the core cluster has 6 actors and the periphery core is smaller with 4 actors. In the second example (C_{55}) both core and periphery cluster have 5 actors, and in the third example (C_{64}) the core cluster is bigger and has 6 actors and the periphery cluster has 4 actors.

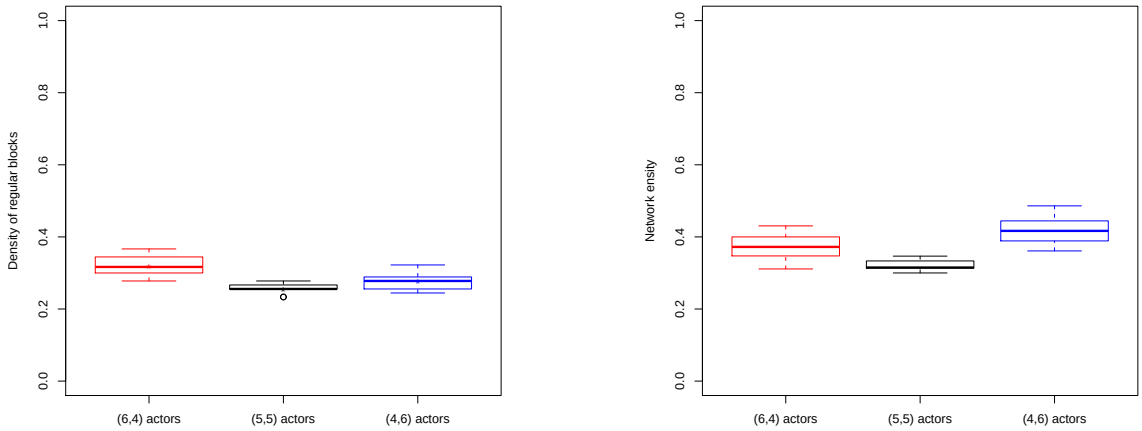
$$IM_3 = \begin{bmatrix} \text{reg} & \text{reg} \\ \text{reg} & \text{null} \end{bmatrix} \qquad IM_4 = \begin{bmatrix} \text{reg} & \text{reg} & \text{reg} \\ \text{reg} & \text{reg} & \text{null} \\ \text{reg} & \text{null} & \text{null} \end{bmatrix} \tag{6.7}$$

For each starting two-cluster partition for the core-periphery model 10 networks were constructed. The density (Figure 6.8(a)) across the 10 whole networks with (6, 4) actors partition ranges from $Min = 0.28$ to $Max = 0.37$, the mean density is 0.321 with standard deviation 0.028. The values for density (see Figure 6.8(a)) across the 10 whole networks with (5, 5) actors partition are lower and ranges from $Min = 0.28$ to $Max = 0.27$ with mean 0.257 (sd=0.015). The mean density across the 10 whole networks with (4, 6) actors partition is 0.279 (sd=0.024, $Min = 0.24$, $Max = 0.32$).

Densities of regular blocks were examined and boxplots are presented on Figure 6.8(b). The mean value of density of regular blocks across 10 whole networks with (6, 4) actors partition is 0.372 (sd=0.036). The minimal regular block density is 0.31, and the maximal value is equal to 0.43. The density of regular blocks across the 10 whole networks with (5, 5) actors partition ranges from $Min = 0.30$ to $Max = 0.35$ ($Q_1 = 0.31, Me = 0.31, Q_3 = 0.33$, mean=0.321, sd=0.016). The mean density of regular blocks with (4, 6) actors partition is 0.417, with standard deviation 0.040 ($Min = 0.36, Q_1 = 0.39, Me = 0.42, Q_3 = 0.44, Max = 0.49$).

Three-clusters partitions

The blockmodel structure with three clusters is presented with IM_4 in Equation (6.7). There are two core clusters, which are internally cohesive and connected with other



(a) Network density

(b) Density of regular blocks

Figure 6.8: Boxplots of density and density of regular blocks for the regular core-periphery two-cluster models

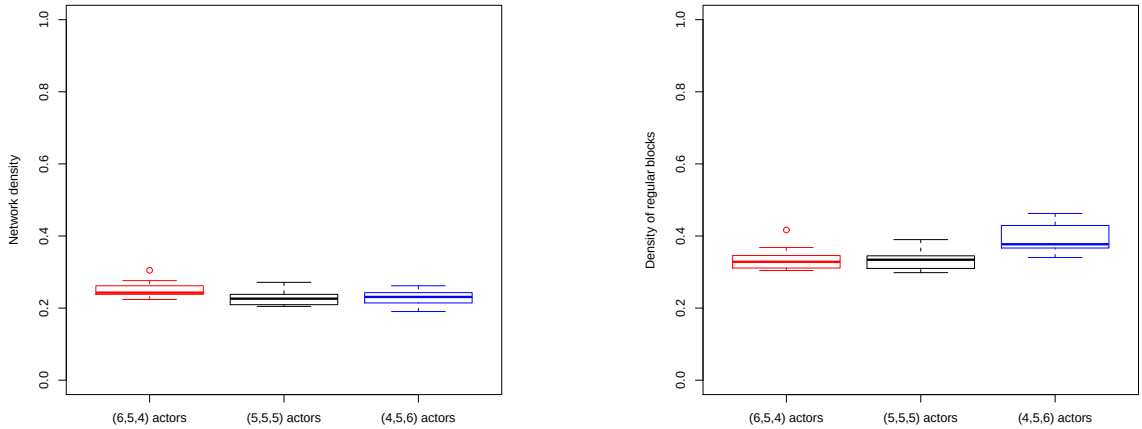
positions and one periphery cluster. The membership of the three-cluster partition is determined with the three different starting partitions:

- (6, 5, 4) actors: $C_{654} = (1,1,1,1,1,1,2,2,2,2,2,3,3,3,3)$,
- (5, 5, 5) actors: $C_{555} = (1,1,1,1,1,2,2,2,2,2,3,3,3,3,3)$,
- (4, 5, 6) actors: $C_{456} = (1,1,1,1,1,2,2,2,2,3,3,3,3,3,3)$.

In the first example (C_{654}) the first core cluster has 6 actors, the second core cluster has 5 actors, and the periphery cluster has 4 actors. In the second example (C_{555}) all three clusters have 5 actors, and in the third example (C_{456}) the first core cluster is the smallest with 4 actors, the second core cluster has 5 actors, and the biggest is the periphery cluster with 6 actors.

The densities of 10 starting networks for C_{654} partition are in range from 0.22 to 0.30 with mean value 0.251. Densities for C_{555} partition are a little bit lower with mean value 0.228 ($Min = 0.20, Q_1 = 0.21, Me = 0.23, Q_3 = 0.24, Max = 0.27$). Mean density for partition with the smallest core cluster (C_{456}) is 0.229 with standard deviation 0.023 and values are in range from 0.19 to 0.27 (Figure 6.9(a)).

The mean density of regular blocks is the highest with C_{456} partition and is equal to 0.393 (sd=0.041). The densities of regular blocks for C_{555} actors partition are in range from 0.30 and 0.39 with mean value 0.333 (sd=0.027). The mean density of regular blocks for C_{654} partition is 0.337 with standard deviation 0.034 and the densities are in range from 0.30 and 0.42 (Figure 6.9(b)).



(a) Network density

(b) Density of regular blocks

Figure 6.9: Boxplots of density and density of regular blocks for the regular core-periphery three-cluster models

7 Evaluation of stability of blockmodeling on design errors

In this chapter the results of simulation studies for estimation 92of the impact of different types of design errors on the blockmodeling are presented. We try to answer our second research question (presented on page 39) *how sensitive is blockmodeling procedure to different types and amounts of errors*.

The stability of blockmodeling solution is evaluated with two indices: the Adjusted Rand index (presented in Section 5.1.1) is used for determining the concordance between two partitions and the proportion of incorrect block types (presented in Section 5.1.2) compares type and position of blocks in two image matrices.

First, the evaluation of blockmodeling stability for errors due to fixed choice instead of free choice design is presented in Section 7.1. The impact of direction of question on the obtained blockmodel in case of two real networks is presented in Section 7.2. The actor non-response and the stability of blockmodeling is extensively presented in Section 7.3. In the last section (Section 7.5) random measurement errors where ties are randomly added or deleted are presented. The amount of changed ties is controlled and their impact on blockmodeling solution is investigated.

7.1 Errors introduced by fixed choice design

The stability of blockmodeling in case of fixed choice design compared to free choice design is presented in this section. The design of simulation study is presented in

Section 7.1.1, while the results are presented in Sections 7.1.2.1 and 7.1.2.2.

7.1.1 The design of simulation studies for fixed choice design

The basic scheme of simulation study is presented on page 70 in Section 6.1. Here we described more detailed construction of errors or measured networks due to fixed choice design from item 3(a) in the basic scheme.

The whole networks are collected with free choice designs. Construction of measured network with fixed number of choices is made by selecting some numbers of limitations of choices - nFixed. The ties were randomly added or deleted (or both) to meet the limitation criteria.

For each network and selection of number of fixed choices (*nFixed*) the blockmodeling procedure was performed 100 times (nGen=100). The established blockmodel from measured network was compared to the blockmodel of real whole network as described in the basic scheme of simulations. Results of simulation studies are presented in next sections.

7.1.2 Results of simulation study of fixed choice design for real networks

The simulation study was performed with two real networks, the boy-girl liking ties network and the note borrowing network.

7.1.2.1 A boy-girl liking ties network

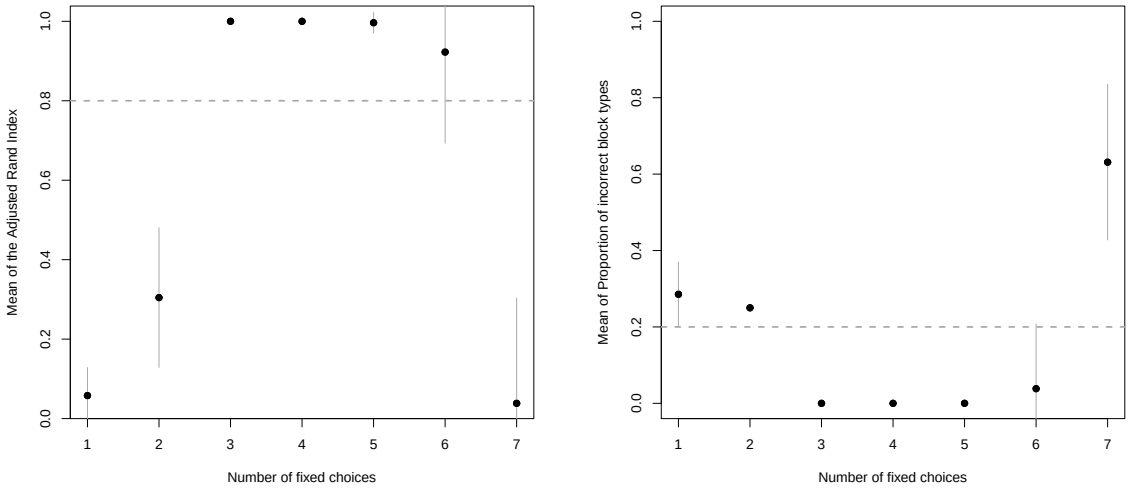
In the collection process of the boy-girl liking ties network data (Section 6.2.1.1) the free choice design was used without limitation of number of actors. The network is presented in Figure 6.1 on page 72 where it can be seen that actors selected from two to four friends.

In the simulation study for each selected number of desired fixed choices hundred networks were generated. The 'measured' network was generated with random addition or deletion of ties, so that the condition concerning the limitation of ties was satisfied. Blockmodels of measured networks were established based on structural equivalence and compared with structure of real network shown in Figure 6.1 with both indices of blockmodeling stability, the Adjusted Rand Index and the proportion of incorrect blocks.

The average number of choices made (average outdegree) in the boy-girl liking ties network is 3.45 (with standard deviation 0.82). Two actors (G3 and G5) made only two choices, three nominations were made by actors B4 and G4, other seven actors selected four other members of a network. Based on those results we may suspect that restriction of number of choices to 3 or 4 actors will not radically change the blockmodel structure of the measured network, because a small proportion of ties is changed.

The fixed choice design was simulated with a range of restriction for nominations from one to seven actors. Stabilitiy of partitions of actors was measured with the Adjusted Rand Index (ARI) and is presented in Figure 7.1(a). Mean values of ARI are presented with black dots and standard deviation is presented with gray error bars. According to simulations of Steinley (2004) we would say that agreement between partition of real and measured network is acceptable if the mean value of ARI is above 0.8 (Section 5.1.1).

As expected, the blockmodeling is stable in terms of partition if the number of choices is set to three or four nominations. In fact, the agreement between partitions is perfect, because mean value of ARI is equal to one (each of hundred measured partitions is the same as the real partition). The agreement between partitions is also almost perfect if the number of fixed choices is equal to five. Acceptable agreement between partition ($mARI > 0.8$) is obtained with random simulation of six choices, while the increase of number of fixed choices to seven choices leads to no agreement between partitions with $mARI$ around zero. If the number of fixed choices is low (one or two nomina-



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.1: Results of the simulation study with the boy-girl liking ties network for simulated fixed choice design

tions), then the ties are deleted from the real boy-girl liking ties network in order to satisfiy the restrictions. Therefore, the measured structure is poorer than the real structure of tie patterns and the agreement between partitions is unacceptable.

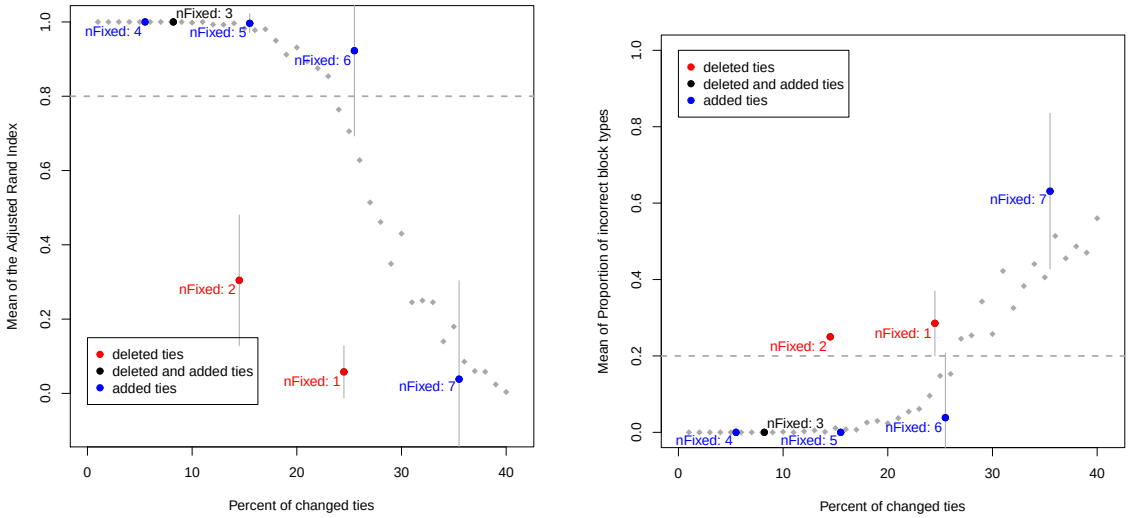
Figure 7.1(b) presents stability of blockmodeling in terms of correctly identified block types. We say that agreement between two blockmodels (or image matrices) is acceptable if the mean values of proportion of incorrectly identified block types ($mErrB$) do not exceed 0.2 (Section 5.1.2). Similar as in comparison of partition, the perfect agreement between image matrices is obtained if fixed choice nominations are restricted to three, four or five choices ($mErrB$=0). Acceptable agreement between block types is also obtained with six choices. On the other hand, small number (one or two) or high number (seven nominations) of fixed choices leads to unacceptable agreement between blockmodels with $mErrB$ higher than 0.2. If the number of choices is restricted to one or two nominations, the proportion of incorrectly identified block types is around 0.25, which indicates that one block in a blockmodel is incorrectly classified.

As noted above, if number of choices is restricted to one or two nominations, then the

ties are deleted from the real boy-girl liking ties network. Actors in the network nominate from two to four friends, therefore the restriction of choices to three nominations will lead to both deletion and addition of ties. The ties are randomly added to the network, if the restriction of choices is set to four or more nominations. In that case, the observed real structure is the subset of measured networks and actors have to add non-existing ties to meet the criteria. Therefore, we try to present the dependence between the percent of ties changed in different restriction rules compared to the whole boy-girl liking ties network and both indices of network stability.

Figure 7.2(a) presents mean values of ARI plotted against the percent of changed ties. Gray points indicate mean values of ARI in dependence to the percent of randomly changed ties in the boy-girl liking ties network. The extensive results of randomly introduced measurement errors are presented in Section 7.5.2.1. With presentation of both error mechanisms in social network study designs on one figure we try to compare both types of errors. If in the restriction of nominations ties were randomly deleted the results are plotted in red, if ties were just randomly added the results are marked with blue, and black color indicates those results where ties were both randomly added and deleted to meet the specified criteria about the number of nominations.

The smallest percent of changed ties (6%) is obtained if the number of nominations is restricted to four choices. As seen above, there is a perfect agreement between measured and real partition and ties were only added to the measured network. With five required nominations 16% of ties were changed (more precisely, ties were added) in the measured network and agreement between partitions is still perfect. The result is similar if 16% of ties are randomly changed (one, a tie, is replaced by zero and and vice versa). More than quarter (26%) of ties were added to the network if the number of fixed choices was equal to six, but the agreement between partitions was still acceptable with $mARI$ values around 0.9. Compared to the same percentage of randomly changed ties, the results for fixed choice mechanism are better. Interesting results are obtained if we compare restrictions to two and five nominations, where 14



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.2: The ercent of changed ties in the simulation study with the boy-girl liking ties network for simulated fixed choice design

and 16 percent of ties were changed, respectively. In the first case (2 fixed choices) ties were deleted from the whole network and the result for $mARI$ is overwhelming ($mARI \approx 0.3$). In the second case (five fixed choices) higher percent of ties was changed, but ties were added to the network. As noted above, the agreement between both partitions in that case is acceptable. Therefore we may conclude that addition of ties in a fixed choice design is less destroyable than deletion of ties. In other words, if the study design requires fixed choice, the restriction of nominations should not be set too low.

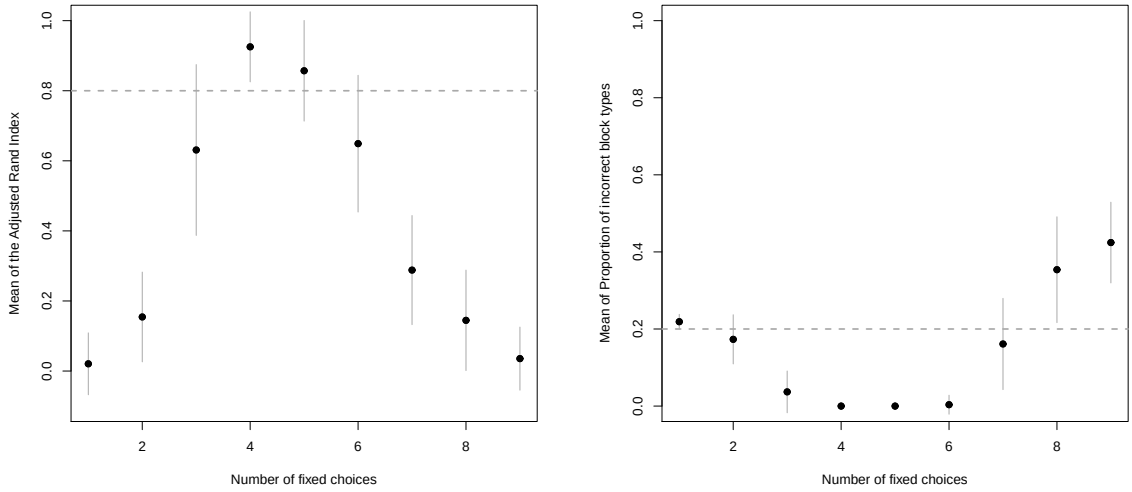
The results for the percent of incorrectly identified block types plotted against the percent of changed ties are presented in Figure 7.2(b). Similar as for $mARI$, the deletion of ties in case when the number of choices is limited to one or two nominations destroys the blockmodel structure, and values of $mErrB$ are higher compared to the same percent of randomly changed ties.

7.1.2.2 The student note borrowing network

The student note borrowing network has 15 actors. The blockmodel based on structural equivalence into three clusters is presented in Figure 6.2 in page 72.

The average outdegree (the average number of nominations) in the note borrowing network is 3.73 (with standard deviation 0.88). One actor nominated one friend, five actors nominated three other members, six actors made four nominations and three actors nominated five other members of a network.

The limitation of number of choices was simulated with a range of restriction from one to nine actors, where a whole note borrowing network has 15 actors. Figure 7.3(a) presents the results for stability of blockmodeling based on structural equivalence in terms of partitions. The mean values of ARI are acceptable for restriction set to four and five nominations, because $mARI$ values are above 0.8. For fixed choice equal six or higher, the $mARI$ decline to zero, which indicates unacceptable agreement between partitions. Mean values of ARI are below 0.8 also for range of limitations from one to three fixed choices.



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

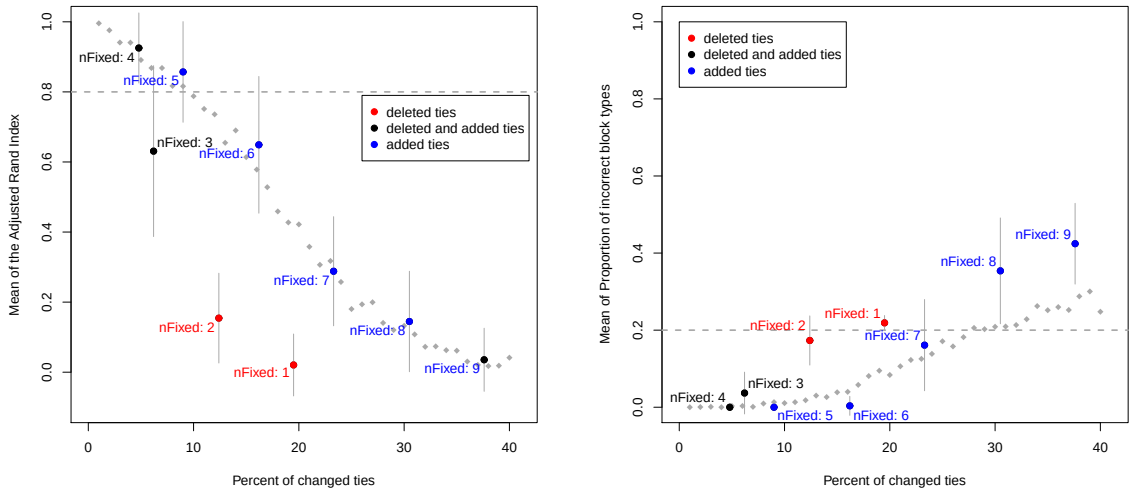
Figure 7.3: Results of the simulation study with the note borrowing network for simulated fixed choice design

Figure 7.3(b) presents the stability of blockmodeling for the note borrowing network in terms of correctly identified block types ($ErrB$). Compared to results for stability of partitions, the stability of blockmodel structure in terms of correctly identified and placed block types is higher. Mean values of $ErrB$ are below 0.2, which indicates less then two incorrectly identified block types in a blockmodel, for whole range of fixed choices from two to seven. The blockmodel or image matrix of measured network compared to whole network is unacceptable if the restriction of number of nominations is set to one or higher than seven.

As noted above, actors made between two and five nominations. Therefore, if the number of choices is restricted to one or two nominations, then ties are deleted from the note borrowing network. If the number of choices is set to three or four, then ties are both added and deleted to satisfy the condition. Ties are randomly added to the network, if the restriction of choices is set to five or more nominated actors.

Figure 7.4(a) presents mean values of ARI plotted against the percent of changed ties. The gray points indicate mean values of ARI when the random measurement errors are introduced to the network (detailed results are presented in Section 7.5.2.2).
The smallest percent of changed ties (5%) is obtained if the number of nominations is restricted to four choices and the agreement between partitions is acceptable. If the choices are limited to three actors, majority of changed ties were deleted tie and only one tie was added, because one actor has just two nominations. In that case 6% of ties were changed and the mean value of ARI indicates unacceptable agreement between both partitions. In case when number of choices was limited to five actors, ties were only added to the network and the agreement between partitions according to $mARI$ is acceptable. In this case, we changed 9% to meet the limitation criteria and the $mARI$ values are higher than in case of three choices restriction. Similar as in boy-girl liking ties network, we may conclude that addition of ties has less destroyable effect than deletion of ties.

The less severe effect of adition than deletion of ties in a fixed choice design can also



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.4: Percent of changed ties in the simulation study with the note borrowing network for simulated fixed choice design

be seen in results for stability of blockmodel in terms of block types (Figure 7.1.2.2). Mean values of $ErrB$ are in range from three to seven, where ties were both added and deleted or just added, fixed choices below 0.2 and have similar pattern as corresponding percent of randomly introduced measurement errors. In case when ties were only deleted to meet the limitation criteria, the $mErrB$ values are noticeably higher (around 0.2) than comparable fixed choice cases (according to the number of changed ties) where ties were added.

7.1.3 Conclusions

According to presented results, we may conclude that limitation of number of choices may destroy the blockmodel structure if the restriction is unrealistic or too far from the true number of desired nominations. Newman (2010, 41) emphasized that "limits are often imposed purely for practical purposes, to reduce the work the experimenter must do". We would like to emphasize that this is not the right reason for selection of fixed choice questionnaire format which has high ability to destroy the underlying true structure and estimates of network statistics (Holland and Leinhardt, 1973; Kossinets, 2006), therefore, several authors warn against its use.

Also from blockmodeling procedure point of view, the questionnaire format in social network studies should not enforce the fixed number of choices. If there is a reasonable argumentation for use of fixed choice design, the limitations should not be set too strict. For example, for establishment of blockmodel it is better that questionnaire format forces the respondents to nominate more friends (than is the real number) than make them impossible to list all their friends.

7.2 Errors caused by direction of questions

The stability of blockmodeling to errors obtained due to direction of questions is presented with two examples of real networks. The first pair of networks is the Student Government recognition networks and the second pair of networks is from extended study of social support dimensions and quality of measurement instrument for social network data (Ferligoj and Hlebec, 1998).

7.2.1 The Student Government recognition networks

Hlebec (1992) collected six networks with three different questions and two methods. All networks are presented on page 73 in Section 6.2.2.1. We try to represent connectedness or agreement between two blockmodels obtained from networks with original and reversed question. The wording of questions was (Hlebec, 1992, 1993):

1. Which members and advisors of the Student Government do you (most often) ask for an opinion?

2. Which of the members and advisors of the Student Government (most often) ask you for an opinion?

We used networks collected with recognition method (Table 6.1 on page 75). The second network, 'ask you for an opinion', was transposed before the analysis and therefore the notation 'being asked for an opinion' will be used.

The blockmodels of the selected networks have not been, according to our knowledge, published. The other network from the collection, the recall discussion network, was presented in (Doreian et al., 2005) in terms of generalized equivalence. Therefore the natural decision would be to examine the selected networks in terms of generalized equivalence, but we decided, due to high instability of blockmodeling according to regular and generalized equivalence to random measurement errors (presented in Section 7.5), to establish blockmodels based on structural equivalence and compare them.

If the ties of both networks are compared, it can be seen that there are 95 agreements about existence of ties (there is a tie in both networks or there is no tie in both networks). 14 ties existing in the 'original' network were not measured in the 'reversed' network and vice versa, 20 ties were measured in the 'reversed' network which do not exist in the 'original' one. The proportion of changed ties in both networks is quite high and equals 0.31.

First, the network for original question ('asking for an opinion') is presented in Figure 7.5. The colors show three clusters based on structural equivalence. In the middle panel the sociomatrix is presented. The colors on the diagonal (red, blue and yellow) represent cluster membership (and not ties). The image matrix shows non-symmetric structure with three complete blocks. The presented partition, which is the best fitting partition according to structural equivalence, has 21 inconsistencies.

In the next step, the network for the reversed question 'being asked for an opinion' was examined in terms of structural equivalence. There were two equally well fitting partitions into three clusters with 27 inconsistenncies (Figure 7.6). The first partition is presented in Figure 7.6(a). The first cluster consist of actors $\{m_1, pm, m_2, m_4, a_2\}$, in the second cluster there are $\{m_3, m_6, m_7\}$ and in the third cluster there are actors $\{m_5, a_1, a_3\}$. The colors on the diagonal represent clusters obtained with 'original' question network presented in Figure 7.5. The cluster membership is quite mixed. The first cluster of 'reversed' blockmodel consists of actors from all clusters from 'original' question blockmodel, and the third cluster preserves the membership but it is reduced. Parti-

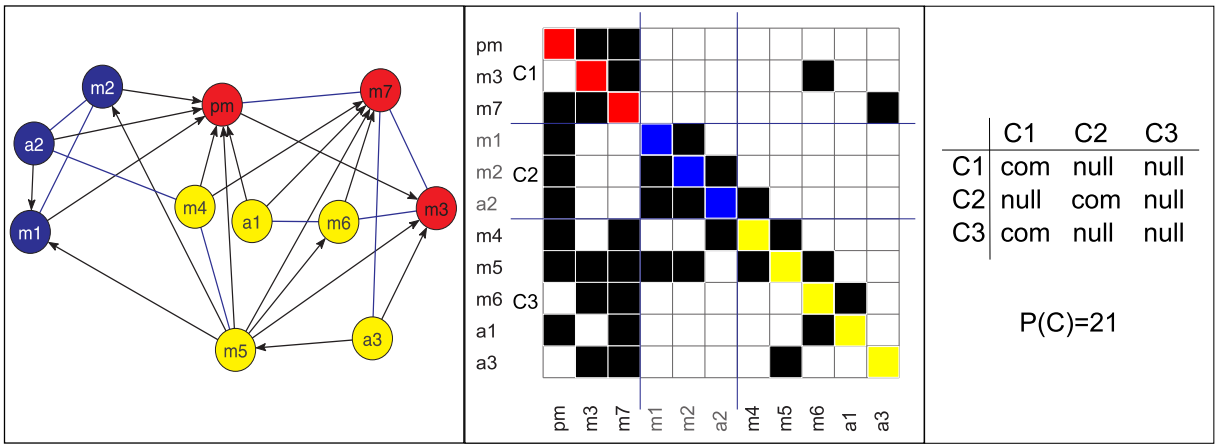
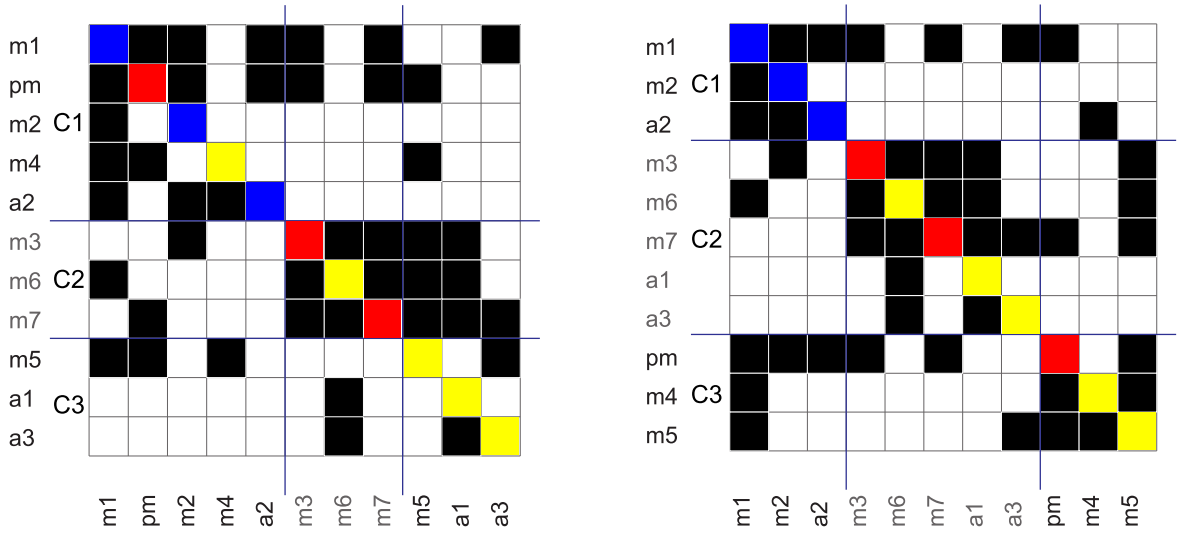


Figure 7.5: The Student Government recognition 'asking for an opinion' network (left), three partitions based on structural equivalence (middle) and image matrix (right)

tions of 'original' question and 'reversed' question blockmodel were compared with the Adjusted Rand Index. The calculated value is 0.21, which indicated poor agreement (Steinley, 2004). According to the image matrix I_1 (Equation (7.1)), the proportion of the differently identified block types in both blockmodels is 0.22 (2 different blocks).



(a) 1st equally well fitting partition

(b) 2nd equally well fitting partition

Figure 7.6: Blockmodels for the Student Government recognition 'being asked for an opinion' network compared to blockmodel from 'asking for an opinion' network

$$I_1 = \begin{bmatrix} \text{com} & \text{null} & \text{null} \\ \text{null} & \text{com} & \text{com} \\ \text{null} & \text{null} & \text{null} \end{bmatrix} \qquad I_2 = \begin{bmatrix} \text{com} & \text{null} & \text{null} \\ \text{null} & \text{com} & \text{null} \\ \text{com} & \text{null} & \text{com} \end{bmatrix} \tag{7.1}$$

The second equally well fitting partition is presented in Figure 7.6(b). The first cluster is the same as in 'original' question blockmodel, and other actors in the second and third cluster are mixed compared to clusters presented in Figure 7.5. The Adjusted Rand Index between partitions of interest is 0.29, which is higher as for the first equally well fitting partition, but still indicates poor concordance. If the blockmodels are compared, the 'reversed' question blockmodel has an additional complete block on the diagonal (I_2 in Equation 7.1). Therefore, the proportion of incorrectly identified block types is equal to 0.11.

7.2.2 Networks of emotional support

The selected networks of emotional support are part of a large study of quality of measurement instrument for social network data (Ferligoj and Hlebec, 1998, 1999; Hlebec, 1999, 2001; Hlebec and Ferligoj, 2001; Zemljič and Hlebec, 2005). Data were collected in eight third grade classes consisting of 30 students on average. They measured four dimensions of social support: instrumental (exchange of study material), informational (informations about important study assignments) and emotional support (discussing important things) and social companionship (invitations to a birthday party). In research design four scales were used: binary, five-point ordinal scale with or without labels and line-production (or line-drawing) scale. Both giving and receiving of social support were measured with so called 'original' and 'reversed' questions. There were also two data collection techniques used, recognition and free recall of actors.

We selected two networks from the first class measuring giving and receiving (original and reversed question) of emotional support with binary scale, without limitation of actors and with recall method. The emotional support was selected, because it was the most stable. It has the lowest proportion of changed ties in the network collected with original question compared to the network collected with reversed question and it is

equal to (0.22). To compare, the instrumental (or material) and informational support (in the same first class) have 0.33 of changed ties and the highest proportion of changed ties between measuring social support dimensions with original and reversed question has companionship dimension (0.38). These results are consistent with findings of Hlebec and Ferligoj (2002, 299-300) that the exchange of study material is the least reliable measure (instrumental support) and that informational and emotional support are the most reliable. They conclude that measurements of social support provided by strong ties are more reliable than social support provided by weak ties.

The exact wording of questions (Hlebec, 2001, 138-139):

1. With which of your classmates would you discuss important things?
2. Which of your classmates would discuss important personal matters with you?[16]

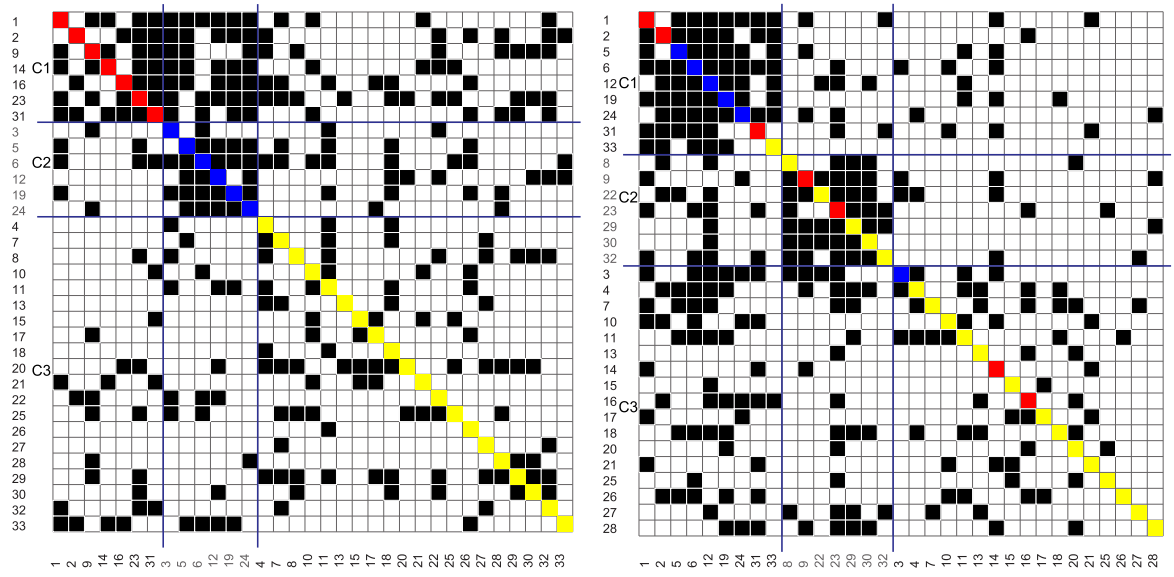
Figure 7.7(a) presents blockmodel into three clusters based on structural equivalence for the 'original' network of emotional support. The network consist of 281 arcs where actors choose on average 8.5 classmates (with standard deviation 4.9). The obtained blockmodel has 231 inconsistencies and three clusters consist of 7, 6 and 20 actors.

The 'reversed' network has 278 arcs and the average outdegree (average number of nominations) is 8.4 (with standard deviation 3.1). The blockmodel of 'reversed' network into three clusters has 210 inconsistencies (Figure 7.7).

Colors on the diagonal (red, blue, and yellow) of the 'reversed' network present the cluster membership of an actor in the 'original' network. Actors from the second blue cluster from the 'original' network remain together (except actor 2) and are joined together with four other actors. The largest yellow group from the 'original' network is mixed with the red group in the 'reversed' network. The Adjusted Rand Index of both partitions is 0.29 which indicates poor agreement between both partitions.

The image matrix of the 'original' network (I_3 in Equation (7.2)) has three complete blocks; within clusters $C1$ and $C2$, and between both mentioned clusters. The image

[16]The obtained network was transposed before the blockmodeling procedure.



(a) 'Original' network (b) 'Reversed' network

Figure 7.7: Blockmodels into three clusters based on structural equivalence for the emotional support 'reversed' network compared to blockmodel from 'original' network

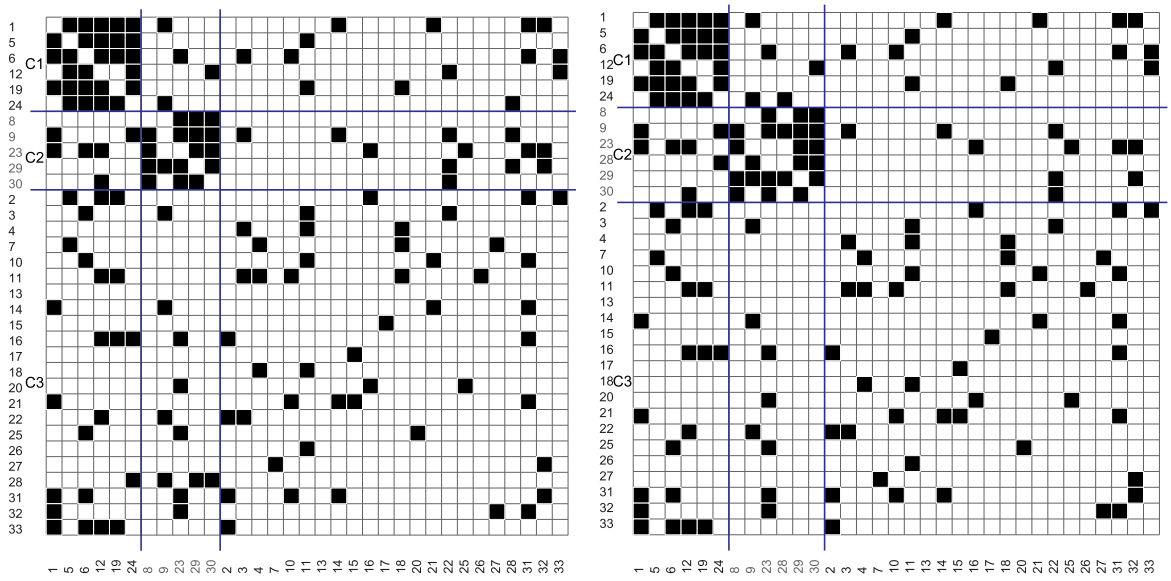
matrix of the 'reversed' network (I_4 in Equation (7.2)) has only two complete blocks on the diagonal, therefore the proportion of incorrectly identified block types is equal to 0.11 which indicates acceptable agreement between image matrices.

$$I_3 = \begin{bmatrix} \text{com} & \text{com} & \text{null} \\ \text{null} & \text{com} & \text{null} \\ \text{null} & \text{null} & \text{null} \end{bmatrix} \qquad I_4 = \begin{bmatrix} \text{com} & \text{null} & \text{null} \\ \text{null} & \text{com} & \text{null} \\ \text{null} & \text{null} & \text{null} \end{bmatrix} \tag{7.2}$$

Another approach of analyzing the 'original' and 'reversed' network would be a combined network. A tie in the 'combined' network exists if a corresponding tie is present in both networks. The new network is in fact intersection of both networks and we will denote it as the 'confirmed' network. The 'confirmed' network has 160 arcs which means that 57% of ties in the 'original' network are confirmed with a tie in the 'reversed' network. The mean outdegree is 4.8 with standard deviation 2.7. The density of the 'original' network is 0.27 and for the 'reversed' network 0.26, while the density of the 'confirmed' network is lower and is equal to 0.15.

The blockmodel of the 'confirmed' network into three clusters based on structural

equivalence was established. There were 122 inconsistencies and two equally well fitting partitions (Figure 7.8). The first cluster is the same in both partitions (actors 1, 5, 6,12, 19, 24), the second cluster in the first partition (Figure 7.8(a)) has 5 actors (actors 8, 9, 23, 29, 30), while the second cluster of the second equally well fitting partition (Figure 7.8(b)) has one actor more (actor 28). Therefore, the agreement between both partitions is high and that was also confirmed with the value of the Adjusted Rand Index which is equal to 0.90. The image matrix of both equally well fitting partitions is the same and has two complete blocks in the diagonal which represent two cohesive subgroups.



(a) 1st equally well fitting partition (b) 2nd equally well fitting partition

Figure 7.8: Blockmodels into three-clusters based on structural equivalence for the 'confirmed' emotional support network

The obtained partitions of the 'confirmed' blockmodel (Figure 7.8) were compared to the partitions of 'original' and 'reversed' blockmodel (Figure 7.7). The agreement between partitions is poor. The Adjusted Rand Index between partition from 'original' blockmodel and both equally well fitting partitions from 'confirmed' blockmodel is 0.35 and 0.31, respectively. Comparison of both 'confirmed' partitions with the partition from 'reversed' blockmodel results in the ARI values equal to 0.47 and 0.41. On the other hand, the image matrix of the 'confirmed' blockmodel is equal to the image

matix I_4 of the 'reversed' blockmodel.

The above example shows that although all three networks, 'original', 'reversed' and 'confirmed', present the relation of giving an emotional support between the same actors, the simplified blockmodel structure is not the same. We have to be aware that this method of gathering the network data, including the direction of question, has a great impact on position membership of actors. The 'confirmed' network could be used to find the most cohesive and stable subgroups of the network.

7.2.3 Conclusions

Although we only presented results on two real networks, we may conclude that direction of question has a great impact on the established blockmodel structure. Both results of the blockmodeling procedure, the position membership and the image matrix, depend on the method used for gathering social network data. Therefore further research should establish if there is a common pattern in the blockmodels obtained with different question in data collection process. The confirmation of ties from the 'original' network from the ties from the 'reversed' one, could probably be used to find the most dense, stable and cohesive subgroups of a network.

7.3 Errors caused by actor non-response

In this section the stability of blockmodeling to actor non-response is presented. The design of a research study with more detailed scheme of simulations is presented in Section 7.3.1, while the results are presented in Sections 7.3.2 and 7.3.3. Based on the obtained results recommendations about the best actor non-response treatment are presented together with an answer to the second thesis presented on page 55.

7.3.1 The design of simulation studies for actor non-response

In order to investigate the vulnerability of blockmodels to different numbers of non-responding actors, along with various ways of treating such missing data, we performed a simulation study of actor non-response where *all* outgoing ties of at least one

actor are missing. We use the following notation: a *whole network* that is known[17]; a *measured network* which is obtained from the whole network by removing all outgoing ties for some actors; and a *treated* network that is obtained by treating a measured network to deal with the introduced non-responses.

Section 7.3.1.1 describes the overall design of the simulations; Section 7.3.1.2 outlines three types of introduced non-response missing data and Section 7.3.1.3 presents five ways of treating the introduced missing data due to actor non-response. Two types of whole networks were included n simulation studies; real networks described in Section 6.2.1 and simulated networks based on structural equivalence in Section 6.2.3.

7.3.1.1 A scheme for simulations for actor non-response

The basic scheme of simulation study from Section 6.1 is supplemented with three different mechanisms for selection of non-respondents (Section 7.3.1.2) and five treatments of missing data presented in (Section 7.3.1.3).

The scheme of simulation study for actor non-response:

1. **Select** a network from the literature or
   generate a whole network under a known starting model.

2. **Establishing a blockmodel of the whole (real) network** that has two parts:
   (i) the known (real) partition of the actors of the whole network into positions; and
   (ii) the image matrix with the known distribution of block types by location.

3. Let *nGen* denote the number of simulations for a given combination of network type, introduced non-response mechanism, number of non-respondents, and missing data treatment regime.
   For i=1:nGen, do the following:

   (a) **Construct the network with non-responses** (the measured networks) by selecting some **proportion of actors** to become non-respondents and deleting

[17] It is a starting network in simulations.

their outgoing ties based on **selected missing mechanism** (MCAR, based on outdegree or based on indegree) described in Section 7.3.1.2.

(b) **Treat the measured network** to substitute for the missing data according to a selected missing data treatment (discussed in Section 4.3.1.1);

(c) **Establish a blockmodel of the measured and treated network** that also has two parts:
(i) the partition of the actors of the measured and treated network into positions; and
(ii) the blockmodel image of the measured and treated network.

(d) **Compare the resulting blockmodels of the whole and the treated networks** using:
(i) the Adjusted Rand Index to compare the two sets of positions; and
(ii) the proportion of incorrect blocks (as described in Section 5.1.1).

4. **Investigate the impact of actor non-response** in terms of the mean of the values of ARI - denoted as $mARI$ - and the mean of the proportion of incorrect blocks - denoted by $mErrB$.

7.3.1.2 Generating non-response missing data

Three different actor non-response mechanisms (or regimes for generating non-response missing data) were used. Each regime defines the probabilities that actors become a non-respondent. One is that these actors are selected at random, the second is that the probability of their selection depends on their outdegree, and the third is that it depends on their indegree. More precisely, the three options are:

(i) Actors are selected at random to become non-respondents.

(ii) The probability of actors becoming non-respondents is proportional to $1/(outdegree + 1)^2$.

(iii) The probability of actors becoming non-respondents is proportional to $1/(indegree + 1)^2$.

Rubin (1976) defined and labeled three types of missing mechanisms: (i) MCAR (missing completely at random), (ii) MAR (missing at random) and (iii) MNAR (missing not at random). Our first non-response mechanism is MCAR, because non-response is unrelated to the network or actor characteristics. Huisman and Steglich (2008, 302) argue that this model for missing data "may be realistic when there is no reason to assume that actors differ in their propensity to fill in network questionnaire".

Our second and third non-response mechanisms depend on the network and the implied network based actor characteristics. Having the non-response probability related to an unknown number of unreported ties implies that data are missing in NMAR way. Mechanisms based on actor degrees where having lower outdegree and lower indegree values implies having higher probabilities of being non-respondents were used. Huisman and Steglich (2008) established that both mechanisms reflect the characteristics of real world networks where popular actors (those with higher indegree) are more willing to participate and are easier to reach than inactive actors with low outdegree. Similarly, Costenbader and Valente (2003) ascertained that actors refusing to participate in surveys, or are missed, are actors more likely to come from the periphery of their network. In critical analysis of their study Borgatti et al. (2006) emphasized that one of the limitations was only randomly introduced errors (e.g. node deletion, edge deletion...) and that in practical studies data collection methods make systematic errors where low degree nodes are more readily lost.

Network data are missing at random (MAR) if the missingness or non-response depends on the observed data (usually some kind of actor's characteristics), but not on missing ties; e.g. Huisman and Steglich (2008) used data about alcohol consumption as additional covariate. This type of non-response is not included in our simulations.

The number of non-respondents for the boy-girl liking ties network and simulated whole networks based on completely symmetric blockmodel structure described in Section 6.2.3.1 ranges from 1 to 5 (with the proportion of non-response taking the values 0.1, 0.2, 0.3, 0.4 and 0.5). For the note borrowing network and the simulated net-

works with three positions (two examples of non-symmetric blockmodel structures described in sections 6.2.3.2 and 6.2.3.3), the number of non-respondents ranges from 1 to 6 (with proportion of non-response taking the values 0.07, 0.13, 0.20, 0.27, 0.33, 0.40).

7.3.1.3 Treatments of missing non-response data

We treated the missing non-response data in five ways, which are described in detail in Section 4.3.1.1. The first is *the complete-case approach* where, in addition to excluding the non-respondents, all incoming ties to them are also removed. The second is *the null tie imputation approach* that keeps the non-respondents, but assigns the value 0 to each of their outgoing ties. Using *reconstruction* is the third approach where the missing r_{ij} tie is replaced by the observed r_{ji} tie, and for ties between two non-respondents a 0 is imputed. In the fourth approach, *the imputations based on the mode*, the set of the incoming ties to actor j is summarized and the modal value is then used to assign values for the r_{ij} tie. Finally, we use *the reconstruction plus imputations based on a mode (for ties between two nonrespondents)*, which is a combination of the third and fourth approaches.

The efficiency of different missing data treatments is examined with real and simulated networks, and obtained results are presented in the following sections.

7.3.2 Results of simulation study of actor non-response for real networks

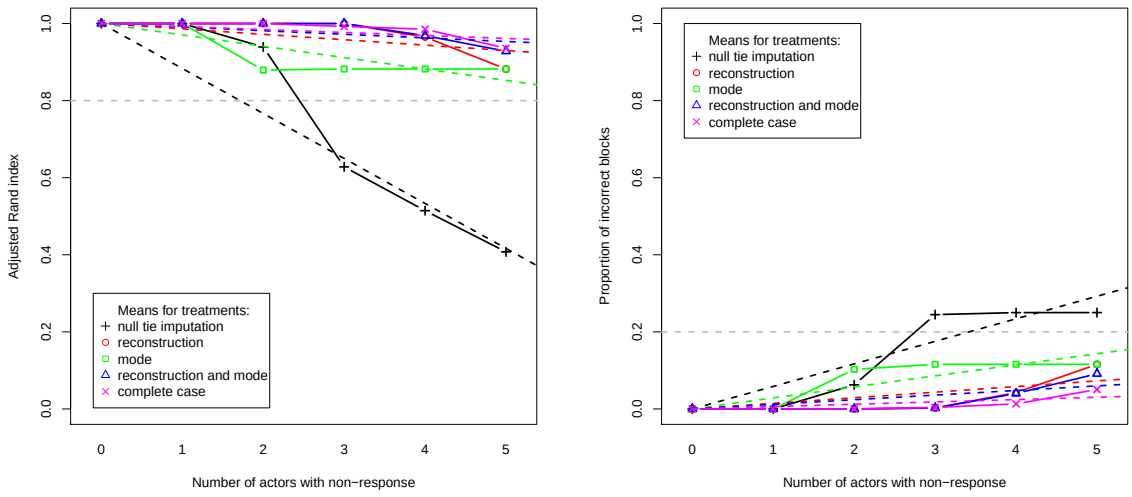
7.3.2.1 A boy-girl liking ties network

In the boy-girl liking ties network (presented in Section 6.2.1.1), non-respondents were selected based on the three missing data mechanisms described in Section 7.3.1.2. Five different treatments of non-response data (described in Sections 4.3.1.1 and 7.3.1.3) were used, and for every new measured and treated network a blockmodel was established and compared with the structure shown in Figure 6.1 on page 72. The resulting factorial design has 75 cells (for the combinations of three missing mechanisms, five treatments of non-response, and five numbers of actors with non-response). Within each cell, the generation of incomplete data was repeated 10 times for networks with

one missing actor, 50 times for combinations of two missing actors and 100 times[18] for combinations of three or more non-respondents. Obviously, the number of generated missing data increases with higher proportions of non-respondents.

Data missing completely at random

Figure 7.9 presents the results when non-responding actors were selected at random. The mean values of ARI are plotted in Figure 7.9(a) and the $mErrB$ values are in Figure 7.9(b). The results are unequivocal when there is only one non-responding actor. For all treatments of non-response missing data, there is perfect agreement with the whole network blockmodel: $mARI = 1$ for all treatments indicating perfect agreement, and $mErrB = 0$ so that all block types are correctly identified and placed. Differences between the ways of treating missing data start to appear when there are at least two non-respondents.



(a) Adjusted Rand Index - ARI (b) Incorrect block types - $ErrB$

Figure 7.9: Results of the simulation study based on the boy-girl liking ties network for data missing completely at random (solid lines) and predictions according to linear regression model (dash lines)

[18]The number of all possible combinations of actors with nonresponse increases. For example, for a network where n = 11 there are: $\binom{11}{1} = 11$ possibilities for selecting one non-respondent, $\binom{11}{2} = 55$ possibilities for selecting two non-respondents; $\binom{11}{3} = 165$ and so on.

The results following the use of null tie imputations and imputations based on mode are the worst. With three non-respondents, $mARI$ for null tie imputations drops below 0.65 indicating poor agreement between the cluster memberships for the whole network and those of the measured and treated network. This performance gets much worse for four and five non-respondents. According to the guidelines of (Steinley, 2004), all of the other missing data treatments perform quite well. Of these four methods, the treatment using modes affected the partitions the most, although $ARI > 0.8$ and the blockmodel is stable for two or more non-respondents ($mErrB \leq 0.2$). The blockmodels for networks treated with the complete-case, reconstruction, and reconstruction coupled to using a mode for ties between non-respondents all lead to excellent agreements with the blockmodel for the whole network. If anything, the complete-case approach fares the best. This ordering also holds true for the identification and location of block types as shown on the right in Figure 7.9(b).

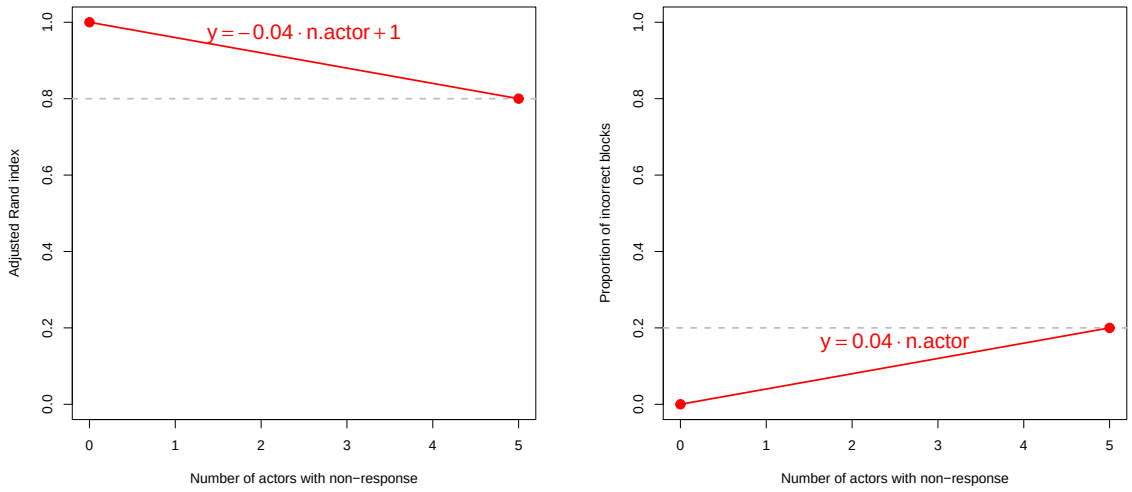
Dash lines in Figures 7.9 present predictions of both indices of blockmodeling stability (ARI and $ErrB$) according to simple linear regression models for each combination of missing data mechanism and non-response data treatment. In case of Adjusted Rand index (Figure 7.9(a)) regression lines are forced through point $(0,\ 1)$, because with zero non-respondents the agreement between partitions is perfect and therefore value of ARI is equal to 1. Therefore linear regression model for ARI is equal to

$$Y_{ARI} = \beta \cdot n.actor + 1\ . \tag{7.3}$$

All dash lines, except for the null tie imputation treatment fit well to the observed data for ARI. The lowest change in values of ARI (Table A.1), when number of non-respondents increases, is obtained with the complete-case approach ($\beta = -0.008$) and with the combination of reconstruction plus imputations based on mode ($\beta = -0.009$). The change in number of non-respondents for one decreases the Adjusted Rand Index for 0.014 when the reconstruction treatment is used, and for 0.030 in the case of imputations based on mode.

The slope coefficients are tested with t-test if they are equal to β_0, instead of usually adopted testing value 0. β_0 is calculated based on the selected criterion for stable block-

modeling according to partition membership (ARI) and number of non-respondents introduced in simulations. As a reminder, according to the simulations of Steinley (2004) we say that blockmodeling is stable in terms of partitions if the mean values of ARI are above 0.8 (presented on page 64 in Section 5.1.1). The boy-girl liking ties network (presented in Figure 6.1) has 11 actors and the maximal number of non-respondents in the simulations was set up to 5 actors (which is equal to 45% of actors in the network). Combining both presented facts we decided to compare the obtained predictions from linear regression models (presented above) with the slope of the line through points $(0,\ 1)$, and $(5,\ 0.8)$. Point $(0,\ 1)$ indicates perfect agreement between two partitions in network without non-respondents and point $(5,\ 0.8)$ indicates acceptable agreement in terms of ARI with five introduced non-respondents. The line is presented in Figure 7.10(a) and its slope coefficient is equal to $\beta_0^{ARI} = \frac{0.2}{-5} = -0.04$.



(a) Adjusted Rand Index - ARI

(b) Incorrect block types - $mErrB$

Figure 7.10: Schematic representation of lines which were used for comparison of linear regression models in smaller networks for both indices of blockmodeling stability

Therefore the slopes of linear regression models (presented with dash lines in Figure 7.9(a)) are tested with one-sided t-test where null and alternative hypotheses are as follows:

$H_0 : \beta_0^{ARI} \geq -0.04$

$H_1 : \beta_0^{ARI} < -0.04$

t-statistic was calculated as

$$t = \frac{\hat{\beta} - \beta_0}{se_{\hat{\beta}}} \tag{7.4}$$

and it follows the t-distribution with $n - 2$ degrees of freedom where n is the number of observations, $\hat{\beta}$ is an estimate of slope coefficient from Equation (7.3), $se_{\hat{\beta}}$ is a standard error of the estimated coefficient and $\beta_0 = \beta_0^{ARI} = -0.04$.

Results are presented in Table A.1 in Appendix A. The prediction for the null tie imputation treatment has statistically significant lower slope than -0.04 (t=-27.009), which confirms that this treatment is the worst and unacceptable for higher numbers of non-respondents. Other four treatments have $mARI$ above 0.8 for whole range of introduced non-respondents (solid lines in Figure 7.9(a)) and also the slope coefficients of predictions (dash lines) are statistically significantly higher than 0.8.

The proportion of incorrectly identified block types ($ErrB$) is equal to zero if we do not have any non-respondents, therefore regression lines in Figure (7.9(b)) are forced through point $(0,\ 0)$. Linear regression model for $ErrB$ according to above restriction is equal to

$$Y_{ErrB} = \beta \cdot n.actor + 0\ . \tag{7.5}$$

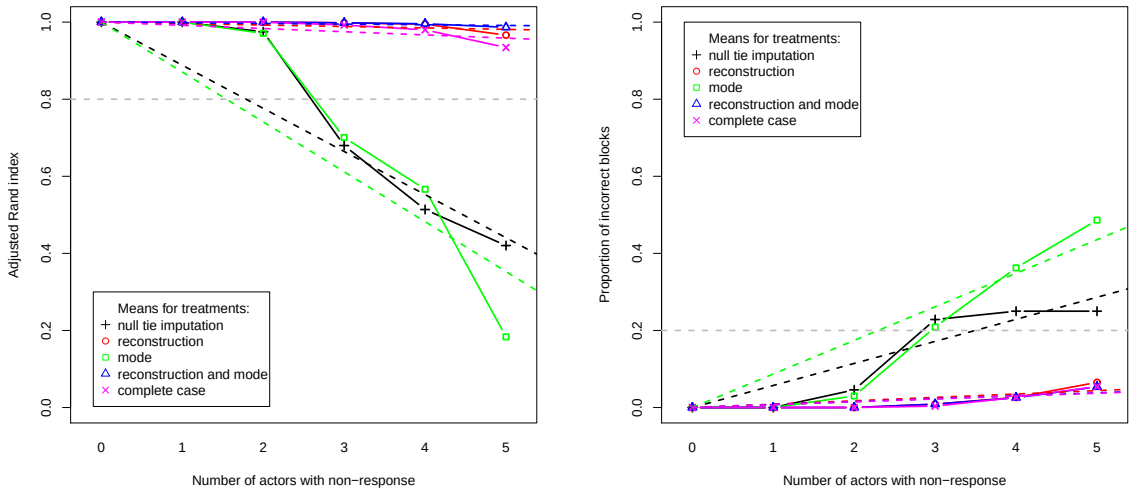
Similar as in the case for ARI we decided to compare slopes of linear predictions with the slope of the line through points $(0,\ 0)$ and $(5,\ 0.2)$. Point $(0,\ 0)$ indicates perfect agreement between positions of blocks in two image matrices without non-respondents and, point $(5,\ 0.2)$ indicates acceptable agreement in terms of $ErrB$ with five introduced non-respondents. The line is presented in Figure 7.10(b) and its slope coefficient is equal to $\beta_0^{ErrB} = \frac{0.2}{5} = 0.04$. Therefore the slopes of linear regression models (presented with dash lines in Figure 7.9(b)) are tested with one-sided t-test where the null and alternative hypotheses are as follows:
$H_0 : \beta_0^{ErrB} \leq 0.04$
$H_1 : \beta_0^{ErrB} > 0.04$

Table A.2 (in Appendix A) presents linear regression models for proportion of incorrect block types. Similar as in models for ARI, changes in values of $ErrB$ are the smallest

if the number of non-respondents is increased for one when the complete-case approach ($\beta = 0.006$), combination of reconstruction and imputations based on mode ($\beta = 0.012$) or simple reconstruction procedure ($\beta = 0.008$) are used. Although the slope coefficients from linear regression models for the null tie imputation and the imputations based on mode are significant ($p - value = 0.000$)), there is not a completely clear linear relationship according to Figure (7.9(b)). Similar as in the case of ARI, the slope of the null tie imputation treatment is statistically significantly higher than 0.04, which confirms that the null tie imputation is unacceptable also in terms of correctly identified blocks in a blockmodel ($ErrB$).

Data missing based on outdegree and indegree

Solid lines in Figures 7.11 and 7.12 display the results when the other two missing (not at random) data regimes (based on outdegree and indegree) are used. The results are similar to those for random missing data for the complete-case, reconstruction and the combination of reconstruction and mean imputation treatments.

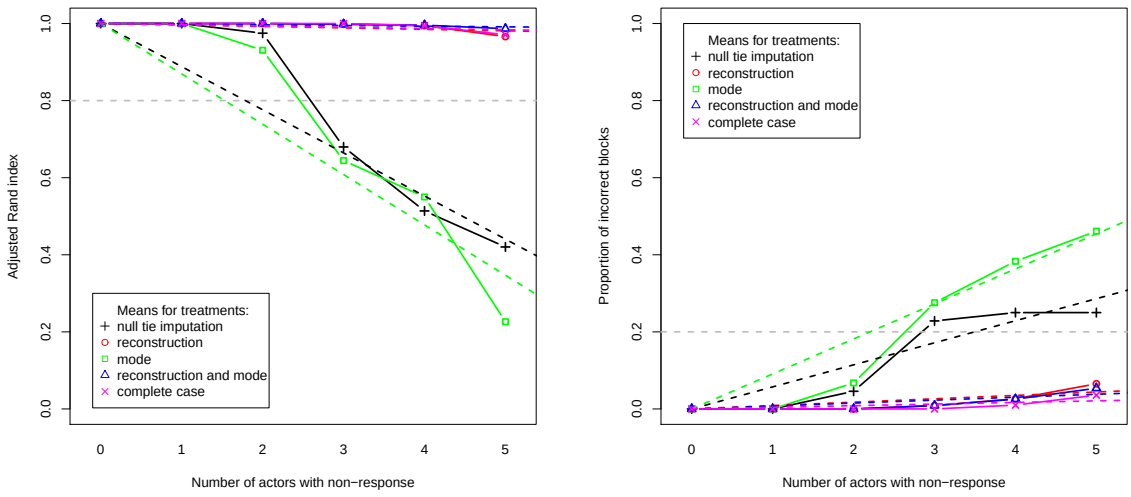


(a) Adjusted Rand Index, $mARI$ (b) Incorrect block types, $mErrB$

Figure 7.11: Results of the simulation study based on the boy-girl liking ties network for missing mechanism based on outdegree (solid lines) and predictions according to linear regression model (dash lines)

The null tie imputation method performs as badly with these two forms of missing

data. The imputation using the mode fares as badly or worse than the null imputation method. The imputation based on mode is a little bit worse, when missing mechanism based on indegree is used compared to missing mechanism based on outdegree. In agreement between partitions (ARI) is the slope coefficient equal to -0.129 when non-respondents are selected based on their outdegree and it case of non-response missing mechanism based on indegree is equal to -0.131 (Table A.1). The highest change in values of $ErrB$ (Table A.2) when number of non-respondents increases for one is obtained with the imputations based on mode ($\beta = 0.087$) when the non-respondents are selected based on outdegree. The slope coefficient is even higher when non-respondents are selected based on their indegree ($\beta = 0.091$). Figures 7.11 and 7.12 Dash lines in Figures 7.11 and 7.12 together with t-test of slope coefficients (Table A.1 and A.2) also indicate that linear models are not appropriate in the case of the null tie imputations and the imputations based on mode.



(a) Adjusted Rand Index, $mARI$

(b) Incorrect block types, $mErrB$

Figure 7.12: Results of the simulation study based on the boy-girl liking ties network for missing mechanism based on indegree (solid lines) and predictions according to linear regression model (dash lines)

This network has a very strong structural signal because the near-complete and null blocks are very clear (Figure 6.1 on page 72). The whole network has high reciprocity (with a reciprocity measure of 0.79). There is little surprise that small amounts of miss-

ing data (one or two non-respondents) do not prevent blockmodeling from identifying the intrinsic network structure in terms of the composition of positions and the identification of blocks. The strong signal also accounts for the poor performance of the null tie imputation treatment because it destroys reciprocity, particularly when there are three or more non-respondents. The very poor performance of the imputation using a mode when non-response is related to indegree and outdegree does have some surprise value at first glance. If we look closer to the image matrix and imagine that two actors are missing from the same cluster, then the imputations based on mode impute exactly the opposite value of a tie from a whole network (e.g. a tie instead of a zero). However, because this whole empirical network has such a clear structure, the results using it are not likely to generalize to other networks. Even so, it also illustrates the general point that the regime generating missing data and the treatments of those data do have the potential to render unstable blockmodeling results.

Tables 7.1 and 7.1 are present the mean values of both indices for comparison of blockmodels together with the standard deviations. The complete-case approach, reconstruction and combination of reconstruction and mode imputation are the best treatments also according to the smallest standard errors.

Establishment of multiple regression models

As we mentioned before, the factorial design has 75 cells (3 missing mechanisms times 5 non-response data treatments times 5 different numbers of non-respondents). First, we try to perform the analysis of variance, but the Levene test for equality of variances revealed significant differences between cells. For the ARI there are 27 cells with variance equal to 0 (cells with zero variance can be found in Table 7.1), and highest variance is 0.073 in cell for missing mechanism based on outdegree with imputations based on mode with 5 non-respondents ($Q_1 = 0.000, Me = 0.002, Q_3 = 0.031$). For the $ErrB$ the variances in 75 cells of factorial design vary from 0.000 to 0.024 ($Q_1 = 0.000, Me = 0.002, Q_3 = 0.009$).

Instead of using the Kruskal-Wallis test as a non-parametric alternative to anova, we

Table 7.1: Mean values and standard deviations for *ARI* for simulations with boy-girl liking ties network

Number of non-respondents		1		2		3		4		5	
Missing mechanism	Treatment	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
	Null tie imputations	1	0	0.939	0.122	0.628	0.202	0.514	0.208	0.408	0.209
Random	Reconstuction	1	0	1	0	1	0	0.965	0.119	0.882	0.182
missing	Mode	1	0	0.879	0.196	0.882	0.182	0.882	0.182	0.882	0.182
mechanism	Reconstruction plus mode	1	0	1	0	1	0	0.969	0.114	0.928	0.149
		1	0	1	0	0.992	0.043	0.985	0.078	0.936	0.175
Missing	Null tie imputations	1	0	0.975	0.089	0.680	0.190	0.514	0.200	0.420	0.148
mechanism	Reconstuction	1	0	1	0	0.998	0.018	0.994	0.041	0.966	0.095
based	Mode	1	0	0.971	0.067	0.701	0.239	0.566	0.248	0.183	0.271
on	Reconstruction plus mode	1	0	1	0	0.998	0.018	0.996	0.030	0.987	0.049
outdegree	Complete Case	1	0	1	0	0.992	0.043	0.980	0.085	0.934	0.177
Missing	Null tie imputations	1	0	0.975	0.089	0.680	0.190	0.514	0.200	0.420	0.148
mechanism	Reconstuction	1	0	1	0	0.998	0.018	0.994	0.041	0.966	0.095
based	Mode	1	0	0.931	0.089	0.644	0.212	0.550	0.176	0.226	0.236
on	Reconstruction plus mode	1	0	1	0	0.998	0.018	0.996	0.030	0.987	0.049
indegree	Complete Case	1	0	1	0	1	0	0.994	0.040	0.971	0.100

Table 7.2: Mean values and standard deviations for *ErrB* for simulations with boy-girl liking ties network

Number of non-respondents		1		2		3		4		5	
Missing mechanism	Treatment	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
	Null tie imputations	0	0	0.063	0.098	0.245	0.022	0.250	0	0.250	0
Random	Reconstuction	0	0	0	0	0.002	0.025	0.042	0.093	0.116	0.119
missing	Mode	0	0	0.103	0.120	0.116	0.119	0.116	0.119	0.116	0.119
mechanism	Reconstruction plus mode	0	0	0	0	0.002	0.025	0.040	0.090	0.091	0.116
	Complete Case	0	0	0	0	0.004	0.021	0.013	0.051	0.050	0.110
Missing	Null tie imputations	0	0	0.046	0.095	0.228	0.046	0.250	0	0.250	0
mechanism	Reconstuction	0	0	0	0	0.009	0.044	0.026	0.076	0.065	0.106
based	Mode	0	0	0.030	0.082	0.209	0.155	0.362	0.152	0.486	0.056
on	Reconstruction plus mode	0	0	0	0	0.009	0.044	0.025	0.074	0.054	0.101
outdegree	Complete Case	0	0	0	0	0.004	0.021	0.026	0.072	0.054	0.116
Missing	Null tie imputations	0	0	0.046	0.095	0.228	0.046	0.250	0	0.250	0
mechanism	Reconstuction	0	0	0	0	0.009	0.044	0.026	0.076	0.065	0.106
based	Mode	0	0	0.068	0.092	0.276	0.099	0.383	0.109	0.461	0.088
on	Reconstruction plus mode	0	0	0	0	0.009	0.044	0.025	0.074	0.054	0.101
indegree	Complete Case	0	0	0	0	0	0	0.010	0.046	0.036	0.083

decided to perform the multiple regression analysis. We run two separate regression analyses, one with values of ARI as dependent variable and the other with $ErrB$ as outcome.

Two factors in our factorial design are categorical variables, so the first stage was to construct dummy variables. The missing mechanism variable has three categories; missing at random, missing based on outdegree and missing based on indegree. The random missing category (MM_random) was selected for the reference or baseline group and two dummy variables were constructed:

- MM_out (ones are assigned for missing mechanism based on outdegree), and
- MM_in (ones are assigned to group with missing mechanism based on indegree).

The second variable 'non-response treatment' has five categories and the null tie imputation group (T_NTI) has been chosen for baseline group. Four dummy variables were constructed:

- T_RE (the value 1 was assigned to the *reconstruction* group),
- T_MO (the value 1 was assigned to the *imputation based on mode* group),
- T_REMO (the value 1 was assigned to the combination of *reconstruction and mode imputations* group), and
- T_CC (the value 1 was assigned to the *complete case* group).

The third predictor in regression analysis is the number of non-respondents which is a ratio variable, therefore no additional recoding was necessary.

The regression model for ARI was set as:

$$\begin{aligned} Y_{ARI} &= \beta_0 + \beta_1 \cdot n.actor + \beta_2 \cdot T\_RE + \beta_3 \cdot T\_MO + \\ &+ \beta_4 \cdot T\_REMO + \beta_5 \cdot T\_CC + \beta_6 \cdot MM\_out + \beta_7 \cdot MM\_in + \varepsilon \,. \end{aligned} \tag{7.6}$$

The model summary in Table 7.3 shows that our regression model predicts values of ARI significantly well (F=866.3; p-value=0.000) and it explains 53% of variation in

ARI. The regression model for ARI form Equation (7.6) can be with estimated (unstandardized) coefficients (Table 7.3) for ARI written as follows:

$$\begin{aligned} \hat{Y}_{ARI} &= 0.900 - 0.0726 \cdot n.actor + 0.3743 \cdot T\_RE + 0.0635 \cdot T\_MO + \\ &+ 0.3829 \cdot T\_REMO + 0.3762 \cdot T\_CC - 0.0488 \cdot MM\_out + \\ &- 0.0483 \cdot MM\_in \ . \end{aligned} \tag{7.7}$$

Table 7.3: Model summary and coefficients of regression analysis for ARI with data from the boy-girl liking ties network

					95% confidence	
	Estimate	Std. Error	t value	Pr(>|t|)	interval for b	
(Intercept)	0.9003	0.0103	87.49	0.0000	0.8801	0.9205
n.actor	-0.0726	0.0022	-32.96	0.0000	-0.0769	-0.0682
T_RE	0.3743	0.0077	48.47	0.0000	0.3592	0.3894
T_MO	0.0635	0.0077	8.22	0.0000	0.0483	0.0786
T_REMO	0.3829	0.0077	49.59	0.0000	0.3678	0.3981
T_CC	0.3762	0.0077	48.72	0.0000	0.3610	0.3913
MM_out	-0.0488	0.0060	-8.16	0.0000	-0.0606	-0.0371
MM_in	-0.0483	0.0060	-8.08	0.0000	-0.0601	-0.0366
Residual standard error: 0.179 on 5392 degrees of freedom						
Multiple R^2: 0.529					Adjusted R^2: 0.529	
F-statistic: 866.253 (on 7 and 5392 df)					p-value: 0.000	

All variables in a model for ARI are significant, because p-values are 0.000 (Table 7.3). The b values (also presented in Figure 7.13) explain to what degree each predictor affects the outcome variable, in our case the values of ARI, if the effects of all other predictors are held constant.

- **n.actor:** ($b = -0.0726$) If the number of non-respondents increases for one non-respondent, the values of ARI decrease for 0.0726. This interpretation is true only if the effects of treatments and missing mechanisms are held constant.

- **T_RE vs. T_NTI:** ($b = 0.3743$) The b value represent the shift in the change of ARI values if the reconstruction treatment is used, compared to null tie imputation. The values of ARI increase for 0.3743 if the reconstruction treatment is used instead of the null tie imputation.

- **T_MO vs. T_NTI:** ($b = 0.0635$) The b coefficient is positive but small if it is compared to other variables for treatments. The values of ARI increase for 0.0635 when imputations based on mode are used instead of the null tie imputation.

- **T_REMO vs. T_NTI:** ($b = 0.3829$) The changes in values of ARI are the highest if the combination of the reconstruction treatment and imputations based on mode is used. In comparison with the null tie imputations, the values of ARI are in that case higher for 0.3829. This means that the combination of reconstruction procedure and imputations based on mode is the best in terms of stability of blockmodel, because the values of ARI are the highest.

- **T_MO vs. T_CC:** ($b = 0.3762$) The use of the complete-case approach instead of the null tie imputations increases values of ARI for 0.3762.

- **MM_out vs. MM_random:** ($b = -0.0488$) The use of the missing mechanism based on outdegree instead of the random missing mechanism has negative effect on values of ARI. In that case values of ARI decrease for 0.0488.

- **MM_in vs. MM_random:** ($b = -0.0483$) The effects of missing mechanism based on indegree are similar to those when missing mechanism based on outdegree is used. The use of the missing mechanism based on indegree instead of the random missing mechanism decreases the value of ARI for 0.0483.

Similarly as for the values of ARI (Equation 7.6) the regression model for $ErrB$ can be written as:

$$\begin{aligned} Y_{ErrB} &= \beta_0 + \beta_1 \cdot n.actor + \beta_2 \cdot T\_RE + \beta_3 \cdot T\_MO + \\ &+ \beta_4 \cdot T\_REMO + \beta_5 \cdot T\_CC + \beta_6 \cdot MM\_out + \beta_7 \cdot MM\_in + \varepsilon \,. \end{aligned} \tag{7.8}$$

The model summary in Table 7.4 shows that our regression model explains 55% of variation in $ErrB$ and predicts values of $ErrB$ significantly well (F=952.2; p-value=0.000).

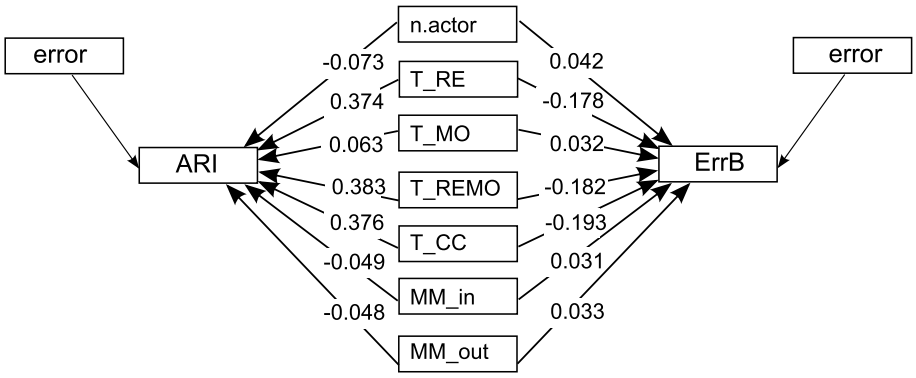


Figure 7.13: Regression models for ARI and $ErrB$ with data from the boy-girl liking ties network

The regression model for $ErrB$ form Equation 7.8 can be with estimated (unstandardized) coefficients (Table 7.4) for ARI written as follows:

$$\begin{aligned}\hat{Y}_{ErrB} &= 0.0373 + 0.0418 \cdot n.actor - 0.1775 \cdot TT\_RE + 0.0321 \cdot TT\_MO \\ &- 0.1927 \cdot TT\_REMO - 0.1927 \cdot TT\_CC + 0.0311 \cdot MM\_out + \\ &+ 0.0335 \cdot MM\_in \,. \end{aligned} \tag{7.9}$$

All variables in a model for $ErrB$ (Equation 7.8) are significant, because p-values are 0.000 (Table 7.4). The model is also presented in Figure 7.13 and the coefficients can be interpreted as follows:

- **n.actor:** ($b = 0.0418$) If the number of non-respondents increases for one non-respondent, the values of $ErrB$ increase for 0.0418.

- **T_RE vs. T_NTI:** ($b = -0.1775$) The shift in the change in $ErrB$ values is negative if the reconstruction treatment is used, compared to the null tie imputation. The value of $ErrB$ decreases for 0.1775 if the reconstruction treatment is used instead of the null tie imputations.

- **T_MO vs. T_NTI:** ($b = 0.0321$) The value of $ErrB$ increases for 0.0321 when imputations based on mode are used instead of the null tie imputations. The absolute value of coefficient b is small compared to other variables for treatments, which means that values of $ErrB$ are the most similar when null tie imputation and imputation based on mode are used. This means that the mode imputation

Table 7.4: Model summary and coefficients of regression analysis for *ErrB* with data from the boy-girl liking ties network

					95% confidence	
	Estimate	Std. Error	t value	Pr(>|t|)	interval for b	
(Intercept)	0.0373	0.0057	6.55	0.0000	0.0261	0.0484
n.actor	0.0418	0.0012	34.39	0.0000	0.0394	0.0442
T_RE vs. T_NTI	-0.1775	0.0043	-41.61	0.0000	-0.1859	-0.1692
T_MO vs. T_NTI	0.0321	0.0043	7.52	0.0000	0.0237	0.0404
T_REMO vs. T_NTI	-0.1823	0.0043	-42.73	0.0000	-0.1907	-0.1739
T_CC vs. T_NTI	-0.1927	0.0043	-45.17	0.0000	-0.2011	-0.1843
MM_out vs. MM_random	0.0311	0.0033	9.41	0.0000	0.0246	0.0376
MM_in vs. MM_random	0.0335	0.0033	10.13	0.0000	0.0270	0.0399
Residual standard error: 0.099 on 5392 degrees of freedom						
Multiple R^2: 0.553					Adjusted R^2: 0.552	
F-statistic: 952.240 (on 7 and 5392 df)					p-value: 0.000	

treatment is the worst in terms of stability of blockmodel (because the values of *ErrB* are the highest).

- **T_REMO vs. T_NTI:** ($b = -0.1823$) In comparison with the null tie imputations, the values of *ErrB* are lower for 0.1823 when combination of reconstruction and the imputations based on mode is used.

- **T_MO vs. T_CC:** ($b = -0.1927$) The changes in values of *ErrB* are the highest if the complete-case approach is used, more precisely the values of *ErrB* decrease for 0.1927 when complete case approach is used instead of the null tie imputations. The values of *ErrB* are the lowest, which indicates that the complete case approach is the best treatment in terms of stability of blockmodels.

- **MM_out vs. MM_random:** ($b = 0.0311$) The use of the missing mechanism based on outdegree instead of the random missing mechanism has positive effect on values of *ErrB*. In that case values of *ErrB* increase for 0.0311.

- **MM_in vs. MM_random:** ($b = 0.0335$) The use of the missing mechanism based on indegree instead of the random missing mechanism increases the value for ARI for 0.033. Both non-random missing mechanisms have similar effects. The results of blockmodel stability are a little bit lower when non-random missing mechanisms are used compared to random missing mechanism (because values of $ErrB$ are higher).

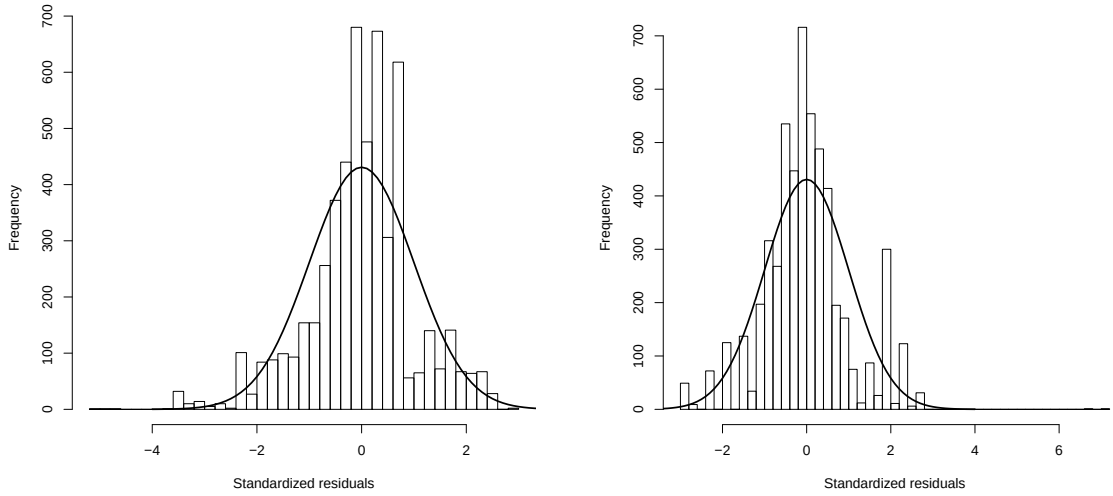
When the regression model is established, two important questions arise (Field, 2009): (i) does the model fit the observed data well or is it influenced by the small number of cases, (ii) can the model be generalized to other samples. If the model is perfectly good for the data (no outliers, influential cases, etc.), then that model can be used to draw conclusions about the sample, even when the assumptions[19] are violated (Field, 2009).

If the model is adequate the residuals should be normally distributed with mean zero and standard deviation σ, $\varepsilon_i \sim N(0, \sigma)$, and if we look at the histograms of standardized residuals it should be normally distributed with mean 0 and standard deviation one ($\varepsilon_i \sim N(0,\ 1)$). Figure 7.14 shows histograms of standardized residuals for both models (Equation (7.7) and (7.9)) and we can conclude that residuals are not normally distributed.

Figure 7.15 shows residuals plotted versus fitted values. The figures show that variances are not constant and esspecially for $ErrB$ the pattern of residuals (Figure 7.15(b)) indicates that probably more adequate model will be obtained with additional variable in linear summand of regression model (Košmelj and Kastelec, 2003).

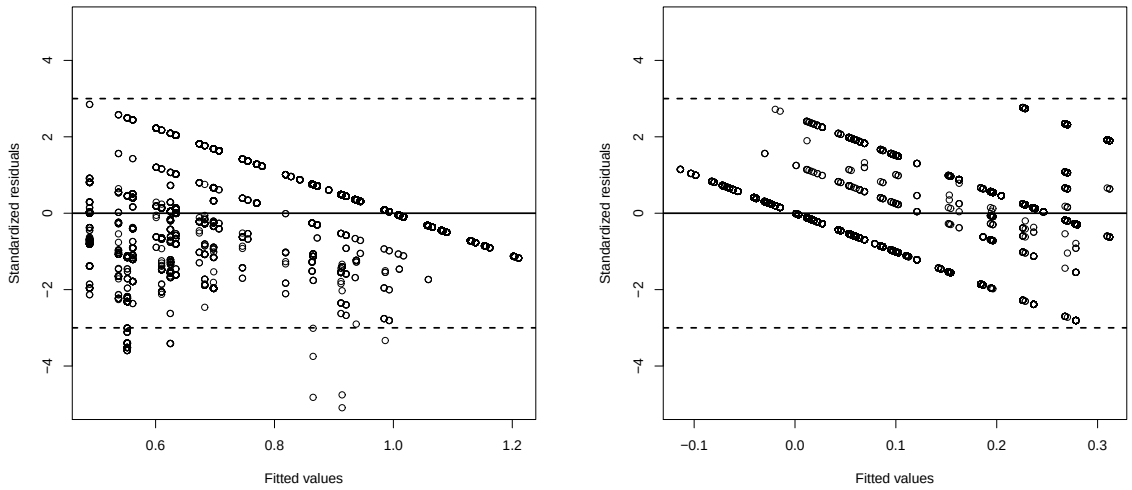
Although our findings based on regression analysis can not be generalized because assumptions are violated, the models can still be used to draw conclusions about the data sample (Field, 2009). The presented regression models for both indices of network stability (ARI and $ErrB$) as outcome confirm that the reconstruction in combination with mode imputations, the complete-case approach and the reconstruction are the

[19] Normality of errors, homoscedasticity, independence of errors, linearity of relations between variables.



(a) Model for ARI (b) Model for $ErrB$

Figure 7.14: Histogram of standardized residuals of regression models with data from the boy-girl liking ties network



(a) Model for ARI (b) Model for $ErrB$

Figure 7.15: Fitted values versus standardized residuals of regression models with data from the boy-girl liking ties network

best treatments. The effects of non-random missing mechanisms (based on indegree and outdegree) compared to random missing mechanism are small, because of similar patterns in outgoing and incoming ties of all actors.

7.3.2.2 The student note borrowing network

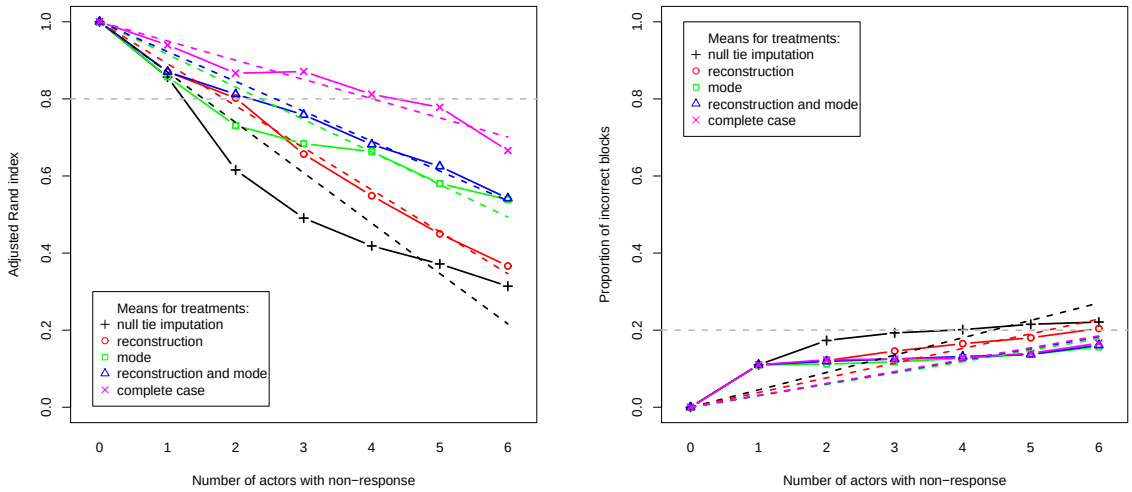
The student note borrowing network has 15 actors and its blockmodel into three clusters based on structural equivalence is presented in Figure 6.2 on page 72. The resulting factorial design has 90 cells (for the combinations of three non-response mechanisms, five treatments of non-response, and six numbers of actors with non-response). Within each cell, the generation of incomplete data was repeated 10 times for networks with one missing actor, 50 times for combinations of two missing actors and 100 times[20] for combinations of three or more non-respondents.

Data missing completely at random

Figure 7.16 (and Table 7.5) presents the results of simulation study for the student note borrowing network when actors are selected randomly as non-respondents. The existence of only one non-respondent has an impact on the established blockmodels and values of $mARI$ are lower than in case of the boy-girl liking ties network for all treatments. For measuring the concordance between positions, the null tie imputation method performs the worst which means that this procedure leads to the most unstable blockmodels.

Overall, using reconstruction comes next with regard to poor performance when there are three or more non-respondents. Use of the mode for imputations and the combination of reconstruction and mode imputation come next. The best performance or the highest stability of blockmodeling in terms of partitions comes with the complete-case approach where high stability of blockmodel with $mARI$ values above 0.8 is obtained also with four non-respondents. If five or six actors refuse to response the actor response rate is equal to 0.67 or 0.6, respectively. The presence of more than five non-respondents shoves off the mean values of ARI below 0.8 indicating that the correspondence of the position memberships is unacceptable.

[20]The number of all possible combinations of actors with non-response increases. For example, for a network where n = 11 there are: $\binom{15}{1} = 15$ possibilities for selecting one non-respondent, $\binom{15}{2} = 105$ possibilities for selecting two non-respondents; $\binom{15}{3} = 455$ and so on.



(a) Adjusted Rand Index, $mARI$

(b) Incorrect block types, $mErrB$

Figure 7.16: Results of the simulation study based on the borrowing network for data missing completely at random (solid lines) and predictions according to linear regression model (dash lines)

Similar as in the case of the boy-girl liking ties network, all five non-response data treatments are indistinguishable when there is one non-respondent (Figure 7.16(b) and Table 7.6). But even here, the agreement between block types and their positions is not perfect, because the average proportion of incorrectly identified blocks ($mErrB$) is 0.11. The starting blockmodel of the note borrowing network has nine blocks and it is presented in the middle panel in Figure 6.2. If one actor refused to respond, the $mErrB = 0.11$ indicates that one block is misrepresented in the treated blockmodel. As the number of non-respondents increases, the performance of blockmodeling under all missing data treatments worsens. For three non-respondents or less the mean of $ErrB$ for all missing data treatment is below 0.2 which indicates acceptable results in regard to blockmodel structure. Consistent with the results for stability of partitions (ARI), both the null imputation and reconstruction treatments lead to the most unstable blockmodels. On average 20% of block types (or two blocks) are identified incorrectly. For six non-respondents values of ARI are the lowest for imputations based on mode with $mErrB$ equal to 0.156, but the standard deviation is the highest among all treatments (Table 7.6).

Table 7.5: Mean values and standard deviations for ARI for simulations with the note borrowing network

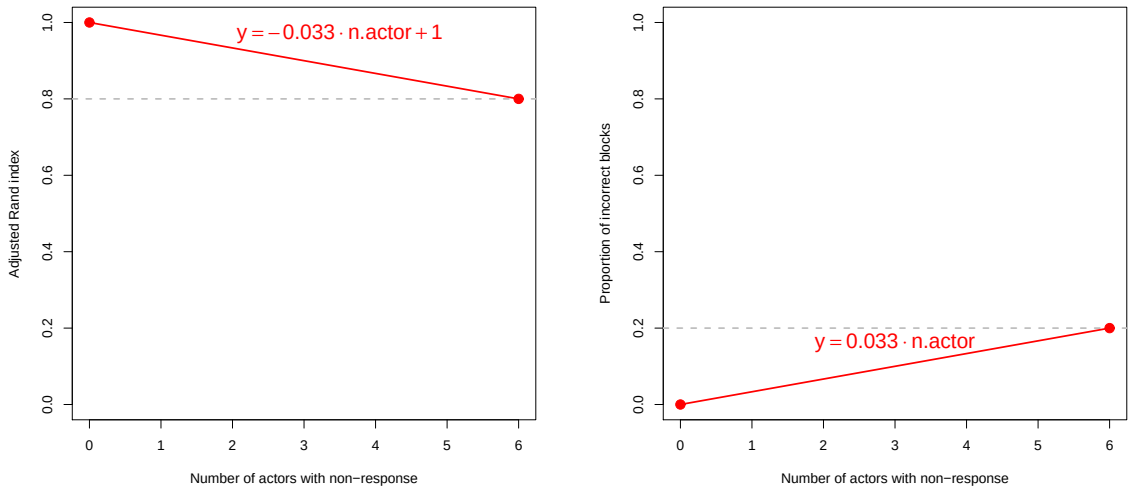
Number of non-respondents		1		2		3		4		5		6	
Missing mechanism		Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
	Null tie imputations	0.857	0.092	0.616	0.177	0.491	0.162	0.419	0.153	0.372	0.155	0.314	0.126
Random	Reconstruction	0.870	0.122	0.802	0.159	0.657	0.200	0.549	0.194	0.450	0.169	0.367	0.153
missing	Mode	0.857	0.126	0.731	0.116	0.684	0.126	0.663	0.158	0.580	0.166	0.539	0.177
mechanism	Reconstr. + mode	0.870	0.122	0.812	0.153	0.759	0.172	0.682	0.178	0.625	0.179	0.542	0.213
	Complete Case	0.940	0.063	0.867	0.203	0.871	0.176	0.812	0.218	0.778	0.225	0.666	0.279
missing	Null tie imputations	0.957	0.075	0.628	0.147	0.510	0.135	0.397	0.130	0.340	0.116	0.301	0.113
mechanism	Reconstruction	0.947	0.091	0.801	0.172	0.708	0.213	0.566	0.199	0.430	0.149	0.382	0.151
based	Mode	0.839	0.170	0.787	0.151	0.711	0.153	0.708	0.160	0.657	0.177	0.603	0.191
on	Reconstr. + mode	0.947	0.091	0.813	0.174	0.762	0.195	0.730	0.201	0.704	0.224	0.658	0.228
outdegree	Complete Case	0.979	0.045	0.923	0.154	0.886	0.161	0.843	0.196	0.809	0.223	0.720	0.270
missing	Null tie imputations	0.957	0.075	0.628	0.147	0.510	0.135	0.397	0.130	0.340	0.116	0.301	0.113
mechanism	Reconstruction	0.857	0.126	0.603	0.151	0.526	0.116	0.574	0.113	0.590	0.114	0.512	0.200
based	Mode	0.839	0.170	0.787	0.151	0.711	0.153	0.708	0.160	0.657	0.177	0.603	0.191
on	Reconstr. + mode	0.857	0.126	0.603	0.151	0.533	0.121	0.589	0.113	0.704	0.224	0.658	0.228
indegree	Complete Case	0.936	0.110	0.928	0.112	0.767	0.119	0.640	0.160	0.578	0.201	0.509	0.201

Table 7.6: Mean values and standard deviations for $ErrB$ for simulations with the borrowing network

Number of non-respondents		1		2		3		4		5		6	
Missing mechanism	Treatment	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
	Null tie imputations	0.110	0.116	0.173	0.119	0.193	0.108	0.201	0.103	0.215	0.094	0.221	0.084
Random	Reconstuction	0.110	0.116	0.121	0.099	0.146	0.096	0.165	0.092	0.180	0.092	0.204	0.098
missing	Mode	0.110	0.116	0.112	0.114	0.118	0.116	0.127	0.115	0.137	0.119	0.156	0.127
mechanism	Reconstr. + mode	0.110	0.116	0.119	0.099	0.125	0.102	0.132	0.096	0.137	0.098	0.161	0.105
	Complete Case	0.110	0.116	0.123	0.117	0.125	0.117	0.128	0.109	0.140	0.113	0.166	0.123
missing	Null tie imputations	0.110	0.116	0.183	0.122	0.196	0.110	0.210	0.104	0.214	0.090	0.211	0.083
mechanism	Reconstuction	0.110	0.116	0.133	0.111	0.148	0.103	0.168	0.097	0.184	0.089	0.192	0.091
based	Mode	0.110	0.116	0.116	0.108	0.121	0.114	0.123	0.111	0.133	0.117	0.137	0.116
on	Reconstr. + mode	0.110	0.116	0.132	0.111	0.136	0.105	0.124	0.101	0.130	0.100	0.133	0.101
outdegree	Complete Case	0.110	0.116	0.121	0.109	0.121	0.110	0.123	0.112	0.130	0.112	0.146	0.113
missing	Null tie imputations	0.110	0.116	0.183	0.122	0.196	0.110	0.210	0.104	0.214	0.090	0.211	0.083
mechanism	Reconstuction	0.110	0.116	0.159	0.116	0.162	0.086	0.137	0.064	0.126	0.052	0.137	0.067
based	Mode	0.110	0.116	0.116	0.108	0.121	0.114	0.123	0.111	0.133	0.117	0.137	0.116
on	Reconstr. + mode	0.110	0.116	0.159	0.116	0.162	0.085	0.137	0.067	0.130	0.100	0.133	0.101
indegree	Complete Case	0.110	0.116	0.110	0.111	0.115	0.111	0.123	0.116	0.130	0.109	0.157	0.106

Dash lines in Figure 7.16 present predictions according to established linear regression models. Similar as for the boy-girl liking ties network the slope coefficients were tested with one sided t-test where testing slope coefficient β_0 are presented in Figure 7.17.

Because the borrowing network has 15 actors and in simulation study maximal 6 non-respondents were selected (40% of actors in a network), we decided to compare the slope coefficients from linear models with the line through points $(0,\ 1)$ and $(6,\ 0.8)$ for ARI and with the line through points $(0,\ 0)$ (Figure 7.17(a)) and $(6,\ 0.2)$ for $ErrB$ (7.17(b)).



(a) Adjusted Rand Index - ARI

(b) Incorrect block types - $mErrB$

Figure 7.17: Schematic representation of lines which were used for comparison of linear regression models in larger networks for both indices of blockmodeling stability

In the linear regression models for ARI the highest slope coefficient has the complete-case approach ($\beta = -0.050$) and it is statistically significantly lower than $\beta_0^{ARI} = \frac{-0.2}{6} = -0.0\bar{3}$ (Table A.1). Other treatments perform worse and they have even lower slope coefficients in linear regression models. In the case of $ErrB$ the slopes are statistically significantly lower than $\beta_0^{ErrB} = \frac{0.2}{6} = 0.0\bar{3}$ when the imputations based on mode ($\beta = 0.030$), reconstruction plus imputations based on mode ($\beta = 0.030$) or the complete-case approach ($\beta = 0.031$) are used (Table A.2).

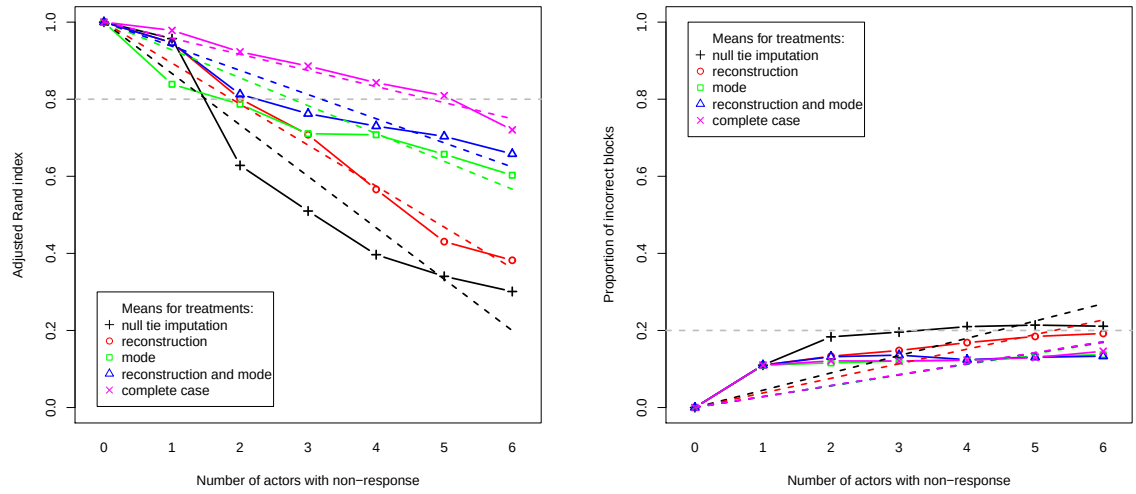
The blockmodel structure on Figure 6.2 has a less clear structure than the one for the boy-girl liking ties network (Figure 6.1) and is less symmetric with reciprocity value equal to 0.46. The two methods that fared the worst (the null tie imputation and the reconstruction) introduce further non-symmetry into the treated network which pre-

vents the revealing of the true blockmodel structure.

.

Data missing based on outdegree

Figure 7.18 presents the simulation results when the chance of being a non-respondent is conditioned by the outdegree of an actor. In practice that means that more active actors (with higher outdegree) have lower probability of being non-respondents. The results with non-random missing mechanism based on outdegree are, in essence, the same as when there is a random selection of non-respondents (Figure 7.18). The main difference is that also the complete-case treatment with five non-respondents leads to acceptable partition agreement, because values of $mARI$ are above 0.8.



(a) Adjusted Rand Index, $mARI$

(b) Incorrect block types, $mErrB$

Figure 7.18: Results of the simulation study based on the borrowing network for data missing based on outdegree (solid lines) and predictions according to linear regression model (dash lines)

Linear regression models show similar patterns as in the case of randomly selected non-respondents. In the case of comparison of two partitions (ARI) the highest slope coefficient has again the complete-case approach ($\beta = -0.042$), but the treatment is unacceptable with six non-respondents. The blockmodeling is more stable in terms of agreement between two image matrices because the slope coefficients for the imputations based on mode ($\beta = 0.028$), the reconstruction plus imputations based on mode

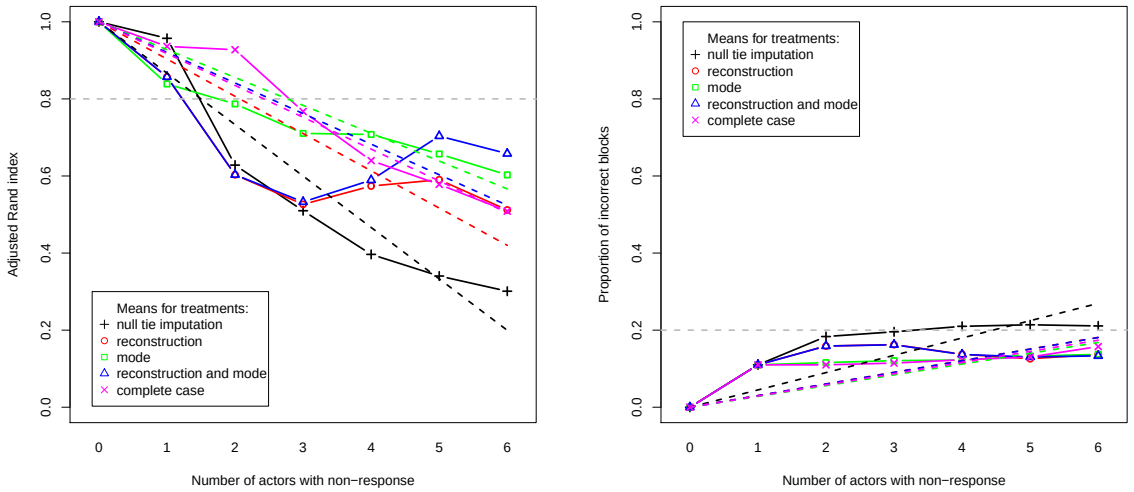
($\beta = 0.028$), and the complete-case approach ($\beta = 0.029$) are statistically significant lower than $\beta_0^{ErrB} = 0.0\bar{3}$ (Table A.2).

Data missing based on indegree

When the chance of being a non-respondent is conditioned by indegree, there are some surprising differences in the values of $mARI$. Thus far in the discussion of results, if there is a degrading of the blockmodeling results, this degradation gets worse as the amount of non-response increases. Some of the $mARI$ plots in Figure 7.18(a) depart from this pattern. As the number of non-respondents increases from 3 to 4 and 5, the $mARI$ values for both reconstruction and combining reconstructing data with use of the mode increase. It seems that if the chance of being a non-respondent is lower as the actor indegree increases, these two treatments of missing data provide some protection as far as blockmodeling results are concerned. Whether it offers enough protection is open to interpretation because the resulting values for $mARI$ measure may be too low to trust in the depiction of position memberships. Thereafter, with further increases in the number of non-respondents to 6 non-respondents the $mARI$ plots resume their downward pattern.

Dash lines in Figure 7.19(a) indicate that the linear models are (conditionally) appropriate for the imputations based on mode and the complete-case approach, while both reconstruction treatments show nonlinear patterns.

The results regarding the proportion of incorrectly identified block types ($mErrB$) are the same as for the other two regimes for generating missing data (Figure 7.19(b)). Similar as in the stability of blockmodeling in terms of partitions where $mARI$ values increased with four and five non-respondents for both reconstruction treatments, values of $mErrB$ are lower for four non-respondents or more compared to two or three non-respondents which indicates higher stability of blockmodeling in terms of block types positions.



(a) Adjusted Rand Index, $mARI$

(b) Incorrect block types, $mErrB$

Figure 7.19: Results of the simulation study based on the borrowing network for data missing based on indegree (solid lines) and predictions according to linear regression model (dash lines)

Establishment of multiple regression models

The factorial design has 90 cells (3 non-response or missing mechanisms times 5 non-response data treatments times 6 different numbers of non-respondents). The Levene tests for equality of variances revealed significant differences between cells. The variances for the ARI are in range from 0.002 to 0.078 ($Q_1 = 0.016,\ Me = 0.024,\ Q_3 = 0.035$). The highest variance (standard deviations are reported in Table 7.5) is in the cell with random missing mechanism for six non-respondents with complete case treatment. The lowest variance for ARI is obtained in cell with missing mechanism based on outdegree with one non-respondent and complete-case approach. The variances for $ErrB$ are in range from 0.003 to 0.016 ($Q_1 = 0.010,\ Me = 0.012,\ Q_3 = 0.013$). The highest variance is in cell of randomly missing data with imputations based on mode for six non-respondents and the lowest variance is in cell with missing mechanism based on indegree with five non-respondents and reconstruction treatment (the standard deviations are reported in Table 7.6).

The multiple regression model was established with dummy variables as described in

Section 7.3.2.1 on page 116. The model summary in Table 7.7 shows that our regression model predicts values of ARI significantly well (F=609.6; p-value=0.000) and it explains 38% of variation in ARI. The regression model for ARI as depednent variable can be with estimated (unstandardized) coefficients (Table 7.7 and Figure 7.20) written as follows:

$$\begin{aligned} \hat{Y}_{ARI} &= 0.6773 - 0.0606 \cdot n.actor + 0.1288 \cdot T\_RE + 0.2409 \cdot T\_MO + \\ &+ 0.2480 \cdot T\_REMO + 0.3348 \cdot T\_CC + 0.0303 \cdot MM\_out + \\ &- 0.0237 \cdot MM\_in \ . \end{aligned} \tag{7.10}$$

All variables in a model for ARI are significant, because p-values are 0.000 (Table 7.3). The regression coefficients b can be interpreted as follows:

- **n.actor:** ($b = -0.0606$) If the number of non-respondents increases for one non-respondent, the values of ARI decrease for 0.0606 when other effects of non-response treatments and missing mechanisms are held constant.

- **T_RE vs. T_NTI:** ($b = 0.1288$) If the reconstruction treatment is used, compared to the null tie imputation the value of ARI increases for 0.1288. In comparison to the regression model for the boy-girl liking ties network (Figure 7.13), the reconstruction treatment in this case has lower effect on values of ARI. The reason for this is not so clear symmetric structure of the starting blockmodel as for the boy-girl liking ties network.

- **T_MO vs. T_NTI:** ($b = 0.2409$) The value of ARI increases for 0.2409 when imputations based on mode are used instead of the null tie imputation.

- **T_REMO vs. T_NTI:** ($b = 0.2480$) In comparison with the null tie imputations, the values of ARI in that case are higher for 0.2480. The reconstruction combined with mode imputation seems to be the second best treatment in terms of ARI values and position agreement between actors in a blockmodel.

- **T_MO vs. T_CC:** ($b = 0.3348$) The complete-case approach is the best in terms of stability of blockmodel according to position membership, because the values of ARI are the highest. The use of the complete-case approach instead of the null tie imputations increases values of ARI for 0.3348.

- **MM_out vs. MM_random:** ($b = 0.0303$) The use of the missing mechanism based on outdegree instead of the random missing mechanism has a little positive effect on values of ARI. In that case values of ARI increase for 0.0303. If the actors with low outdegree (inactive actors) have higher probabilities to be non-respondents, the partitions of actors in blockmodel are recovered more accurately.

- **MM_in vs. MM_random:** ($b = -0.0237$) The use of the missing mechanism based on indegree instead of the random missing mechanism decreases the value of ARI for 0.0237. When the actors with low indegree (less popular actors) have higher probabilities of being non-respondents, the stability of blockmodel in terms of partitions of actors is a little bit lower compared to random non-respondents.

Table 7.7: Model summary and coefficients of regression analysis for ARI with data from the note borrowing network

					95% confidence	
	Estimate	Std. Error	t value	Pr(>|t|)	interval for b	
(Intercept)	0.6773	0.0088	76.53	0.0000	0.6599	0.6946
n_actor	-0.0606	0.0016	-37.89	0.0000	-0.0637	-0.0575
T_RE	0.1288	0.0070	18.42	0.0000	0.1151	0.1425
T_MO	0.2409	0.0070	34.45	0.0000	0.2272	0.2546
T_REMO	0.2480	0.0070	35.47	0.0000	0.2343	0.2617
T_CC	0.3348	0.0070	47.88	0.0000	0.3211	0.3485
MM_out	0.0303	0.0054	5.59	0.0000	0.0197	0.0409
MM_in	-0.0237	0.0054	-4.38	0.0000	-0.0344	-0.0131
Residual standard error: 0.1837 on 6892 degrees of freedom						
Multiple R^2: 0.382					Adjusted R^2: 0.382	
F-statistic: 609.584 (on 7 and 6892 df)					p-value: 0.000	

In our model there is 3.86% of residuals higher than 2 or lower than -2, 1.16% of residuals are higher than 2.5 or lower than -2, and 0.51% of standardized residuals have

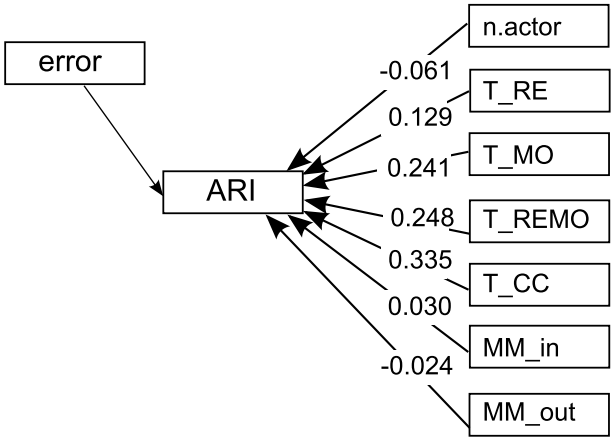
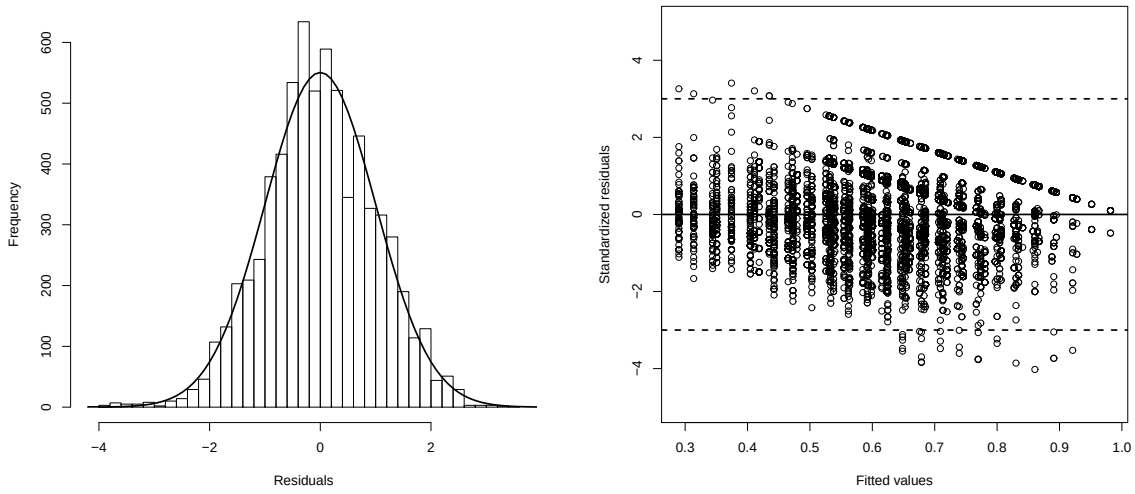


Figure 7.20: Regression models for ARI with data from the note borrowing network

absolute value greater than 3. Histogram of standardized residual looks approximately normal (Figure 7.21(a)). Figure 7.21(b) shows residuals plotted versus fitted values where the pattern of residuals indicates that probably more adequate model will be obtained with inclusion of additional variable in the linear summand of the model (Košmelj and Kastelec, 2003). Although the assumptions of regression analysis for the ARI are not completely satisfied, the above conclusions about our sample are valid (Field, 2009).



(a) Histogram of standardized residuals (b) Fitted values versus standardized residuals

Figure 7.21: Reisduals from model for ARI with data from the note borrowing network

We also try to set up the regression model for $ErrB$. It turns out that it explains just 7.5%

of variance (it is not reported here). One reason why the model is bad is that there is no linear relationship between number of non-respondents and values of *ErrB* index. In Figure 7.22 the radius of the circles is proportional to the number of cases with the same value of *ErrB*. For one non-respondent there are just two possible values of *ErrB*, 0 and 0.22, which indicates that blockmodels obtained with treated data are exactly the same as the whole blockmodel or that treated blockmodels have two different block types. The distribution of *ErrB* for three to six non-respondents shows similar pattern where the majority of *ErrB* occupy four values. Therefore, the number of non-respondents is obviously inadequate to predict the proportion of incorrect block types (*ErrB*) in the linear regression model.

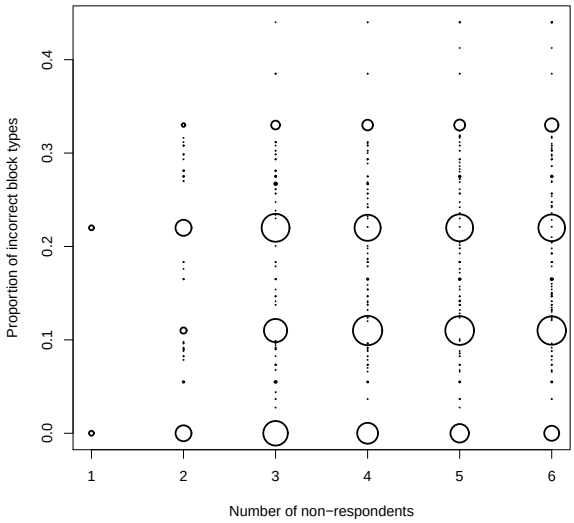


Figure 7.22: The relationship between the number of non-respondents and values of *ErrB* index for the note borrowing network

7.3.3 Results of simulation study of actor non-response for simulated networks

Study of empirical networks is one way of considering the potential consequences of the presence of non-repondents in the network data. While the two real networks that were examined in Section 7.3.2 provide some clues about these consequences, they do not provide an adequate foundation for assessing the general impact of the presence

of non-respondents and, more importantly, the impact that treatments of missing data may have on the results produced by blockmodeling. For that we turn to simulating whole networks with known properties such as number of actors, number of clusters and type together with position of blocks in a blockmodel (described in Section 6.2.3).

7.3.3.1 Results for the completely symmetric blockmodel structure

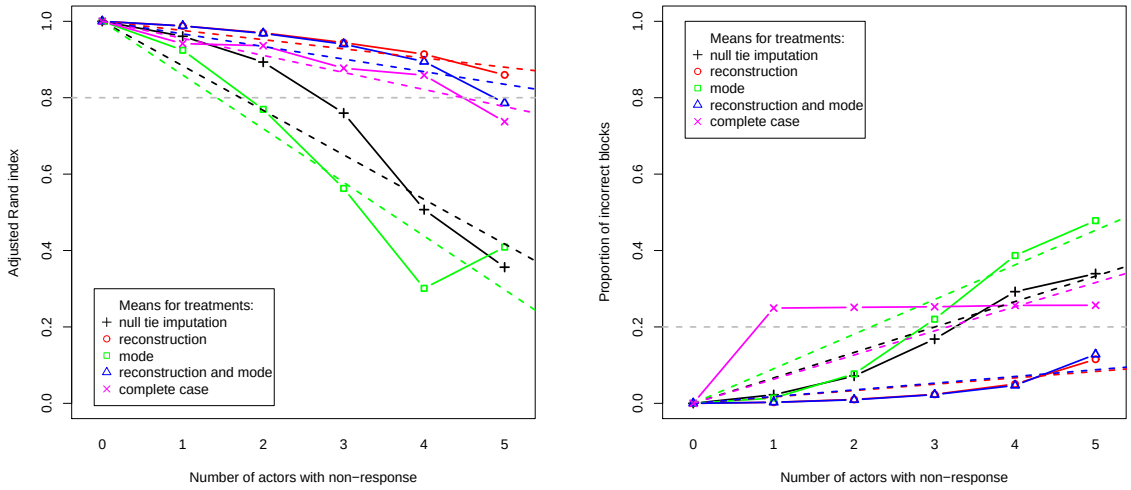
The prototype for simulated networks for the completely symmetric blockmodel structure is the boy-girl liking ties network. We generated 140 whole starting networks with different combinations of probabilities of ties within null and complete blocks. The construction of networks is described in detail in Section 6.2.3.1 together with properties of simulated networks.

The factorial design for this blockmodel has 75 cells which arise from the combination of 3 non-response data mechanisms, 5 non-response (or missing) data treatments, and 5 numbers of actors with non-response. Within each cell, the generation of incomplete data was repeated 10 times for one missing actor, 30 times for combination of two missing actors and 100 times for combinations of three or more missing actors.

Data missing completely at random

Solid lines in Figure 7.23(a) show the mean values for the ARI. In general, as the number of non-respondents increases, the $mARI$ values decline for all missing data treatments. There is one exception in that the $mARI$ means for the imputation based on mode increase from four to five non-respondents. However, these values are all in the unacceptable region because values of $mARI$ are below 0.8. When there is one non-respondent in the network the results from all treatment methods are acceptable (and are in the excellent range with $mARI$ above 0.9). However, for more than one non-respondent differences in the results emerge for all treatments.

If we have two non-respondents in the network, the $mARI$ drops below 0.8 for imputations based on the mode and its agreement is unacceptable for all higher numbers on non-responses. The results following the other four treatments are acceptable. For



(a) Adjusted Rand Index, $mARI$

(b) Incorrect block types, $mErrB$

Figure 7.23: Results of the simulation study based on the completely symmetric block-model structure for data missing completely at random (solid lines) and predictions according to linear regression model (dash lines)

three non-respondents, the results from the null tie imputation treatment drops into the unacceptable range and remains there for higher numbers of non-respondents. There is a modest decline in the $mARI$ for the remaining three treatments as number of non-respondents increases from 1 to 4, but all $mARI$ values are acceptable. When five non-respondents are present in the network, the actor response rate is equal to 50% and the agreement for the complete-case treatment also becomes unacceptable. Over the full range non-respondents, there are two treatments that permit acceptable identification of position memberships. They are reconstruction and the combined use of reconstruction and imputation based on mode for ties between non-respondents. Of the two, the former performs slightly better.

Dash lines present predictions from linear regression models. The slope coefficient for the reconstruction procedure is indeed the highest ($\beta = -0.024$) and it is statistically significant higher than $\beta_0^{ARI} = -0.04$ (Table A.1).

Figure 7.23(b) shows results for incorrectly identified block types. The results from

using the complete-case approach are insensitive to the number of non-respondents where the $mErrB$ is equal to 0.25. This indicates that in every simulation of non-response one block type in a blockmodel was misspecified. However, $mErrB > 0.2$ makes the results following this treatment unacceptable. For the remaining four treatments, having one non-respondent in the network implies acceptable results with near-zero mean values for the proportion of incorrectly identified blocks. As for $mARI$, the results from treating missing data with either reconstruction or the combination of reconstruction with using the mode are acceptable over the full range of non-respondents considered here. However, the mean values of $mErrB$ do increase slightly as the number of non-respondents increases. The mean value for this index for both the null tie imputation and the imputation based on mode becomes unacceptable for three and four non-respondents, respectively.

Similar as in the case of ARI the linear models are appropriate for both reconstruction procedures (dash lines in Figure 7.23(b)). The slope coefficents are in both cases statistically significant lower than $\beta_0^{ErrB} = 0.04$ (Table A.2). On the other hand, the slope coefficient for the imputations based on mode treatment is the highest ($\beta = 0.091$) and is obviously statistically significant higher than $\beta_0^{ErrB} = 0.04$.

Considering values of $mARI$ and $mErrB$, when data are missing at random, only two treatments for non-response pass muster over the range of the number of non-respondents considered here: reconstruction and the combination of reconstruction with using the mode permit the return of accurate blockmodels. While the resulting blockmodels are acceptable (except when we have 5 non-respondents) for the complete-case approach as far as $mARI$ is concerned, its performance in terms of $mErrB$ is never acceptable regardless of the number of non-respondents. It appears that null tie imputation and reconstruction based on the mode are acceptable *only* when the number of non-respondents is very small (one or two actors at most).

We kept track of reciprocity values calculated for each simulated whole network in Section 6.2.3.1 and presented in Figure 6.3(b). The data underlying Figure 7.23 can

be presented also in terms of the reciprocity of the starting whole networks. Figure 7.24 presents values of $mARI$ and $mErrB$ plotted against reciprocity values of starting whole networks. The reciprocity values of networks generated based on the completely symmetric blockmodel structure are in range from 0.5 to 1.00. The curves for all treatments in all panels are fitted to the data based on the smallest mean squared errors between real data values and the fitted values for a selected function. The set of curves is only used to provide a visual pattern of the performance of the two criteria for evaluating the extent to which treated blockmodels are close to (or far from) the known blockmodels of the whole networks. The functions that were fitted were selected from the following: a linear function $f(x) = a \cdot x + b$, a quadratic function $f(x) = a \cdot x^2 + b \cdot x + c$, an exponential function $f(x) = e^{a \cdot x + b}$, a logarithmic function $f(x) = \log(a \cdot x + b)$, and a logistic function $f(x) = \frac{c}{1 + b \cdot e^{-a \cdot x}}$.

The above findings are reinforced by reciprocity values. We can add that higher values of reciprocity lead to more stable blockmodels in terms of $mARI$. The mean of the Adjusted Rand Index tends to increase with higher values of network reciprocity for all five missing data treatments. This general improvement in stability of blockmodeling is also true for $mErrB$ where proportion of incorrectly identified block types decreases with higher reciprocity values.

The summary of means and standard deviations of ARI and $ErrB$ values for the completely symmetric blockmodel structure are presented in Tables 7.8 and 7.9. Not only the reconstruction and the reconstruction together with imputation based on mode are the best treatments according to the $mARI$ and $mErrB$, they also have the smallest standard deviations for ARI and $ErrB$, which can be another indicator of lower blockmodel instability. For example, the standard deviation for ARI for the reconstruction treatment is 0.062 when there is one non-respondent, and for the imputation based on mode it is more than two times larger (0.145). The relative differences between standard deviations for ARI become smaller with higher number of non-respondents.

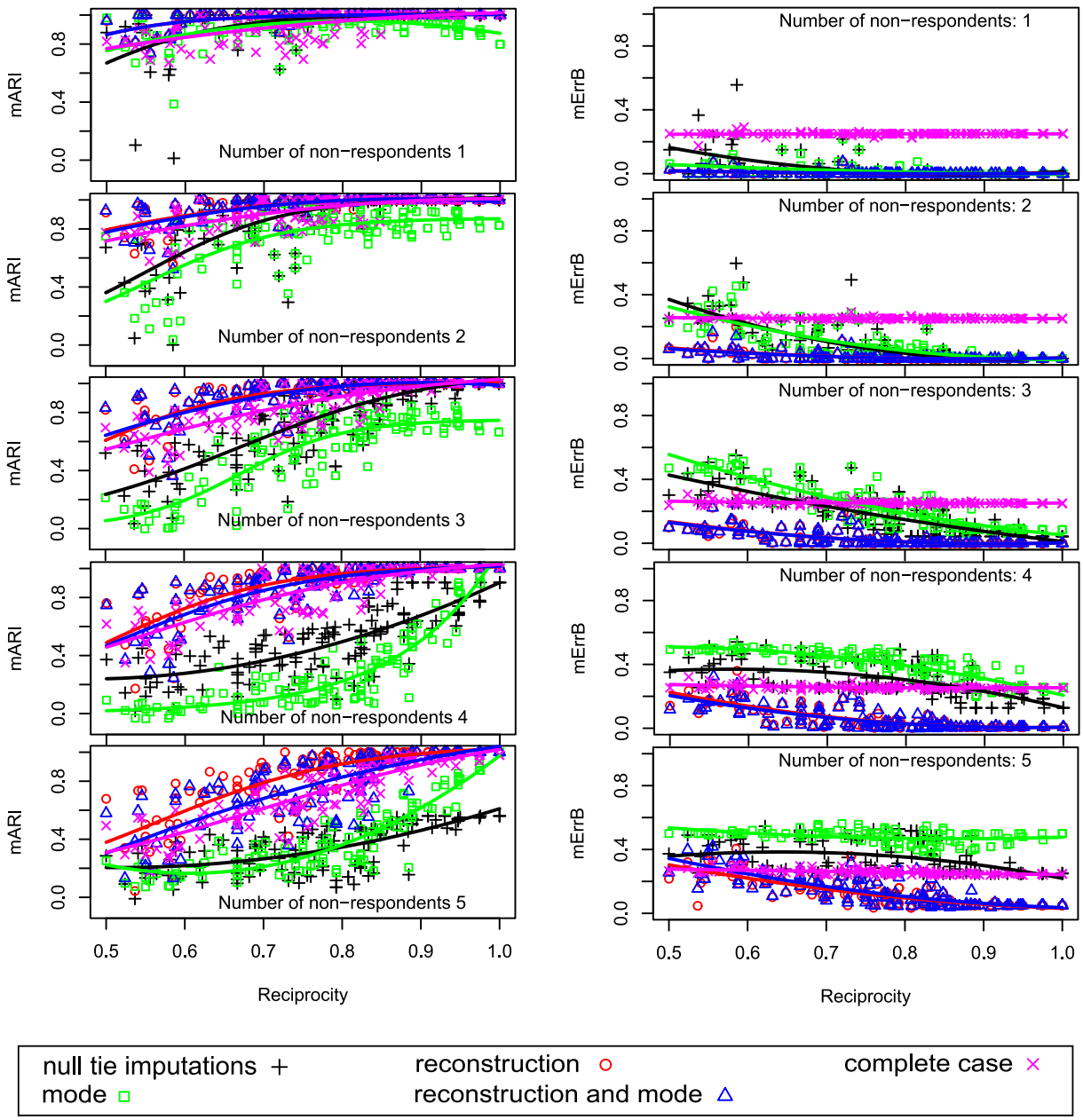


Figure 7.24: The mean of the Adjusted Rand Index, $mARI$ (left), and the mean of Proportion of Incorrect block types, $mErrB$ (right), for completely symmetric blockmodel structure and random missing mechanism

Randomly missing data based on outdegree

In the first non-random missing mechanism actors were selected to be non-respondents based on its outdegree. That means that actors with lower outdegree have higher probability to be selected as non-respondents. The results which are practically the same as for randomly selected non-respondents are presented in Figure 7.25. The recon-

Table 7.8: Mean values and standard deviations for ARI for simulations for the completely symmetric blockmodel structure

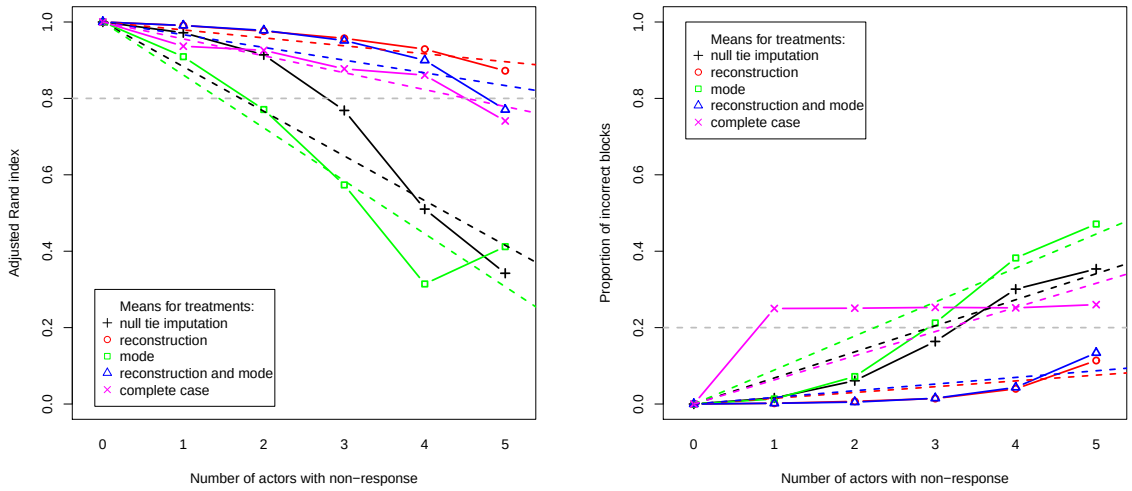
Number of non-respondents		1		2		3		4		5	
Missing mechanism	Treatment	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
	Null tie imputations	0.961	0.148	0.893	0.239	0.760	0.322	0.507	0.336	0.356	0.280
Random	Reconstuction	0.988	0.062	0.970	0.114	0.944	0.167	0.914	0.211	0.860	0.262
missing	Mode	0.924	0.145	0.770	0.295	0.563	0.352	0.301	0.372	0.409	0.375
mechanism	Reconstruction plus mode	0.988	0.062	0.968	0.121	0.940	0.170	0.895	0.234	0.786	0.328
	Complete Case	0.942	0.126	0.936	0.185	0.877	0.225	0.859	0.304	0.737	0.360
Missing	Null tie imputations	0.972	0.132	0.914	0.219	0.768	0.326	0.510	0.355	0.342	0.291
mechanism	Reconstuction	0.991	0.053	0.976	0.112	0.957	0.153	0.929	0.204	0.872	0.275
based	Mode	0.909	0.162	0.771	0.291	0.573	0.352	0.314	0.382	0.412	0.379
on	Reconstruction plus mode	0.991	0.053	0.978	0.102	0.952	0.160	0.900	0.242	0.771	0.356
outdegree	Complete Case	0.937	0.130	0.926	0.196	0.877	0.223	0.861	0.297	0.741	0.354
Missing	Null tie imputations	0.959	0.138	0.896	0.230	0.759	0.311	0.528	0.324	0.366	0.262
mechanism	Reconstuction	0.984	0.081	0.964	0.128	0.940	0.167	0.907	0.212	0.852	0.258
based	Mode	0.908	0.160	0.766	0.291	0.589	0.346	0.346	0.378	0.419	0.374
on	Reconstruction plus mode	0.984	0.081	0.965	0.123	0.937	0.171	0.894	0.226	0.794	0.312
indegree	Complete Case	0.935	0.133	0.927	0.199	0.847	0.261	0.792	0.372	0.270	0.416

Table 7.9: Mean values and standard deviations for $ErrrB$ for the simulations of the completely symmetric blockmodel structure

Number of non-respondents		1		2		3		4		5	
Missing mechanism	Treatment	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
	Null tie imputations	0.022	0.083	0.071	0.140	0.168	0.172	0.292	0.167	0.339	0.164
Random	Reconstuction	0.003	0.023	0.010	0.060	0.024	0.094	0.051	0.132	0.115	0.171
missing	Mode	0.014	0.059	0.077	0.144	0.220	0.218	0.387	0.187	0.478	0.092
mechanism	Reconstruction plus mode	0.003	0.023	0.009	0.056	0.023	0.090	0.047	0.130	0.128	0.190
	Complete Case	0.249	0.249	0.251	0.247	0.253	0.239	0.256	0.231	0.257	0.208
Missing	Null tie imputations	0.016	0.073	0.061	0.131	0.164	0.175	0.301	0.173	0.354	0.168
mechanism	Reconstuction	0.002	0.023	0.007	0.060	0.014	0.085	0.040	0.132	0.114	0.197
based	Mode	0.013	0.067	0.072	0.142	0.212	0.213	0.382	0.183	0.471	0.089
on	Reconstruction plus mode	0.002	0.023	0.005	0.045	0.015	0.084	0.044	0.143	0.134	0.210
outdegree	Complete Case	0.250	0.249	0.251	0.247	0.253	0.238	0.252	0.227	0.260	0.204
Missing	Null tie imputations	0.027	0.083	0.075	0.133	0.164	0.158	0.276	0.145	0.321	0.143
mechanism	Reconstuction	0.009	0.050	0.019	0.079	0.031	0.098	0.062	0.129	0.123	0.162
based	Mode	0.020	0.075	0.091	0.153	0.222	0.216	0.391	0.181	0.483	0.083
on	Reconstruction plus mode	0.009	0.050	0.018	0.075	0.032	0.099	0.059	0.137	0.133	0.184
indegree	Complete Case	0.250	0.246	0.249	0.241	0.251	0.220	0.257	0.220	0.333	0.162

struction and the combination of reconstruction and imputations based on mode are the only two missing data treatments that permit the establishment of accurate blockmodels (according to position membership and positions of blocks in the blockmodel)

across all range of introduced non-respondents. However, for one non-respondent there are slightly larger differences in the mean values of the Adjusted Rand Index for the five missing data treatments.



(a) Adjusted Rand Index, $mARI$ (b) Incorrect block types, $mErrB$

Figure 7.25: Results of the simulation study based on the completely symmetric blockmodel structure for data missing based on outdegree (solid lines) and predictions according to linear regression model (dash lines)

The extended figures with reciprocity values are presented in Figure 7.26. According to the above findings, the figures on all panels are similar to those for randomly selected missing actors presented in Figure 7.24. The high symmetry according to reciprocity values leads to stable blockmodels for all missing data treatments except for imputations based on mode. Where there is only one non-respondent in the network and the reciprocity of the whole network is 1, the values of $mARI$ are approximately 0.8 for the imputations based on mode, while the $mARI$ values for all other treatments are equal to one.

Randomly missing data based on indegree

The results for simulations when the probability of an actor being a non-respondent depends on its indegree are shown in Figure 7.3.3.1. For one non-respondent there are, again, slightly more differences between the five treatments compared to randomly

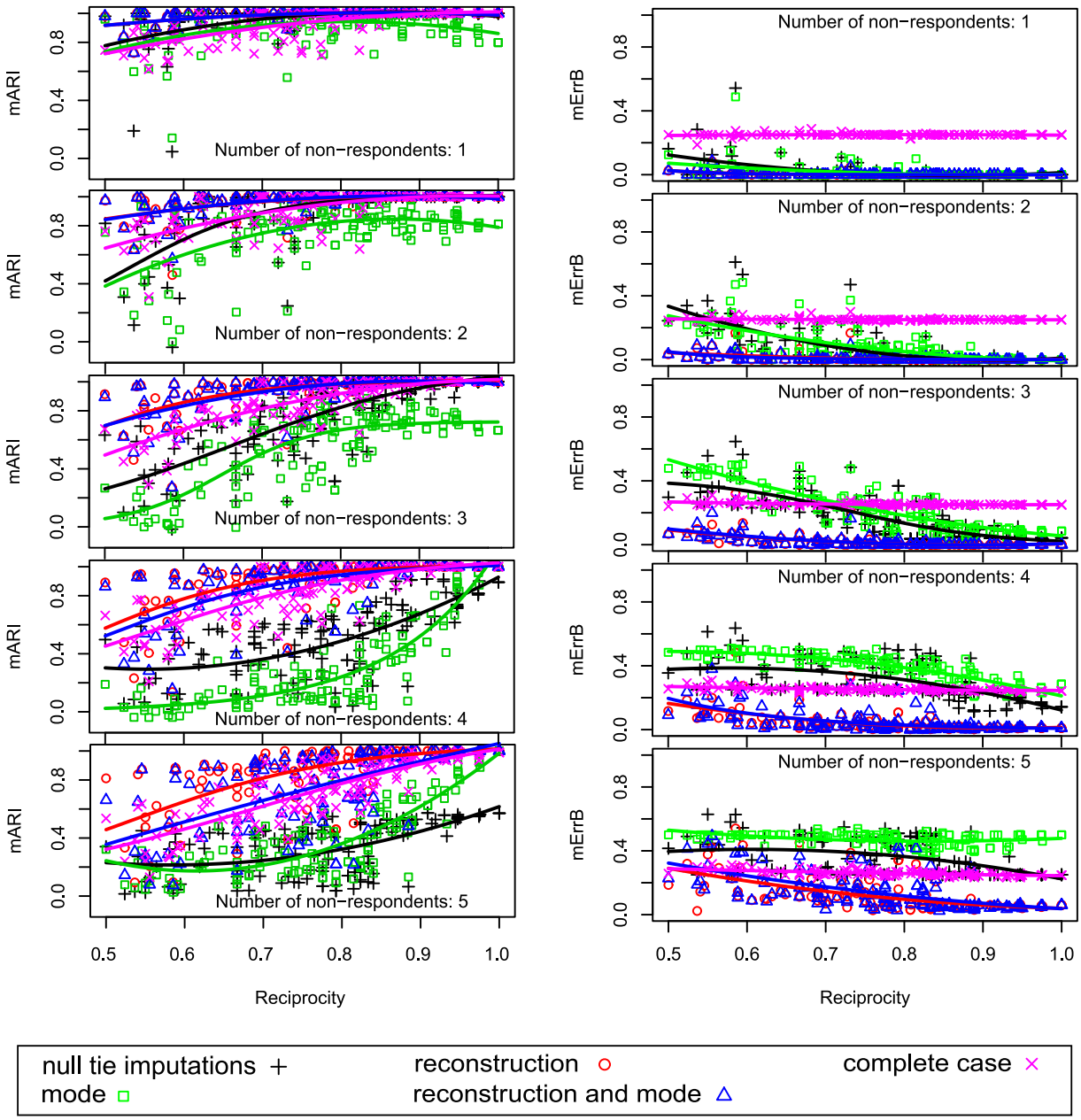
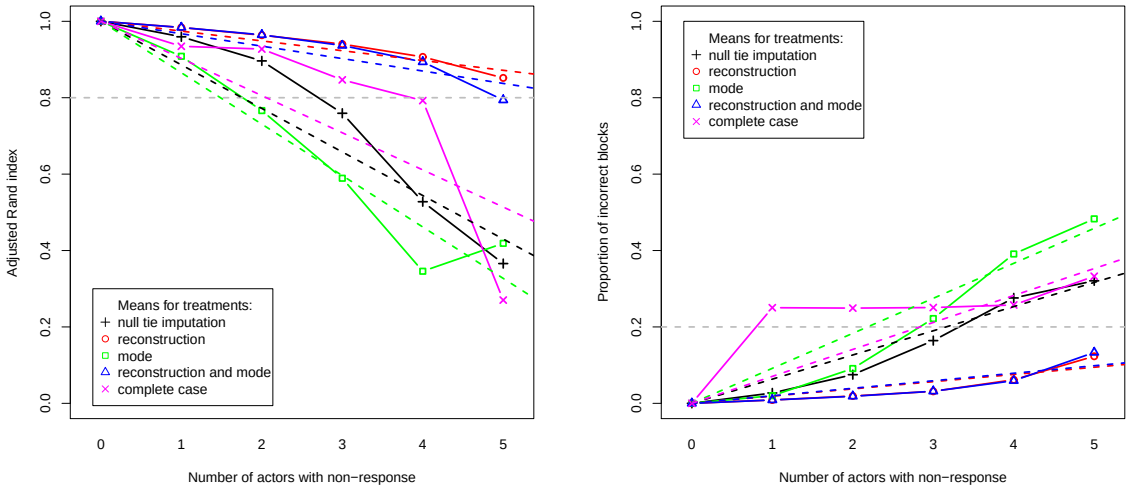


Figure 7.26: The mean of the Adjusted Rand Index, $mARI$ (left) and the mean of the Proortion of Incorrect block types, $mErrB$ (right) for completely symmetric blockmodel structure and missing mechanism based on outdegree

missing data. Taken as a whole the patterns in the results are similar as for the MCAR case. The biggest differences are in the case of complete-case approach where its performance is the worst for five non-respondents. The complete-case approach is the worst treatment also in terms of correctly identified block types (Figure 7.27(b)).



(a) Adjusted Rand Index, $mARI$

(b) Incorrect block types, $mErrB$

Figure 7.27: Results of the simulation study based on the completely symmetric blockmodel structure for data missing based on indegree (solid lines) and predictions according to linear regression model (dash lines)

Figure 7.28 presents both indices of blockmodel stability plotted versus the reciprocity values. In general, lower reciprocity values of real whole networks lead to less stable blockmodels with lower values of $mARI$. The complete-case approach for five non-respondents is the worst treatment for symmetrical whole starting networks (with reciprocity values higher than 0.8). For less symmetrical whole starting networks with five non-respondents the worst treatment in terms of $mARI$ is imputation based on mode, which is the worst treatment also for lower number of non-respondents. The impact of reciprocity values for correctly identified block types is the smallest for complete-case approach where $mErrB$ values are around 0.33 for all range of reciprocity values for all treatments.

It seems that having a very clear structural signal of the completely symmetric blockmodel structure (with regard to reciprocity values and also starting blockmodel structure presented in Equation 6.1 on page 77) is the main reason for the similarity of the results for the three ways of generating non-response missing data. The impact of non-response mechanisms and also different treatments to the indices of network stability

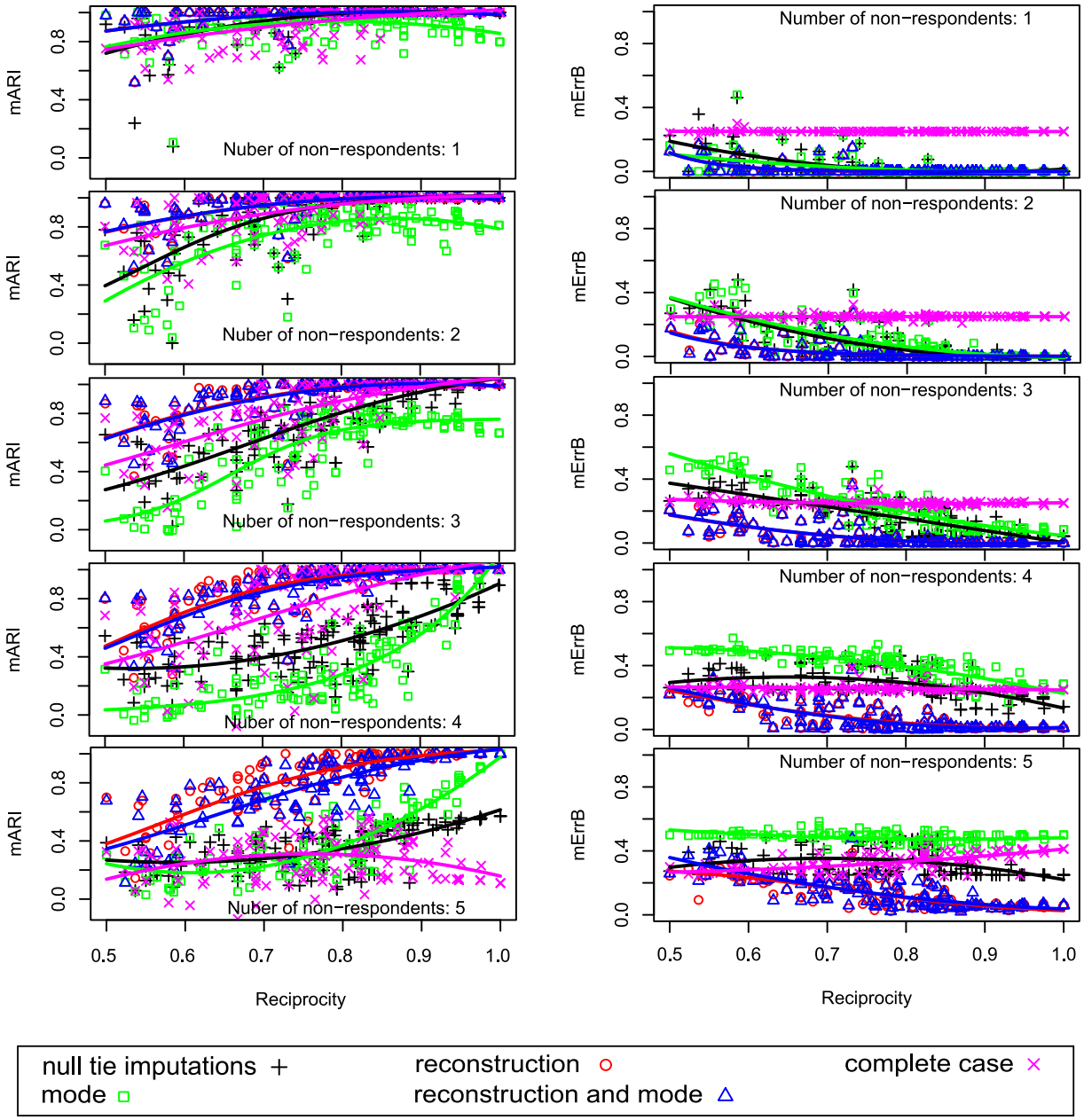


Figure 7.28: The mean of the Adjusted Rand Index, $mARI$ (left), and the mean of the Proportion of Incorrect block types, $mErrB$ (right), for completely symmetric block-model structure and missing mechanism based on indegree

was investigated also with established multiple regression models which are presented below.

Establishment of multiple regression models

The factorial design has 75 cells, similarly as for the boy-girl liking ties network (Sec-

tion 7.3.2.1). The multiple regression models for ARI and $ErrB$ were established due to unequal variances in cells instead of anova. The model summary in Table 7.10 shows that our regression model explains 30% of variation in ARI. The regression model for ARI can be with estimated (unstandardized) coefficients (Figure 7.29) written as follows:

$$\begin{aligned} \hat{Y}_{ARI} &= 0.9775 - 0.1022 \cdot n.actor + 0.3282 \cdot T\_RE - 0.1083 \cdot T\_MO + \\ &+ 0.2980 \cdot T\_REMO + 0.1944 \cdot T\_CC + 0.0047 \cdot MM\_out + \\ &- 0.0281 \cdot MM\_in \ . \end{aligned} \tag{7.11}$$

All variables in a model for ARI are significant, because p-values are 0.000 (Table 7.10). The established model for completely symmetric blockmodel structure is similar to the regression model of the boy-girl liking ties network (Figure 7.13). This is not a surprise, because the boy-girl liking ties network was in fact the base for simulated whole networks of completely symmetric blockmodel structure.

Table 7.10: Model summary and coefficients of regression analysis for ARI with data for the completely symmetric blockmodel structure

					95% confidence	
	Estimate	Std. Error	t value	Pr(>|t|)	interval for b	
(Intercept)	0.9775	0.0016	622.80	0.0000	0.9744	0.9806
n_actor	-0.1022	0.0003	-304.75	0.0000	-0.1029	-0.1015
T_RE	0.3282	0.0011	290.34	0.0000	0.3259	0.3304
T_MO	-0.1083	0.0011	-95.85	0.0000	-0.1105	-0.1061
T_REMO	0.2980	0.0011	263.63	0.0000	0.2958	0.3002
T_CC	0.1944	0.0011	172.00	0.0000	0.1922	0.1966
MM_out	0.0047	0.0009	5.41	0.0000	0.0030	0.0065
MM_in	-0.0281	0.0009	-32.08	0.0000	-0.0298	-0.0264
Residual standard error: 0.302 on 713992 degrees of freedom						
Multiple R^2: 0.310					Adjusted R^2: 0.310	
F-statistic: 45833.3 (on 7 and 71399 df)					p-value: 0.000	

The regression coefficients b (Table 7.10) can be interpreted as follows:

- **n.actor:** ($b = -0.1022$) If the number of non-respondents increases for one non-respondent, the values of ARI decrease for 0.1022 if other effects are held constant. This result is expected; the higher number of non-respondents leads to less stable blockmodel in terms of restored partitions of actors.

- **T_RE vs. T_NTI:** ($b = 0.3282$) If the reconstruction treatment is used, compared to the null tie imputation the value of ARI increases for 0.3282. In comparison to the combination of reconstruction and mode imputation treatment, which are practically undistinguishable on previous figures, the reconstruction treatment has a little higher effect on values of ARI according to established regression model.

- **T_MO vs. T_NTI:** ($b = -0.1083$) The mode treatment turns out to be the worst treatment. The values of ARI decrease for 0.1083 when imputations based on mode are used instead of the null tie imputation.

- **T_REMO vs. T_NTI:** ($b = 0.2980$) In comparison with the null tie imputations, the values of ARI in that case are higher for 0.2980. The reconstruction combined with mode imputation seems to be the second best treatment in terms of ARI values and position agreement between actors in a blockmodel.

- **T_MO vs. T_CC:** ($b = 0.1944$) The major difference compared to the regression model for the data for the the boy-girl liking ties network (Figure 7.13) is that the complete-case approach is not the best treatment. The use of the complete-case approach instead of the null tie imputations increases values of ARI for 0.1944.

- **MM_out vs. MM_random:** ($b = 0.0047$) The use of the missing mechanism based on outdegree has practically no effect. The values of ARI increase for just 0.0047, if the missing mechanism based on outdegree is used instead of randomly selected non-respondents. If the actors with low outdegree (inactive actors) have higher probabilities to be non-respondents, the partitions of actors in blockmodel are recovered a little more accurately.

- **MM_in vs. MM_random:** ($b = -0.0281$) The use of the missing mechanism based on indegree instead of the random missing mechanism decreases the value for ARI for 0.0281.

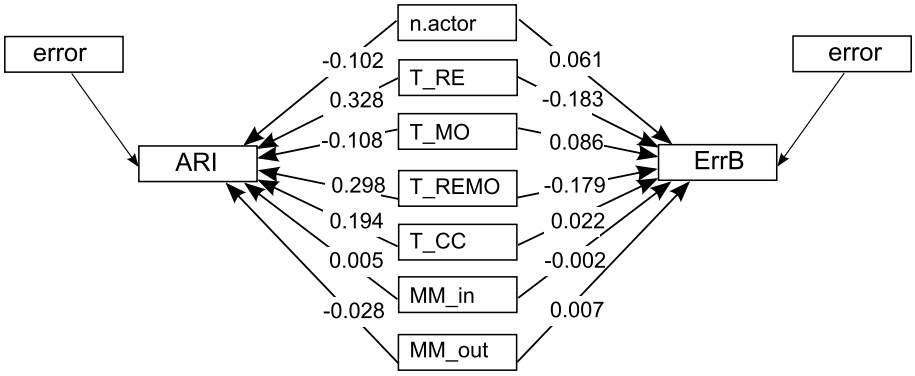


Figure 7.29: Regression models for ARI and $ErrB$ with data for the completely symmetric blockmodel structure

The regression model for $ErrB$ can be with estimated (unstandardized) coefficients (Figure 7.29) written as follows:

$$\begin{aligned} \hat{Y}_{ErrB} &= 0.0112 + 0.0609 \cdot n.actor - 0.1825 \cdot T\_RE + 0.0859 \cdot T\_MO + \\ &- 0.1786 \cdot T\_REMO + 0.0220 \cdot T\_CC - 0.0022 \cdot MM\_out + \\ &+ 0.0066 \cdot MM\_in \ . \end{aligned} \tag{7.12}$$

All variables in a model for $ErrB$ are significant, because p-values are 0.000 (Table 7.11). The regression coefficients b can be interpreted as follows:

- **n.actor:** ($b = 0.0609$) If the number of non-respondents increase for one non-respondent, the values of $ErrB$ increase for 0.0609. Higher number of non-respondents leads to less stable blockmodel structure according to types and positions of blocks.

- **T_RE vs. T_NTI:** ($b = -0.1825$) The value of $ErrB$ decreases for 0.1825 if the reconstruction treatment is used instead of the null tie imputation. The negative effect among all treatments is the lowest, which indicates that the reconstruction treatment is the best one also in terms of correctly identified block types in the blockmodel.

- **T_MO vs. T_NTI:** ($b = 0.0859$) Compared to the null tie imputations the shift in the change in $ErrB$ values is positive if the imputations based on mode is used. The mode imputations are the worst treatment according to $mErrB$.

- **T_REMO vs. T_NTI:** ($b = -0.1786$) In comparison with the null tie imputations, the values of $ErrB$ are lower for 0.1786 when combination of reconstruction and imputations based on mode are used. This indicates that the combination of reconstruction and mode imputation is the second best treatment.

- **T_MO vs. T_CC:** ($b = 0.0220$) The absolute value of b coefficient is small compared to other variables for treatments, which means that values of $ErrB$ are the most similar when the null tie imputation and complete-case approach are used.

- **MM_out vs. MM_random:** ($b = -0.0022$) The use of the missing mechanism based on outdegree instead of the random missing mechanism has a small negative effect on values of $ErrB$. In that case values of $ErrB$ decrease for 0.0022.

- **MM_in vs. MM_random:** ($b = 0.0066$) The use of the missing mechanism based on indegree instead of the random missing mechanism increases the value of $ErrB$ for 0.0066. Both non-random missing mechanisms have similar effects.

The histograms in Figure 7.30 show that the assumption of normally distributed residuals is violated. Therefore, the results can not be generalized beyond our sample.

7.3.3.2 Results for the first non-symmetric blockmodel structure

The first non-symmetric blockmodel has 9 blocks with three complete blocks on the diagonal and one complete block in the lower left corner. According to selected combination of probabilities of ties in (near) complete and null blocks we generated 80 whole starting networks with 15 actors. The simulation of networks together with their main characteristics (density and reciprocity) is presented in Section 6.2.3.2.

Similarly, as for the note borrowing network the factorial design has 90 cells (3 different non-response mechanisms times 5 non-response data treatments times 6 different

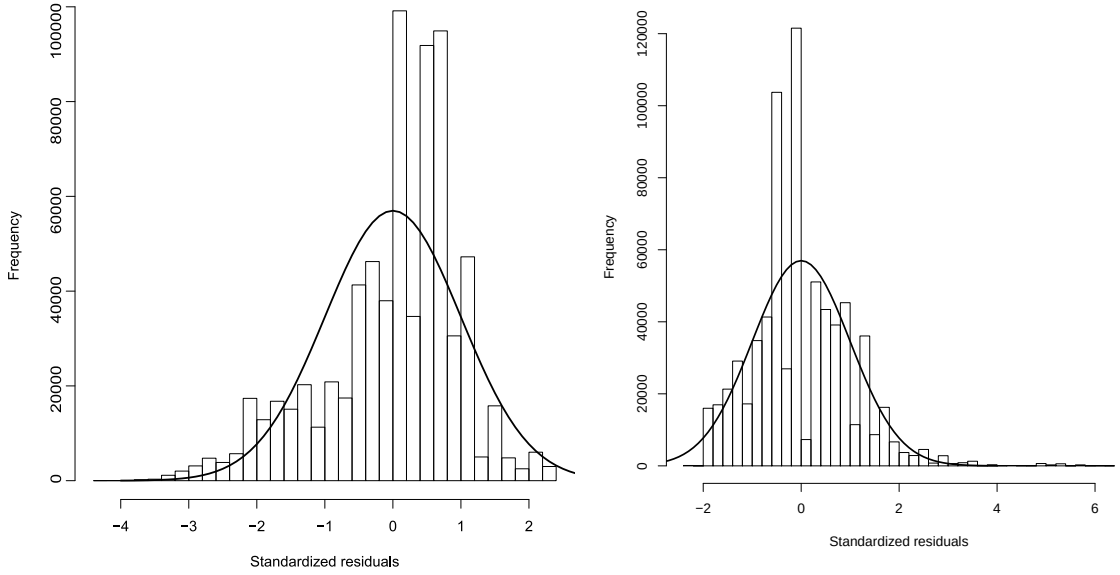
Table 7.11: Model summary and coefficients of regression analysis for $ErrB$ with data for the completely symmetric blockmodel structure

	Estimate	Std. Error	t value	Pr(>|t|)	95% confidence interval for b	
(Intercept)	0.0112	0.0009	12.36	0.0000	0.0094	0.0129
n_actor	0.0609	0.0002	315.49	0.0000	0.0605	0.0612
T_RE	-0.1825	0.0007	-280.73	0.0000	-0.1838	-0.1812
T_MO	0.0859	0.0007	132.10	0.0000	0.0846	0.0872
T_REMO	-0.1786	0.0007	-274.73	0.0000	-0.1799	-0.1773
T_CC	0.0220	0.0007	33.85	0.0000	0.0207	0.0233
MM_out	-0.0022	0.0005	-4.33	0.0000	-0.0032	-0.0012
MM_in	0.0066	0.0005	13.08	0.0000	0.0056	0.0076
Residual standard error: 0.174 on 713992 degrees of freedom						
Multiple R^2: 0.350					Adjusted R^2: 0.350	
F-statistic: 54992.4 (on 7 and 71399 df)					p-value: 0.000	

numbers of non-respondents). The results are presented below.

Data missing completely at random

Figure 7.31 (solid lines) present results of $mARI$ and $mErrB$ for the first non-symmetric blockmodel structure where non-respondents were simulated at random. The complete-case approach is the best treatment and it is acceptable with regard to both position membership identification and block type identification for whole range of introduced non-respondents. Agreement between partitions of real whole blockmodel and treated blockmodel is acceptable ($mARI \geq 0.2$) for all treatments except the null tie imputations if the number of non-respondents is two or less. In contrast, all treatments are acceptable in terms of $mErrB$ for one to five non-respondents. If there are six non-respondents in the network, the actor response rate is 60%. In that case both reconstruction and the imputation based on mode have $mErrB$ values above 0.2 which means that almost 2 blocks out of 9 are incorrectly identified. The null tie imputation and imputa-



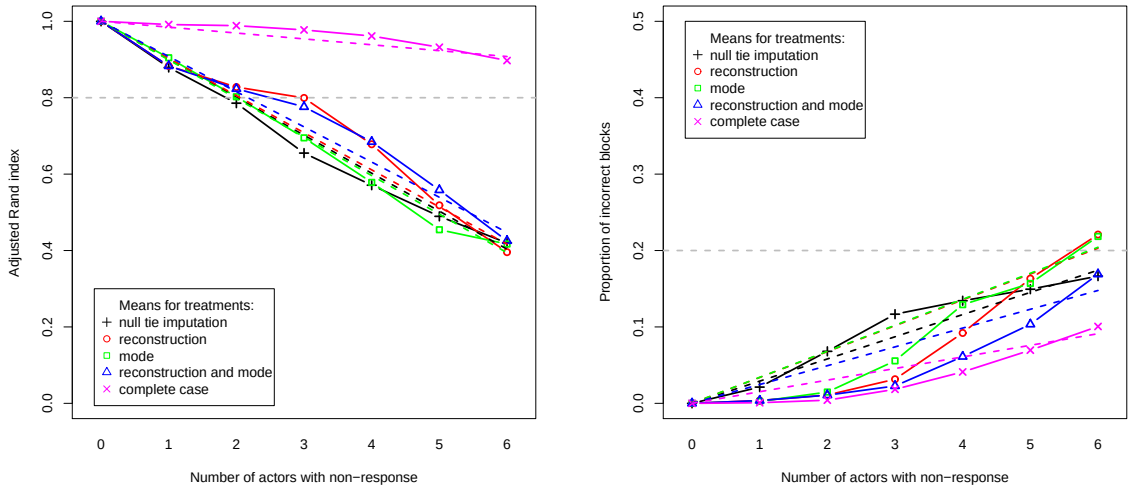
(a) Model for ARI (b) Model for $ErrB$

Figure 7.30: Histogram of standardized residuals of regression models with data for the completely symmetric blockmodel structure

tion based on the mode have the worst performance regarding $mARI$ for whole range of introduced non-response. For one to four non-respondents they are also the worst regarding to proportion of incorrectly identified block types.

Dash lines (Figure 7.31) present predictions from linear regression models. The prediction of the Adjusted Rand Index for the complete-case approach fits well to the observed data. The slope coefficient is highest for the complete-case approach ($\beta = -0.015$) and it is statistically significant higher than $\beta_0^{ARI} = -0.0\bar{3}$ (Table A.1). As written above other four treatments perform worse and the slope coefficients from linear regression model are statistically significant lower than $\beta_0^{ARI} = -0.0\bar{3}$. The lowest slope coefficient among all linear regresion models for $ErrB$ (A.2) also has the complete-case approach ($\beta = 0.015$).

Figure 7.32 where both indices of blockmodeling stability are plotted versus the reciprocity values shows that the complete-case approach is the best treatment irrespective of the level of symmetry of the whole network or number of actors. The differences between complete-case approach and other treatments are higher at position



(a) Adjusted Rand Index, $mARI$

(b) Incorrect block types, $mErrB$

Figure 7.31: Results of the simulation study based on the first non-symmetric block-model structure for data missing completely at random (solid lines) and predictions according to linear regression model (dash lines)

membership agreement ($mARI$) than at proportion of incorrectly identified block types ($mErrB$).

Randomly missing data based on outdegree

The results of the simulation for selection of non-respondents based on their outdegree are presented in Figure 7.33(a). Regarding the identification of position memberships ($mARI$), all five treatment methods are acceptable for one non-respondent and all are unacceptable for six non-respondents. The complete-case approach is the best treatment according to the position membership with acceptable values of $mARI$ for five non-respondents or less. Perhaps somewhat surprisingly, the null tie imputation is acceptable for four non-respondents or less and is the second best treatment in the stability of blockmodeling according to partitions. Having the complete-case dominate is consistent with its performance thus far, and the improved performance of the null tie imputation seems due to the increased presence of reciprocal null ties. For more than three non-respondents all of the other three treatment regimes lead to unacceptable results. Both treatments with reconstruction procedure show similar patterns and

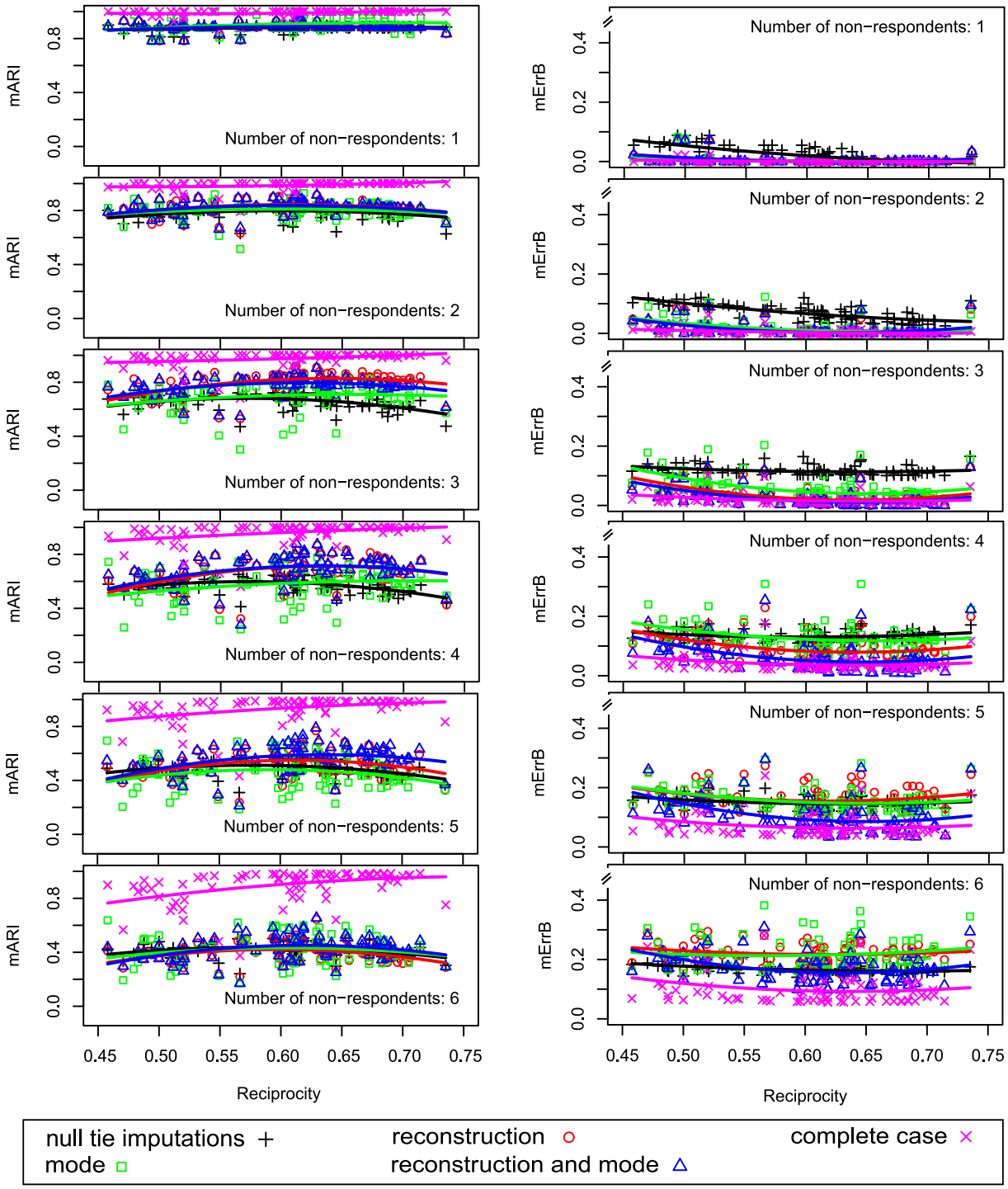
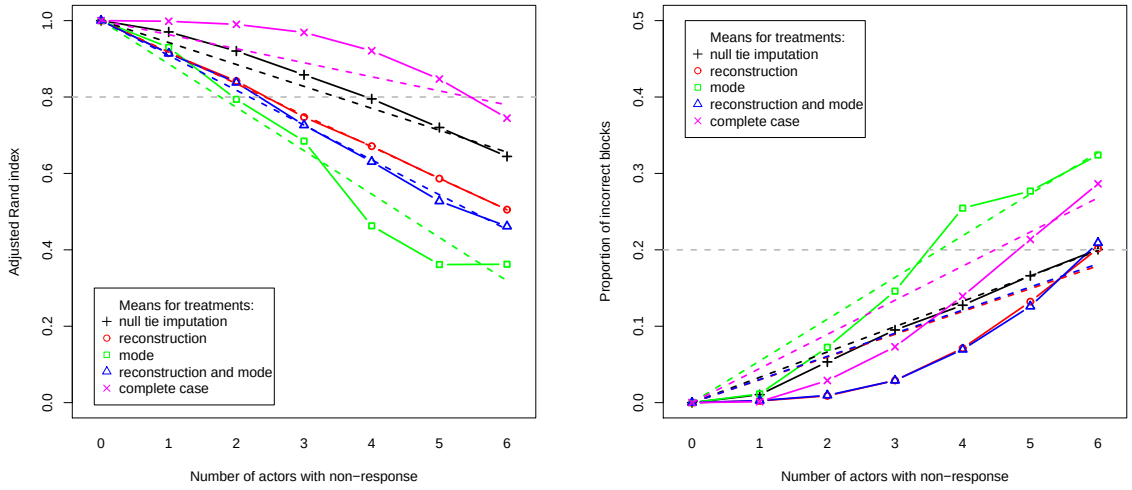


Figure 7.32: The mean of the Adjusted Rand Index, $mARI$ (left), and the mean of the Proportion of Incorrect block types, $mErrB$ (right), for the first non-symmetric block-model structure and random missing mechanism

the imputation based on mode is the worst treatment according to $mARI$ values. Dash lines in Figure 7.33(a) indicate that linear regression models for ARI fit well to the

observed data in the case of null tie imputation and both reconstruction procedures. The best treatment according to the highest slope coefficient is again the complete-case approach ($\beta = -0.037$), although the slope coefficient is statistically significant lower than $\beta_0^{ARI} = -0.0\bar{3}$ (Table A.1).



(a) Adjusted Rand Index, $mARI$

(b) Incorrect block types, $mErrB$

Figure 7.33: Results of the simulation study based on the first non-symmetric block-model structure for data missing based on outdegree (solid lines) and predictions according to linear regression model (dash lines)

The worst performance according to incorrectly identified blocks (Figure 7.33(b)) is the imputation based on mode. For four non-respondents $mErrB$ exceeds 0.2 and reaches 0.3 for six non-respondents for which 3 out of 9 blocks are identified incorrectly. All treatment methods lead to unacceptable results when six non-respondents are introduced in the network (Table 7.13). The complete-case approach becomes unacceptable with five non-respondents. The best performances regarding $mErrB$ follow the use of the reconstruction and the combined use of reconstruction and imputations based on mode. It is of interest that these two methods are among the worse treatments when looking at $mARI$.

Dash lines in Figure 7.33(b) represent predictions from the linear regression models for

$ErrB$. The complete-case approach and both reconstruction procedures indicate that linear model is not the best one, because mean values of Incorrect block types have nonlinear pattern. The complete-case approach is not the best treatment as in the case of ARI, because the slope coefficient from the linear model is equal to $\beta = 0.045$ and it is statistically significant higher that testing value $\beta_0^{ErrB} = 0.0\bar{3}$ (Table A.2).

Table 7.12: Mean values and standard deviations for ARI for the first non-symmetric blockmodel structure

Number of non-respondents		1		2		3		4		5		6	
Missing mechanism		Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
	Null tie imputations	0.879	0.120	0.786	0.154	0.655	0.175	0.571	0.177	0.489	0.171	0.420	0.158
Random	Reconstuction	0.883	0.121	0.828	0.161	0.799	0.207	0.678	0.262	0.518	0.267	0.396	0.204
missing	Mode	0.905	0.125	0.803	0.173	0.695	0.206	0.578	0.244	0.454	0.218	0.417	0.223
mechanism	Reconstr. plus mode	0.883	0.121	0.823	0.163	0.776	0.196	0.685	0.251	0.558	0.253	0.425	0.241
	Complete Case	0.992	0.035	0.989	0.060	0.977	0.091	0.961	0.127	0.932	0.171	0.897	0.213
missing	Null tie imputations	0.970	0.077	0.920	0.130	0.858	0.172	0.795	0.192	0.720	0.208	0.644	0.215
mechanism	Reconstuction	0.915	0.120	0.842	0.150	0.747	0.183	0.671	0.202	0.586	0.208	0.505	0.199
based	Mode	0.930	0.111	0.794	0.204	0.685	0.264	0.463	0.229	0.361	0.174	0.362	0.148
on	Reconstr. plus mode	0.914	0.122	0.838	0.155	0.726	0.186	0.631	0.211	0.527	0.204	0.462	0.203
outdegree	Complete Case	0.998	0.016	0.990	0.061	0.969	0.106	0.921	0.164	0.847	0.218	0.745	0.262
missing	Null tie imputations	0.909	0.116	0.833	0.139	0.724	0.158	0.637	0.161	0.550	0.149	0.464	0.135
mechanism	Reconstuction	0.971	0.079	0.930	0.127	0.882	0.169	0.804	0.223	0.683	0.257	0.553	0.241
based	Mode	0.910	0.114	0.838	0.169	0.740	0.211	0.618	0.234	0.471	0.181	0.438	0.175
on	Reconstr. plus mode	0.971	0.079	0.929	0.128	0.879	0.165	0.809	0.207	0.716	0.233	0.587	0.242
indegree	Complete Case	0.995	0.030	0.988	0.059	0.965	0.104	0.924	0.145	0.873	0.177	0.815	0.205

The performances of missing data treatments according to number of non-respondents and reciprocity of the whole network are presented in Figure 7.34. Compared to the random missing mechanism the difference between complete-case approach and other treatments is smaller.

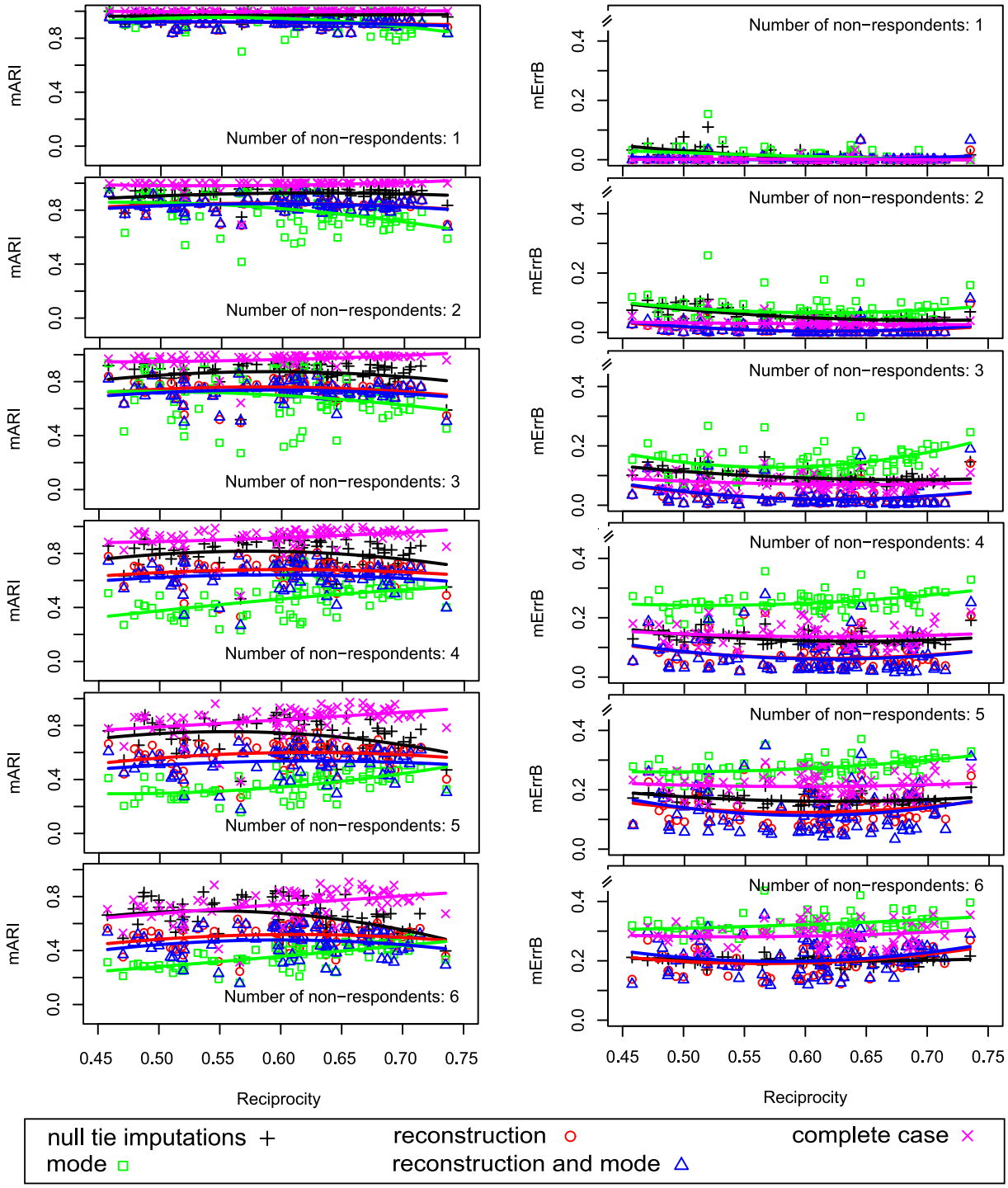


Figure 7.34: The mean of the Adjusted Rand Index, *mARI* (left), and the mean of the Proportion of the Incorrect block types, *mErrB* (right), for the first non-symmetric blockmodel structure and missing mechanism based on outdegree

Table 7.13: Mean values and standard deviations for $ErrB$ for the first non-symmetric blockmodel structure

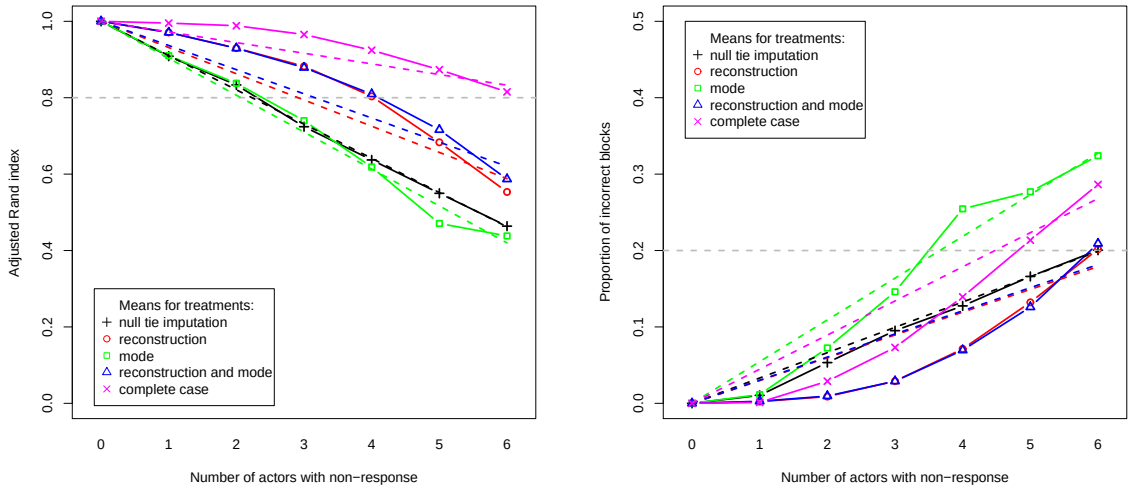
Number of non-respondents		1		2		3		4		5		6	
Missing mechanism	Treatment	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
	Null tie imputations	0.021	0.043	0.068	0.064	0.117	0.047	0.134	0.047	0.149	0.055	0.167	0.059
Random	Reconstuction	0.004	0.021	0.011	0.042	0.032	0.073	0.092	0.110	0.163	0.113	0.221	0.091
missing	Mode	0.003	0.016	0.015	0.052	0.056	0.101	0.129	0.135	0.157	0.129	0.218	0.129
mechanism	Reconstr. plus mode	0.004	0.021	0.011	0.044	0.023	0.065	0.061	0.103	0.103	0.124	0.169	0.120
	Complete Case	0.001	0.007	0.004	0.026	0.018	0.056	0.041	0.081	0.070	0.113	0.101	0.131
missing	Null tie imputations	0.010	0.032	0.053	0.057	0.095	0.054	0.128	0.043	0.166	0.059	0.200	0.048
mechanism	Reconstuction	0.002	0.022	0.009	0.038	0.029	0.070	0.071	0.098	0.132	0.104	0.202	0.092
based	Mode	0.011	0.044	0.072	0.118	0.146	0.137	0.255	0.120	0.277	0.109	0.324	0.093
on	Reconstr. plus mode	0.003	0.025	0.010	0.041	0.029	0.075	0.070	0.110	0.126	0.129	0.209	0.125
outdegree	Complete Case	0.001	0.012	0.029	0.057	0.073	0.107	0.139	0.139	0.214	0.159	0.286	0.157
missing	Null tie imputations	0.020	0.042	0.094	0.052	0.123	0.043	0.137	0.049	0.157	0.056	0.181	0.056
mechanism	Reconstuction	0.004	0.021	0.016	0.044	0.036	0.066	0.075	0.090	0.132	0.104	0.187	0.105
based	Mode	0.013	0.040	0.081	0.119	0.141	0.129	0.234	0.126	0.252	0.105	0.303	0.103
on	Reconstr. plus mode	0.004	0.021	0.016	0.045	0.034	0.064	0.065	0.087	0.103	0.101	0.160	0.107
indegree	Complete Case	0.002	0.016	0.031	0.058	0.072	0.108	0.136	0.142	0.198	0.158	0.252	0.157

Randomly missing data based on indegree

The results of simulations for probabilities of non-response being conditioned by indegree and outdegree show considerable similarities (Figure 7.35). Again, the complete-case approach leads to the best performances regarding position membership identification ($mARI$) for whole range of non-respondents. Both the reconstruction and the combination of reconstruction and the imputations based on mode lead to acceptable results for four non-respondents or less. If the number of non-respondents is higher than two, both the null tie imputation and the imputation based on the mode lead to unacceptable agreement between partitions.

Similarly as in the case of randomly selected non-respondents, the complete-case approach is the only acceptable treatment according to the slope coefficient from the linear models ($\beta = -0.028$). As written above, the worst treatments are the null tie imputations ($\beta = -0.090$) and the imputations based on mode with slope coefficient from linear model equal to $\beta = -0.097$ (Table A.1).

Results of correctly identified block types ($mErrB$) are also very similar to those when the probabilities for non-response were conditioned by outdegree (Figure 7.35(b)). The



(a) Adjusted Rand Index, $mARI$

(b) Incorrect block types, $mErrB$

Figure 7.35: Results of the simulation study based on the first non-symmetric blockmodel structure for data missing based on indegree (solid lines) and predictions according to linear regression model (dash lines)

combination of reconstruction and imputations based on mode for ties between non-respondents performs the best for whole range of non-respondents. Even though the null tie imputation is among the worst performers when $mARI$ is considered, it is the worst only for one and two non-respondents when correctly identified block types and positions are studied. The imputations based on mode are the worst treatment (for more than two non-respondents) also when $MErrB$ is considered.

If we compare mean values of $ErrB$ (solid lines) with the predictions (dash lines) we could say that the linear regression model is suitable only for the null tie imputations. The imputations based on mode has the highest slope coefficient ($\beta = 0.051$) and is therefore the worst treatment (Table A.2). The lowest $mErrB$ values and also the lowest slope coefficients of predictions have reconstruction and reconstruction plus imputations based on mode ($\beta = 0.028$ and $\beta = 0.024$, respectively) and are therefore the best treatments according to the correctly identified block in the blockmodels.

The redrawn figures with reciprocity values for missing mechanism based on indegree

are presented in Figure 7.36. Again, the complete-case approach is the best treatment according to values of $mARI$ regardless to the level of the symmetry in the network. The proportion of incorrectly identified block types in the blockmodel for one non-respondent shows that treated blockmodels with the null tie imputation or the imputations based on mode established from less symmetric networks (with lower reciprocity values) are less stable.

Establishment of multiple regression models

The factorial design has 90 cells, similarly as for the note borrowing network (Section 7.3.2.2). Due to unequal variances in cells for $mARI$ and $mErrB$ (Tables 7.12 and 7.13), multiple regression models were established instead of anova. The model summary for ARI in Table 7.14 shows that our regression model explains 35% of variation in ARI. The regression model for ARI can be with estimated (unstandardized) coefficients (Figure 7.44) written as follows:

$$\begin{aligned} \hat{Y}_{ARI} &= 0.9845 - 0.0849 \cdot n.actor + 0.0028 \cdot T\_RE - 0.1096 \cdot T\_MO + \\ &+ 0.0047 \cdot T\_REMO + 0.2843 \cdot T\_CC + 0.0240 \cdot MM\_out + \\ &+ 0.0624 \cdot MM\_in \, . \end{aligned} \tag{7.13}$$

All variables in a model for ARI are significant (p-values are 0.000 in Table 7.14). The established model for the first non-symmetric blockmodel structure is similar to the regression model of borrowing network (Figure 7.20). The regression coefficients b can be interpreted as follows:

- **n.actor:** ($b = -0.0849$) If the number of non-respondents increases for one non-respondent, the values of ARI decrease for 0.0849 if all other variables are held constant.

- **T_RE vs. T_NTI:** ($b = 0.0028$) If the reconstruction treatment is used, compared to the null tie imputation the values of ARI increase for 0.0028. The reconstruction procedure is better than the null tie imputations, but it is not the best treatment.

- **T_MO vs. T_NTI:** ($b = -0.1096$) The imputation based on mode is the worst treatment. The values of ARI decrease for 0.1096 when the imputations based on mode are used instead of the null tie imputations.

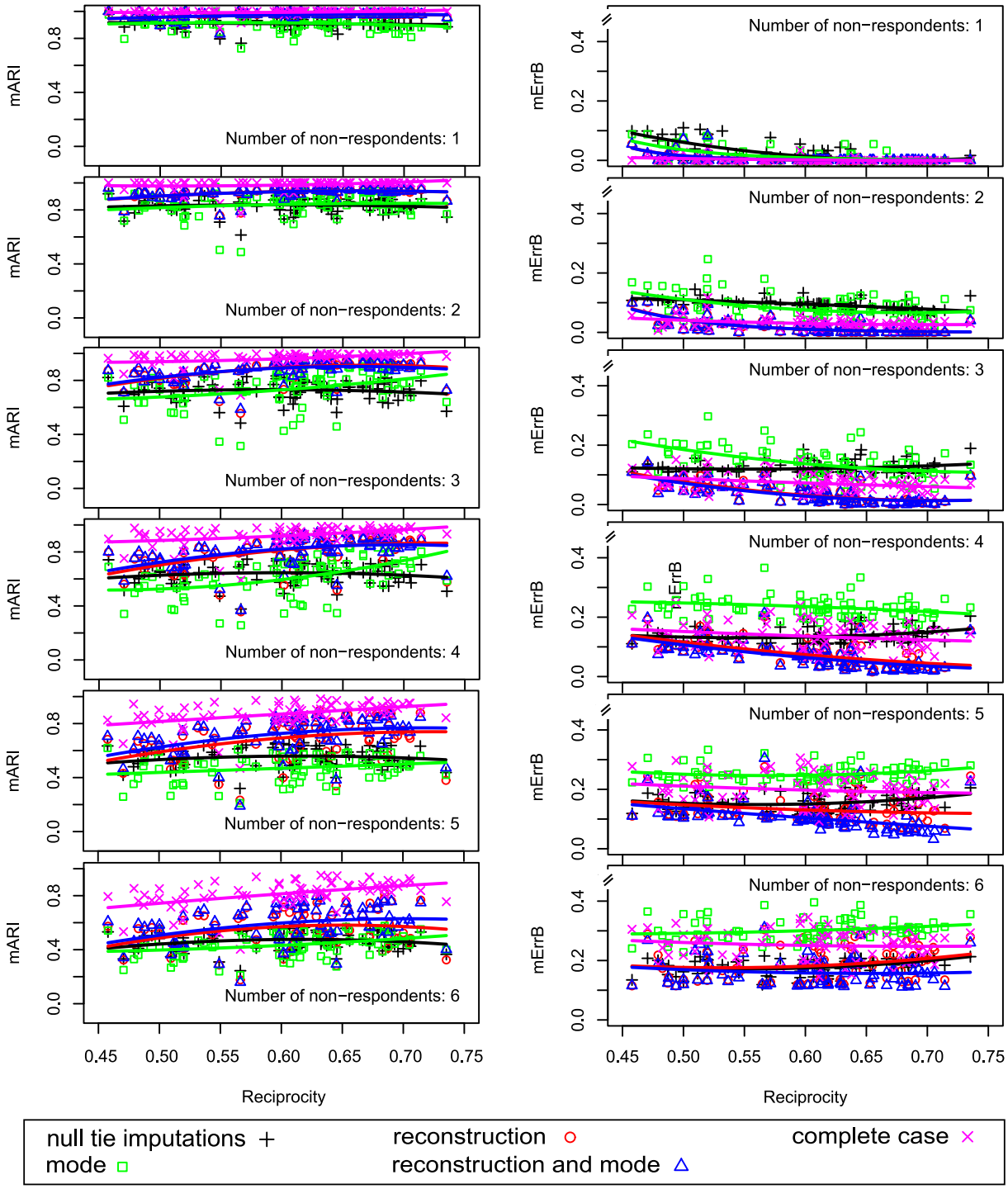


Figure 7.36: The mean of the Adjusted Rand Index, $mARI$ (left), and the mean of the Proportion of Incorrect block types, $mErrB$ (right), for the first non-symmetric block-model structure and missing mechanism based on indegree

- **T_REMO vs. T_NTI:** ($b = 0.0047$) The combination of the reconstruction procedure and the imputations based on mode performs slightly better than reconstruction itself. In comparison with the null tie imputations, the values of ARI in that case are higher for 0.0047.

- **T_MO vs. T_CC:** ($b = 0.2843$) The complete-case approach is the best treatment in terms of partition agreement in a blockmodeling. If the complete-case approach is used instead of the null tie imputations, the values of ARI increase for 0.2843.

- **MM_out vs. MM_random:** ($b = 0.0240$) Use of the missing mechanism based on outdegree has little positive effect on stability of blockmodeling in terms of partitions. The values of ARI increase for just 0.0240, if the missing mechanism based on outdegree is used instead of random selection of non-respondents.

- **MM_in vs. MM_random:** ($b = 0.0624$) Use of the missing mechanism based on indegree instead of the random missing mechanism increases the values of ARI for 0.0624. Both non-random missing mechanisms lead to a little bit more stable blockmodeling according to identification of position membership.

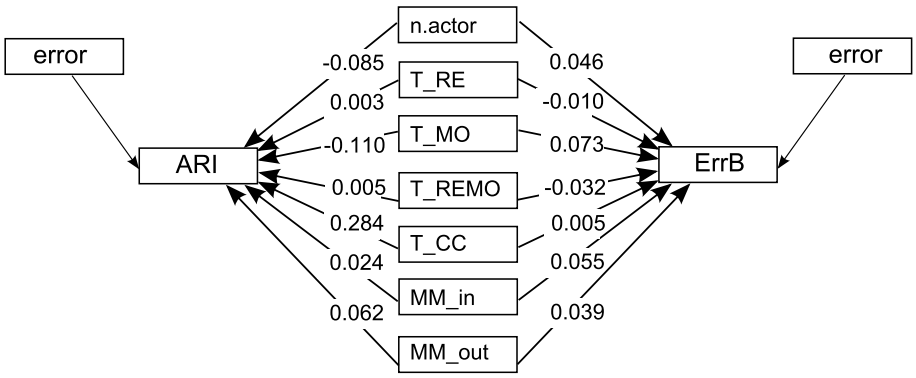


Figure 7.37: Regression models for ARI with data for the first non-symmetric blockmodel stucture

Table 7.14: Model summary and coefficients of regression analysis for ARI with data for the first non-symmetric blockmodel structure

	Estimate	Std. Error	t value	Pr(>|t|)	95% confidence interval for b	
(Intercept)	0.9845	0.0009	1074.93	0.0000	0.9827	0.9863
n_actor	-0.0849	0.0002	-553.81	0.0000	-0.0852	-0.0846
T_RE	0.0028	0.0006	4.71	0.0000	0.0016	0.0040
T_MO	-0.1096	0.0006	-183.58	0.0000	-0.1108	-0.1085
T_REMO	0.0047	0.0006	7.79	0.0000	0.0035	0.0058
T_CC	0.2843	0.0006	476.02	0.0000	0.2831	0.2854
MM_out	0.0240	0.0005	51.84	0.0000	0.0231	0.0249
MM_in	0.0624	0.0005	134.86	0.0000	0.0615	0.0633
Residual standard error: 0.213 on 1271992 degrees of freedom						
Multiple R^2: 0.3887					Adjusted R^2: 0.3887	
F-statistic: 115542.1 (on 7 and 1271992 df)					p-value: 0.000	

In the next step, the regression model for the proportion of incorrectly identified block types and positions was examined. Model summary in Table 7.15 shows that all variables in the model are significant (p-values are 0.000) and that our regression model explains 29% of variation in $ErrB$.

The regression model for the first non-symmetric blockmodel structure can be with estimated unstandardized coefficients (Figure 7.37) written as follows:

$$\begin{aligned} \hat{Y}_{ErrB} &= -0.1025 + 0.0456 \cdot n.actor - 0.0101 \cdot T\_RE + 0.0731 \cdot T\_MO + \\ &- 0.0319 \cdot T\_REMO + 0.0047 \cdot T\_CC + 0.0549 \cdot MM\_out + \\ &+ 0.0388 \cdot MM\_in \ . \end{aligned} \tag{7.14}$$

All variables in a model for $ErrB$ are significant (p-values are 0.000 in Table 7.10). The established model for the first non-symmetric blockmodel structure is similar to the regression model of borrowing network (Figure 7.20). The regression coefficients b can be interpreted as follows:

- **n.actor:** ($b = 0.0456$) If the number of non-respondents is increased for one non-respondent, the values of $ErrB$ increase for 0.0456 if all other variables are held constant.

- **T_RE vs. T_NTI:** ($b = -0.0101$) If the reconstruction treatment is used instead of the null tie imputation, the values of $ErrB$ increase for 0.0101. The reconstruction procedure is the second best treatment according to block structure recovery.

- **T_MO vs. T_NTI:** ($b = 0.0731$) The imputation based on mode is the worst treatment. The values of $ErrB$ increase for 0.0731 when imputations based on mode are used instead of the null tie imputations.

- **T_REMO vs. T_NTI:** ($b = -0.0319$) The combination of reconstruction procedure and imputations based on mode performs the best. In comparison with the null tie imputations the values of $ErrB$ in that case are higher for 0.0319.

- **T_MO vs. T_CC:** ($b = 0.0047$) The effects of the complete-case approach are the most similar to the null tie imputations. If the complete-case approach is used instead of null tie imputations, the values of $ErrB$ increase for 0.0047.

- **MM_out vs. MM_random:** ($b = 0.0549$) The use of the missing mechanism based on outdegree has little positive effect on stability of block types and positions in the blockmodel. The values of $ErrB$ increase for 0.0549, if the missing mechanism based on outdegree is used instead of randomly selected non-respondents.

- **MM_in vs. MM_random:** ($b = 0.0388$) The use of the missing mechanism based on indegree instead of the random missing mechanism increases the values for $ErrB$ for 0.0388. Both non-random missing mechanisms leads to a little bit more stable blockmodeling according to correctly identified blocks in the image matrix.

7.3.3.3 Results for the second non-symmetric blockmodel structure

The 80 whole networks with 15 actors were generated based on blockmodel structure presented in Equation 6.4 on page 80. The prototype for the generation of whole starting networks was the note borrowing network. The impact of non-response with data from that network is presented in Section 7.3.2.2. Similarly, as for the note borrowing

Table 7.15: Model summary and coefficients of regression analysis for $ErrB$ with data for the first non-symmetric blockmodel structure

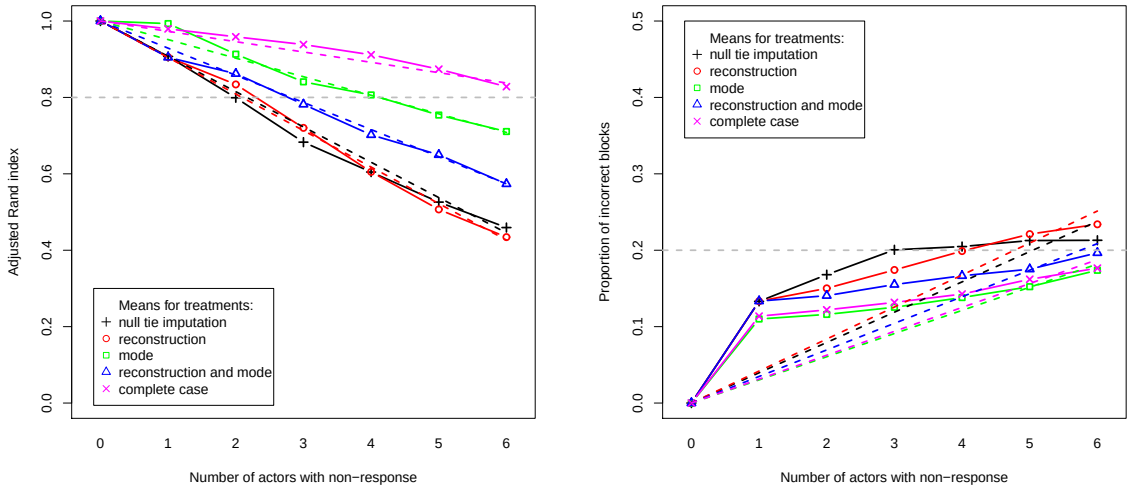
	Estimate	Std. Error	t value	Pr(>|t|)	95% confidence interval for b	
(Intercept)	-0.1025	0.0005	-215.10	0.0000	-0.1034	-0.1016
n_actor	0.0456	0.0001	572.51	0.0000	0.0455	0.0458
T_RE	-0.0101	0.0003	-32.47	0.0000	-0.0107	-0.0095
T_MO	0.0731	0.0003	235.31	0.0000	0.0725	0.0737
T_REMO	-0.0319	0.0003	-102.62	0.0000	-0.0325	-0.0313
T_CC	0.0047	0.0003	14.97	0.0000	0.0040	0.0053
MM_out	0.0549	0.0002	228.32	0.0000	0.0545	0.0554
MM_in	0.0388	0.0002	161.02	0.0000	0.0383	0.0392
Residual standard error: 0.111 on 1271992 degrees of freedom						
Multiple R^2: 0.2869					Adjusted R^2: 0.2869	
F-statistic: 73125.9 (on 7 and 1271992 df)					p-value: 0.000	

network and the first non-symmetric blockmodel structure, the factorial design has 90 cells (3 different non-response mechanisms times 5 non-response data treatments times 6 different numbers of non-respondents).

Data missing completely at random

For the random missing mechanism the graphical display of results for stability of partitions is provided in Figure 7.38(a). All of the trajectories for the mean values of the Adjusted Rand Index decline as the number of non-respondents increases. Among all five non-response treatments only the complete-case approach provides acceptable position membership identification for whole range of non-respondents. For one non-respondent, all missing data treatments permit acceptable identification of the composition of positions. However, the five treatments form two groups. The first has the complete-case approach and the imputations based on the mode, and the second group has the null tie imputation, reconstruction and the combination of reconstructions with imputations based on mode. The first group of treatments performs better

than the second group and the differences are more obvious when the number of non-respondents increases.



(a) Adjusted Rand Index, $mARI$

(b) Incorrect block types, $mErrB$

Figure 7.38: Results of the simulation study based on the secon non-symmetric block-model structure for data missing completely at random (solid lines) and predictions according to linear regression model (dash lines)

The slope coefficient of the complete-case approach treatment for prediction of ARI (dash line in Figure 7.38(a)) is equal to $\beta = -0.027$ and it is statistically significant higher than testing value $\beta_0^{ARI} = -0.0\bar{3}$ (Table A.1). Slope coefficients for other four treatments are statistically significant lower than $\beta_0^{ARI} = -0.0\bar{3}$.

The null tie imputation treatment is the first for which $mARI$ values drop below 0.8 (Table 7.16), and this happens for two non-respondents. With three non-respondents in the network both the reconstruction and the combination of reconstruction and mode imputations drop below this threshold for acceptable blockmodels according to agreement between partitions. Thereafter their values for $mARI$ fall further. The imputation based on the mode drops below the threshold of 0.8 for five non-respondents in the network. This leaves the complete-case approach as the only treatment which is able to reveal the position membership acceptably for the whole range of introduced

non-respondents. If there are four non-respondents in the network, the actor response rate is 63% and, up to this value, the imputation based on the mode permits the return of acceptable blockmodels.

If the results are compared to results of the completely symmetric blockmodel structure (Section 7.3.3.1) there is one potentially consequential difference. Rather than having $mErrB$ start near zero for one non-respondent, the trajectories for all five treatments for this measure start above 0.11 in case of second non-symmetric blockmodel structure (Figure 7.38 and Table 7.17). This implies that at least one of nine blocks is incorrectly identified. When two non-respondents are introduced to the network all the $mErrB$ values are below the threshold of 0.2 and imply that acceptable blockmodels are returned. The null tie imputation treatment becomes unacceptable for three non-respondents. The reconstruction treatment returns the unacceptable blockmodels for five and six non-respondents. The $mErrB$ trajectories for the other three treatments approach the 0.2 threshold and stay below as the number of non-respondents increases to six non-respondents.

Table 7.16: Mean values and standard deviations for ARI for the second non-symmetric blockmodel structure

Number of non-respondents		1		2		3		4		5		6	
Missing mechanism	Treatment	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Random missing mechanism	Null tie imputations	0.907	0.133	0.799	0.174	0.683	0.166	0.605	0.168	0.526	0.171	0.459	0.169
	Reconstuction	0.905	0.138	0.834	0.179	0.720	0.197	0.605	0.186	0.507	0.176	0.434	0.163
	Mode	0.993	0.042	0.913	0.126	0.841	0.152	0.807	0.156	0.754	0.166	0.711	0.164
	Reconstr. plus mode	0.905	0.138	0.862	0.178	0.782	0.198	0.702	0.219	0.650	0.238	0.574	0.259
	Complete Case	0.979	0.051	0.959	0.107	0.938	0.136	0.912	0.173	0.874	0.212	0.828	0.253
missing mechanism based on outdegree	Null tie imputations	0.934	0.125	0.836	0.167	0.746	0.176	0.673	0.177	0.596	0.181	0.525	0.181
	Reconstuction	0.945	0.115	0.877	0.170	0.781	0.203	0.676	0.210	0.569	0.192	0.483	0.178
	Mode	0.898	0.124	0.827	0.137	0.777	0.148	0.725	0.154	0.678	0.153	0.626	0.150
	Reconstr. plus mode	0.946	0.114	0.889	0.166	0.812	0.199	0.737	0.217	0.653	0.232	0.541	0.236
	Complete Case	0.969	0.092	0.948	0.121	0.917	0.162	0.876	0.196	0.826	0.229	0.756	0.261
missing mechanism based on indegree	Null tie imputations	0.859	0.142	0.652	0.120	0.548	0.095	0.493	0.086	0.481	0.112	0.504	0.176
	Reconstuction	0.873	0.139	0.688	0.157	0.562	0.121	0.496	0.099	0.483	0.122	0.527	0.185
	Mode	0.965	0.092	0.925	0.128	0.874	0.158	0.849	0.163	0.778	0.172	0.747	0.174
	Reconstr. plus mode	0.870	0.142	0.692	0.159	0.570	0.133	0.507	0.120	0.494	0.138	0.530	0.193
	Complete Case	0.980	0.069	0.967	0.102	0.950	0.123	0.936	0.143	0.915	0.172	0.878	0.215

Because the values of $mErrB$ are around 0.11 for one non-respondent for all treat-

Table 7.17: Mean values and standard deviations for $ErrB$ for the second non-symmetric blockmodel structure

Number of non-respondents		1		2		3		4		5		6	
Missing mechanism	Treatment	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
	Null tie imputations	0.133	0.106	0.168	0.118	0.201	0.109	0.205	0.103	0.213	0.094	0.213	0.086
Random	Reconstuction	0.133	0.114	0.150	0.124	0.174	0.123	0.199	0.113	0.221	0.103	0.234	0.099
missing	Mode	0.110	0.110	0.116	0.114	0.125	0.119	0.138	0.125	0.152	0.123	0.174	0.129
mechanism	Reconstr. plus mode	0.133	0.114	0.141	0.117	0.155	0.121	0.167	0.121	0.175	0.116	0.197	0.121
	Complete Case	0.114	0.115	0.122	0.117	0.132	0.121	0.143	0.125	0.162	0.124	0.177	0.125
missing	Null tie imputations	0.129	0.112	0.177	0.122	0.203	0.114	0.214	0.106	0.218	0.098	0.220	0.088
mechanism	Reconstuction	0.124	0.115	0.139	0.128	0.168	0.132	0.195	0.120	0.220	0.103	0.238	0.096
based	Mode	0.115	0.114	0.129	0.121	0.145	0.127	0.171	0.134	0.195	0.138	0.226	0.140
on	Reconstr. plus mode	0.124	0.115	0.133	0.120	0.156	0.127	0.174	0.128	0.192	0.127	0.227	0.123
outdegree	Complete Case	0.119	0.116	0.134	0.122	0.151	0.129	0.178	0.133	0.205	0.137	0.231	0.134
missing	Null tie imputations	0.157	0.125	0.215	0.114	0.220	0.109	0.218	0.107	0.213	0.104	0.204	0.099
mechanism	Reconstuction	0.149	0.126	0.192	0.125	0.209	0.115	0.214	0.109	0.207	0.104	0.188	0.097
based	Mode	0.113	0.112	0.116	0.111	0.121	0.113	0.129	0.116	0.138	0.120	0.161	0.128
on	Reconstr. plus mode	0.150	0.127	0.190	0.125	0.207	0.116	0.211	0.111	0.204	0.107	0.191	0.101
indegree	Complete Case	0.112	0.111	0.117	0.111	0.123	0.113	0.131	0.114	0.143	0.115	0.164	0.119

ments, the linear predictions forced through point $(0, 0)$ visually do not fit well to the observed data. Slope coefficients are below testing value $\beta_0^{ErrB} = 0.0\bar{3}$ only for the complete-case approach and the imputations based on mode (Table A.2).

Therefore, the mode imputations performs the best for whole range of non-respondents for $mErrB$, while for the mean value of ARI it is the best treatment only for one non-respondent. The complete-case approach fares better over all values of introduced non-respondents according to stability of position membership. Overall, the complete-case performs the best, because the imputation using the mode does not return acceptable partitions of actors when there are more than four non-respondents.

Figure 7.39 shows indices of blockmodel stability plotted versus the reciprocity values of the whole networks. The reciprocity is in range from 0.26 to 0.57. The values of $mARI$ are almost invariant regarding to different reciprocity values of starting whole network. According to Figure 7.38 we wrote that treatments form two groups. According to additional figures with reciprocity values the combination of reconstruction and imputations based on mode for ties between non-respondents can be classified into its own group. For $mARI$ and $ErrB$ values this treatment shows interesting behavior. For

high symmetry in the network (high reciprocity values) the combination of reconstruction and imputations based on mode has higher values of $mARI$ and is similar to the complete-case approach and mode imputations. In networks with lower symmetry the $mARI$ values of that treatment decline and are similar to the null imputation and simple reconstruction treatment. For values of $mErrB$ similar pattern is shown also with reconstruction treatment which performs better for more symmetric networks. The complete-case approach and mode imputations are the best treatments, irrespective of reciprocity values of whole starting network.

Randomly missing data based on outdegree

Compared to random missing mechanism for this whole network structure, the same broad conclusions hold, albeit with some interesting differences. The values of $mARI$ drop in the same way as the number of non-respondents increases (Figure 7.40(b)). The trajectory for the complete-case treatment is the best for whole range of non-respondents, but it does not permit the return of the correct position memberships for six non-respondents. For one and two introduced non-respondents all treatments assure the return of acceptable partitions of actors. For three non-respondents, reconstruction plus imputations based on mode remains, as well as the complete-case approach, above the 0.8 threshold. The slope coefficients for linear predictions (dash lines in Figure 7.40(a)) are for all treatments below the testing value $\beta_0^{ARI} = -0.0\bar{3}$ (Table A.1).

Results for the identification of block types are presented in Figure 7.40(b). The three treatments that perform the best are the complete-case approach, the mode imputations and the combination of reconstruction and mode imputations. For less than five non-respondents also the reconstruction treatment performs well. The null tie imputations are unacceptable already for three non-respondents with $mErrB$ values above 0.2. The slope coefficients of linear predictions (dash lines in Figure 7.40(b)) are for all treatments statistically significant above the testing value $\beta_0^{ErrB} = 0.0\bar{3}$ (Table A.2).

Combining the use of these two criteria, the complete-case treatment dominates all

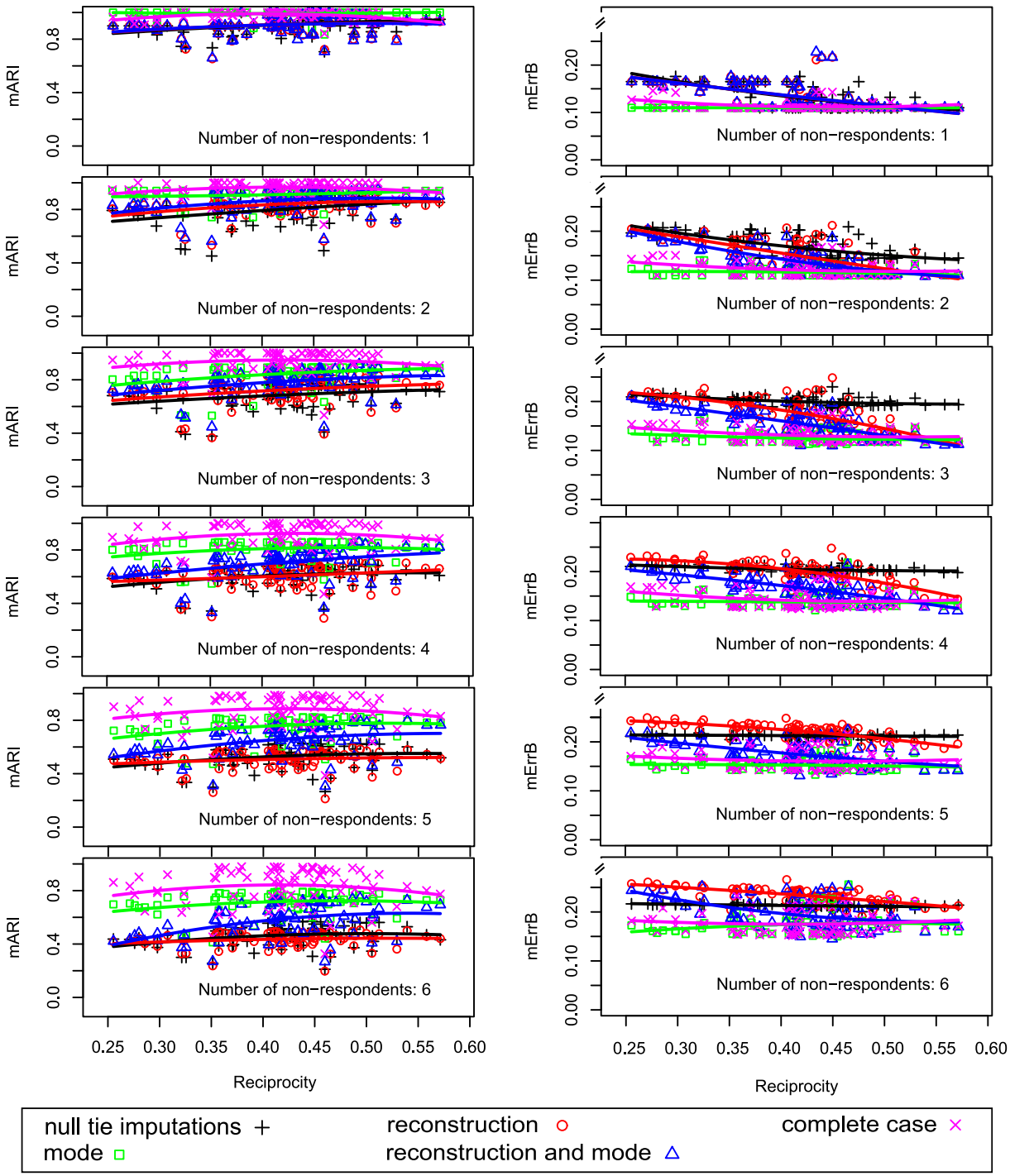
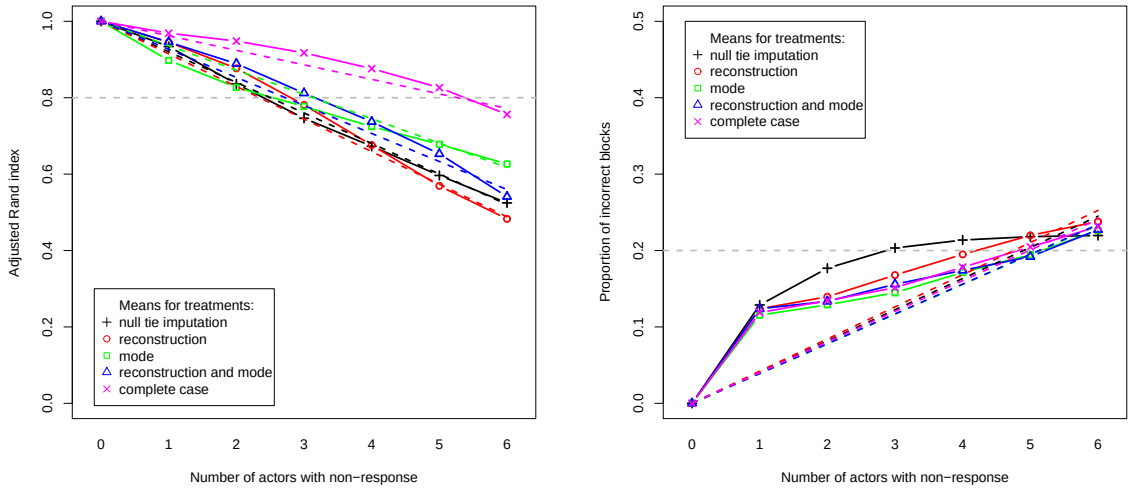


Figure 7.39: Mean of the Adjusted Rand Index, *mARI* (left), and the Mean of Incorrect block types, *mErrB* (right), for second non-symmetric blockmodel structure and random missing mechanism

other treatments except for six non-respondents when *mARI* is too low and *mErrB* to high. Even though the imputation based on the mode and combination of recon-



(a) Adjusted Rand Index, $mARI$

(b) Incorrect block types, $mErrB$

Figure 7.40: Results of the simulation study based on the first non-symmetric blockmodel structure for data missing based on outdegree (solid lines) and predictions according to linear regression model (dash lines)

struction and mode imputation lead to correctly identified block types for less than six non-respondents, it performs poorly with regard to position membership identification for more than three non-respondents. But, to the extent that having the correct blockmodel seems more important with regard to representing network structure, then the blockmodel types identified after using this treatment are acceptable for whole range of introduced non-respondents, even though the membership of identification of positions is not acceptable for all non-respondents.

The reciprocity values of whole networks reveal no special patterns (Figure 7.41). The complete-case approach and imputations based on mode remains the bets treatments irrespective of symmetry of the starting networks.

Randomly missing data based on indegree

The trajectories of $mARI$ for all five treatments are shown in Figure 7.42. In case of agreement between partitions ($mARI$) the complete-case treatment stays well above the 0.8 threshold for all values of introduced non-respondents. The imputation based

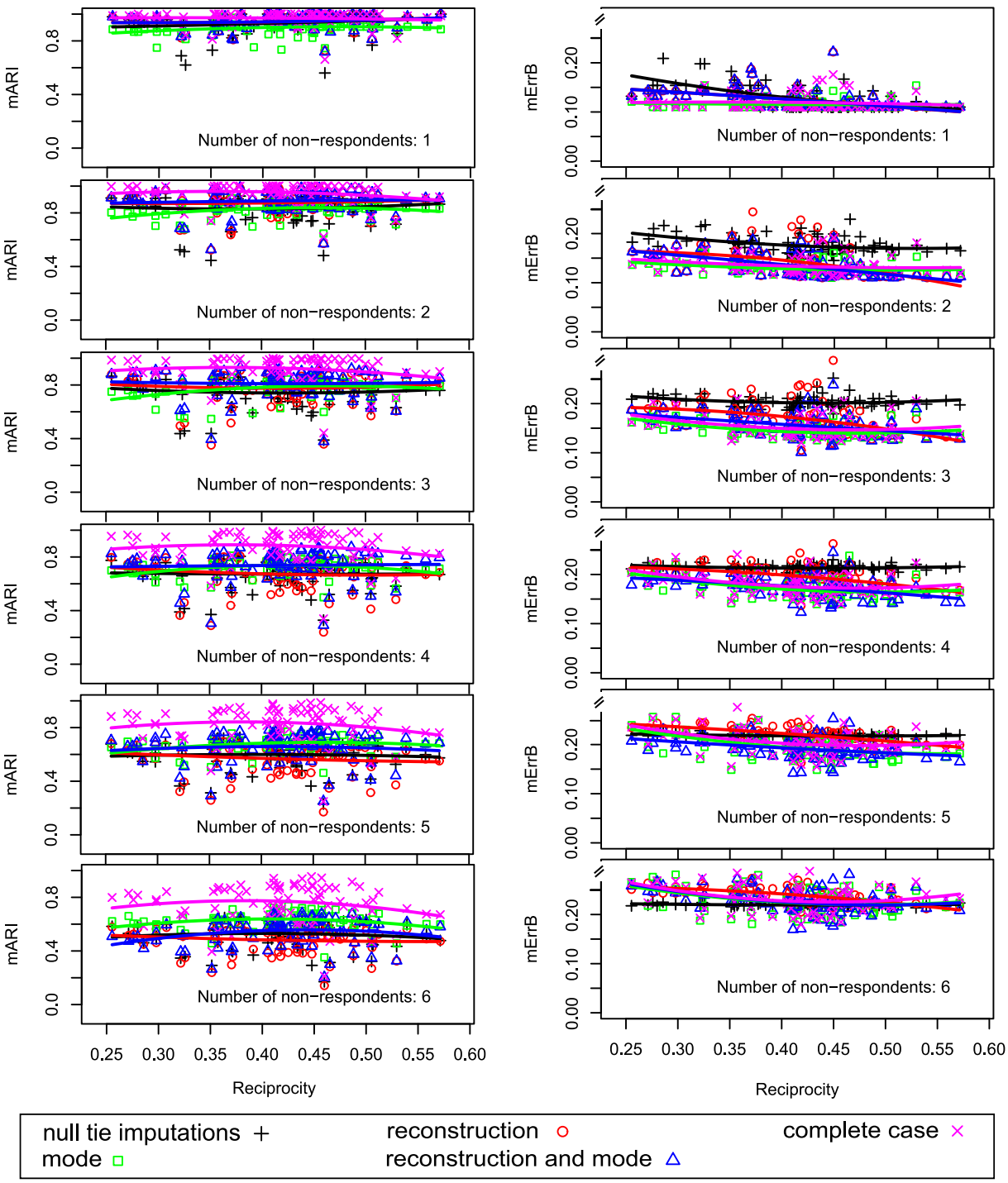
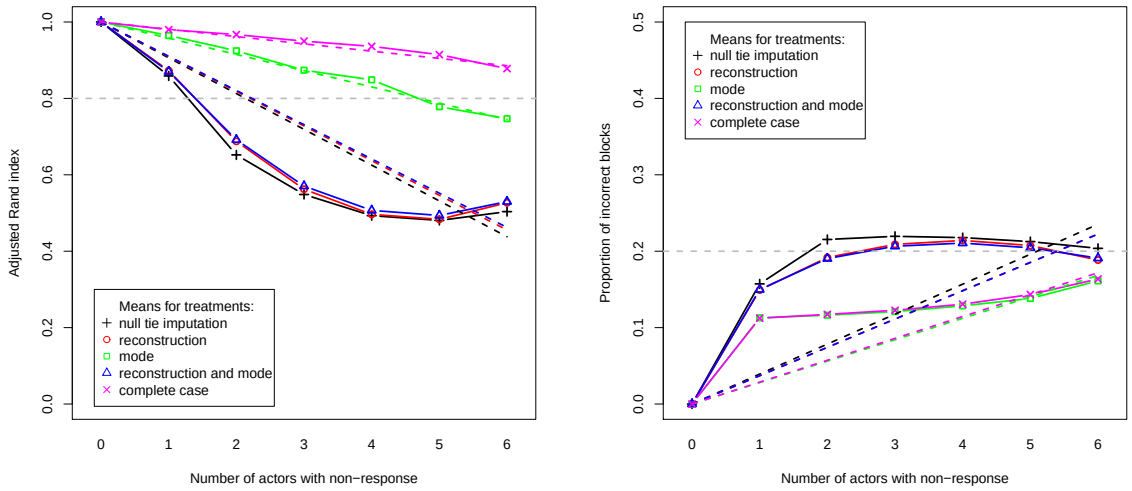


Figure 7.41: The mean of the Adjusted Rand Index, $mARI$ (left) and the mean of the Proportion of Incorrect block types, $mErrB$ (right) for the second non-symmetric block-model structure and missing mechanism based on outdegree

on the mode comes next in the identification of positions, but its $mARI$ drops below 0.8 for five non-respondents. The other three treatments permit acceptable identifi-

cations of position memberships only with one non-respondent. When the number of non-respondents is increased all three of these treatments fail to recover position membership.



(a) Adjusted Rand Index, $mARI$

(b) Incorrect block types, $mErrB$

Figure 7.42: Results of the simulation study based on the first non-symmetric blockmodel structure for data missing based on indegree (solid lines) and predictions according to linear regression model (dash lines)

The trajectories for $mErrB$ are presented in Figure 7.42(b). The two treatments that perform the best are again the complete-case approach and the imputation based on the mode. With one non-respondent in the network the $mErrB$ is slightly above 0.11 (Table 7.17). For higher number of non-respondents the mean proportion of incorrectly identified block types increases, but stays well below 0.2 for the whole range of non-respondents. For these two treatments, the block type identification is acceptable and on average there is only one misspecified block in the blockmodel. The three other treatments have higher values for $mErrB$ (around 0.15) for one non-respondent and tend to increase when the number of non-respondents goes from 1 to 4. As the number of non-respondents increases beyond four non-respondents, the trajectories flatten and begin to decline very slightly. The mean of $mErrB$ for the null tie imputation treatment is above 0.2 when there are two or more non-respondents in the network and therefore this treatment is not acceptable according to agreement between both blockmodels.

While the trajectories for both reconstruction and the combination of reconstruction with using the mode move closely together, they also move above the 0.2 threshold for three non-respondents. The $mErrB$ values drop below the 0.2 threshold for six non-respondents, but it seems the safest to conclude that only the complete-case and the imputation based on the mode permit correctly identified blockmodels.

The reciprocity values of whole starting networks in Figure 7.43 confirm the above findings and reveal the additional pattern that the reconstruction and reconstruction combined with mode imputations for $mErrB$ values depend on the symmetry of the whole networks. More symmetrical networks tend to have lower $mErrB$ values, which means that block agreement is better. This pattern is visible in smaller extent (especially for one and two non-respondents) also in case of the Adjusted Rand Index where more symmetric networks (with higher reciprocity values) have higher $mARI$ values.

Establishment of multiple regression models

According to the simulation of actor non-response the factorial design has 90 cells (3 non-response mechanisms, five non-response treatments and six different numbers of introduced non-respondents). Multiple regression models for $mARI$ and $mErrB$ were established instead of anova due to unequal variances in cells (Tables 7.16 and 7.17). The model summary ARI in Table 7.18 shows that our regression model explains 35% of variation in ARI.

The established model for ARI for the second non-symmetric blockmodel structure is similar to the regression model of the borrowing network (Figure 7.20) and can be with estimated (unstandardized) coefficients (Figure 7.44) written as follows:

$$\begin{aligned} \hat{Y}_{ARI} &= 0.8338 - 0.0561 \cdot n.actor - 0.0062 \cdot T\_RE + 0.1928 \cdot T\_MO + \\ &+ 0.0568 \cdot T\_REMO + 0.3147 \cdot T\_CC - 0.0039 \cdot MM\_out + \\ &- 0.0119 \cdot MM\_in \, . \end{aligned} \tag{7.15}$$

All variables in a model for ARI are significant (p-values are 0.000 in Table 7.18). The regression coefficients b for the second non-symmetric blockmodel structure can be interpreted as follows:

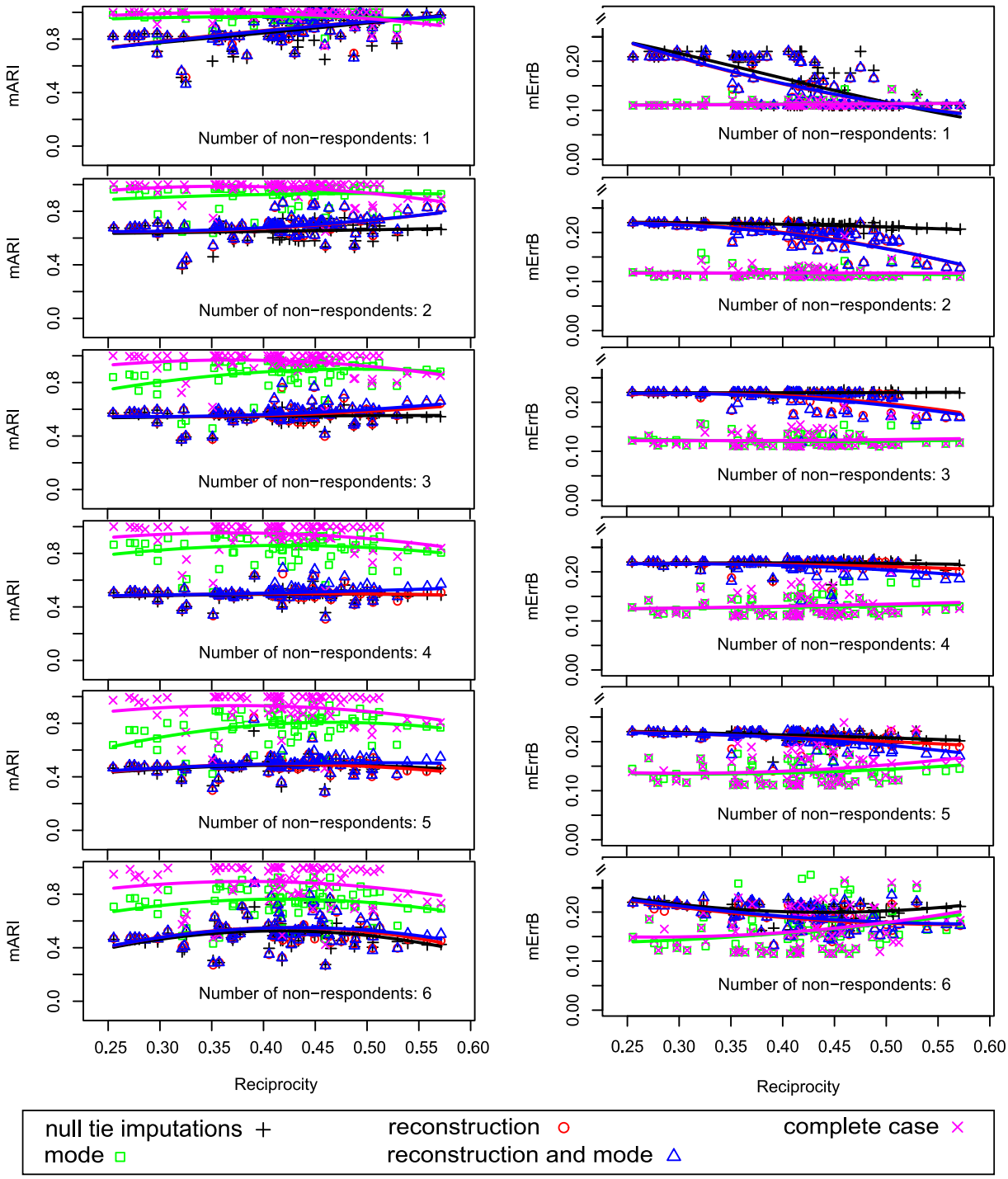


Figure 7.43: The mean of the Adjusted Rand Index, $mARI$ (left), and the mean of the Proportion of Incorrect block types, $mErrB$ (right), for second non-symmetric block-model structure and missing mechanism based on indegree

- **n.actor:** ($b = -0.0561$) If the number of non-respondents increases for one non-respondent, the values of ARI decrease for 0.0561.

- **T_RE vs. T_NTI:** ($b = -0.0062$) The reconstruction is the worst treatment. If the reconstruction treatment is used, compared to the null tie imputation the value of ARI decreases for 0.0062. Because of small absolute value of this coefficient, the reconstruction procedure is the most similar to the null tie imputations.

- **T_MO vs. T_NTI:** ($b = 0.1928$) The imputation based on mode treatment turns out to be the second best treatment. The value of ARI increases for 0.1928 when imputations based on mode are used instead of the null tie imputation.

- **T_REMO vs. T_NTI:** ($b = 0.0568$) The reconstruction combined with the mode imputations performs slightly better then reconstruction itself. In comparison with the null tie imputations, the values of ARI in that case are higher for 0.0568.

- **T_MO vs. T_CC:** ($b = 0.3147$) The complete-case approach is the best treatment in terms of partition agreement in a blockmodel. The values of ARI increase for 0.3147 if the complete-case approach is used instead of the null tie imputations.

- **MM_out vs. MM_random:** ($b = -0.0039$) The use of the missing mechanism based on outdegree has little negative effect. The values of ARI decrease for just 0.0039, if the missing mechanism based on outdegree is used instead of randomly selected non-respondents.

- **MM_in vs. MM_random:** ($b = -0.0119$) The use of the missing mechanism based on indegree instead of the random missing mechanism decreases the value of ARI for 0.0119. Both non-random missing mechanisms lead to a little bit unstable blockmodels according to partitions.

We also try to set up the regression model for $ErrB$. Similar as for the note borrowing network, it turns out that it explains just 5.1% of variance in proportion of incorrectly identified block types (and it is not reported here). One reason why the percent of explained variance is low is that there is no clear linear relationship between the number of non-respondents and values of $ErrB$ index. In Figure 7.45 the radius of the circles is proportional to the number of cases with the same value of $ErrB$. For one non-respondent the majority of $ErrB$ values is equal to 0 and 0.22, and for two to six non-respondents the majority of $ErrB$ values occupy four values; 0, 0.11, 0.22 and 0.33,

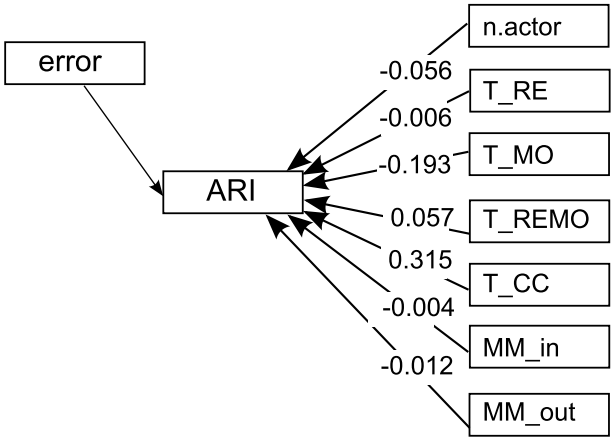


Figure 7.44: Regression models for ARI with data for the Ssecond non-symmetric blockmodel stucture

Table 7.18: Model summary and coefficients of regression analysis for ARI with data for the second non-symmetric blockmodel structure

	Estimate	Std. Error	t value	Pr(>|t|)	95% confidence interval for b	
(Intercept)	0.8338	0.0008	999.88	0.0000	0.8322	0.8355
n_actor	-0.0561	0.0001	-402.24	0.0000	-0.0564	-0.0558
T_RE	-0.0062	0.0005	-11.45	0.0000	-0.0073	-0.0052
T_MO	0.1928	0.0005	354.54	0.0000	0.1917	0.1938
T_REMO	0.0568	0.0005	104.48	0.0000	0.0557	0.0579
T_CC	0.3147	0.0005	578.77	0.0000	0.3136	0.3158
MM_out	-0.0039	0.0004	-9.23	0.0000	-0.0047	-0.0031
MM_in	-0.0119	0.0004	-28.20	0.0000	-0.0127	-0.0111
Residual standard error: 0.194 on 1271992 degrees of freedom						
Multiple R^2: 0.350					Adjusted R^2: 0.350	
F-statistic: 97815.7 (on 7 and 1271992 df)					p-value: 0.000	

which indicates zero to three incorrectly identified block types. Therefore, the number of non-respondents is obviously inadequate to predict the proportion of incorrect block types ($ErrB$) in the linear regression model.

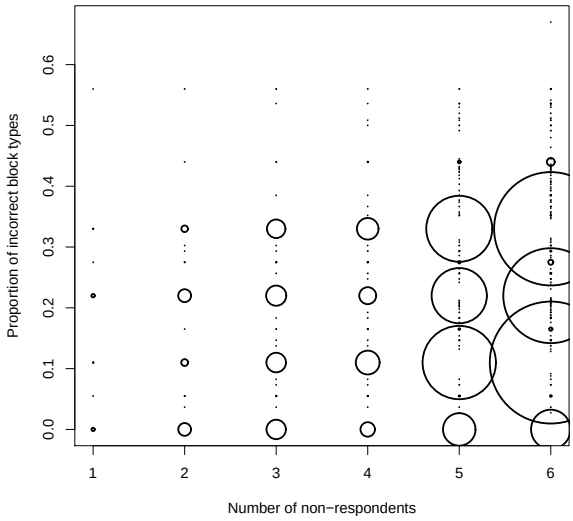


Figure 7.45: The relationship between the number of non-respondents and values of $ErrB$ index for the second non-symmetric blockmodel structure

7.3.4 Conclusions

The summary of results from simulation studies of two real networks and three artificial network structures is presented in Table 7.19. For each type of network, the best overall treatment based on both the Adjusted Rand Index ($mARI$) and proportion of incorrect block types ($mErrB$) is presented by the + sign. The worst overall performance is represented by the - (minus) sign, and the moderate performances by the ∘ sign. We separate our summary report for the networks with a symmetric block structure from those whose underlying block structure departs from this form of symmetry.

The main conclusion is that the performance of the missing data treatments for nonresponse in social networks depends on the symmetry of the networks. The symmetry of the network refers to reciprocity value and also to the blockmodel structure. The treatments that are the best for symmetric networks perform the worst in the case of non-symmetric networks and vice versa. More exactly, the best treatments for symmetric networks are reconstruction and combination of reconstruction and mode imputations. For the non-symmetric network the best treatments are the imputations based on mode and the complete-case approach.

Therefore, the Thesis 2 about the best non-response treatment presented on page 55 can be only partially confirmed. The stability of blockmodeling is higher when the reconstruction is used compared to imputations based on mode if the whole network is symmetric. Otherwise, the opposite is true. In the non-symmetric networks the imputations based on mode are more preferable treatment than the reconstruction.

Table 7.19: Impact of the non-response treatments on the stability of blockmodeling

Blockmodel	Symmetric				Non-symmetric					
					Simulated				Simulated	
	Real		Simulated		First		Real		Second	
Treatment	ARI	ErrB	ARI	ErrB	ARI	ErrB	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 always performs 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.

Simulation studies for not at random deletion of actors based on indegree and outdegree show no major differences in performance patterns for different treatments compared to randomly selected non-respondents. One of the reasons for this is also the small size of networks in the studies where actors are similar to each other according to their indegree and outdegree. Therefore, future work on actor non-response should include larger networks and wider set of blockmodel structures.

The above findings are confirmed with simulation study on two real networks from baseball Little League (Žnidaršič et al., 2011b). Networks consist of boys from two teams and were first reported by (Fine, 1987) and extensively studied in terms of generalized blockmodeling by (Doreian et al., 2005). The best treatments for actor non-

response in case of symmetric network (the Transatlantic Industries network) are reconstruction and combination of reconstruction and modes imputations. Results of simulation study with an example of non-symmetric network (the Sharpstone Auto network) show that the best treatments are imputations based on mode and the complete-case approach. Again, the null tie imputation is not advisable.

We used ANOVA to investigate the effects of the number of non-respondents, treatment of non-response data, non-response missing mechanism (MM) and type of the symmetry of the network (Žnidaršič et al., 2011b). We established that the largest effect on both indices of blockmodeling stability (ARI and $ErrB$) has the number of non-respondents. The second largest effect in the case of the Adjusted Rand Index has the treatment, while in the case of the Proportion of incorrect blocks the second largest effect has the interaction of the treatment and the symmetry of the network. The lowest main effect has in both case the non-response missing mechanism.

7.3.5 An example of generalized type of equivalence - A review of the Student Government discussion network

In Section 6.2.2.1 we described a set of six networks which are an example of use of complete-case approach as a non-response treatment. In early work of Hlebec (1992) we found description of the partially reported ties between respondents and non-respondent. Obtained blockmodels (Doreian et al., 2005, pg. 228-233) of the Student Government recall discussion network will be compared to blockmodels of 'treated' networks based on structural and also generalized type of equivalence.

As described above, the Student Government recall discussion data was used as complete-case approach of a network with one non-respondent. The non-respondent R (or the refusal actor) received two ties from respondents, namely a tie from prime minister (pm) and from minister 2 (m_2). Three different missing data treatments, described in section 4.3.1.1, can be used: reconstruction, imputations based on mode and null tie imputations. Combination of reconstruction and imputations based on mode in that case is the same as reconstruction treatment alone, because there is only one non-respondent

and no additional imputations for ties between non-respondents are needed.

The network with 11 actors presents the complete-case approach and is presented on the top panel of Figure 7.46. Colors of vertices show two-clusters partition based on structural equivalence. In blockmodeling procedure into two clusters we obtain two equally well fitting partitions with 29 inconsistencies. In the first obtained partition actors pm, m_3, m_6 and m_7 form one cluster and the second cluster consist of actors $\{m_1, m_2, m_4, m_5, a_1, a_2, a_3\}$. The second equally well fitting partition has in first cluster three actors m_3, m_6, m_7, and the prime minister is moved to the second cluster. Both partitions formed the same image matrix of centralized model with ties to the core position (complete block on the diagonal for the first cluster and between second and first cluster). Table 7.20 shows the value of the Adjusted Rand Index between both partitions which is equal to 0.64.

Table 7.20: The ARI between partitions obtained in blockmodeling procedure of Student Government discussion into two-clusters with structural equivalence for four different non-response data treatments

Index		ARI						ErrB					
Treatment		CC		RE		MI	NTI	CC		RE		MI	NTI
		1st	2nd	1st	2nd			1st	2nd	1st	2nd		
Complete case	1st part.	1.00	0.64	1.00	0.13	1.00	0.13	0.00	0.00	0.00	0.25	0.00	0.25
	2nd part.		1.00	0.64	0.35	0.64	0.35		0.00	0.00	0.25	0.00	0.25
Reconstruction	1st part.			1.00	0.18	1.00	0.18			0.00	0.25	0.00	0.25
	2nd part.				1.00	0.18	1.00				0.00	0.25	0.00
Mode imputations						1.00	0.18					0.00	0.25
Null tie imputations							1.00						0.00

In that case, we do not know the real whole network, therefore the comparison of whole and treated blockmodel (as in previous sections) is not possible. However, we can estimate the reciprocity of the real unknown network. If the reconstruction procedure is introduced to network with one non-respondent, the outgoing ties from non-respondent R are added in a way that they are symmetrical to the ingoing ties to that actor (a row of missing ties is replaced with the column for actor R). In that case the

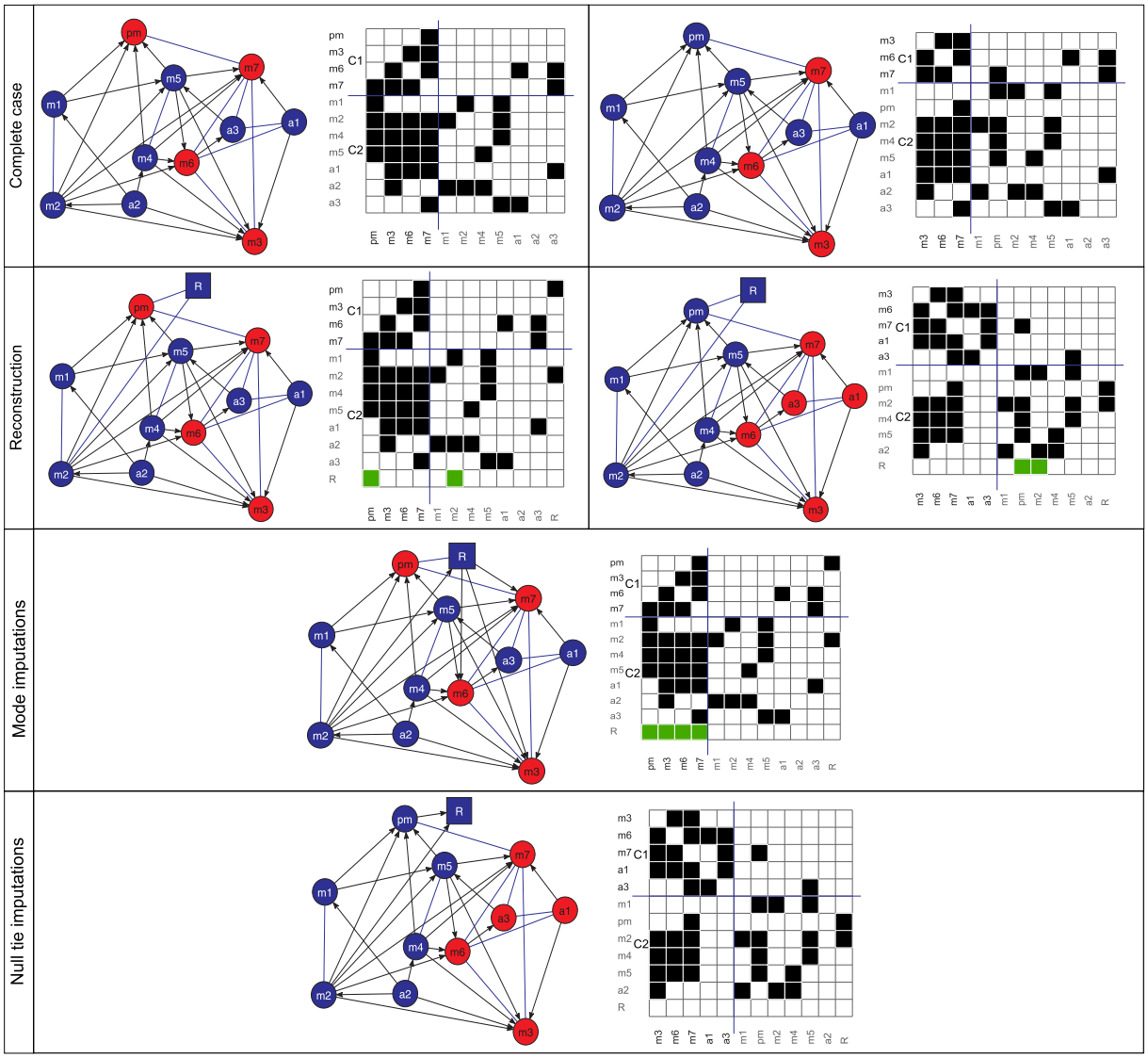


Figure 7.46: Blockmodeling into two clusters based on structural equivalence of Student Government discussion network and non-response data treatments

reciprocity of the real whole network will be the highest and will be equal to 0.49. If the ties will be added in the opposite way as in reconstruction procedure (instead of ties zeros will be imputed and otherwise, instead of zeros ties (ones) will be added), the network will have the largest possible amount of asymmetric ties, therefore the reciprocity will be the lowest and will be equal to 0.35. In previous sections we try to determine the best non-response data treatment. The main conclusion was that the selection of the best missing data treatment depends on symmetry of the whole network. According to possible values of reciprocity, which are between 0.35 and 0.49 for real Student Government discussion network, the network is an example of non-symmetric

network. According to the range of reciprocity values, the Student Government discussion network is similar to the networks generated for the second non-symmetric blockmodel structure, where reciprocity values are in range from 0.25 to 0.58 (Figure 7.38(a) in Section 7.3.3.3). The summary of the results in Table 7.19 shows that the best treatments for non-symmetric networks are the complete-case approach and mode imputations.

If imputation based on mode is used, four outgoing ties for actor R are added (shown in green in sociomatrix in Figure 7.46). We get one best partition with 31 inconsistencies. The partition is the same as the first partition in the complete-case approach. In addition to the complete-case approach, we know that the refusal actor R belongs to the larger second cluster.

Reconstruction is not the advised treatment for non-symmetric networks. With reconstruction procedure two best fitting partitions with 25 inconsistencies are obtained. The first obtained partition is the same as the first partition in the complete-case approach. The second obtained partition is different from partitions obtained with complete-case or mode imputations. In the first cluster are actors m_3, m_6, m_7, a_1 and a_3 and the second cluster consist of actors $\{m_1, pm, m_2, m_4, m_5, a_2, R\}$. The described partition is the same as partition obtained with null tie imputations, which turns out to be the worst possible missing data treatment. With this partition we get different image matrix with complete block on the diagonal and three null blocks.

Without knowledge about the best treatments, the conclusion can be made only about 'stable' members of each cluster. Irrespective of the chosen treatment, actors $\{m_3, m_6, m_7\}$ and $\{m_1, m_2, m_4, m_5, a_2, R\}$ are always in the same cluster, while the other actors $pm, a1, a3$ are changing their position between both clusters.

According to above qualitative analysis of obtained partitions, the preferable partition is $\{pm, m_3, m_6, m_7\}$ $\{m_1, m_2, m_4, m_5, a_1, a_2, a_3\}$, which is obtained by both preferable treatments, complete-case approach and mode imputations. When the mode imputa-

tions are used, we also know the position of the refusal actor R, which belong to the second (larger) cluster.

Different selection of equivalence leads to different partitions. Doreian et al. (2005) investigated the generalized equivalence with allowed block types restricted to {null, com, rdo, cdo, reg} into two to five clusters for the Student Government recall discussion network (the complete case approach). The summary of their findings is presented in Table 6.2 on page 76. The obtained results for four-cluster blockmodels were compared with 'treated' blockmodels.

Table 7.21 presents obtained blockmodels for different treatments of non-response for the Student Government discussion network into four clusters with allowed block types {null, com, rdo, cdo, reg}. With complete-case approach three equally well fitting partitions are obtained. Again, quantitative comparison of obtained partitions with real starting network is not possible. In the complete-case approach four 'stable' pairs of actors are established. In all equally well fitting partitions, the following pairs of actors are always placed together: $\{m_1, m_2\}$, $\{pm, m_4\}$, $\{m_3, m_5\}$ and $\{a_1, a_3\}$. Comparison of all three equally well fitting partitions with value of criterion function 0 with the Adjusted Rand Index is presented in Table 7.22[21].

The above conclusion, that preferable treatments are complete-case approach and mode imputations have been made on the basis of simulations for structural equivalence. The findings can not be generalized without additional simulations. The presented example and qualitative analysis of obtained results can be viewed just as a starting point for planning future simulations.

Based on ARI values presented in Table 7.22 the average values of ARI between dif-

[21] The presented partitions are obtained with the constraint that each cluster must contain at least two vertices. Without this constraint we obtain in Pajek with 10000 repetitions seven additional partition with just one actor in a cluster for complete-case approach, three additional equally well fitting partitions for reconstruction, six partitions for mode imputations and seven partitions with one actor in a cluster for null tie imputations.

Table 7.21: Optimal partitions for Student Government recall discussion network and different non-response treatments and allowed block types {null, com, rdo, cdo, reg}

Treatment		Partition	I_{min}
Complete	1st part.	$\{m_1, m_2\}\{pm, m_4\}\{m_3, m_5, m_6, m_7, a_2\}\{a_1, a_3\}$	0
case	2nd part.	$\{m_1, m_2, a_2\}\{pm, m_4\}\{m_3, m_5, m_6, m_7\}(a_1, a_3\}$	0
	3rd part.	$\{m_1, m_2, a_2\}\{pm, m_4, m_6, m_7\}\{m_3, m_5\}\{a_1, a_3\}$	0
Reconstruction	1st part.	$\{m_1, a_2\}\{pm, m_2, m_5, m_6, a_3, R\}\{m_3, m_4\}\{m_7, a_1\}$	1
	2nd part.	$\{m_1, R\}\{pm, m_3, m_5, m_7, a_1, a_3\}\{m_3, a_2\}\{m_3, m_6\}$	1
	3rd part.	$\{m_1, R\}\{pm, m_4\}\{m_2, m_3, m_5, m_6, m_7, a_2\}\{a_1, a_3\}$	1
	4th part.	$\{m_1, R\}\{pm, m_2, m_3, m_4, m_5, m_7\}\{m_6, a_3\}\{a_1, a_2\}$	1
	5th part.	$\{m_1, R\}\{pm, m_5, m_6\}\{m_2, a_2\}\{m_3, m_4, m_7, a_1, a_3\}$	1
	6th part.	$\{m_1, R\}\{pm, m_4, m_5, m_6, m_7\}\{m_2, a_2\}\{m_1, a_1, a_3\}$	1
	7th part.	$\{m_1, a_2\}\{pm, m_4\}\{m_2, m_3, m_5, m_6, m_7, R\}\{a_1, a_3\}$	1
Mode	1st part.	$\{m_1, m_2, a_2\}\{pm, m_4\}\{m_3, m_5, m_6, m_7, R\}\{a_1, a_3\}$	0
imputations	2nd part.	$\{m_1, m_4\}\{pm, R\}\{m_2, m_3, m_5, m_6, m_7, a_2\}\{a_1, a_3\}$	0
	3rd part.	$\{m_1, m_2, a_2\}\{pm, m_4, m_6, m_7, R\}\{m_3, m_5\}\{a_1, a_3\}$	0
	4th part.	$\{m_1, m_2\}\{pm, m_4\}\{m_3, m_5, m_6, m_7, a_2, R\}\{a_1, a_3\}$	0
Null tie	1st part.	$\{m_1, m_4\}\{pm, R\}\{m_2, m_3, m_5, m_6, m_7, a_2\}\{a_1, a_3\}$	1
imputations	2nd part.	$\{m_1, m_2, m_3, m_4, m_5, m_7\}\{pm, R\}\{m_6, a_3\}\{a_1, a_2\}$	1
	3rd part.	$\{m_1, a_2\}\{pm, m_4\}\{m_2, m_3, m_5, m_6, m_7, R\}\{a_1, a_3\}$	1

ferent treatments can be computed. This values can rawly reveal which treatments produces the most similar partitions. The highest average value of ARI values is between complete-case approach and the mode imputations (0.59). The average values of ARI between other pairs of treatments are lower and are in range from 0.19 (average values of ARI between the reconstruction and the null tie imputations) and 0.37 (average values of ARI between the mode imputations and the null tie imputations). Therefore, the complete-case approach and the imputations based on mode are the best non-response treatments for non-symmetric networks and blockmodeling based on structural equivalence. Our example shows that both treatments produce the most similar solutions also in case of generalized equivalence. Additional simulations are

necessary to make conclusions about best treatment in that case.

The complete-case approach and the mode imputations also produce the most similar equally well fitting partitions. The average value of ARI among three comparisons of pairs of partitions for complete-case approach is equal to 0.50. The average value of ARI for comparison of four equally well fitting partitions for imputations based on mode is 0.44. The averages of ARI values for the remaining two treatments are rather low; 0.14 for the reconstruction procedure and 0.32 for the null tie imputations.

The same pair of 'stable' actors, as described in complete-case approach, can also be found in imputations based on mode. In three of four partitions, the refusal actor R is placed in the same cluster as actors $\{m_6, m_7\}$ and in one case it is placed in a cluster with prime minister (pm).

Table 7.22: The ARI between partitions obtained in blockmodeling procedure of the Student Government discussion into four-clusters with generalized equivalence and allowed block types {com, reg, null, rdo, cdo} for four different non-response data treatments

Treatment		Complete case			Reconstruction							Mode imputations				Null tie imputations		
	Partition	1st	2nd	3rd	1st	2nd	3rd	4th	5th	6th	7th	1st	2nd	3rd	4th	1st	2nd	3rd
Complete case	1st part.	1	0.68	0.26	-0.21	0	0.73	0	-0.03	0.17	0.50	0.68	0.64	0.26	1	0.64	0	0.50
	2nd part.		1	0.55	-0.06	0.15	0.53	0.06	0.12	0.33	0.68	1	0.43	0.55	0.68	0.43	0.06	0.68
	3rd part.			1	-0.06	0.15	0.15	0.06	0.12	0.54	0.26	0.55	0.06	1	0.26	0.06	-0.04	0.26
Reconstruction	1st part.				1	0	-0.09	0.09	0.17	-0.03	0.09	-0.06	-0.09	-0.06	-0.21	-0.09	-0.09	0.09
	2nd part.					1	-0.02	0.06	0.33	0.33	0	0.15	-0.02	0.15	0	-0.02	-0.19	0
	3rd part.						1	0.15	-0.03	0.15	0.73	0.53	0.91	0.15	0.73	0.91	0.06	0.73
	4th part.							1	-0.03	0.15	0.27	0.06	0.06	0.06	0	0.06	0.57	0.27
	5th part.								1	0.43	-0.03	0.12	-0.03	0.12	-0.03	-0.03	-0.12	-0.03
	6th part.									1	0.17	0.33	0.06	0.54	0.17	0.06	-0.12	0.17
	7th part.										1	0.68	0.64	0.26	0.50	0.64	0.18	1
Mode	1st part.											1	0.43	0.55	0.68	0.43	0.06	0.68
imputations	2nd part.												1	0.06	0.64	1	0.15	0.64
	3rd part.													1	0.26	0.06	-0.04	0.26
	4th part.														1	0.64	0	0.50
Null tie	1st part.															1	0.15	0.64
imputations	2nd part.																1	0.18
	3rd part.																	1

Conclusions

Based on the above example, we could speculate that in case of non-symmetric network the complete-case approach and the mode imputations are the best treatments also when generalized equivalence is used. It should be noted that this is only an assumption, which should be tested in extended simulation study, where the obtained results can be compared with real starting blockmodels. The results of simulations studies with random measurement errors and different types of equivalence (presented in Section 7.5) suggest that the determination of the best non-response treatment for regular and generalized equivalence will be difficult or maybe even impossible because of high instability of blockmodeling in that case. High number of equally well fitting partitions, which are frequently result of generalized blockmodeling, even more aggravates the decision about the best missing data treatment.

7.4 Errors caused by item non-response

Logical continuation of the studies about the impact of actor non-response on the stability of blockmodeling presented in the previous section is the investigation of tie non-response. The theoretical background about tie (or item) non-response is presented in Section 4.3.2. All concepts of non-response treatments (except the complete-case approach if the tie non-response is high) used for an actor non-response can be used also for replacement of non-reported or absent ties.

The results of simulation studies on tie non-response were preliminary presented in QMSS2 2011 Workshop about Social Network Data Collection and the detailed version can be found in Žnidaršič et al. (2011a). Only outline of simulations with main conclusions are presented here.

In the tie non-response we have no information regarding the nature of a tie regardless of whether it is a tie or a null tie. We called such non-reported ties the *absent* ties. The absent tie could be every tie in the adjacency matrix and the researchers are often inattentive to the presence of absent ties and record them as null ties (zero). With the

simulations we tried to establish if reporting of absent ties with zeros could be acceptable and if this problem could be reduced with different non-response treatments.

The simulation of tie non-response was performed with four real whole networks: boy girl-liking ties network, note borrowing network, and two networks of Little League, Transatlantic Industries and Sharpstone Auto. Different amounts of tie non-response were introduced to the whole networks which were then treated with four tie non-response treatments: reconstruction, imputations based on the mode values of incoming ties, a combination of reconstruction with imputations based on the mode, and the null tie imputation. For all networks, whole and treated, blockmodels using structural equivalence were established and compared. For every combination of a real whole network, amount[22] of introduced tie non-response, and tie non-response treatment our simulations were based on 100 repetitions.

Both factors from the simulation design, the amount of tie non-response and the treatment non-response, together with the blockmodel structure of a network and the level of reciprocity all have an impact on the results of blockmodeling. We draw the following conclusions. First, the combination of reconstruction and imputation based on mode is the best overall treatment method for tie non-response according to both correctly revealed position membership and blockmodel structure. Second, both reciprocity and block model structure matter in systematic ways. The results of blockmodeling following the use of imputation based on mode are good when reciprocity is low, but they are unacceptable for networks with high reciprocity[23]. Imputation based on modes fares badly for core-periphery structures, while reconstruction works well for them. Third, the null tie imputation is the worst treatment for tie non-response and its use never succeeds with regard to correctly obtaining the membership of positions. Therefore, the simple recording of zeros instead of absent ties is the worst solution, although it is frequently used in network data collection process. Forth, the criteria of getting the position membership correct and the blockmodel structure correct do not always lead to the same implications with regard to blockmodeling outcomes. In

[22]The percent of tie-nonresponse varies from 1% to 50%.

[23]The same conclusion was made for the actor-nonresponse case.

general, performances are better for the blockmodel structure than for position membership. Put differently, performance is better with regard to the macrostructure of the networks (the image matrix) and worse with regard to micro-structural details (the position membership of actors).

The described simulations should be extended to larger networks, different blockmodel structures or block patterns, different types of equivalence, non-random patterns of tie non-response, different treatments of non-response which also consider actor's characteristics, and valued networks.

7.5 Random measurement errors

In this section the stability of blockmodeling to randomly changed ties will be presented and we try to confirm our Thesis 1 (from page 36) that *structural equivalence gives more stable results than regular (or other generalized types) equivalence*.

The definition of random measurement error is presented in Section 4.3.3. The design of simulation studies is presented in Section 7.5.1. The simulations were run on both real and simulated networks with different types of equivalence. The results for real and simulated networks based on structural equivalence are presented in Section 7.5.2 and Section 7.5.3, respectively. The impact of random measurement error to blockmodels established based on regular equivalence is presented in Section 6.2.4 with extensive set of simulated networks according to cohesive subgroups model and core-periphery model. The stability of blockmodeling established based on generalized type of equivalence is presented in Section 7.5.5.

7.5.1 The design of our simulation studies for random measurement errors

The basic scheme of simulations is presented on page 70 in Section 6.1. More detailed construction on random measurement errors (or measured network due randomly introduced errors) from item 3 (a) in the basic scheme is described here.

The **measured networks** were constructed from real or simulated whole networks by introducing controlled amount of random measurement errors. According to the definition of Holland and Leinhardt (1973) presented on page 57 in Section 4.3.3, the measurement error occurs when there is no tie recorded for underlying relation or opposite, or when tie is recorded in the network for which there is no corresponding relation in the true underlying structure. Therefore, to imitate the random measurement error the ties were randomly selected to be changed, and if there was a tie in the whole network we replaced it with zero in measured network and vice versa. The amount of randomly changed ties was controlled in the studies.

7.5.2 Real networks based on structural equivalence

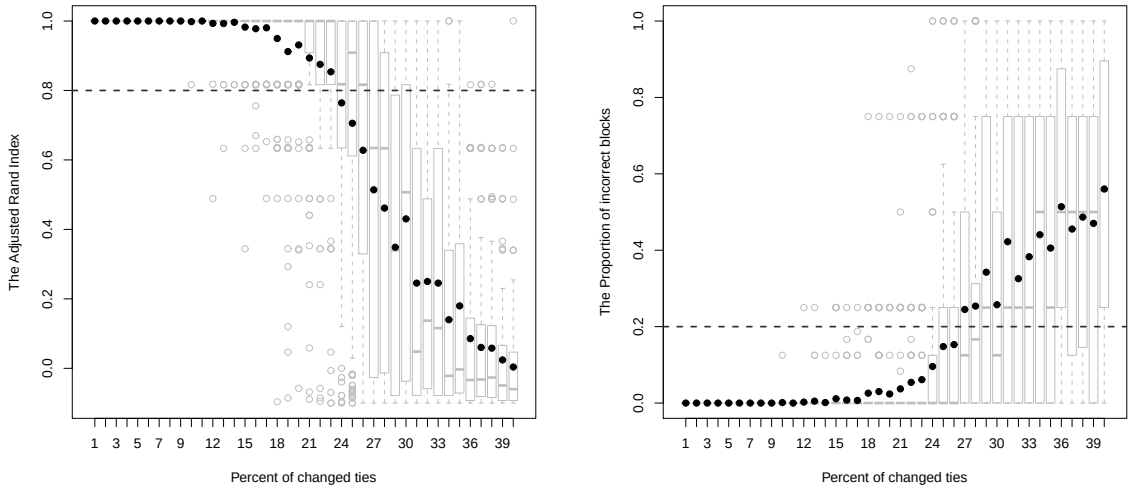
The simulation study was performed with two real networks, the boy-girl liking ties network and the note borrowing network.

7.5.2.1 Results for the boy-girl liking ties network

In the boy-girl liking ties network (Section 6.2.1.1), the percent of random measurement errors was selected in range from 1 to 40. The number of possible ties in the network is $11 \cdot 10 = 110$, for the network consisting of 11 actors. The number of all possible combinations for random selection of ties increases with higher number (or proportion) of ties to be changed. For example, there is $110 = \binom{110}{1}$ possibilities to randomly change one tie, $\binom{110}{2} = 5995$ possible combinations how to change two ties, $\binom{110}{3} = 215820$ combinations for selecting three ties,... Because the blockmodeling algorithm is very time consuming, the simulations were run 50 times for one percent of random errors and 100 times for higher percent or random measurement errors. In the whole simulation study, 3950 new measured networks with different percent of errors were constructed and for every new measured network a blockmodel based on structural equivalence was established and compared with structure shown in Figure 6.1 on page 72.

Figure 7.47(a) presents mean values for ARI together with boxplots. For 11% of changed

ties the $mARI$ is 1, which indicates perfect agreement between real starting partition and measured partitions after introducing random measurement errors into boy-girl liking ties network. For higher percent of changed ties the $mARI$ values start to decline and exceed the value 0.8 at 23% of changed ties. For 23% of changed ties or less there are three-quarters of ARI values above 0.8 according to boxplots. For 35% of changed ties the values of $mARI$ approach 0, which indicates that there is absolutely no agreement between real and measured partitions. Figure 7.47(b)presents results of stability of blockmodeling in terms of correctly identified block types.



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.47: Results of the simulation study based on the boy-girl liking ties network with random measurement errors

Values of $mErrB$ increase, when the percent of changed ties increases. There is perfect agreement between blocks positions for 15% of introduced measurement errors or less and acceptable agreement for 26% of introduced random measurement errors or less ($mErrB$ values below 0.2).

7.5.2.2 Results for the student note borrowing network

The student note borrowing network with blockmodeling structure based on structural equivalence is presented in Figure 6.2 in Section 6.2.1.2. The percent of random measurement errors was selected in range from 1 to 40. The network is larger than

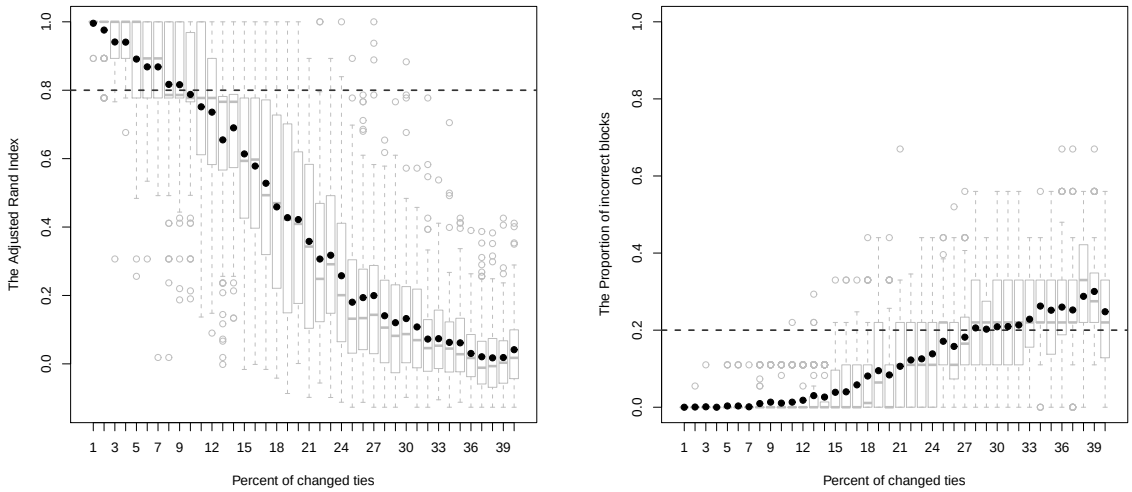
the boy-girl liking ties network, so the number of combinations for selecting different number of changed ties increases even faster. The number of ties in the network with 15 actors, because loops are not allowed, is 210. This means that we have 210 possibilities how one tie can be changed, $\binom{210}{2} = 23871$ combinations for selecting two ties to be changed, $\binom{120}{3} = 1726669$ possible combinations for selection of three ties,... Similarly as for the boy-girl liking ties network (Section 7.5.2.1) we decided to run the simulations 50 times for 1% of random error and for higher percent of random measurement errors the simulations were run 100 times. In the simulation study 3950 new measured networks with different percent of errors were constructed and measured established blockmodel was compared with structure shown in Figure 6.2 in page 72.

Figure 7.48 presents results of simulation study with introduced random measurement errors to the student note borrowing network (Figure 6.2 in Section 6.2.1.2). In Figure 7.48(a) dots represent mean values of ARI and boxplots are drawn in gray. The mean values for ARI decline almost linearly when the percent of introduced error increases. According to determined boundaries for acceptable values of $mARI$, only 9% or less of changed ties lead to satisfying measured partition of actors.

If we set up more restrictive rules that also rectangle of boxplots should be above 0.8, then maximally 4% of introduced errors lead to acceptable agreement between partitions. The blockmodeling seems to be more stable in terms of block agreement (Figure 7.48(b)). The mean values of $ErrB$ are below 0.2 for 27% of changed ties or less. For 17% of introduced random measurement errors or less the Err values above 0.2 in almost three-quarters of simulations. The note borrowing network shows similar stability in terms of correctly identified block types as the boy-girl liking ties network and far higher instability in terms of partitions (Figure 7.47).

7.5.3 Studies of simulated networks based on structural equivalence

Studies of empirical networks provide some clues about sensitivity of blockmodels to random measurement errors. The extension of our study to simulated whole networks with known structure and properties will provide us an adequate foundation for assessing the general impact of amount of random measurement on the results of



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

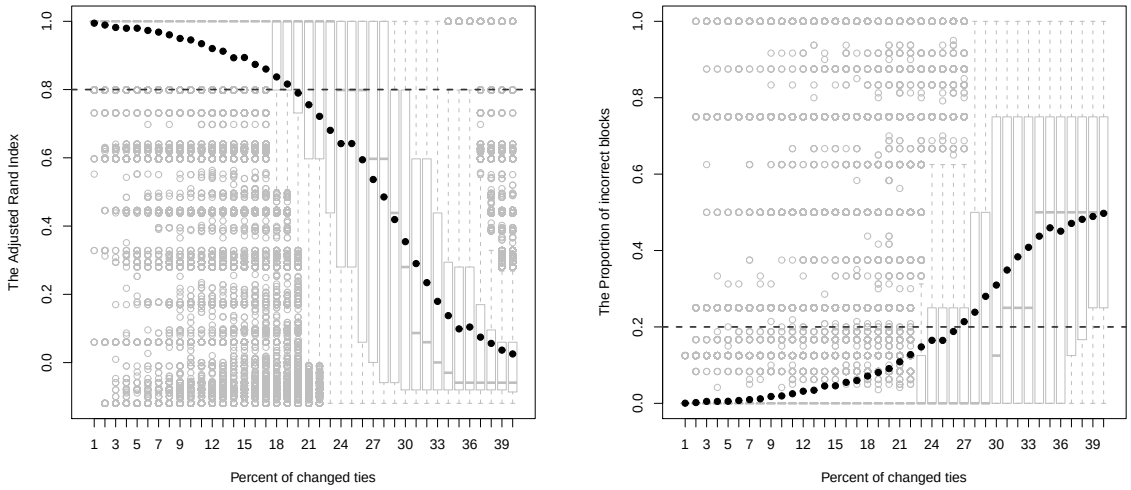
Figure 7.48: Results of the simulation study based on the note borrowing network with random measurement errors

blockmodeling procedure. The strategies for simulated whole networks are described in Section 6.2.3.

7.5.3.1 Results for the completely symmetric blockmodel structure

The whole networks for completely symmetric blockmodel structure follow the blockmodel structure of the boy-girl liking ties network (Figure 6.1 in Section 6.2.1.1) and generation of amount of random measurementerrors described in Section 7.5.1.

Figure 7.49 presents results of the simulation study with introduced random measuremet errors. The mean values of ARI (Figure 7.49(a)) decrease with higher proportion of measurement errors. The decline is not linear, but it is bigger with higher proportion of randomly changed ties. The mean values of the Adjusted Rand Index suggest that the blockmodel is stable in terms of preserving the starting partition if the amount of randomly introduced errors is lower or equal to 19%. As a reminder, according to (Steinley, 2004) we decided that the correspondence of the position memberships is acceptable if values of $mARI \geq 0.8$.



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.49: Results of the simulation study based on the completely symmetric blockmodel structure with random measurement errors

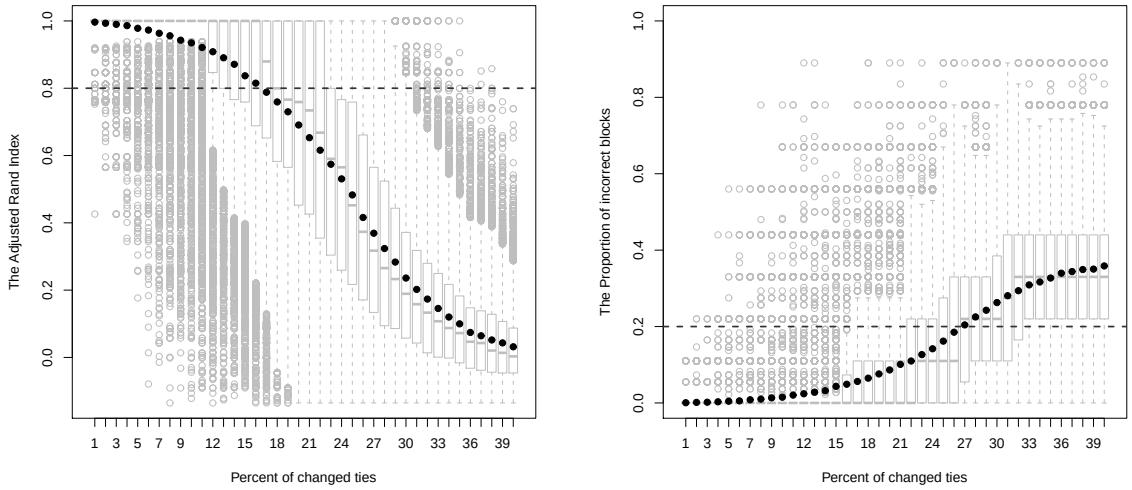
On the other hand, the blockmodel structure, or position and type of blocks in a blockmodel, is acceptable if the mean value of incorrectly identified block types ($mErrB$) is lower than 0.2. The stability of blockmodels of completely symmetric blockmodel structure is higher in terms of block types than in terms of concordance between partitions. If 26% or less of random measurement errors are introduced, the mean values of $ErrB$ suggest that the blockmodel is stable. The rectangles of boxplots or three-quarters of $ErrB$ values are below 0.2 for 23% of introduced errors or less.

7.5.3.2 Results for the first non-symmetric blockmodel structure

The whole networks for the first non-symmetric blockmodel structure (presented in Section 6.2.3.2) is similar to blockmodel structure of the note borrowing network (Figure 6.2 in Section 6.2.1.2) with additional complete block on the diagonal.

Figure 7.50(a) presents results of the simulation study with randomly introduced errors. The mean values of the Adjusted Rand Index decrease with higher percent of introduced measurement errors. The agreement between partitions from the whole and measured network is acceptable for 16% of introduced errors or less, because mean

ARI values are above 0.8.



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.50: Results of the simulation study based on the first non-symmetric blockmodel structure with random measurement errors

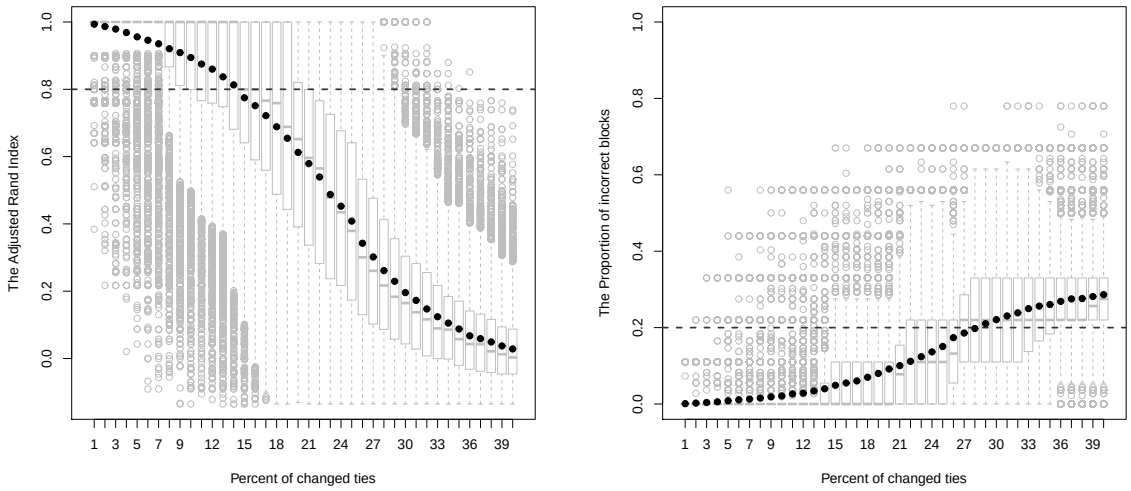
The stability of blockmodel in terms of correctly identified block types for the first non-symmetric blockmodel structure is even higher than in terms of partitions. The mean values of proportion of incorrect block types ($mErrB$) are below 0.2, which indicates acceptable blockmodels for 26% of introduced measurement errors or less (Figure 7.50(b)).

7.5.3.3 Results for the second non-symmetric blockmodel structure

The whole networks for the second non-symmetric blockmodel structure follow the blockmodel structure of the student note borrowing network (Figure 6.2 in Section 6.2.1.2).

Figure 7.51(a) presents impact of randomly introduced errors to the stability of blockmodeling in terms of partitions. The agreement between partitions from the whole and measured network is acceptable, if the mean values of ARI are above 0.8. In case of the second non-symmetric blockmodel structure, the blockmodel is stable in terms of partitions for introduced 19% of random errors or less. If we compare those results to

results for the first non-symmetric blockmodel structure (Figure 7.50(a)), we can conclude that the second non-symmetric structure is more stable in terms of agreement between partitions.



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.51: Results of the simulation study based on the second non-symmetric blockmodel structure with random measurement errors

The second non-symmetric blockmodel structure is stable in terms of correctly identified block types for 28% of introduced errors or less (Figure 7.51(b)). If we compare this result with the first non-symmetric blockmodel structure we can say that the second non-symmetric blockmodel structure is more stable also in terms of correctly identified block types. Therefore, we can conclude that the second non-symmetric blockmodel structure is more stable than the first one on both, micro and macro level of the blockmodel.

7.5.4 Simulated whole networks based on regular equivalence

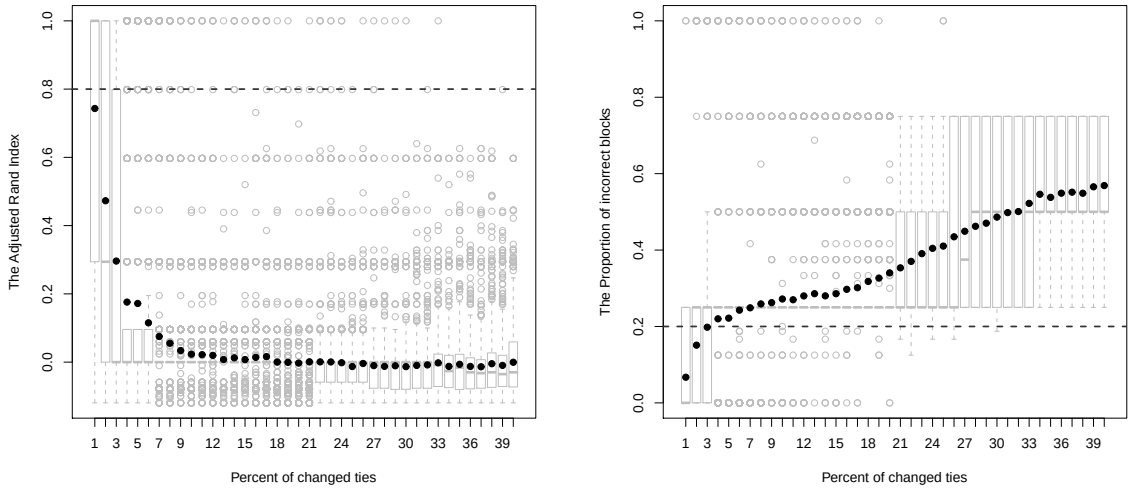
Two blockmodeling structures, the cohesive subgroups model and core-periphery model, were used to generate the starting whole networks based on regular equivalence. Networks are presented in Section 6.2.4.

7.5.4.1 Results for the cohesive subgroup model

Two-clusters partitions

First, the results with smaller network with 10 actors and two-cluster partition will be presented (Section 6.2.4.1). Figure 7.52 presents results with randomly introduced measurement errors in model with two equally large clusters (5 actors in each cluster).

One percent of randomly changed ties (or one tie) leads to unstable blockmodeling in terms of partitions because mean value of the Adjusted Rand Index is below 0.8 (Figure 7.52(a)). When 2% of ties were randomly changed, the $mARI$ is below 0.5, which according to Steinley (2004) indicates poor agreement between partitions. Agreement between correctly identified block types and therefore stability of blockmodeling in terms of block types is just a little bit better.

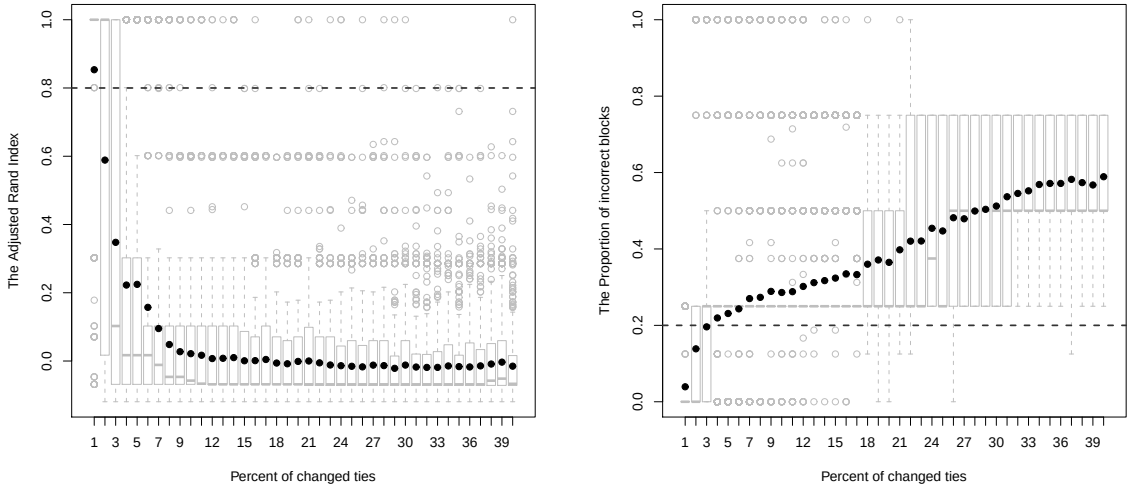


(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.52: Results of the simulation study based on two clusters (5,5) regular cohesive subgroups model with random measurement errors

Mean values of proportion of incorrect blocks in blockmodels are above 0.2 for 4% of randomly introduced errors or more (Figure 7.52(b)). When $mErrB$ is around 0.25 that means that on average one block was incorrectly classified. In more detailed investigation we observed that one null block is changed to regular one and this happened first with 7% of introduced random errors.

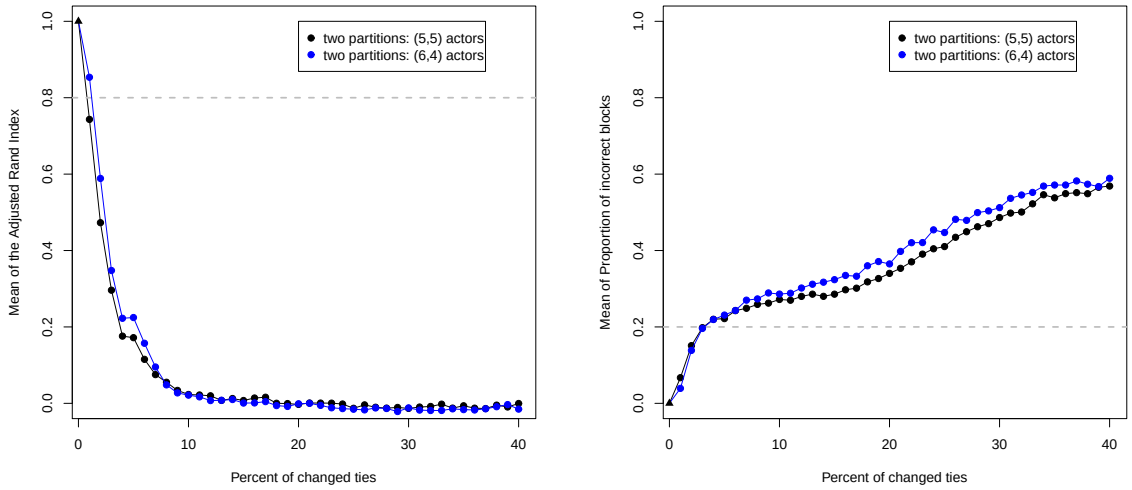
Figure 7.53 shows the results of randomly introduced errors to two-cluster regular cohesive subgroups model with 6 and 4 actors in clusters. For 1% of introduced random measurement errors the mean values of ARI are above 0.8, which indicates good agreement between partitions (Figure 7.53(a)). Higher percent of introduced errors leads to unstable blockmodel with $mARI$ around 0 for 8% of introduced errors or more, which indicates that partition from whole network and from measured network are in fact two random partitions (Figure 7.53(a)). Mean proportion of incorrectly identified block types ($mErrB$) is below 0.2 for at least 3% of introduced errors (Figure 7.53(b)).



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.53: Results of the simulation study based on two clusters (6,4) regular cohesive subgroups model with random measurement errors

Figure 7.54 compares results for two starting partitions in simulation of two-cluster regular cohesive subgroups model. Values of $mARI$ are practically the same, except for 1% of introduced errors, where (6,4) partition is stable and (5,5) is not (Figure 7.54(a)). The mean values of $ErrB$ show a little differentiation between both partitions, where partition with equal membership in both clusters (5,5) has lower values of $mErrB$. Despite these differences, both starting partitions (with (6,4) and (5,5) actors in clusters) are unstable in terms of correctly recovered blockmodel for 4% of randomly introduced errors or more (because $mErrB$ values are above 0.2).



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

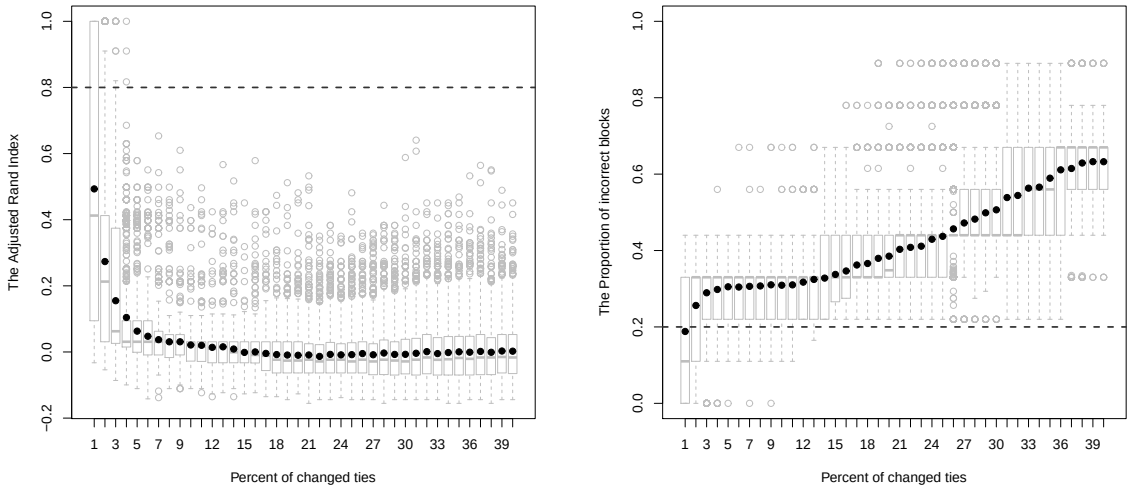
Figure 7.54: Comparison of results of two-clusters regular cohesive subgroups model with random measurement errors

Three-clusters partitions

Starting networks with 15 actors and two three-clusters partitions were used in simulation. One partition has 5 actors in each cluster and the other partition has 4, 5, and 6 actors in separate clusters. The blockmodel of the whole network is presented with image matrix IM_2 from Equation (6.6) in Section 6.2.4.1.

The established blockmodels are even less stable in three-clusters case than in two-clusters examples. Figure 7.55(a) shows that mean values of the Adjusted Rand Index are below 0.5 if just 1% of ties is randomly changed. For higher percent of introduced errors the $mARI$ values are around 0.

Mean values of proportion of incorrect block types are below 0.2 only for 1% of changed ties (Figure 7.55(b)). In range from 3 to 15% of changed ties, the $mErrB$ values are approximately 0.3, which indicates that one third of blocks (3 blocks out of nine) are incorrectly identified in an established measured blockmodel. For 16% of introduced errors or more the $ErrB$ values start to increase linearly.



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.55: Results of the simulation study based on three-clusters (4,5,6) regular cohesive subgroups model with random measurement errors

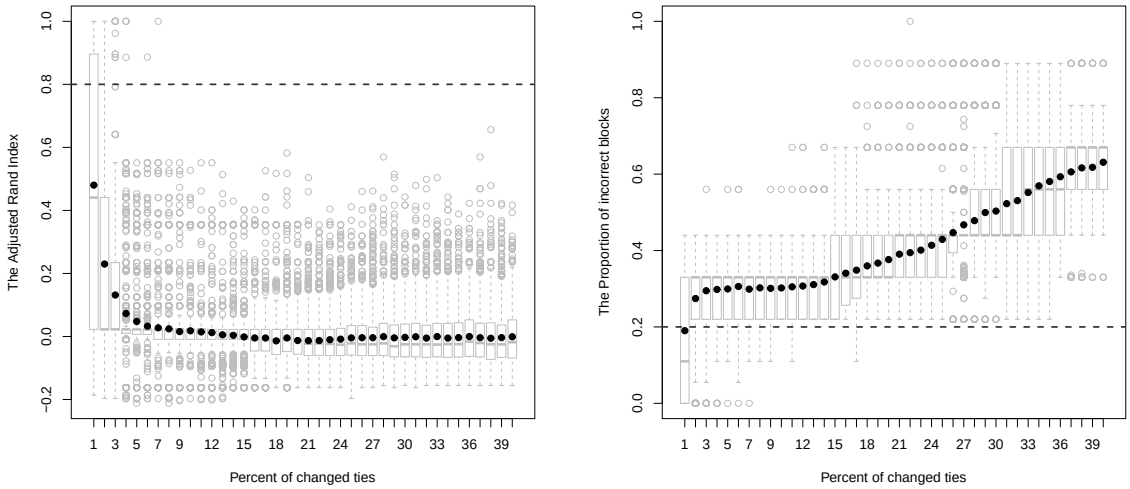
The three-cluster partition with 5 actors in each cluster shows that partition is unstable also if only one percent of ties is changed (Figure 7.56(a)). The stability of blockmodel in terms of correctly identified block types ($mErrB$) is also poor with values of $ErrB$ below 0.2 just for 1% of randomly changed ties (Figure 7.56(b)).
Results for both three-clusters partitions together are presented in Figure 7.57. There are no major differences neither in stability of partition ($mARI$) nor in stability of block types ($mErrB$).

7.5.4.2 The core-pheriphery model

The scheme of core-periphery model and image matrices of starting whole networks in simulations are presented in Section 6.2.4.2. Results with two-cluster partitions with 10 actors and three-clusters partitions with 15 actors are presented below.

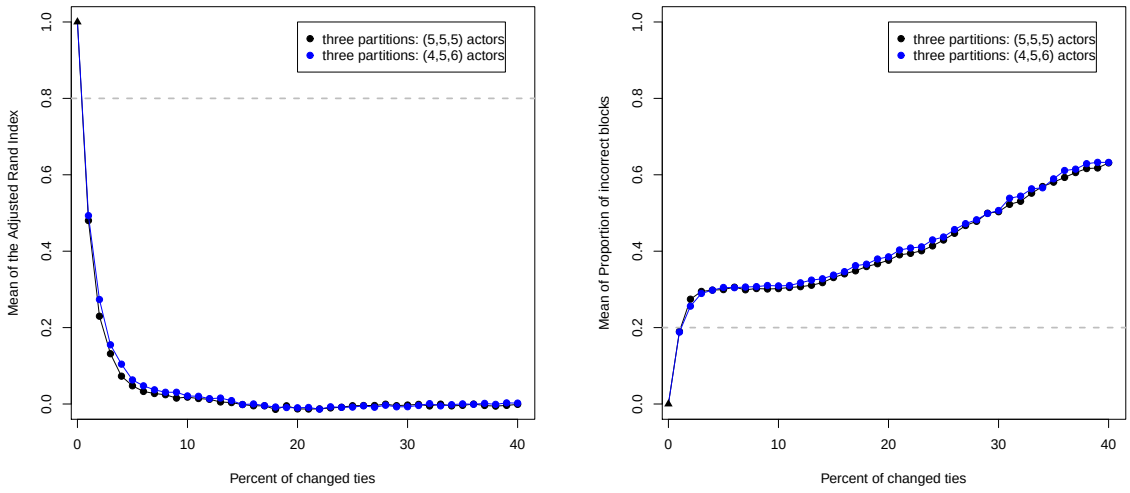
Two-clusters partitions

Results of introduced random errors to regular core periphery structure with 6 actors in core cluster and 4 in periphery are presented in Figure 7.58(a). The mean values of Adjusted Rand index are above 0.8 only for 1% of changed ties, which indicates good



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

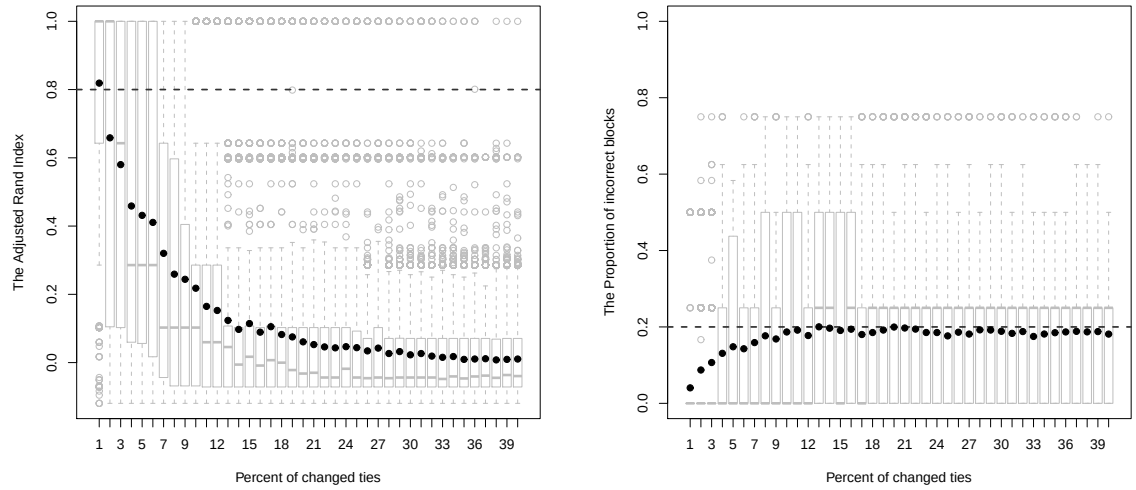
Figure 7.56: Results of the simulation study based on three-clusters (5,5,5) regular cohesive subgroup model with random measurement errors



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.57: Comparison of results of three-clusters regular cohesive subgroups model with random measurement errors

agreement between real and measured partitions. For higher percent of changed ties values of $mARI$ decrease to 0, but this fall is a little bit slower than in regular cohesive subgroup model (Figure 7.58(a)).



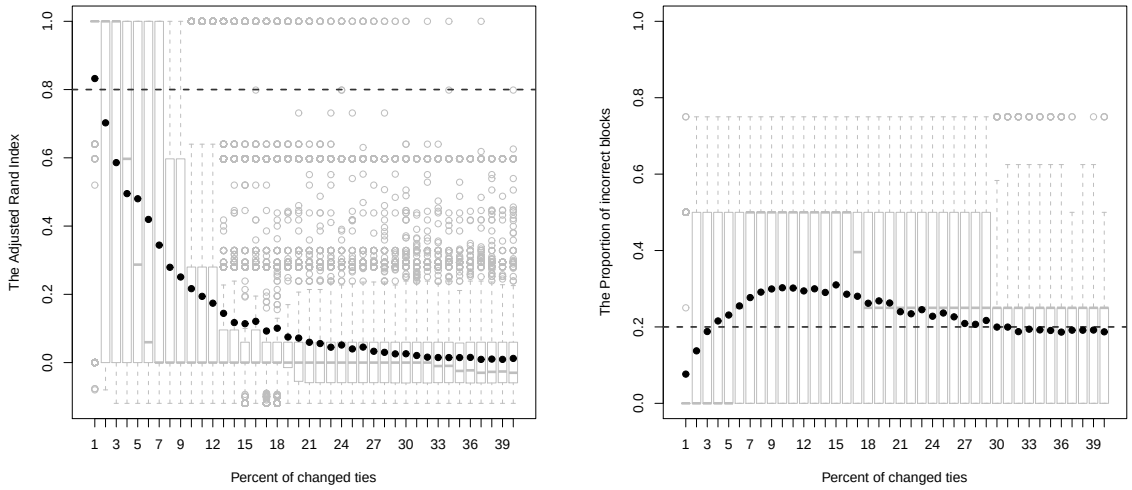
(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.58: Results of the simulation study based on two clusters (6,4) regular core-periphery model with random measurement errors

The mean proportion of incorrectly identified block types ($mErrB$) is less or equal to 0.2 for all range of introduced errors (Figure 7.58(b)). As written in Section 5.1.2, we could say that blockmodel is stable in terms of correctly identified block types in a blockmodel. Boxplots show that for higher percentages of introduced errors (17% or more) three-quarters of $ErrB$ values are below 0.25 which indicates that one out of 4 blocks is incorrectly identified.

The mean values of ARI are above 0.8 only for 1% of randomly changed ties, for higher percentage of changed ties the $mARI$ values exponentially drop to zero (Figure 7.59(a)). This pattern in case of equally large core and periphery clusters (5 actors in both of them) is similar to previously presented partition with larger core cluster (Figure 7.58(a)) and is also observed in the following figure (Figure 7.60(a)) with larger periphery cluster.

The values of $mErrB$ show stirred patterns. First, values quickly increase to 15% of changed ties, where $mErrB$ is around 0.3 and then values slowly decrease to 0.2 (Figure 7.59(b)). The measured blockmodels are stable for 3% or less of randomly introduced



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

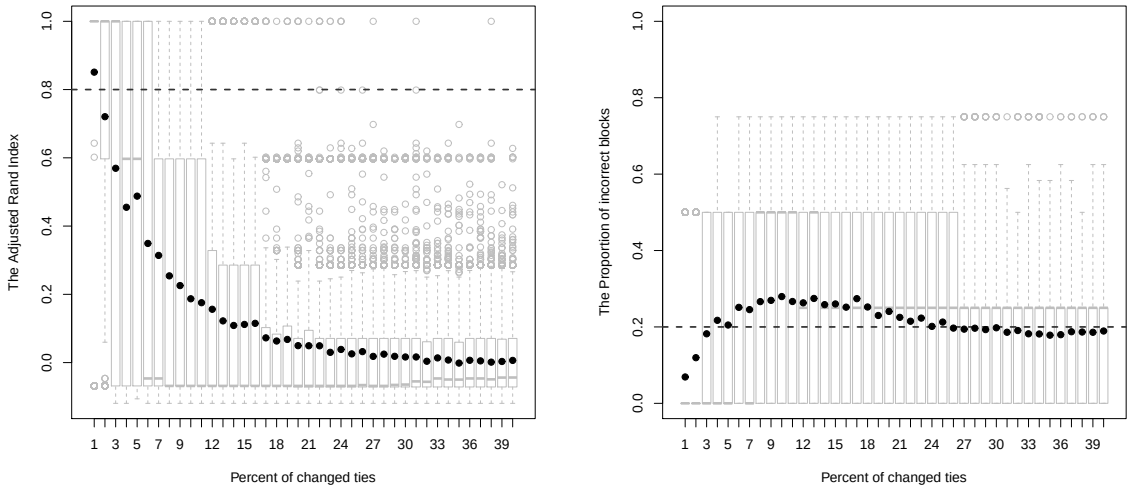
Figure 7.59: Results of the simulation study based on two-clusters (5,5) regular core-periphery model with random measurement errors

errors. Between 30% and 40% of changed ties the $mErrB$ values are again below 0.2, but if we also take into consideration the results for stability of partitions, we can not conclude that blockmodel is unstable for higher percentages of change ties.

As mentioned before, the results for the Adjusted Rand Index with smaller core cluster with four actors and periphery cluster with six actors are similar to other two starting partitions (five actors in both, core and periphery, clusters and bigger core cluster with six actors). Only 1% of changed ties leads to $mARI$ values above 0.8 (Figure 7.60(a)).

The pattern of correctly identified block types is similar to the above example of equally large core and periphery clusters. The blockmodel is stable in terms of $mErr$ for 3% of randomly changed ties or less (Figure 7.60(b)). For 26% of changed ties or more, the mean proportion of incorrectly identified block types is again below 0.2, but the stability of partitions ($mARI$ values) indicates that there is no satisfied agreement between both blockmodels.

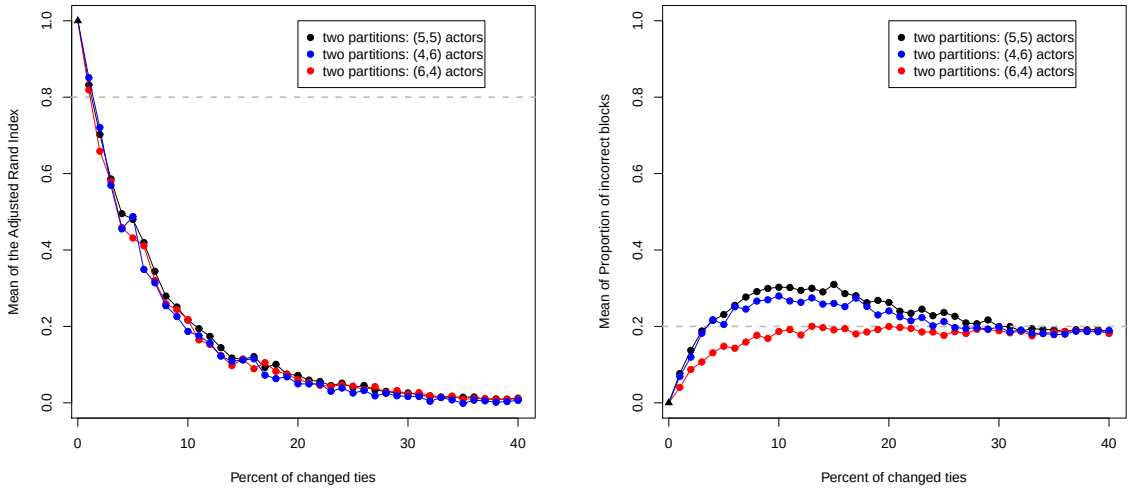
Combined results for all three real partitions with different number of actors in each



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.60: Results of the simulation study based on two-clusters (4,6) regular core-periphery model with random measurement errors

cluster in Figure 7.61(a) indicate that there is no clear differentiation between partitions in terms of $mARI$ values. As observed above, the agreement between three image matrices is the best, when the core cluster is larger than the periphery one (Figure 7.61(b)).



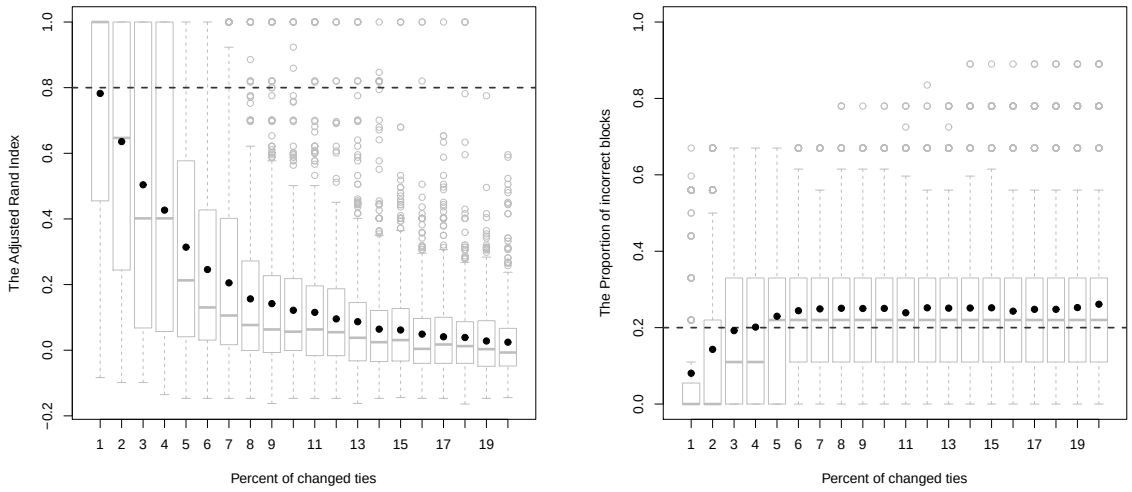
(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.61: Comparison of results of two-clusters regular core-periphery models with random measurement errors

Three-clusters partitions

Three types of three-clusters partitions with regular core-periphery models (presented with image matrix IM_4 in Section 6.2.4.2) were used in the next simulations. Because of unstable results for low percent of introduced random measurement errors in previous examples, the maximal amount of introduced random errors in these examples is 20%. First, the results for starting whole blockmodel with larger first core cluster with six actors, second core cluster with 5 actors and periphery cluster with 4 actors are presented.

The $mARI$ value for one percent of introduced measurement errors is a little bit below 0.8, which indicates moderate agreement between 'real' and 'measured partition' (Figure 7.62(a)). The mean values of incorrectly classified block types ($mErrB$) are below 0,2 for maximally 4% of introduced errors. For higher percentages of introduced random errors, $mErrB$ values are around 0.25 (Figure 7.62(b)).

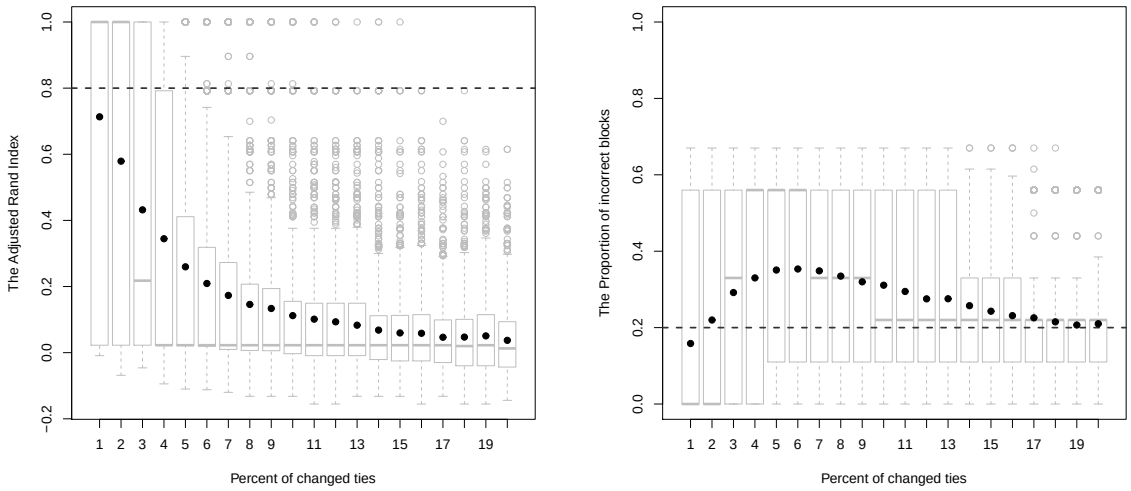


(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.62: Results of the simulation study based on three-clusters (6,5,4) regular core-periphery model with random measurement errors

In three-cluster partition with equally large cores and periphery clusters (5 actors in each of them) the $mARI$ values are below 0.8 for the whole range of introduced random measurement errors and exponentially decline with higher percent of introduced errors (Figure 7.63(a)). The mean proportion of incorrectly identified block types is be-

low 0.2 only for 1% of changed ties which indicates acceptable agreement between both image matrices (Figure 7.63(b)). For 5% of introduced errors $mErrB$ values increase to 0.4 and then decrease slowly back to 0.2, but stay above this threshold.

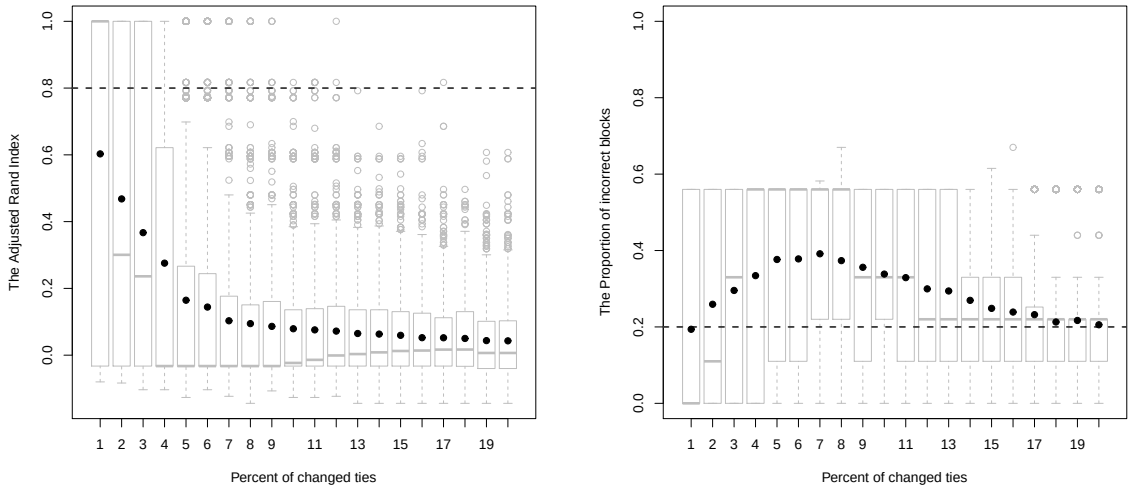


(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.63: Results of the simulation study based on three-clusters (5,5,5) regular core-periphery model with random measurement errors

The third partition with regular core-periphery model has the smallest first core cluster with 4 actors, second core cluster has 5 actors and the periphery cluster is the biggest with 6 actors. The $mARI$ value for one percent of introduced measurement errors is below 0.65, which indicates poor agreement between starting and measured partitions (Figure 7.64(a)). The $mErrB$ values are below 0.2 only for one percent of randomly introduced measurement errors and the pattern is similar as in the above example. Values for $mErrB$ increase to 5% of introduced errors and then linearly decline and approach to 0.2 with 20% of changed ties (Figure 7.64(b)).

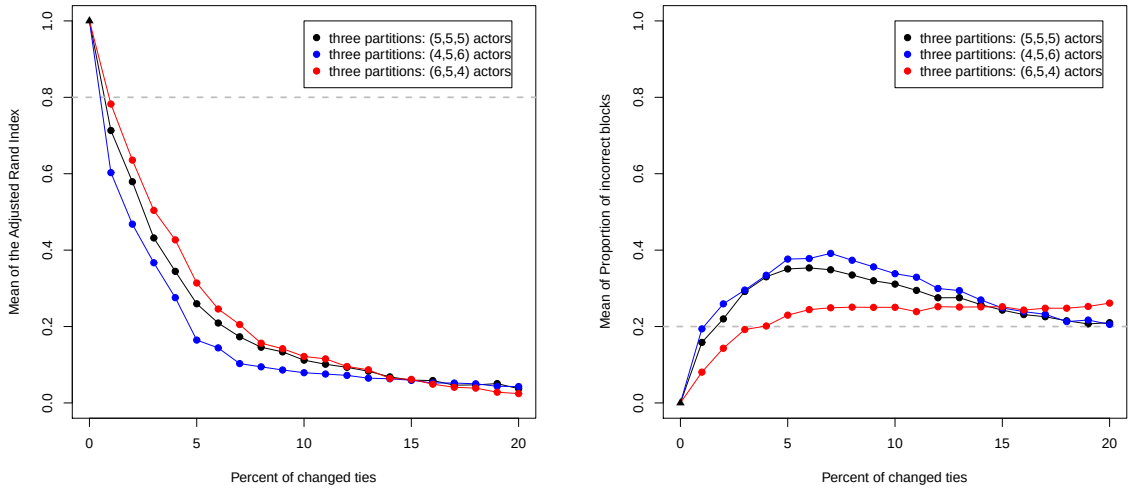
Comparison between all three types of starting partitions is presented in Figure 7.65. The highest values for $mARI$ has C_{654} partition with biggest core cluster. Values of $mARI$ for all three partitions exponentially decrease and are for the whole range below 0.3, which indicates that blockmodeling is unstable in terms of agreement between membership of actors (Figure 7.65(a)). The patterns in values of $mErrB$ (Figure 7.65(b))



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.64: Results of the simulation study based on three-clusters (4,5,6) regular core-periphery model with random measurement errors

reveal that partition with biggest core cluster (C_{654}) is the most stable also in terms of correct block types in a blockmodel. Acceptable agreement between starting and measured image matrices is obtained with at most 4% of changed ties.



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.65: Comparison of results of three-clusters regular core-periphery model with random measurement errors

If we compare results for $mErrB$ from both, two-clusters (Figure 7.61) and three-clusters partitions (Figure 7.65(b)) for regular core-periphery models, we observe that partitions with biggest core cluster have the lowest values of $mErrB$. Patterns for other two partitions are similar to each other and values of $ErrB$ are higher.

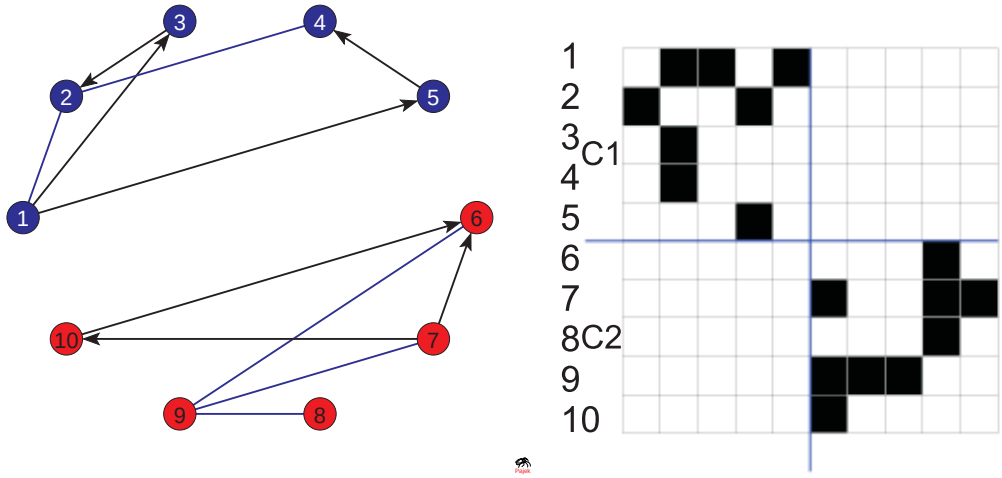
7.5.4.3 Detailed view on regular equivalence

We try to ascertain why regular equivalence is extremely unstable and what happens with blockmodel when random measurement errors are introduced. Therefore, some particular whole starting networks with known blockmodel structure based on regular equivalence and with small amount of randomly generated errors are presented below.

Example of a network with calculated probability $pTie_{reg}$ of a tie in a regular block
First, an example of a network for regular cohesive subgroup model with probability of ties in a regular block ($pTie_{reg}$) calculated from Equation (6.5) was examined. The network is presented in Figure 7.66(a). The starting partition for generation of network is two-cluster partition with 5 actors in each cluster (C_{55}) and the starting image matrix has regular blocks on the diagonal. According to Equation (6.5) the probability for generation of ties in a regular blocks is $\frac{1}{4}$. As described in Section 6.2.4.1, regular blocks are checked for regularity and ties are enforced if the regularity condition of at least one 1 in each row and each column is not satisfied. The density of presented network is 0.19 and the mean density of regular blocks is 0.425.

Figure 7.66(b) presents the same network in matrix format and it reveals that the best fitting partition with zero inconsistencies is the same as the partition used in generation of network. In network representation clusters are presented with different colors, blue and red.

Random measurement errors were introduced to the presented network. First, we randomly introduced 1% of errors and Figure 7.67 shows that a tie from actor 3 to actor 2 was deleted. This deletion of a tie causes that actor 3 has no outgoing ties and in matrix representation in Figure 7.66(b) this results in an empty row. The blockmodeling pro-



(a) Network for cohesive subgroups model

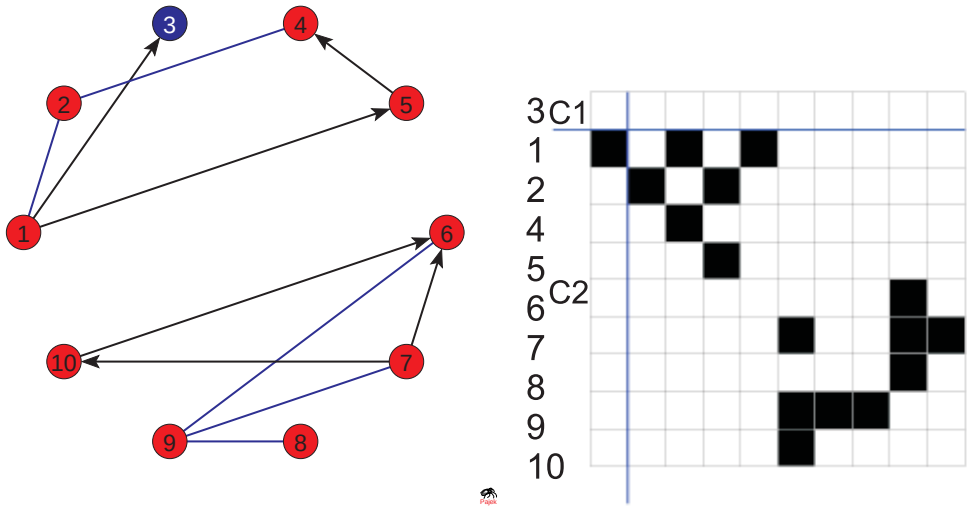
(b) Sociomatrix with two clusters based on regular equivalence

Figure 7.66: Example of a network for cohesive subgroups model with probability of ties in regular block $pTie_{reg}$ and corresponding sociomatrix with best fitting two-cluster partition

cedure into two clusters was run on the measured network and results are presented in Figure 7.67. We obtained one cluster with just one actor (actor 3) and second cluster with nine actors. As described in Section 5, the agreement between two blockmodels is measured with two indices. The Adjusted Rand Index between real starting partition C_{55} and measured partition $c(1,1,2,1,1,1,1,1,1,1)$ is 0, which indicates absolutely no agreement between them.

Sociomatrix in Figure 7.67(b) shows that the image matrix has three null blocks and one large regular block within the second cluster. Therfore the proportion of incorrectly identified block types ($ErrB$) is 0.75. The value of criterion function is 1 (one present tie in a null block). This example shows that one changed tie can completely destroy the obtained blockmodel based on regular equivalence.

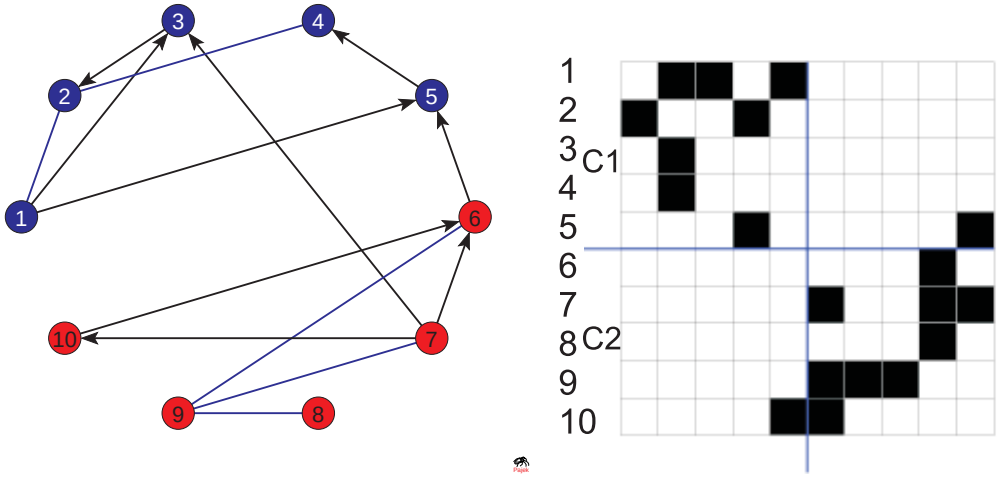
The second question was what happens if larger amount of random measurement errors is introduced. 2% of random measurement errors were introduced to the presented network in Figure 7.66. Figure 7.68(a) shows that two ties were added; a tie



(a) Network (b) Sociomatrix

Figure 7.67: Measurement network with one changed tie and corresponding sociomatrix with best fitting two-cluster partition

from actor 6 to actor 5 and a tie from actor 7 to actor 3.



(a) 'Measured' network with two changed ties (b) Sociomatrix with two-cluster partition C_{55}

Figure 7.68: Measurment network with two changed ties and corresponding sociomatrix with two-cluster partition C_{55}

The blockmodels based on regular equivalence was established for measured network with 2 changed ties. As mentioned above, sociomatrix of real network (Figure 7.66(b))

reveals that the best fitting partition of starting real network with zero inconsistencies is the same as the partition used in generation of network. The next step was the calculation of criterion function if the starting partition with 5 actors in each cluster (C_{55}) is enforced to measured network (Figure 7.68(a)). Figure 7.68 reveals that image matrix represents two cohesive subgroups and that value of criterion function is 2.

According to the results presented in Section 7.5.4.1, we made a conclusion that blockmodeling based on regular equivalence is extremely unstable. Therefore, there is high chance that blockmodel of measured network presented above is not the best one. The blockmodeling procedure was run again on measured network, and results in Figure 7.69 reveal that partition C_{55} is one of three equally well fitting partitions. We obtained two partitions with one actor in a cluster (actor 10 or actor 8) and with the same image matrix as in 'the one changed tie example' presented above. There is no accordance between first two partitions and partition C_{55} (ARI=0) and the image matrix is highly deformed with three quarters of incorrectly identified block types. The third solution (the right panel in Figure 7.69) presents the same partition C_{55} (and image matrix) as was used in generation of a starting network. Because there is no objective criteria for selection between equally well fitting partitions, the mean values of indices of stability (ARI and $ErrB$) were calculated. Consequently, the mean value of the Adjusted Rand Indices of three equally well fitting partitions is 0.33 and the mean value of the proportion of incorrectly identified block types is 0.5. The above results help to understand the high instability of blockmodeling based on regular equivalence. In many examples of established measured blockmodels (from networks with randomly introduced errors) based on regular equivalence equally well fitting partitions are obtained. They can be completely different to the real starting partition or just the opposite, they could be identical. Because at the moment we have no objective quantitative criteria for selection between well fitting partitions, the regular equivalence should be used with extra caution and combined with additional knowledge of the researchers.

The starting network (Figure 7.66) with calculated probability of ties in regular blocks $pTie_{reg}$ from Equation (6.5) has the mean regular block density equal to 0.425. The

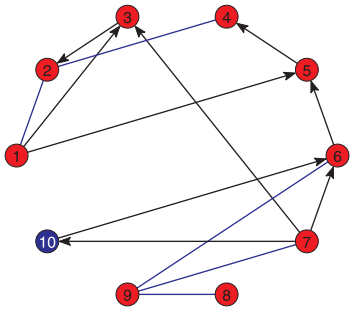
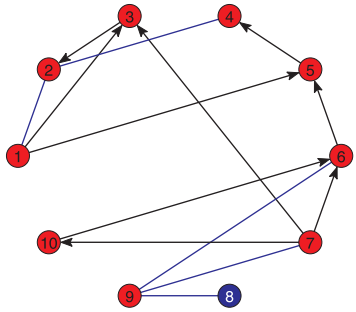
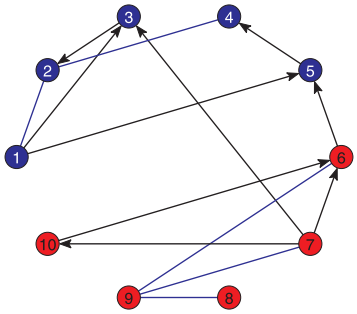
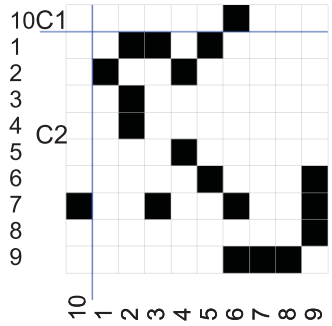
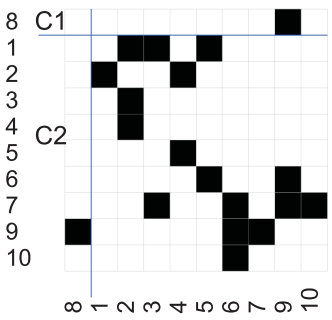
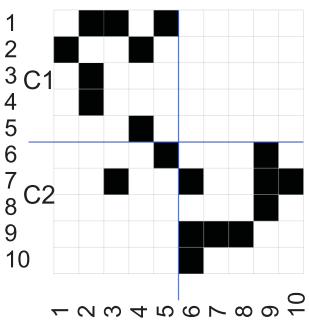
par	(2,2,2,2,2,2,2,2,2,1)	(2,2,2,2,2,2,2,2,1,2,2)	(1,1,1,1,1,2,2,2,2,2)
IM	C1 C2 C1 null null C2 null reg	C1 C2 C1 null null C2 null reg	C1 C2 C1 reg null C2 null reg
P(C)	2	2	2
network			
matrix			
ARI	0.0000	0.0000	1.0000
ErrB	0.75	0.75	0.00

Figure 7.69: Results ob blockmodeling procedure based on regular equivalence with three equally well fitting partitions for 'measured' network with two changed ties

ties were enforced to the network during the generation process to satisfy the regular equivalence condition of at least one tie in each row and column. Therefore, our suspicion was that the main reason for high instability of blockmodeling based on regular equivalence were this 'barely' regular starting blockmodels. The next step was to examine what effect has higher density of regular blocks to stability of blockmodeling procedure.

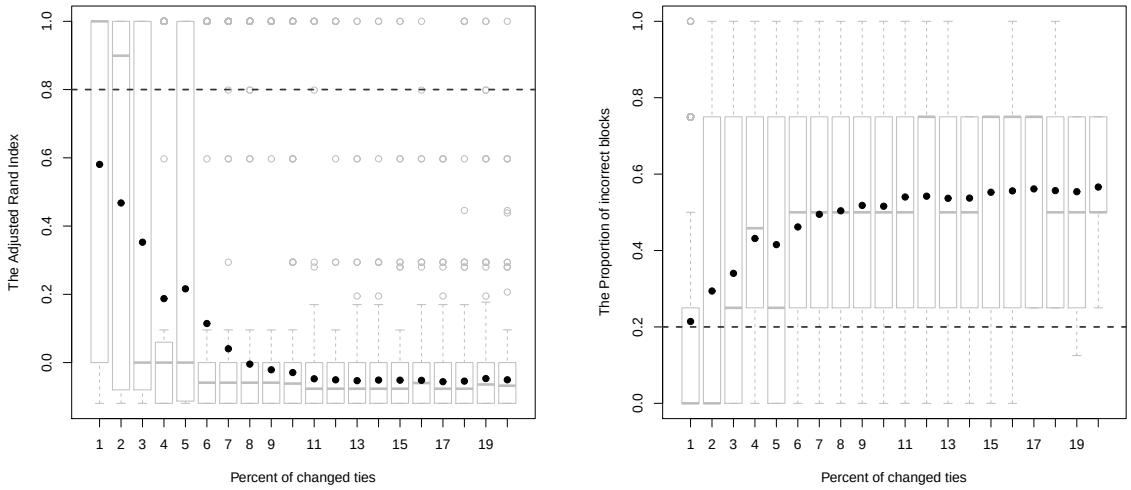
Example of a network with probability of a tie in a regular block equal to 0.6
In the algorithm for generating starting networks on page 81 in Section 6.2.4.1 the probability of ties in regular blocks was set to 0.6 instead of calculated probability from Equation 6.5. The network was still checked for regularity and additional ties were added if necessary.

Similarly as for regular cohesive subgroups model with calculated density of regular blocks (Section 6.2.4.1), ten starting networks were generated. The densities of networks were in range from 0.23 and 0.31 with mean network density equal to 0.28 and standard deviation 0.03. The densities of regular blocks are in range from 0.52 to 0.70 with mean density of regular blocks equal to 0.62 (sd=0.07). In comparison, the mean density of regular blocks in networks with computed probability of regular ties based on block size is 0.29.

The results of stability of blockmodeling are presented in Figure 7.70(a) and are almost the same as in Figure 7.52. The blockmodeling results are extremely unstable in terms of agreement between partitions and image matrices already for 1% of introduced measurement errors. The agreement between block types in two image matrices for five or more percent of introduced errors leads on average to half of incorrectly identified block types ($mErrB \approx 0.5$).

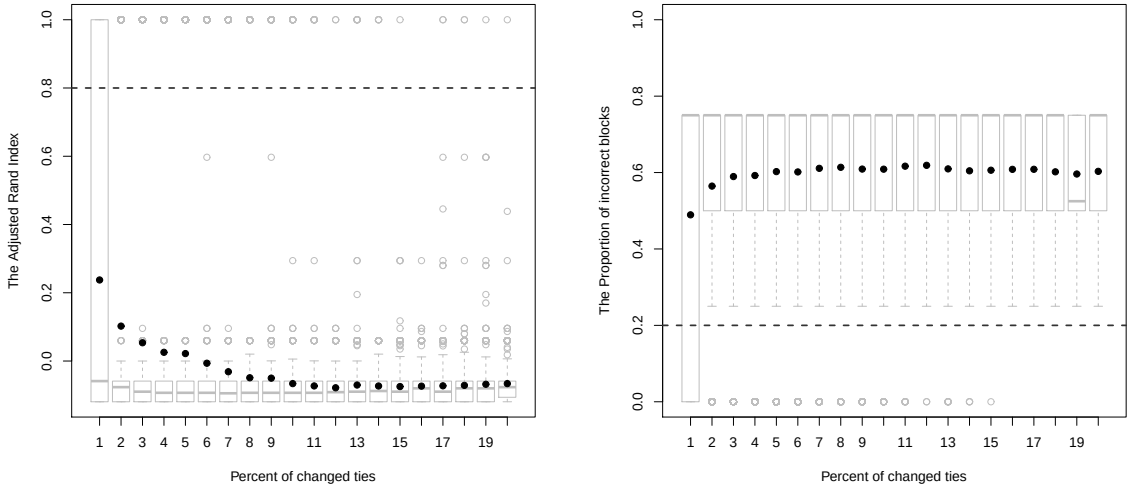
In the next step, we set the probability of ties in regular blocks even higher, $pTie_{reg} = 0.8$. The mean density of 10 starting real networks is 0.33 and standard deviation is 0.02. The differences from previously introduced networks (with $pTie_{reg} = 0.6$) are even higher in mean density of regular blocks. Density of regular blocks ranges from 0.68 to 0.82 with mean value 0.75 (sd=0.046).

Figure 7.71 shows the results of blockmodeling procedure based on regular equivalence. The blockmodeling results show similar, but also a little more extreme patterns than those presented in Figure 7.70. The mean values of proportion of incorrectly identified block types are on average equal to 0.6.



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.70: Results of the simulation study based on two-clusters (5,5) regular cohesive subgroup model with $pTie_{reg} = 0.6$ and introduced random measurement errors

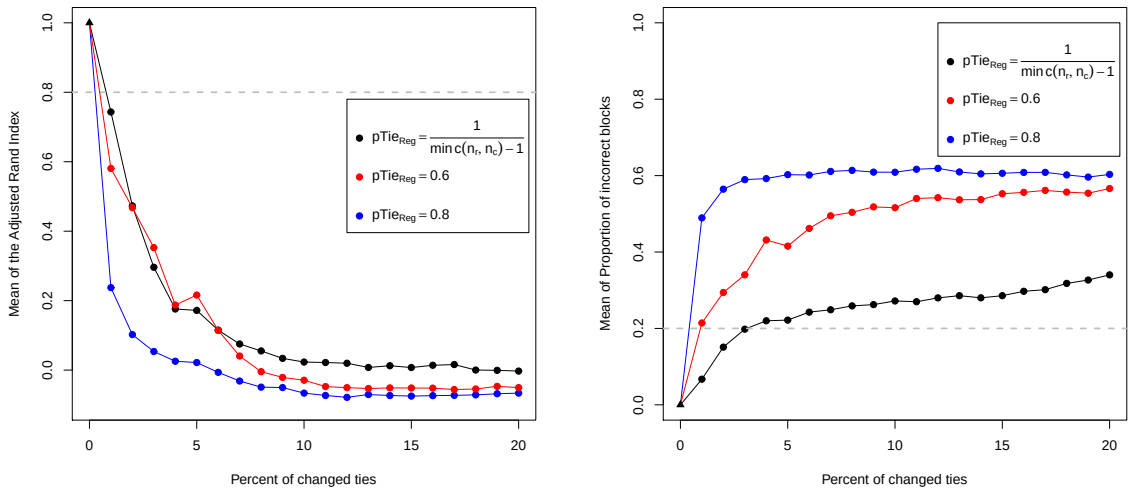


(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.71: Results of the simulation study based on two-clusters (5,5) regular cohesive subgroup model with $pTie_{reg} = 0.8$ and introduced random measurement errors

Comparison of stability of blockmodeling for regular cohesive subgroup model with different probabilities of ties in regular blocks is presented in Figure 7.72. As inter-

preted above, the higher densities of regular blocks lead to even poorer agreement between partitions. Differences are higher in comparison of index of agreement between two image matrices (Figure 7.72(b)). The $mErrB$ for networks with probabilities of ties in regular blocks calculated based on Equation (6.5) increases with higher percent of introduced errors to 0.3. Networks with mean density of regular blocks around 0.6 have on average proportion of incorrectly identified block types equal to 0.5 (for five percent or introduced errors on more). Mean values of $ErrB$ are constant around all range of introduced errors and are approximately equal to 0.6 for networks with density of regular blocks 0.8.



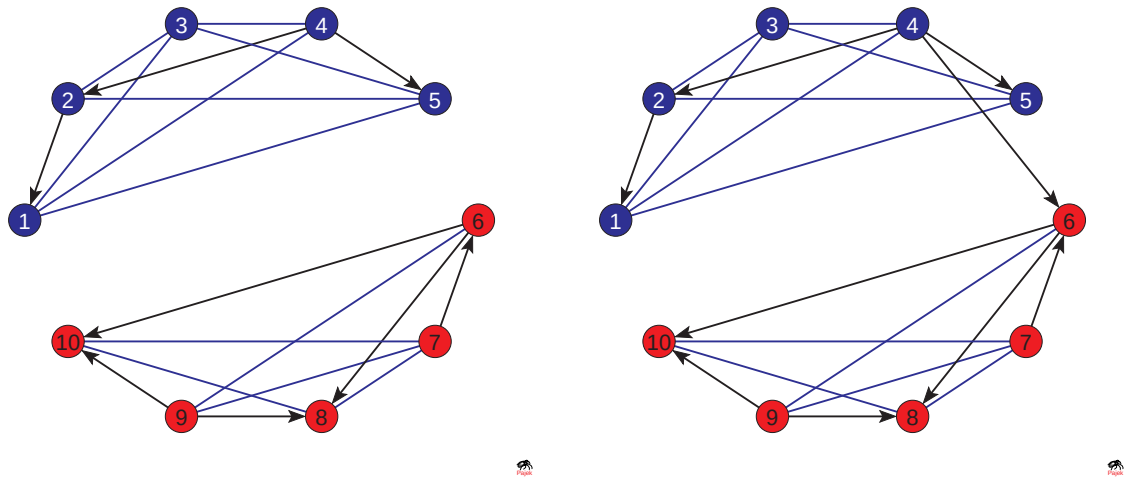
(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.72: Comparison of results for the simulation studies based on two-clusters (5,5) regular cohesive subgroup model with different probabilities of ties in regular blocks and introduced random measurement errors

Therefore, our suspicion that models with higher density of regular blocks could lead to more stable blockmodeling results turn out to be incorrect. On concrete example of a network we tried to examine what happens to blockmodel when small amount of errors is introduced.

Figure 7.73(a) presents network with regular cohesive subgroup model with probability of ties in regular blocks ($pTie_{reg}$) equal to 0.8. Blue and red color present two clusters

for starting partition, which was used in the simulation of a network. We randomly changed one tie in a network and we got an example of measured network. Figure 7.73(b) presents measured network where a tie is added from actor 4 to actor 6.

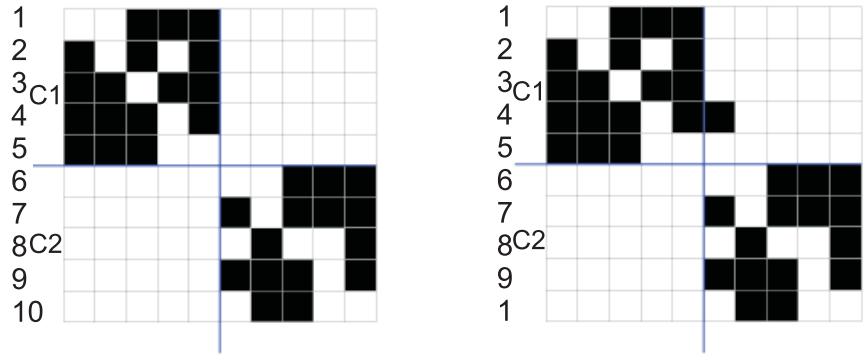


(a) Network for cohesive subgroups model (b) 'Measured' network with one changed tie

Figure 7.73: Example of a network for cohesive subgroups model with probability of ties in regular block $pTie_{reg} = 0.8$ and measured network with one changed tie

The blockmodeling procedure based on regular equivalence was run with both networks from Figure 7.73. Patterns in a sociomatrix in Figure 7.74(a) show that with blockmodeling procedure (for real starting network) we get the same partition which was used in simulation of a network (7.73(a)). The blockmodel has two regular blocks in the diagonal and null blocks out of diagonal. Value of criterion function is 0, which means that we get perfectly fitting partition. If the measured blockmodel is stable in terms of partition, then the partition of real starting network should be the best fitting partition also for the measured network. We imposed the partition $C_{55} = (1,1,1,1,1,2,2,2,2,2)$ to the measured network and we observed that the image matrix is the same (Figure 7.74(b)), but the value of criterion function is one (a tie from actor 4 to actor 6 in a null block).

The next step was to examine if the forced measured blockmodel with C_{55} partition is the best one. The answer to that question is presented in Figure 7.75. With block-



(a) Sociomatrix for cohesive subgroups model

(b) Matrix of 'measured' network with one changed tie

Figure 7.74: Sociomatrices for cohesive subgroups model with probability of ties in regular block $pTie_{reg} = 0.8$ and measured network with one changed tie

modeling procedure of measured network we get three best fitting partitions (value of criterion function is 0). In the first case we get partition with six and four actors which are completely disarranged compared to original C_{55} partition. This is also confirmed with value of the Adjusted Rand Index which is equal to -0.1194. Image matrix shows four regular blocks, therefore the value of proportion of incorrectly identified block types is equal to 0.5. The second and the third blockmodel show equal image matrix with half of correctly identified block types in a blockmodel. Both ARI indices are around 0, which indicates no agreement between real starting and measured partition.

When the results of two changed ties in the network with calculated probability $pTie_{reg}$ (from Equation 6.5) of a tie in a regular block were presented, we suggested that because of obtained equally well fitting partition, the addition expertise knowledge of researchers should be used in blockmodeling procedure based on regular equivalence. Remember, one of three equally well fitting paritions was the real one and the other two partitions had (usually) undesired pattern of just one actor in a cluster. In this situation an experienced researcher could be able to select the right partition. The networks with calculated $pTie_{reg}$ have low regular blocks density (Section 6.2.4.1) and the first explanation was that this is the main cause for instability of blockmodels based on regular equivalence. The simulations with higher density of regular blocks show even higher instability of blockmodeling procedure in terms of agreement between parti-

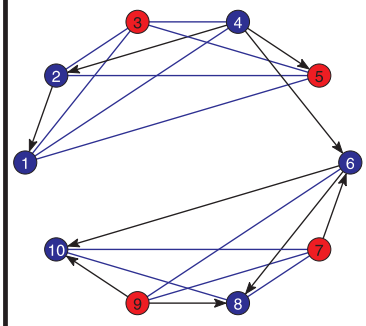
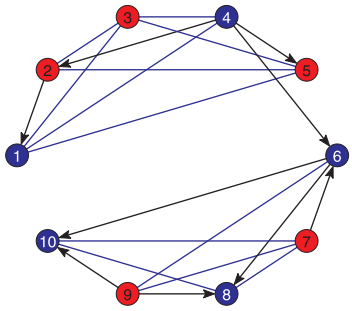
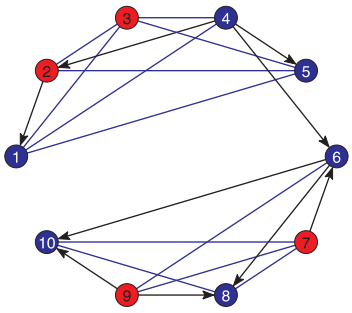
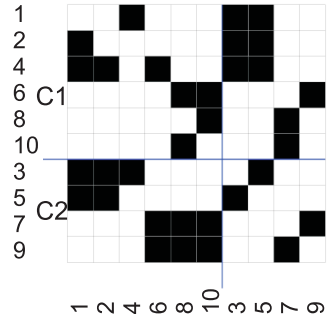
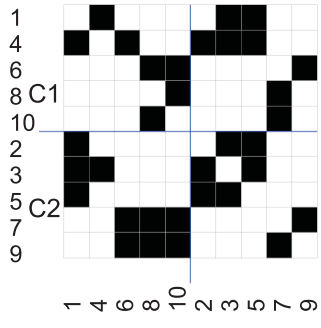
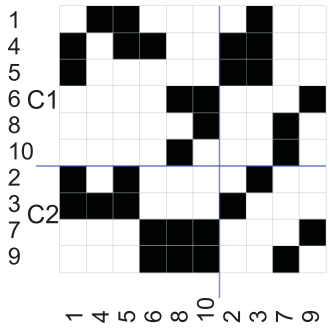
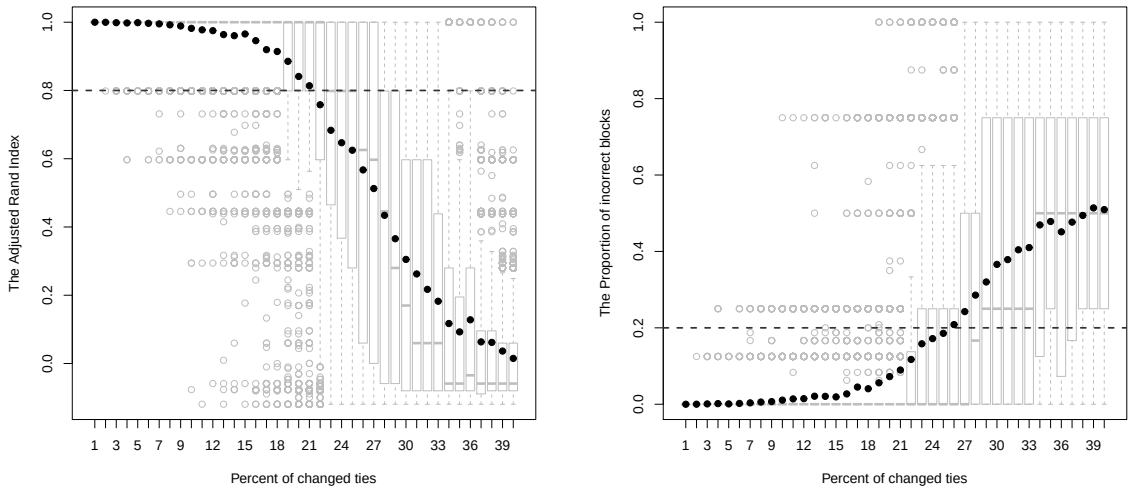
par	(1,1,2,1,2,1,2,1,2,1)	(1,2,2,1,2,1,2,1,2,1)	(1,2,2,1,1,1,2,1,2,1)
IM	C1 C2 C1 reg reg C2 reg reg	C1 C2 C1 reg reg C2 reg reg	C1 C2 C1 reg reg C2 reg reg
P(C)	0	0	0
network			
matrix			
ARI	-0.1194	-0.0800	-0.1194
ErrB	0.50	0.50	0.50

Figure 7.75: Results ob blockmodeling procedure based on regular equivalence with three equally well fitting partitions for 'measured' network with one changed ties

tions and block types in image matrices. But unfortunately, as shown in particular example, none of three equally well fitting partitions are accordant with the real one and there are no objective criteria for selection between them.

One solution could be use of structural equivalence instead or regular one, even if the blocks meet the criteria for regular equivalence. Figure 7.76 shows results of blockmodeling procedure based on structural equivalence with networks of regular cohesive subgroups models with probability of ties in regular blocks equal to 0.8 ($pTie_{reg} = 0.8$).

The real partition has, as in previous examples, five actors in each cluster (C_{55}) and the image matrix has two complete blocks on the diagonal and null blocks out of diagonal.



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.76: Results of the simulation study with two-clusters (5,5) regular cohesive subgroup networks with $pTie_{reg} = 0.8$ with introduced random measurement errors and blockmodeling procedure based on structural equivalence

The obtained blockmodel is stable in terms of partitions for 21% of randomly introduced measurement errors or less ($mARI$ values are above 0.8). The agreement between real and established blockmodel is acceptable for 25% of measurement errors or less, because mean values of incorrectly identified block types ($mErrB$) are lower than 0.2. Similar results were obtained also with boy-girl liking ties network (Section 7.5.2.1) and in simulation study of randomly introduced errors to completely symmetric blockmodel structure presented in Section 7.5.3.1.

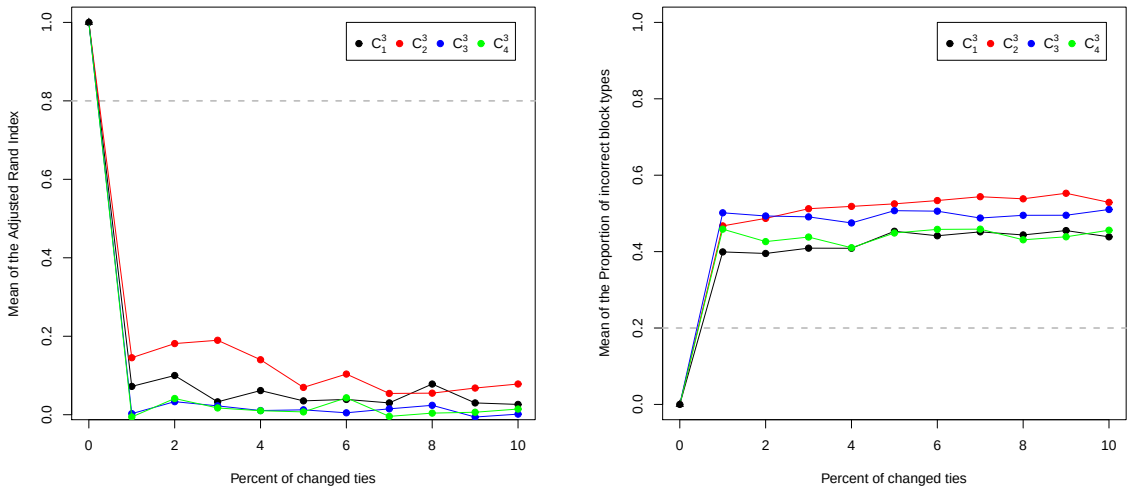
7.5.5 Real networks partitioned based on generalized types of equivalence

In this section the impact of randomly introduced errors on blockmodels established with generalized equivalence will be examined. Only one real network will be used, the Student Government discussion network, therefore the obtained results are used

only for illustrative purposes. Following the results from the previous section about regular equivalence, we expect high instability of blockmodeling also in case of extended selection of block types based on philosophy of generalized equivalence.

The Student Government data was in terms of generalized equivalence extensively studied by Doreian et al. (2005) and the main results of obtained blockmodel are presented in Section 6.2.2.1.

The restriction of block types to {null, com, rdo, cdo, reg} and applied blockmodeling procedure into three clusters produce three equally well fitting partitions. Therefore, in the simulation of randomly introduced errors the measured networks were compared to each of those four equally well fitting solutions. Figure 7.77 shows the obtained results. The colors on the figure distinguish between equally well fitting partitions of the whole network (presented in Table 6.2 on page 76). Despite of small differences the results are overwhelming. One percent of randomly introduced errors completely destroys the position membership (regardless of the starting partition), because $mARI$ values are around zero (Figure 7.77(a)).

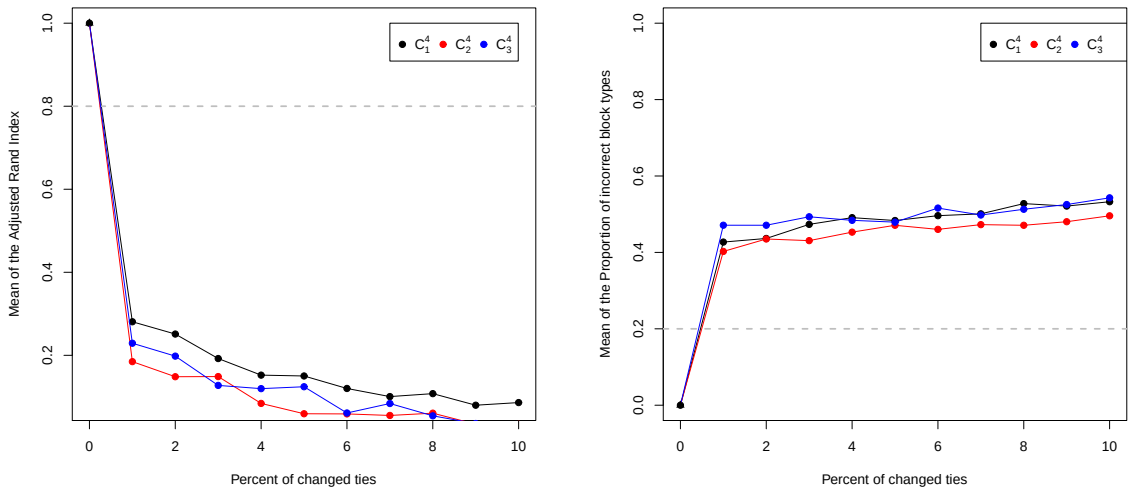


(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.77: Results of the simulation study with Student Government discussion network with introduced random measurement errors and blockmodeling procedure into three clusters based on generalized equivalence with {null, com, rdo, cdo, reg} blocks

The agreement between whole and measured blockmodel in terms of correct block types and positions is also poor. With one percent of randomly introduced errors the $mErrB$ values are around 0.5 which indicates that on average half of blocks is incorrectly identified (Figure 7.77(a)) which is unacceptable.

The blockmodeling with the same blocks types as above into four clusters also results in criterion function with zero inconsistencies. The results for all three equally well fitting starting partitions from the whole network are practically the same (Figure 7.78) and are also similar to the results with three-clusters partitions from the previous figure. One percent of introduced errors completely destroy both, position membership of actors ($ARI \approx 0.2$) and also composition of blocks in the image matrix ($mErrB \approx 0.5$).

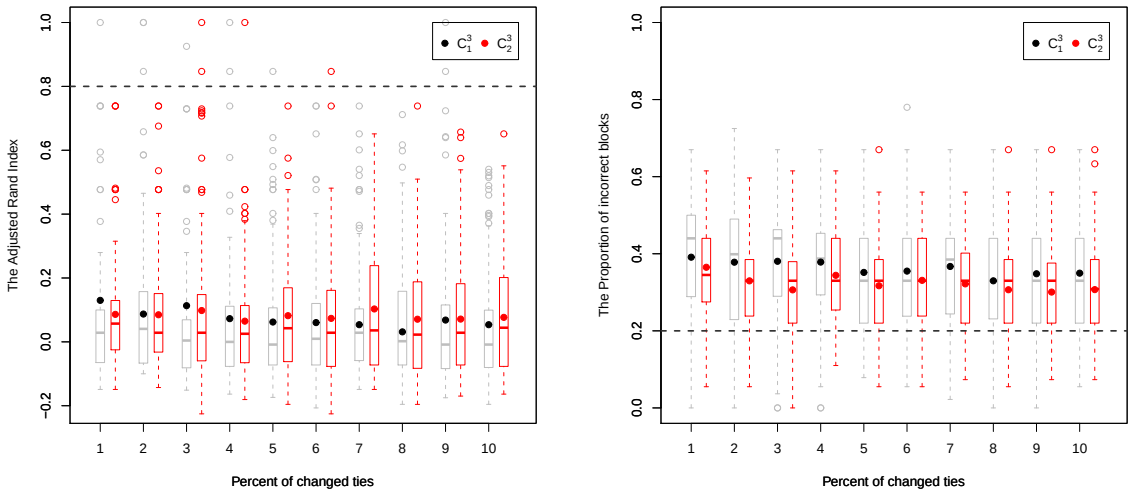


(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.78: Results of the simulation study with Student Government discussion network with introduced random measurement errors and blockmodeling procedure into four clusters based on generalized equivalence with {null, com, rdo, cdo, reg} blocks

In the next step the block types were restricted to {null, rdo, cdo}. There are two equally well fitting partitions of the whole network into three clusters with zero inconsistencies (Table 6.3 on page 76). The results of randomly introduced errors (Figure 7.79) show that the obtained blockmodels are extremely unstable. One percent of randomly changed ties completely destroys the position membership of the actors and

causes incorrect identification of half of blocks in a blockmodel. Practically the same results as described above are also obtained with four-clusters partition based on generalized blockmodeling with block types restricted to {null, rdo, cdo} (Figure 7.80).



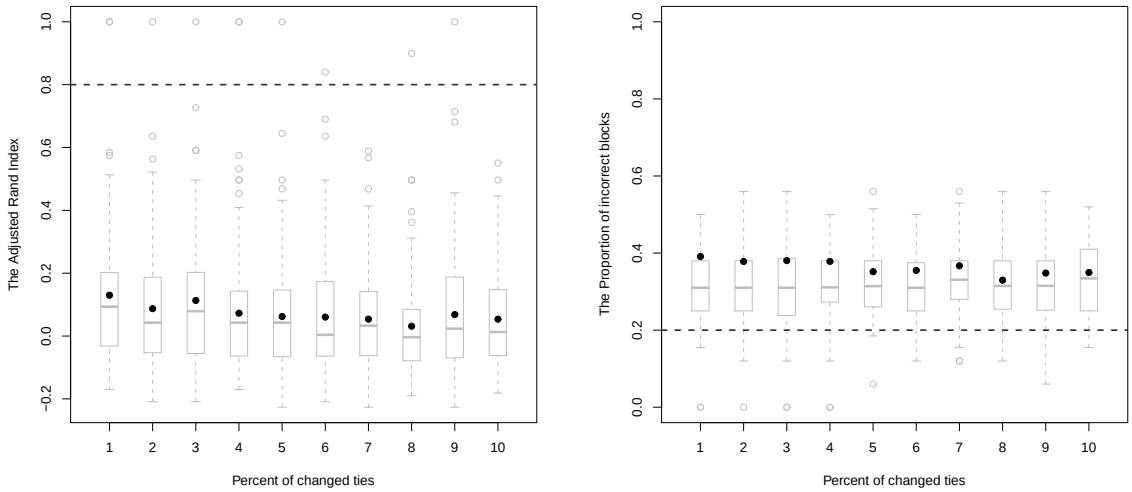
(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.79: Results of the simulation study with Student Government discussion network with introduced random measurement errors and blockmodeling procedure into three clusters based on generalized equivalence with {null, rdo, cdo} blocks

According to the above results we can conclude that blockmodels established based on generalized equivalence (at least those for the Student Government data) are extremely sensitive to the minimal changes in composition of ties. Another interesting result is that with randomly introduced errors equally well fitting partitions behave almost identically. Therefore, randomly introduced errors can not help to distinguish between them in order to find the most stable one.

7.6 Conclusions

Based on the described simulations two main conclusions can be drawn. First, the blockmodeling based on structural equivalence is highly stable. According to the presented results with real and simulated networks, blockmodels are a little bit more stable in terms of blockmodel structure than in terms of position membership of actors.



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $mErrB$

Figure 7.80: Results of the simulation study with Student Government discussion network with introduced random measurement errors and blockmodeling procedure into four clusters based on generalized equivalence with { null, rdo, cdo } blocks

Therefore, the blockmodels based on structural equivalence (regardless of the number of clusters and the symmetry of the blockmodel) are more stable on macro than the micro level of the network.

Second, blockmodels based on regular or generalized equivalence are extremely sensitive to the minor changes in network ties. One randomly changed tie could completely destroy both, position membership and the blockmodel structure. The importance of guidelines made by Doreian et al. (2005) that in generalized blockmodeling the prior knowledge of the researcher should be incorporated in prespecified blockmodels prior to blockmodeling analysis is confirmed with our results. We could probably aggravate this guideline to 'use the generalized equivalence only with previous knowledge about the desired model'.

The suspicion that high density of regular blocks leads to more stable blockmodeling based on regular equivalence turns out to be incorrect. One solution could be the use of structural equivalence even if the networks posses patterns of regular equivalence.

Therefore, our Thesis 1 (from page 36) that *Structural equivalence gives more stable results than regular (or other generalized types) equivalence* can be confirmed. The unstable blockmodeling based on regular equivalence demands the precise revision of the definition of regular equivalence. Based on that the detailed guideline should be made about when to use the regular equivalence and how to correctly interpret the obtained blockmodels.

8 The impact of differences in network characteristics on the stability of blockmodeling

In this chapter we tried to answer the first research question (presented on page 37) about impact of network characteristics and properties on stability of blockmodeling. In more detail, we want to determine to what extent the relative differences in network characteristics and correlation and/or Euclidean distance between vectors of vertex properties are able to predict the results of blockmodeling.

First, the short review of methods used to investigate the impact of network characteristics on the stability of blockmodeling is presented in Section 8.1. The impact of network characteristics on the stability of blockmodeling is then presented with real data (8.2.2) and afterwards our study is expanded to simulated networks (Section 8.3). The network characteristics and properties used in studies are presented in Section 2.2.

8.1 Methods used to investigate the impact of differences in networks characteristic on the stability of blockmodeling

As presented in Chapter 5, the stability of blockmodeling can be measured with two indices. The Adjusted Rand index is used for measuring the agreement between two partitions and the proportion of incorrect block types is used for investigating the amount

of correctly identified block types in a blockmodel.

8.1.1 Relative differences between network characteristics

In Section 2.2 main network characteristics and actor properties are presented. According to second research question, our goal was to examine the impact of relative differences in network characteristics to prediction of blockmodeling result. If the characteristic of a network is expressed by a single number (e.g., network density), then the relative difference between two networks, whole one and measured one, was calculated. The relative difference of two networks according to density *Dens* was calculated as

$$Dens = \frac{|\Delta(\text{whole network}) - \Delta(\text{measured network})|}{\Delta(\text{real network})}\,, \tag{8.1}$$

where $\Delta(\text{whole network})$ indicates the density (Equation 2.1) of whole starting network and $\Delta(\text{measured network})$ indicates the density of measured network with introduced errors.

Similarly, the relative difference in number of null dyads (D_Null), the relative difference in number of asymmetric dyads (D_Asymm), the relative difference in number of mutual dyads (D_Mut), and the relative difference in reciprocity index (*Rec*) can be computed.

8.1.2 Pearson correlation coefficient and Euclidean distances between vectors of actor properties

On the other hand, measures of centrality and prestige (Section 2.2.2) are calculated on an actor level, which means that mentioned measures are calculated for each vertex separately and presented for whole network as a vector. For comparison of two vectors two dissimilarity measures were selected; Euclidean distance, and Pearson correlation coefficient.

Let's say that we have two vectors of actor properties $y_{WN} = (y_1, y_2, \ldots, y_n)$ and $y_{MN} = (y'_1, y'_2, \ldots, y'_n)$ from the whole network and from the measured network, respectively. The Euclidean distance is defined as

$$d_E(y_{WN}, y_{MN}) = \sqrt{\sum_{i=1}^{n} \sqrt{(y_i - y'_i)^2}}\,. \tag{8.2}$$

Another used measure is Pearson correlation coefficient, which measures the strength and linear relationship between two vectors. It is defined as

$$r(y_{WN}, y_{MN}) = \frac{\sum_{i=1}^{n} \sqrt{(y_i - \mu_y)\left(y'_i - \mu_{y'}\right)}}{\sqrt{\sum_{i=1}^{n} \sqrt{(y_i - \mu_y)^2 \left(y'_i - \mu'_y\right)^2}}}\,, \tag{8.3}$$

where $\mu_y = \sum_{i=1}^{n} y_i$ and $\mu_{y'} = \sum_{i=1}^{n} y'_i$.
Values of Pearson correlation coefficient are in range between -1 and 1.

Ferligoj (1989) argued that we should be careful when selecting measure of dissimilarity. First of all, we have to know what kind of dissimilarity we would like to measure. The Pearson correlation coefficient between vector $u = (u_1, u_2, \ldots, u_n)$ and vector $v = (u_1 + a, u_2 + a, \ldots, u_n + a)$, which is obtained from vector u with addition of constant a to each component of vector u (it means that vectors u and v are parallel), is equal to 1. The 'profile' of those two vectors is translated, but equal. Values of both vectors are different, which can be revealed with use of Euclidean distance as dissimilarity measure[24]. (Kang, 2007, 140) noted that in such example the correlation coefficient indicates proximity, but not similarity.

The correlations and Euclidean distances between both indices of blockmodeling stability (ARI and $ErrB$) and indices of network characteristics presented in Section 2.2. In this chapter, the notation presented in Table 8.1 will be used beside the standard notation ARI and $ErrB$.

[24]The Euclidean distance between vectors u and v is $d_E(u, v) = a\sqrt{n}$

Label	Meaning
p.changed	Percent of randomly **changed ties** in a network.
	Relative difference in ...
Dens	... **density** between whole and measured network.
Rec	... **reciprocity** between whole and measured network.
D_Mut	... number of **mutual dyads** between real and measured network.
D_Asymm	... number of **asymmetric dyads** between real and measured network.
D_Null	... number of **null dyads** between real and measured network.
	Euclidean distance between vectors of...
PP_e	... **proximity prestige** for real and measured network.
CCout_e	... **closeness centrality based on outdegree** for real and measured network.
CCin_e	... **closeness centrality based on indegree** for real and measured network.
Dall_e	... **all-degree centrality** for real and measured network.
Dout_e	... **outdegree centrality** for real and measured network.
Din_e	... **indegree centrality** for real and measured network.
B_e	... **betweenness centrality** for real and measured network.
A_e	... **authority weights** for real and measured network.
H_e	... **hub weights** for real and measured network.
	Correlation between vectors of...
PP_cor	... **proximity prestige** for real and measured network.
CCout_cor	... **closeness centrality based on outdegree** for real and measured network.
CCin_cor	... **closeness centrality based on indegree** for real and measured network.
Dall_cor	... **all-degree centrality** for real and measured network.
Dout_cor	... **indegree centrality** for real and measured network.
Din_cor	... **outdegree centrality** for real and measured network.
B_cor	... **proximity prestige** for real and measured network.
A_cor	... **authority weights** for real and measured network.
H_cor	... **hub weights** for real and measured network.

Table 8.1: Notation used in studies of impact of network characteristics on results of blockmodeling

8.1.3 Linear regression models

Beside the correlation coefficients, the linear relationships between indices of network characteristics as predictors and indices of stability as outcome variables are examined. The amount of variance explained is reported with R^2, and if linear relationship is strong enough, the linear model is drawn together with corresponding data. In the literature different interpretations of strength of correlation coefficients can be found. Cohen (1988, 79-81) and Field (2009, 170) divided the values of Pearson correlation coefficient into three categories: ± 0.1 represents small effect, ± 0.3 a medium effect and ± 0.5 large effect. Cohen (1988) pointed out that this criteria are in fact arbitrary and should be redefined, if necessary, for particular problem. Beside the scatterplots, the linear models are drawn if the correlation coefficient is higher or equal to 0.5.

Because the amount of data for each simulation study is large (e.g. for the the boy-girl liking ties network we have 3950 comparisons of measured networks to the whole one, and with simulated data we have 316000 comparisons to the whole starting network within each starting blockmodel structure), the *'aggregated' scatterplots* are drawn. The values of two indices of interest are rounded to one decimal place. On obtained grid circles are drawn, where the radius of the circle is proportional to the number of data in the corresponding grid point. The artificial example of 'aggregated' scatterplot is presented in Figure 8.1. This is just the rough picture of the data and is used just as first insight to the distribution of data. All regression models in this chapter are calculated on original non-aggregated data.

In Section 7.5 the impact of random measurement error on stability of blockmodeling is investigated. In addition to regression models with indices of network characteristics as predictor variables, models with number of randomly changed ties (*p.changed*) as independent variable are established.

8.1.4 Generalized linear models

In cases where distribution of indices suggested the exponential functional dependency, generalized linear models (GLM) are used instead of linear ones. The restrictive

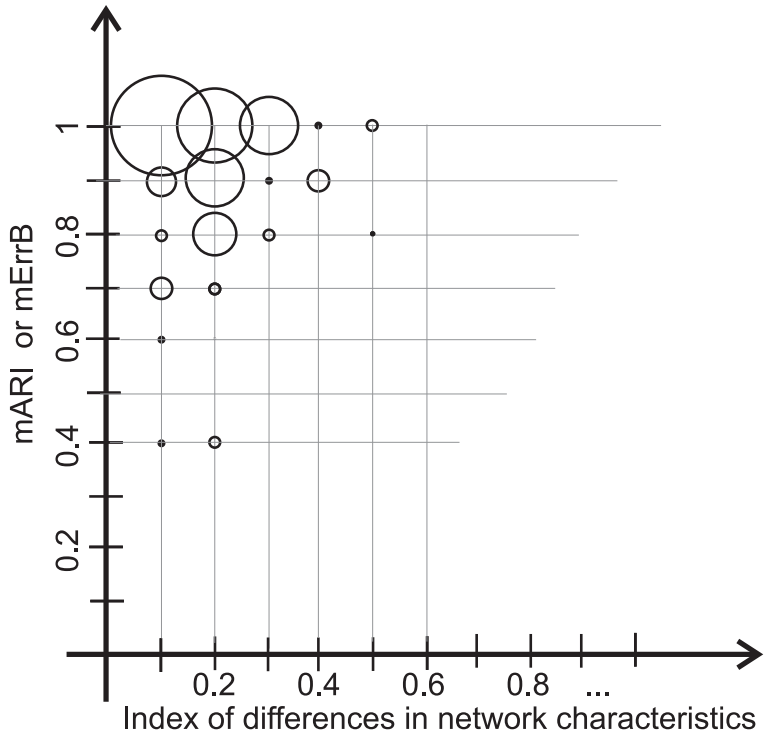


Figure 8.1: The 'aggregated' scatterplot

assumption from linear models that the variance should be constant is loosened in GLM. It could be proportional to a function of the mean (Friedl, 2010), e.g. when the mean increases also the variance is increasing. A generalized linear model procedure consists of three-part specification (McCullagh and Nelder, 1989, 27):

(i) The random component: $y_1, \ldots, y_n$ are independently distributed from a member of exponential family with $E(y_i) = \mu_i$ $(i = 1, \ldots, n)$ and $var(y_i) = \phi V(\mu_i)$, where ϕ is a dispersion parameter.

(ii) The systematic component: fixed covariates $x_{i1}, \ldots, x_{ip}$ define a linear predictor η;

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j \,, \tag{8.4}$$

where x_{ij} are values of p explanatory and β_j are unknown parameters, which should be estimated from the data.

(iii) The link function between random and systematic components, $g(\mu_i) = \eta_i$. It provides the relationship between the linear predictor and the mean of y.

"To determine the fit of a given model, a GLM evaluates the linear predictor for each value of the response variable, then compares the predicted value with a transformed value of y. The transformation to be employed is specified in the link function" (Crawley, 2007, 513). The measure of goodness of fit of our model is called the deviance. The scaled deviance compares the maximum of the log-likelihood under our current model with its maximum under the best, saturated, model. If the value of scaled deviance is similar to the degrees of freedom (number of observations n minus number of explanatory variables p) of our model, then the model is expectable.

The GLM requests the precisely established relationship between the variance and the mean. When the precise form of the error distribution is hard to determine, the robust alternative known as quasi-likelihood can be used (Crawley, 2007, 516-517). In our models, where the exponential dependency was present, we used quasi-Poisson errors, which can compensate also the overdispersion (an extra, unexplained variance than assumed) together with the *log* link.

In order to compare the linear models with the generalized linear ones in some way, the pseudo R^2 measures can be used. They are based on the concept of deviance and can be defined as (Mittlbock, 2004):

$$R_D^2 = 1 - \frac{the\ scaled\ deviances\ of\ the\ full\ model}{the\ scaled\ deviances\ of\ the\ null\ model}\ . \tag{8.5}$$

The adjustment of R^2 (Equation 8.5) for GLM can be made with consideration of the number of parameters fitted as

$$R_{D,df}^2 = 1 - \frac{the\ scaled\ deviances\ of\ the\ full\ model \cdot (n-1)}{the\ scaled\ deviances\ of\ the\ null\ model \cdot (n-k-1)}\ , \tag{8.6}$$

where n is the number of observations and k the number of fitted covariates.

8.2 The impact of differences in network characteristic on the stability of blockmodeling in case of real networks

In this section the impact of differences in network characteristics and properties is investigated in case of two real networks; the boy-girl liking ties network and the note borrowing network.

8.2.1 The boy-girl liking ties network

First, we investigated the impact of different network characteristics to the stability of blockmodeling with data from the boy-girl liking ties network (Figure 6.1 in Section 6.2.1.1). The network has symmetric structure and its blockmodel based on structural equivalence has two clusters. The image matrix has two complete blocks on diagonal and null blocks out of diagonal. The stability of blockmodeling will be examined with two indices; with the Adjusted Rand Index (ARI) which measures agreement between two partitions and with percent (or proportion) of correctly identified block types in blockmodel (ErrB).

8.2.1.1 Stability of partitions

First, the impact of differences in network characteristics to stability of blockmodels in terms of partitions agreement with data from the simulation of randomly introduced measurement errors to the boy-girl liking network (Figure 6.1 in Section 6.2.1.1) is examined.

The correlation coefficient between ARI and other network characteristics indices are presented in Table 8.2. The highest correlation coefficient is between the Adjusted Rand Index (ARI) and the proportion of changed ties ($r = -0.773$). The proportion of changed ties is not an index of network characteristic known from social network data analysis. It is a parameter from the simulation process of networks with introduced errors. We already know that it has a great impact on values of ARI and $ErrB$

from figures in Section 7.5. We will try to compare its impact with other 'more standard' indices in the following sections.

Table 8.2: Correlations and results of fitted linear models for ARI with data for the boy-girl liking ties network

index	r	R^2	b_0	b_1
p.changed	-0.773	0.598	1.298	-0.030
Dens	-0.484	0.235	0.981	-1.556
Rec	-0.555	0.309	1.178	-1.718
D_Mut	-0.326	0.106	0.873	-1.035
D_Asymm	-0.642	0.412	1.235	-0.371
D_Null	-0.622	0.386	1.178	-1.668
PP_e	-0.563	0.317	1.203	-0.899
CCout_e	-0.573	0.329	1.170	-1.137
CCin_e	-0.571	0.326	1.170	-1.157
Dall_e	-0.572	0.327	0.993	-2.06
Dout_e	-0.593	0.351	1.047	-1.518
Din_e	-0.589	0.347	1.041	-1.448
B_e	0.370	0.137	0.474	1.956
A_e	-0.460	0.211	1.118	-0.966
H_e	-0.445	0.198	1.103	-0.947
PP_cor	0.206	0.042	0.540	0.378
CCout_cor	0.185	0.034	0.560	0.347
CCin_cor	0.208	0.043	0.537	0.384
Dall_cor	0.475	0.226	0.301	0.775
Dout_cor	0.406	0.165	0.383	0.677
Din_cor	0.452	0.205	0.319	0.735
B_cor	0.086	0.007	0.631	0.153
A_cor	0.451	0.203	0.397	0.660
H_cor	0.424	0.179	0.415	0.624

Legend:

r - Pearson correlation coefficient

R^2 - variance explained

b_0 - the intercept parameter in regression model

b_1 - the slope parameter in regression model

Among differences in network characteristic, ARI correlates the highest with differences in number of asymmetric dyads ($r = -0.642$) and null dyads ($r = -0.622$). Indices based on Euclidean distance have higher correlation coefficients with ARI than corresponding indices of correlation between two vectors of network properties. There

are the highest correlations between closeness centrality based on outdegree obtained with Euclidean distance ($r = -0.593$) and closeness centrality based on indegree with Euclidean distance ($r = -0.589$). The smallest correlation coefficient among indices of Euclidean distance between vectors of network properties and ARI index is obtained by betweenness centrality ($r = 0.370$). Similarly, betweenness centrality index based on correlation among vectors has the smallest correlation coefficient ($r = 0.086$) among all indices based on correlations.

The Pearson correlation coefficients among indices of network properties themselves are presented in Table B.1 in Appendix B. Correlation coefficients between corresponding indices calculated with correlation or Euclidean distance (e.g. between PP_e and PP_cor,...) are in range from -0.86 to -0.07. The negative sign is expected, because higher Euclidean distance indicates bigger difference between two vectors. Low absolute values of correlation coefficients between corresponding indices indicate that Euclidean distance and correlation between vectors reveal different patterns in the data.

Table 8.2 presents results of fitted linear regression models to the data for ARI. As written in Section 8.1 the linear models are drawn only if the model explains at least 25% of variation in ARI. Figure 8.2 presents scatterplots for indices from the boy-girl liking ties network data. The linear regression models are fitted to indices of differences in reciprocity, the number of asymmetric dyads and the number of null dyads. Although these linear regression models are able to explain between 30% and 40% of variation in ARI, there is no clear linear relationship. The majority of ARI values is equal to 1, regardless of the values of explanatory indices. The reason of high stability of blockmodeling (Section 7.5.2.1) is probably the completely symmetrical blockmodel structure (Figure 6.1). On the other hand, this high stability prevents the correct prediction of ARI from differences in network characteristics.

Almost all linear regression models with indices based on Euclidean distance as predictors are able to explain more than 30% of variance, exceptions are betweenness centrality, authority weights and hub weights. The highest percent of variance (35%) in

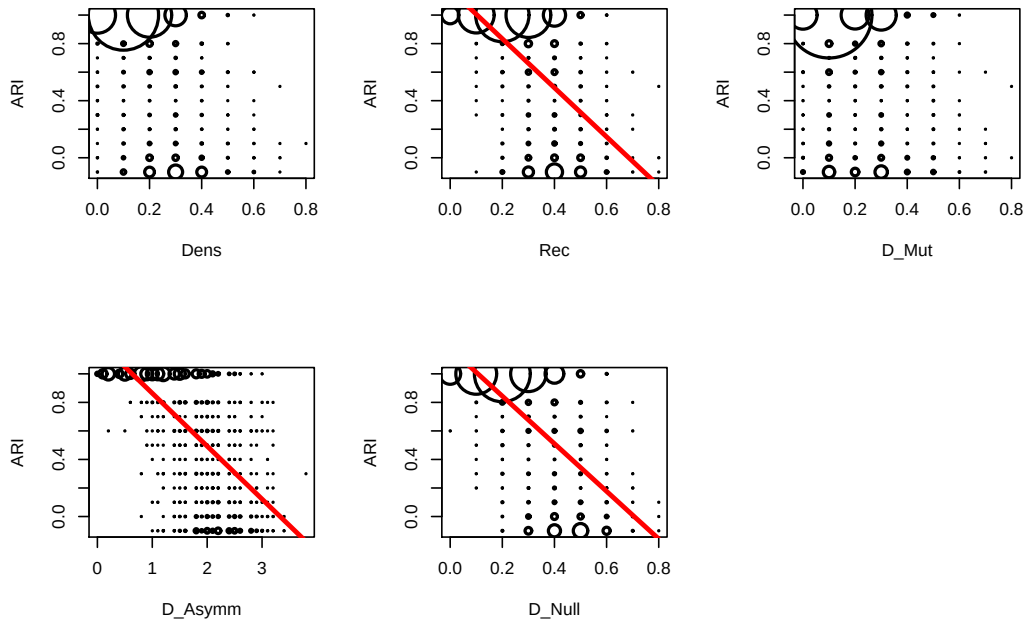


Figure 8.2: Impact of differences in network characteristics to values of ARI with data for the boy-girl liking ties network

ARI can be explained with closeness centrality based on outdegree based on Euclidean distance (model is drawn in Figure 8.3).

Models based on correlation between two vectors of vertex properties are obviously not the best models, because the percentages of explained variance are low. All the 'aggregated' scatterplots presented in Figure 8.4 show the same pattern; different values of ARI are almost equally distributed accross the range of correlation values[25]. These models are able to explain at most 23% of variance in ARI (in case of all-degree centrality).

The percent of changed ties has the greatest impact on values of ARI among all indices of network characteristics (Table 8.2). Figure 8.5 present different fitted models; linear, two piecewise linear and quadratic regression models. Linear regression model

[25]In all cases where vectors of network properties are compared with correlation, only the strength of correlation is taken into account and not the sign of the correlation.

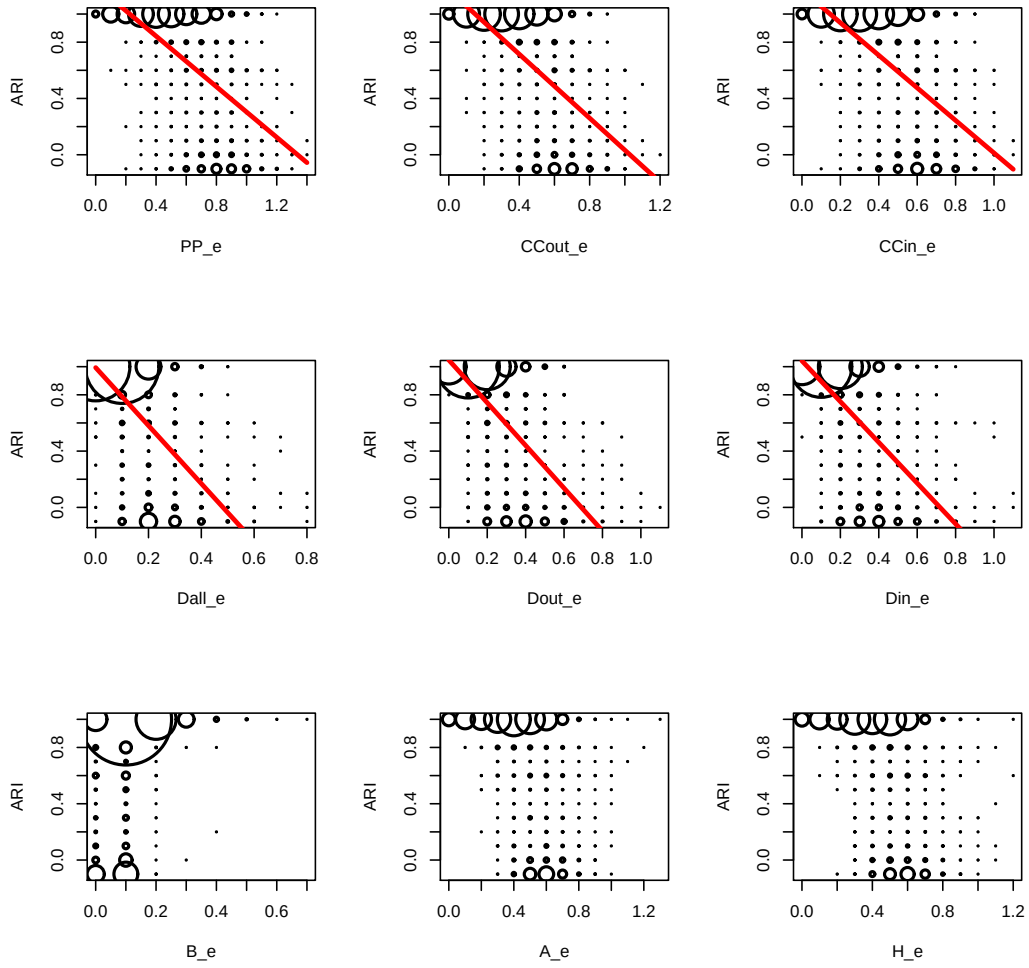


Figure 8.3: Impact of differences in network properties based on Euclidean distance to values of ARI with data for the boy-girl liking ties network

is presented as red straight line. Table 8.3 shows that this model explains just 59.8% of variance in values of ARI. The data in Figure 7.47(a), where also error bars of one standard deviation are presented, suggest that piecewise linear regression model should be more appropriate.

In piecewise regression modeling two questions should be answered (Crawley, 2007):

(i) how many segments to choose to divide the line into, and

(ii) where to position the break points on the x axis (predictor values).

In our example the data suggest two lines. On the left side of Figure 7.47(a) where there

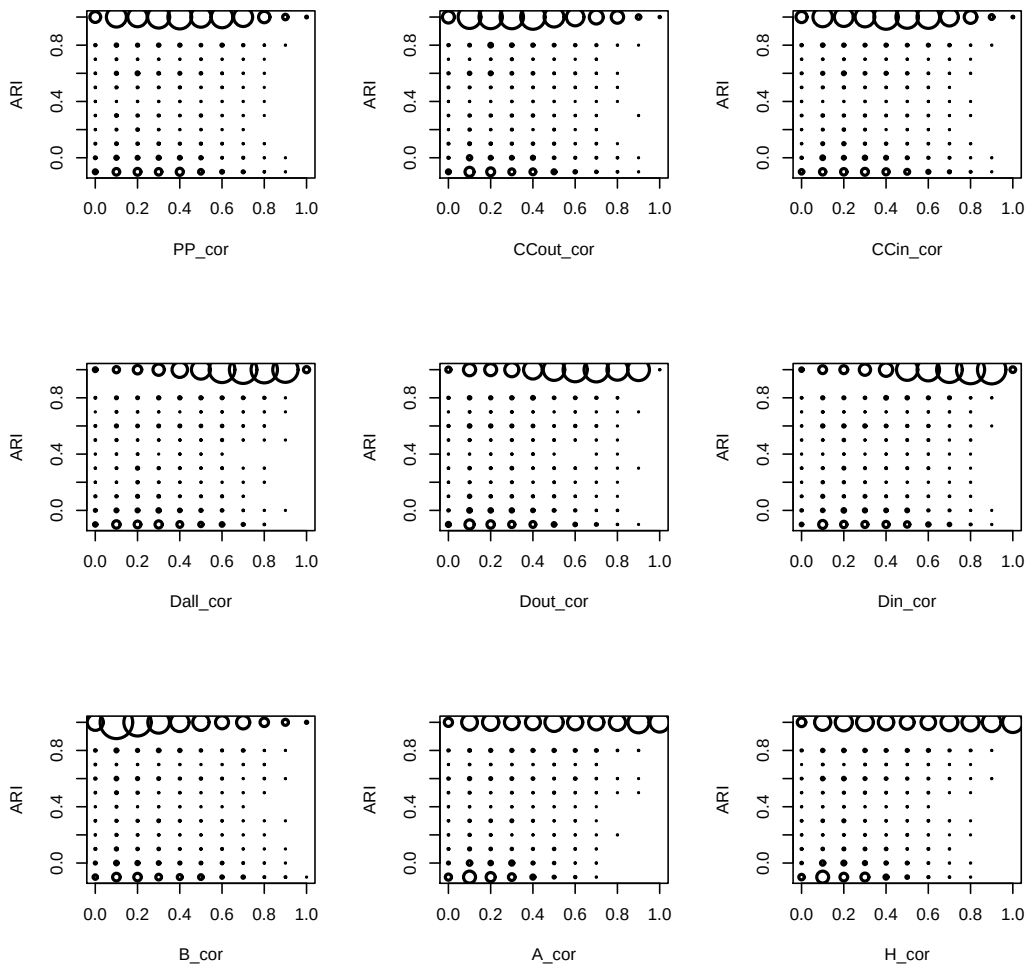


Figure 8.4: Impact of differences in network properties based on correlations to values of ARI with data for the boy-girl liking ties network

are standard deviations equal to zero (or very small) should be (almost) horizontal line and the second line on the right should be steeper (with negative slope coefficient). The answer to the second question can be simply computed. For each value on x-axes a two-segment piecewise regression model is estimated. The best piecewise model is defined as the model with the minimum deviance. The plot of residual standard errors according to different breaks (or percents of changed ties) is presented on the right part of Figure C.1 in Appendix C. It suggests that break should be made at 24% of changed ties, with slightly worse results by breaks between 18 and 23% of changed ties.

Figure 8.5 presents two piecewise models. The first one has break at 18% of changed ties because standard deviations are visibly smaller (the blue lines) and the second one (the green lines) is determined analytically at 24% of changed ties (see above explanation and Figure C.1(a)). The percent of explained variance with two piecewise models was compared also to quadratic regression model, where the quadratic term of the *p.changed* is added to the regression equation. The fit of the quadratic model is a little bit worse than both piecewise linear regression models, because the percent of explained variance in quadratic model is approximately for one percent lower.

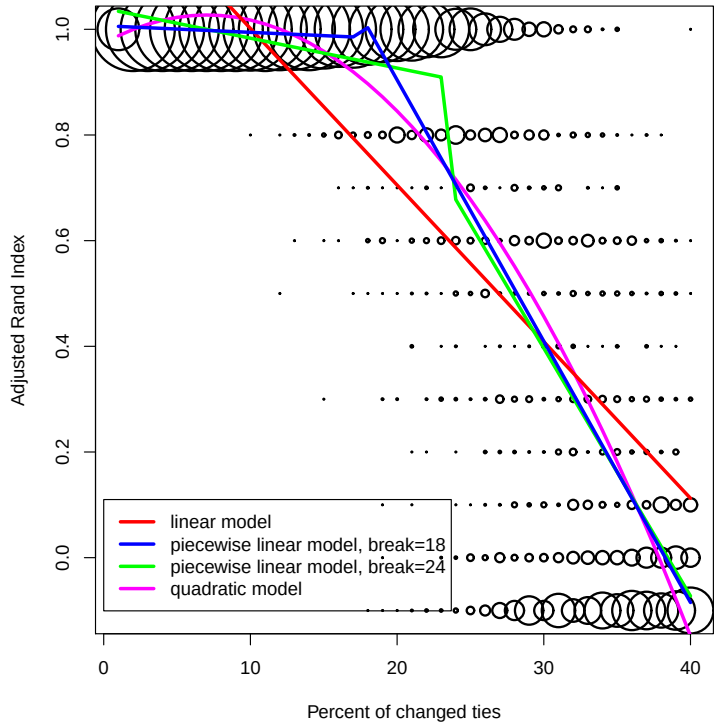


Figure 8.5: Impact of percent of changed ties on values of ARI with data for the boy-girl liking ties network

8.2.1.2 Stability of block types

The stability of blockmodel in terms of correctly identified blocks is measured by $ErrB$. The correlation coefficients between differences in network characteristics and $ErrB$ are all lower than 0.5, which indicates medium linear relationship effect according to Cohen (1988). In comparison to Pearson correlation coefficient for ARI (Table 8.2), all

Table 8.3: Different fitted models for ARI with $p.changed$ ties as a predictor with data from the boy-girl liking ties network

Name of the model	Formula	R^2
Linear model	$\hat{y}_{ARI} = 1.298 - 0.030 \cdot p.changed$	0.598
Piecewise linear models where		
break=18	$\hat{y}_{ARI} = \begin{cases} 1.007 - 0.001 \cdot p.changed; & p.changed < 18 \\ 1.893 - 0.049 \cdot p.changed; & p.changed \geq 18 \end{cases}$	0.694
break=24	$\hat{y}_{ARI} = \begin{cases} 1.040 - 0.006 \cdot p.changed; & p.changed < 24 \\ 1.803 - 0.047 \cdot p.changed; & p.changed \geq 24 \end{cases}$	0.695
Quadratic model	$\hat{y}_{ARI} = 0.974 - 0.015 \cdot p.changed - 0.001 \cdot p.changed^2$	0.681

corresponding correlations are lower.

The highest correlation coefficient is between the proportion of incorrect blocks and relative difference in number of asymmetric dyads (D_Asymm) which is equal to 0.495. Between all indices of vertex properties based on Euclidean distance and $ErrB$, there is the highest correlation coefficient for closeness centrality based on outdegree ($r = 0.443$). The correlation coefficient is positive, indicating that the larger Euclidean distances between whole and measured vector of vertex outdegree lead to lower values of $ErrB$ index and lower blockmodeling stability in terms of correctly identified block types. There is the highest negative correlation coefficient between $ErrB$ and indices calculated based on correlations between two vectors in case of all-degree centrality (-0.331). All indices based on correlation have negative sign, indicating that larger values on those indices lead to lower values of $ErrB$. Correlations between corresponding indices calculated with Euclidean distance and correlation are in range from -0.63 to -0.96 (Table B.2 in Appendix B).

The low or medium linear effect of indices to values of $ErrB$ shown with correlation is visible also on 'aggregated scatterplots'. Figure 8.6 shows that the majority of $ErrB$ values are equal to 0 and are distributed across the whole range of possible values of a predictor (for all indices of differences in network characteristics; $Dens$, Rec, D_Mut,

Table 8.4: Correlations and results of fitted linear models for *ErrB* with data for the boy-girl liking ties network

index	r	R^2	b_0	b_1
p.changed	0.580	0.337	-0.155	0.015
Dens	0.351	0.123	0.011	0.764
Rec	0.437	0.191	-0.106	0.917
D_Mut	0.274	0.075	0.05	0.589
D_Asymm	0.495	0.245	-0.131	0.194
D_Null	0.467	0.218	-0.095	0.851
PP_e	0.416	0.173	-0.103	0.45
CCout_e	0.424	0.18	-0.087	0.571
CCin_e	0.422	0.178	-0.087	0.581
Dall_e	0.41	0.168	0.007	1.001
Dout_e	0.443	0.196	-0.027	0.77
Din_e	0.44	0.194	-0.024	0.734
B_e	-0.268	0.072	0.261	-0.963
A_e	0.339	0.115	-0.06	0.483
H_e	0.324	0.105	-0.05	0.468
PP_cor	-0.143	0.02	0.225	-0.178
CCout_cor	-0.12	0.014	0.212	-0.153
CCin_cor	-0.145	0.021	0.227	-0.182
Dall_cor	-0.333	0.111	0.339	-0.369
Dout_cor	-0.283	0.08	0.299	-0.32
Din_cor	-0.327	0.107	0.336	-0.36
B_cor	-0.056	0.003	0.18	-0.068
A_cor	-0.331	0.109	0.3	-0.328
H_cor	-0.3	0.09	0.287	-0.299

Legend:

r - Pearson correlation coefficient

R^2 - variance explained

b_0 - the intercept parameter in regression model

b_1 - the slope parameter in regression model

D_Asymm and *D_Null*).

The linear regression models with indices based on Euclidean distance as predictors are able to explain from 10% to 20% of variance in *ErrB*. Despite of medium linear effect shown in correlation coefficients, there is no clear linear relationship visible in Figure 8.7.

Linear models based on correlation between two vectors of vertex properties are able

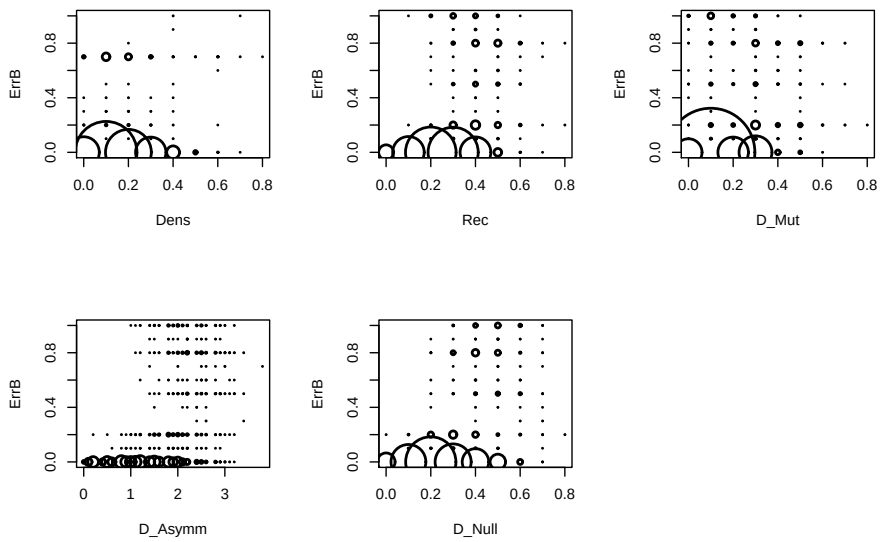


Figure 8.6: Impact of differences in network characteristics to values of ARI with data for the boy-girl liking ties network

to explain maximally 11% of variance in $ErrB$. Therefore just 'aggregated' scatterplots without fitted linear models are presented in Figure 8.8. Similarly as in scatterplots for ARI there are no clear nonlinear functional relationships between indices based on correlation between two vectors of vertex properties and $ErrB$ values.

The next step was to find out which function best describes the relationship between percent of changed ties and Err values. Beside the linear regression model which explains 33.7% of variance in $ErrB$ we also try to fit two-piecewise linear model.

In order to find the best two tailed linear model, the models were fitted to all possible break points from 1 to 39 percent of changed ties. The best model was determined based on the small residual error. Figure C.1 in Appendix C suggests that the break at 25% of changed ties leads to the best piecewise model. It explains 40% of variance in $ErrB$ and is drawn in blue in Figure 8.9. The scatterplot of data also suggest that non-linear model could be suitable. Another step was to fit quadratic model and generalized linear model, because Figure 7.47(b) (in Section 7.5.2.1) shows that the standard deviation and/or variation increases with higher percent of changed ties. The

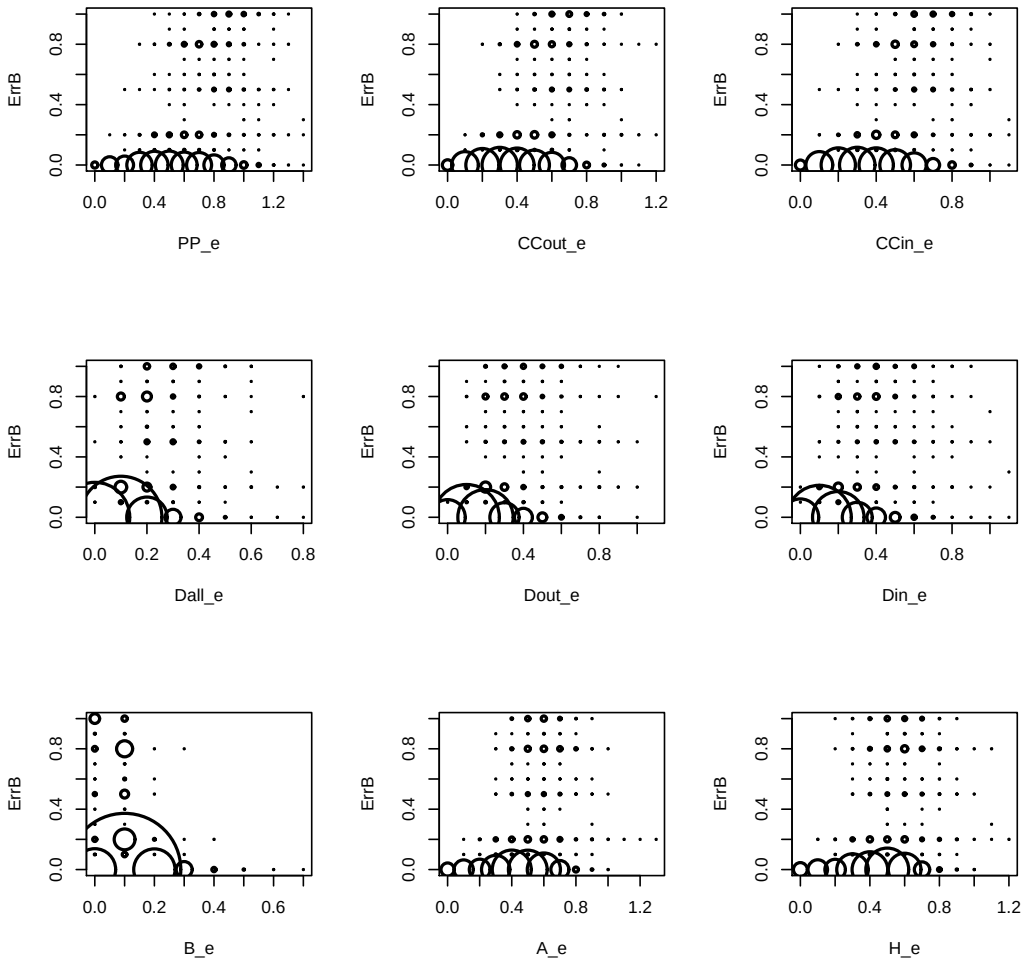


Figure 8.7: Impact of differences in network properties based on Euclidean distance to values of *ErrB* with data for the boy-girl liking ties network

quadratic model explains 39% of variance and is therefore similar to piecewise linear model. The exponential generalized linear model is the best one, because it is able to explain almost 50% of variance in *ErrB*.

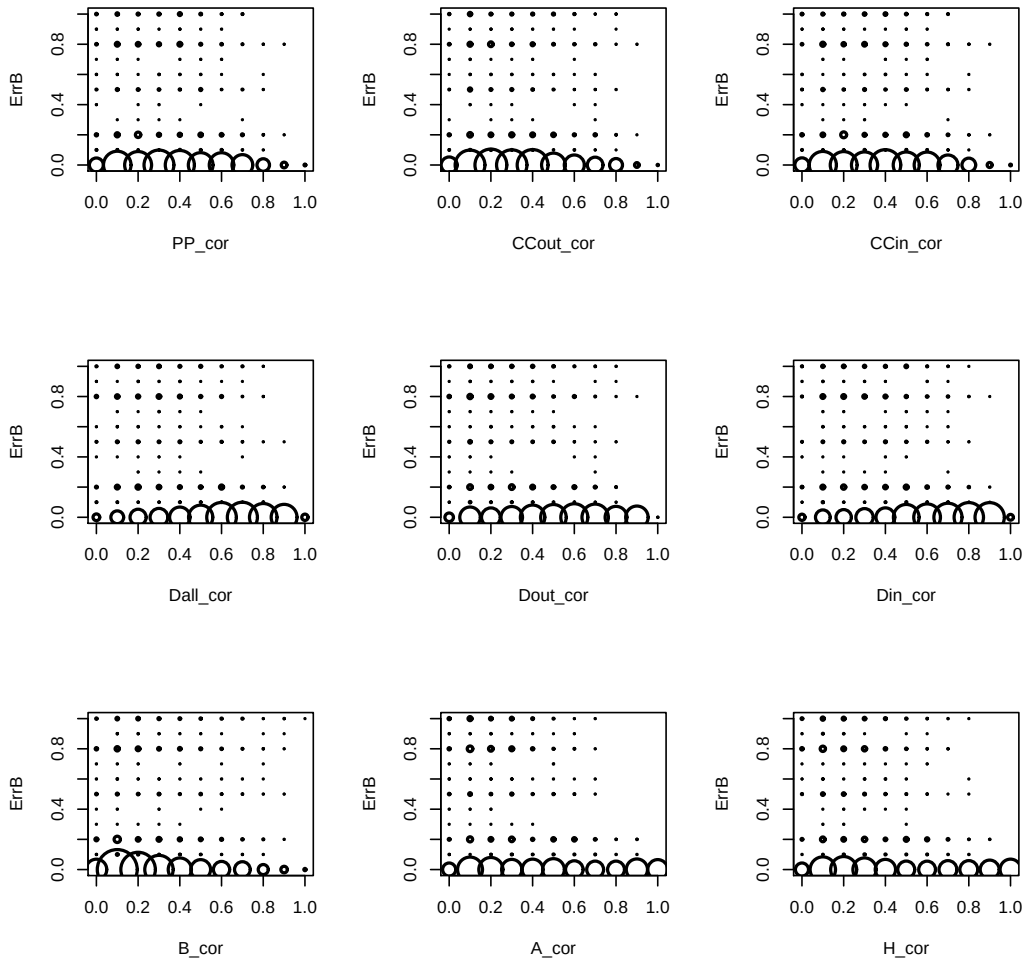


Figure 8.8: Impact of differences in network properties based on correlations to values of *ErrB* with data for the boy-girl liking ties network

Table 8.5: Different fitted models for *ErrB* with *p.changed* ties as a predictor with data for the boy-girl liking ties network

Name of the model	Formula	R^2
Linear model	$\hat{y}_{ErrB} = -0.155 + 0.015 \cdot p.changed$	0.337
Exponential model	$\hat{y}_{ErrB} = e^{-5.580+0.133 \cdot p.changed}$	0.491
Piecewise linear models where break=25	$\hat{y}_{ErrB} = \begin{cases} -0.021 + 0.003 \cdot p.changed; & p.changed < 24 \\ -0.438 - 0.025 \cdot p.changed; & p.changed \geq 24 \end{cases}$	0.400
Quadratic model	$\hat{y}_{ErrB} = 0.025 - 0.010 \cdot p.changed - 0.0006 \cdot p.changed^2$	0.392

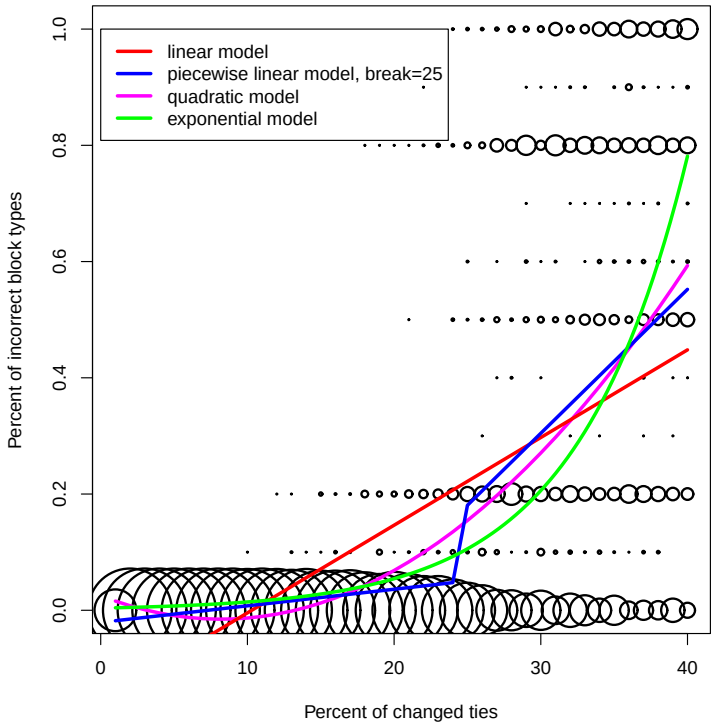


Figure 8.9: Impact of percent of changed ties on values of $ErrB$ with data for the the boy-girl liking ties network

8.2.2 The note borrowing network

The second real network used in the study of impact of network characteristics to stability of blockmodeling was the student note borrowing network (Figure 6.2 in Section 6.2.1.2). With blockmodeling procedure based on structural equivalence three clusters were obtained with non-symmetric structure (image matrix).

8.2.2.1 Stability of partitions

First, we examined the correlations between ARI and indices of network characteristics (Table 8.6). The highest negative correlation among network characteristics is between ARI and relative difference in null dyads ($r = -0.758$) and relative difference in network density ($r = -0.746$).

Table 8.6: Correlations and results of fitted linear models for ARI with data for the note borrowing network

index	R	R^2	b_0	b_1
p.changed	-0.848	0.718	1.023	-0.029
Dens	-0.746	0.556	0.909	-1.325
Rec	-0.097	0.009	0.479	-0.392
D_Mut	-0.514	0.264	0.651	-0.750
D_Asymm	-0.672	0.451	0.913	-1.055
D_Null	-0.758	0.575	0.965	-1.957
PP_e	-0.794	0.630	1.054	-0.610
CCout_e	-0.786	0.619	0.975	-0.786
CCin_e	-0.794	0.630	1.055	-0.611
Dall_e	-0.745	0.555	0.782	-1.137
Dout_e	-0.735	0.541	0.814	-1.084
Din_e	-0.785	0.617	0.825	-0.773
B_e	-0.460	0.211	0.977	-3.752
A_e	-0.833	0.693	0.926	-1.664
H_e	-0.734	0.539	0.908	-3.002
PP_cor	0.667	0.445	-0.398	1.224
CCout_cor	0.539	0.291	0.133	0.805
CCin_cor	0.668	0.446	-0.399	1.225
Dall_cor	0.736	0.542	-0.374	1.169
Dout_cor	0.567	0.322	0.069	0.879
Din_cor	0.738	0.545	-0.549	1.31
B_cor	0.462	0.213	0.175	0.702
A_cor	0.758	0.575	-0.491	1.266
H_cor	0.742	0.551	-0.022	0.95

Legend:

r - Pearson correlation coefficient

R^2 - variance explained

b_0 - the intercept parameter in regression model

b_1 - the slope parameter in regression model

If we look at Euclidean distances between network properties, the highest correlations are in proximity prestige and closeness centrality based on indegree ($r = -0.794$), closeness centrality based on outdegree ($r = -0.786$), and degree centrality based on indegree ($r = -0.785$). The smallest correlation is obtained in case of betweenness centrality ($r = -0.46r$). All correlations between ARI and Euclidean distances have negative sign, indicating that the higher differences between vectors of network properties mean lower values of ARI, or more precisely lower blockmodeling stability in terms

of partitions. On the other hand, correlations between ARI and differences between network properties based on correlation have positive sign. That means that higher correlations between vectors of whole and measured network properties, indicate the higher values of the Adjusted Rand Index. The highest correlations between ARI and correlations among corresponding vectors are by authority weights ($r = 0.753$), degree centrality based on indegree ($r = 0.735$), and hub weights ($r = 0.730$). The highest correlations among all indices and ARI are obtained with percent of changed ties ($r = -0.848$).

The second step was to fit linear regression models. Table 8.6 present proportion of variance explained (R^2, which is in fact the square of correlation coefficient), and both parameters b_0 and b_1 in fitted linear function ($ARI = b_0 + b_1 \cdot index$). Models in below figures are fitted just to predictors with correlation higher or equal to 0.5.

Figure 8.10 shows values of ARI plotted against differences in network characteristics. Linear trend is present in case of differences in network density, number of asymmetric dyads, and number of null dyads. The linear regression model explains 56% of variance in values of ARI, when the predictor is difference in network density. A little bit more variance can be explained (58%) if predictor is difference in number of null dyads. The difference in number of asymmetric dyads explains (45%) of variance in values of ARI.

Figure 8.11 presents linear models for ARI and Euclidean distances between vectors of network properties. There is no correlation or linear trend in the case of betweenness centrality, this is why the linear function is not fitted to the data. The linear regression model for authority weights explains 69% of variance in values of ARI. More than 60% of variance is also explained with use of proximity prestige (63%), closeness centrality based on indegree (63%) and closeness centrality based on outdegree (62%).

Figure 8.12 presents linear regression models for ARI with predictors obtained by correlation between vectors of network properties for whole and measured network. The

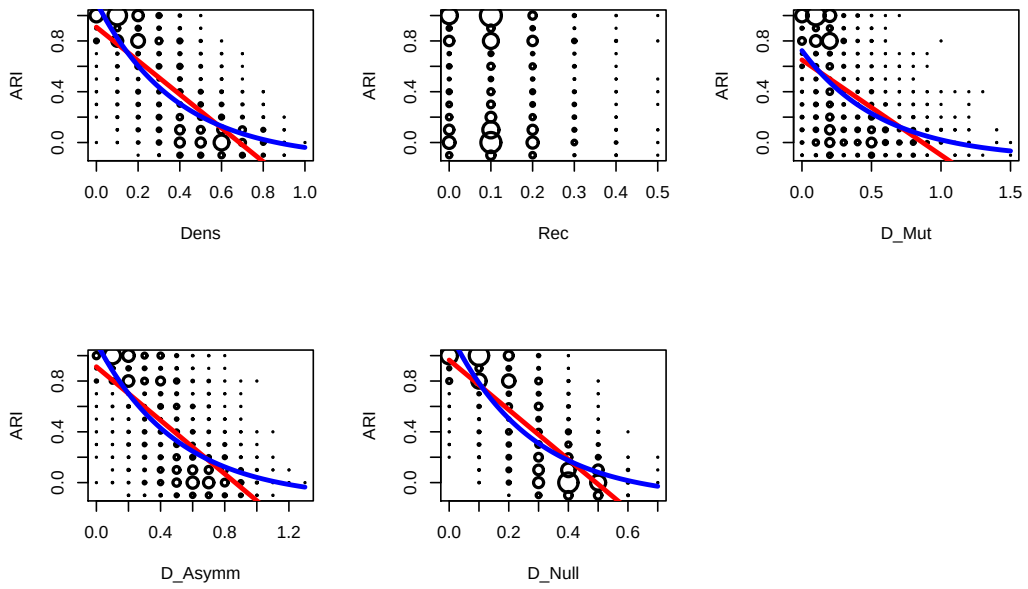


Figure 8.10: Impact of differences in network characteristics to values of ARI with data for the note borrowing network

linear models are not fitted to the figure where betweenness centrality should be a predictor, because there is no clear linear trend. The highest percent of explained variance (58%) we get in case where correlation between authority weights between networks is used for predictor.

Table 8.7 shows results of fitted generalized linear models with exponential dependency. The most predictive power has percent of changed ties *p.changed*, which is able to explain 67.3% variance in values of ARI. Between indices of network characteristics the highest percent of explained variance in values of ARI is obtained with relative difference in number of null dyads (53.3%), and relative difference in network density (53.2%) as a predictor. The 'aggregated' scatterplot for relative difference in reciprocity values as a predictor in Figure 8.10 shows no functional relationship, which is also confirmed with very low percent of explained variance in ARI values with both, linear (0.9%) and GLM model (0.8%). All GLMs with indices calculated with Euclidean distance between vectors of network properties (except B_e) as predictors are able to explain between 51.7% and 67.4% of variance in ARI. The GLM models are drawn

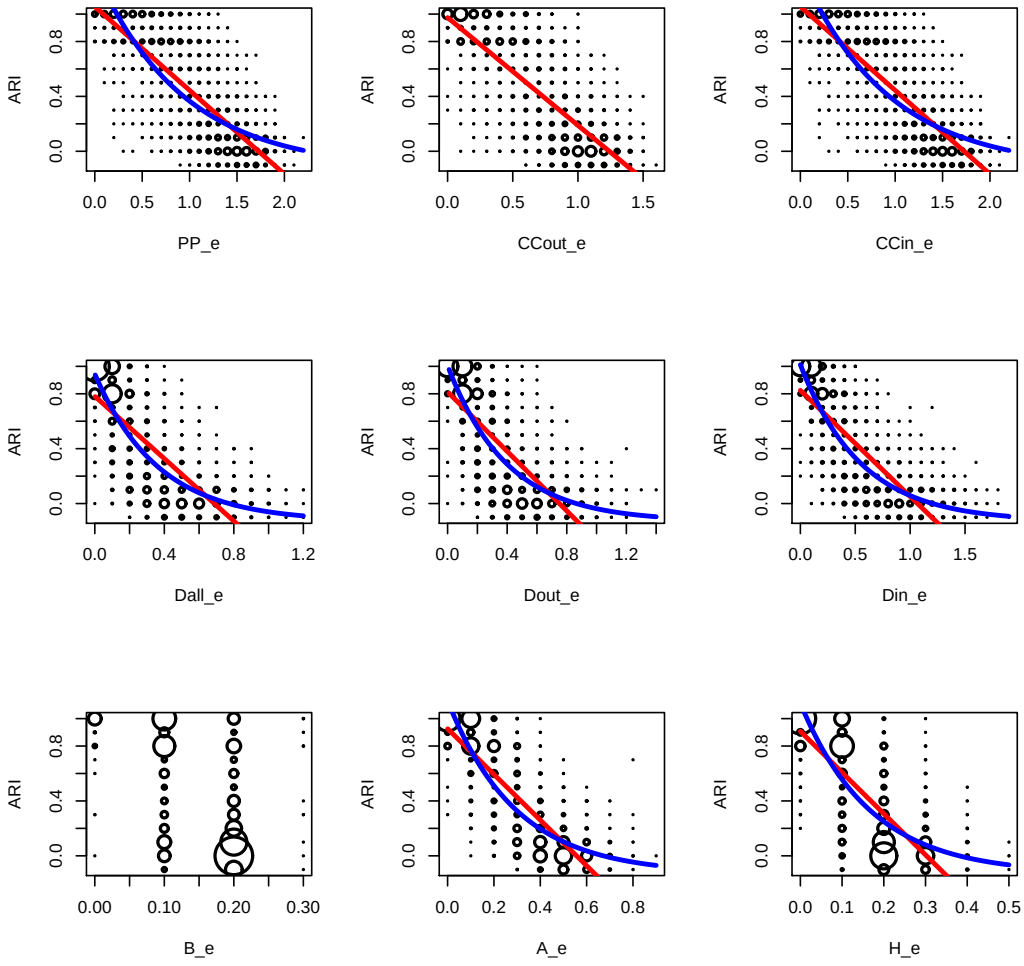


Figure 8.11: Impact of differences in network properties based on Euclidean distance to values of ARI with data for the note borrowing network

in blue in Figures 8.10, 8.11, and 8.12. Models with indices based on correlation as predictors explain a little less variance in ARI than corresponding indices based on Euclidean distance. The most variance in values of ARI (75.8%) can be explained with use of correlation between vectors of authority scores (A_cor) as a predictor.

Although the percent of changed ties is not a network characteristic in the narrower sense, it turns out that it predicts the values of ARI best. Figure 7.48 from Section 7.5.2.2 is supplemented with regression model which explains 72% of variance in ARI. The linear model is presented in Figure 8.13 with red line.

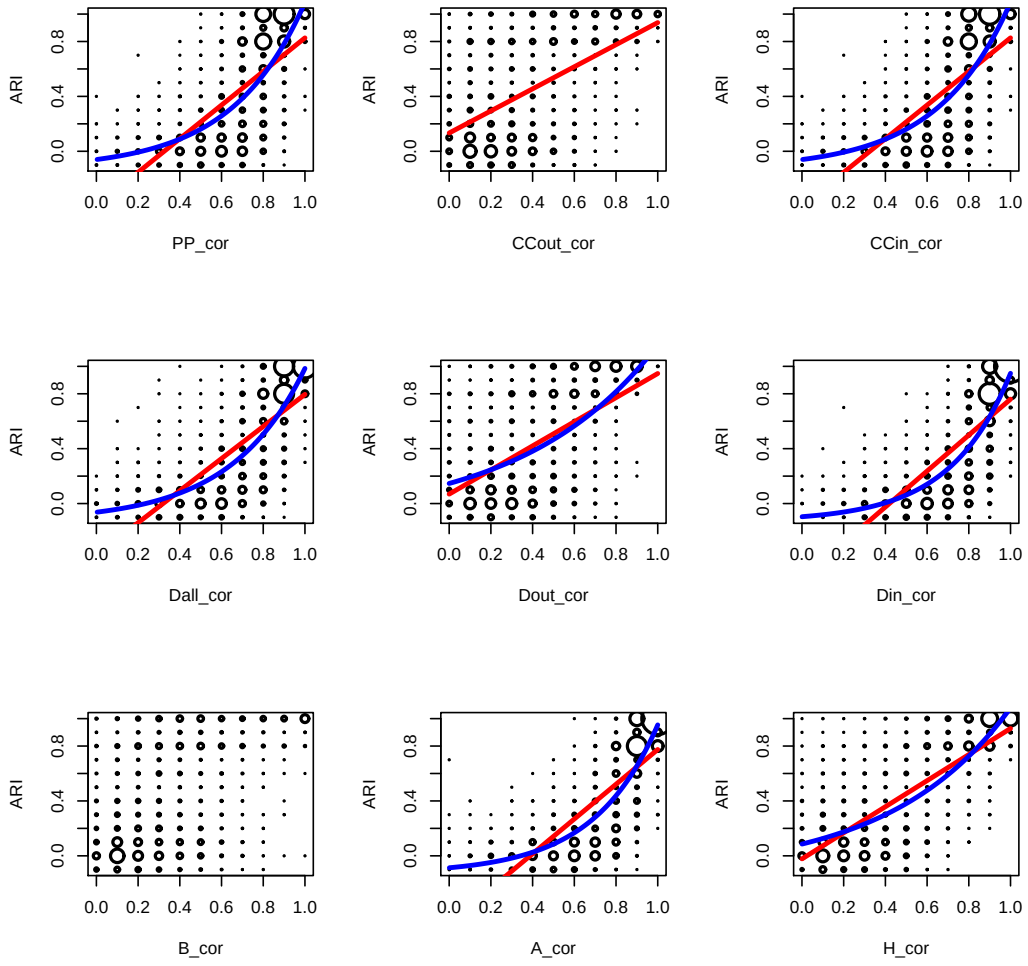


Figure 8.12: Impact of differences in network properties based on correlations to values of ARI with the note borrowing network

Beside the linear regression model, the quadratic and the exponential models were also fitted to the data. The exponential model performs worse than the linear model. The quadratic model (drawn in magenta in Figure 8.13) is a little bit better than the simple linear one, because it can explain 72.9% of variance in ARI values (Table 8.8).

Table 8.7: Results of fitted generalized linear models for ARI with data for the note borrowing network

index	a	b	Dispersion	Residual deviance	Scaled deviance	R^2
p.changed	0.37	-0.055	0.094	376.497	4011.563	0.673
Dens	0.212	-2.606	0.131	539.976	4116.673	0.532
Rec	-0.492	-0.724	0.261	1142.877	4374.543	0.008
D_Mut	-0.158	-1.732	0.191	828.687	4327.669	0.281
D_Asym	0.21	-1.976	0.163	669.556	4097.426	0.419
D_Null	0.278	-3.67	0.132	538.387	4078.645	0.533
PP_e	0.372	-1.074	0.125	508.033	4055.633	0.559
CCout_e	0.264	-1.435	0.125	504.156	4030.381	0.563
CCin_e	0.372	-1.075	0.125	507.496	4054.424	0.560
Dall_e	0.072	-2.735	0.114	459.955	4036.938	0.601
Dout_e	0.127	-2.505	0.122	492.345	4042.77	0.573
Din_e	0.133	-1.804	0.099	397.673	4034.164	0.655
B_e	0.23	-5.834	0.233	952.412	4079.693	0.174
A_e	0.232	-3.38	0.095	375.637	3957.501	0.674
H_e	0.217	-5.927	0.141	556.151	3943.213	0.517
PP_cor	-2.662	2.857	0.145	603.181	4153.529	0.477
CCout_cor	-1.129	1.323	0.21	867.388	4134.981	0.247
CCin_cor	-2.665	2.86	0.145	602.223	4153.074	0.478
Dall_cor	-2.703	2.812	0.121	475.941	3924.844	0.587
Dout_cor	-1.286	1.537	0.202	823.563	4086.217	0.285
Din_cor	-3.376	3.452	0.11	432.548	3931.494	0.625
B_cor	-1.041	1.146	0.226	943.753	4181.282	0.181
A_cor	-3.147	3.228	0.107	408.557	3832.663	0.646
H_cor	-1.533	1.722	0.148	578.478	3902.915	0.498

Degrees of freedom: 3949 for the null model and 3948 for the residual model

Null deviance of all models: 1152.917

All models are significant, p-value is 0.000

Legend:
R^2 - deviance explained
a, b - parameters in exponential glm $\hat{y}_{ARI} = e^{a+b \cdot index}$

Table 8.8: Different fitted models for ARI with *p.changed* ties as a predictor with data for the note borrowing network

Name of the model	Formula	R^2
Linear model	$\hat{y}_{ARI} = 1.022 - 0.029 \cdot p.changed$	0.718
Exponential model	$\hat{y}_{ARI} = e^{0.369 - 0.055 \cdot p.changed}$	0.673
Quadratic model	$\hat{y}_{ARI} = 1.235 - 0.043 \cdot p.changed - 0.0003 \cdot p.changed^2$	0.729

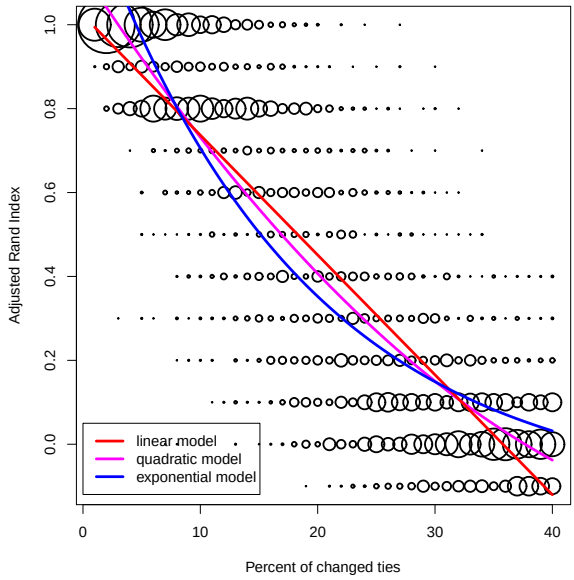


Figure 8.13: Impact of percent of changed ties on values of ARI with the note borrowing network

8.2.2.2 Stability of block types

Another index for estimation of stability of blockmodeling is proportion of incorrectly identified block types ($ErrB$). Table 8.9 presents the correlations between differences in network characteristics and properties, and values of $ErrB$.

The highest correlation is similar as in case for ARI; the Pearson correlation coefficient between $ErrB$ and percent of changed ties is $r = 0.702$. Among indices calculated with Euclidean distance we get the highest correlation for closeness centrality based on indegree ($r = 0.681$). Among all indices based on correlations between network properties the highest correlation is obtained with authority weights ($r = -0.629$). Similarly as for ARI values there is no correlation between $ErrB$ and relative differences in network reciprocity.

Figure 8.14 shows values for percent of incorrectly identified block types ($ErrB$) plotted against differences in network characteristics. Linear trend (correlation coefficient above 0.5) is present in case of differences in network density and number of null

Table 8.9: Correlations and results of fitted linear models for *ErrB* with data for the note borrowing network

index	R	R^2	b_0	b_1
p.changed	0.702	0.492	-0.06	0.009
Dens	0.643	0.413	-0.032	0.414
Rec	0.044	0.002	0.110	0.064
D_Mut	0.471	0.221	0.045	0.249
D_Asym	0.551	0.304	-0.026	0.314
D_Null	0.642	0.413	-0.047	0.601
PP_e	0.658	0.432	-0.070	0.183
CCout_e	0.663	0.440	-0.049	0.240
CCin_e	0.658	0.433	-0.070	0.184
Dall_e	0.665	0.443	0.004	0.369
Dout_e	0.661	0.437	-0.007	0.354
Din_e	0.681	0.464	-0.006	0.243
B_e	0.338	0.114	-0.028	1.001
A_e	0.677	0.458	-0.028	0.490
H_e	0.569	0.324	-0.016	0.845
PP_cor	-0.566	0.320	0.372	-0.376
CCout_cor	-0.387	0.150	0.195	-0.210
CCin_cor	-0.566	0.320	0.373	-0.377
Dall_cor	-0.607	0.368	0.358	-0.350
Dout_cor	-0.418	0.175	0.215	-0.235
Din_cor	-0.621	0.385	0.416	-0.399
B_cor	-0.334	0.112	0.185	-0.184
A_cor	-0.629	0.396	0.395	-0.381
H_cor	-0.564	0.318	0.243	-0.262

Legend:

r - Pearson correlation coefficient

R^2 - variance explained

b_0 - the intercept parameter in regression model

b_1 - the slope parameter in regression model

dyads. The linear regression model explains 41% of variance in values of *ErrB*, when the predictor is difference in network density or number of null dyads.

Figure 8.15 presents linear models for *ErrB* and indices obtained by Euclidean distance from vectors of network properties. Similarly as for *ARI* values, there is no linear trend in the case of betweenness centrality. The fitted model is also not drawn for hubs weight because of correlation coefficient below 0.5. The linear regression models for

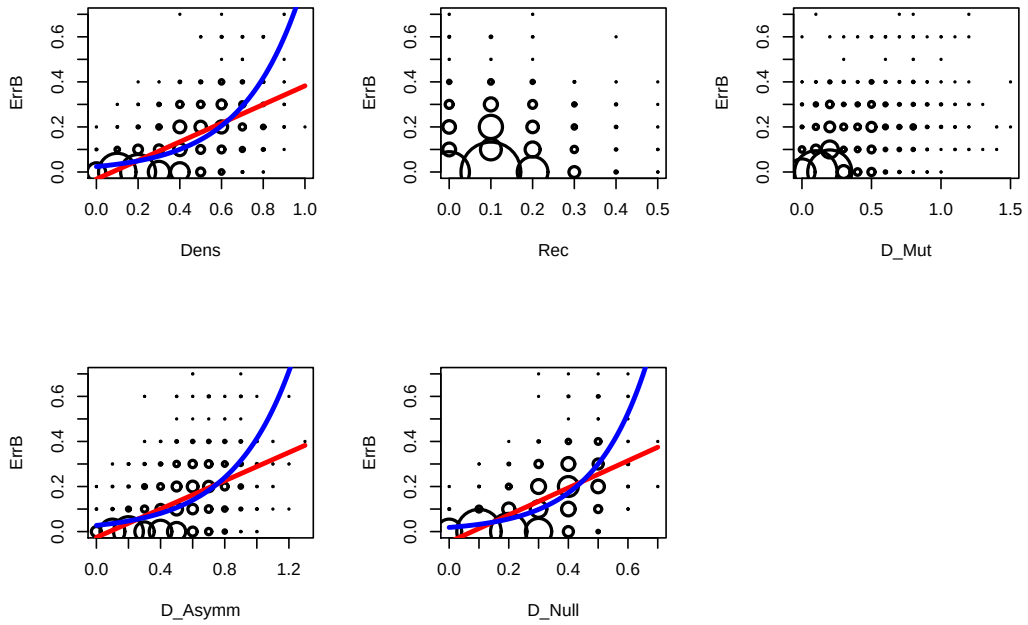


Figure 8.14: Impact of differences in network characteristics to values of *ErrB* with the note borrowing network

closeness centrality based on indegree and authority weights explain 46% of variance in values of *ErrB*. Both indices based on outdegree, closeness centrality and degree centrality are able to explain 44% of variance.

In Figure 8.16 linear regression models for *ErrB* with predictors obtained by correlation between vectors of network properties for whole and measured network are presented. The linear models are fitted for six indices with correlation coefficient higher than 0.5. Among them, the highest percent of variance (40%) can be explained with use of authority weights between networks. In general, the linear models for *ErrB* explain less variance than models for *ARI*.

The generalized linear models for values of *ErrB* are presented in Table 8.10. Similarly as in linear models, the most variance can be explained with use of percent of changed ties (*p.changed*) as a predictor (48.5%). Similar pattern as for values of *ARI* among indices of differences in network characteristics can be observed. The most variance can

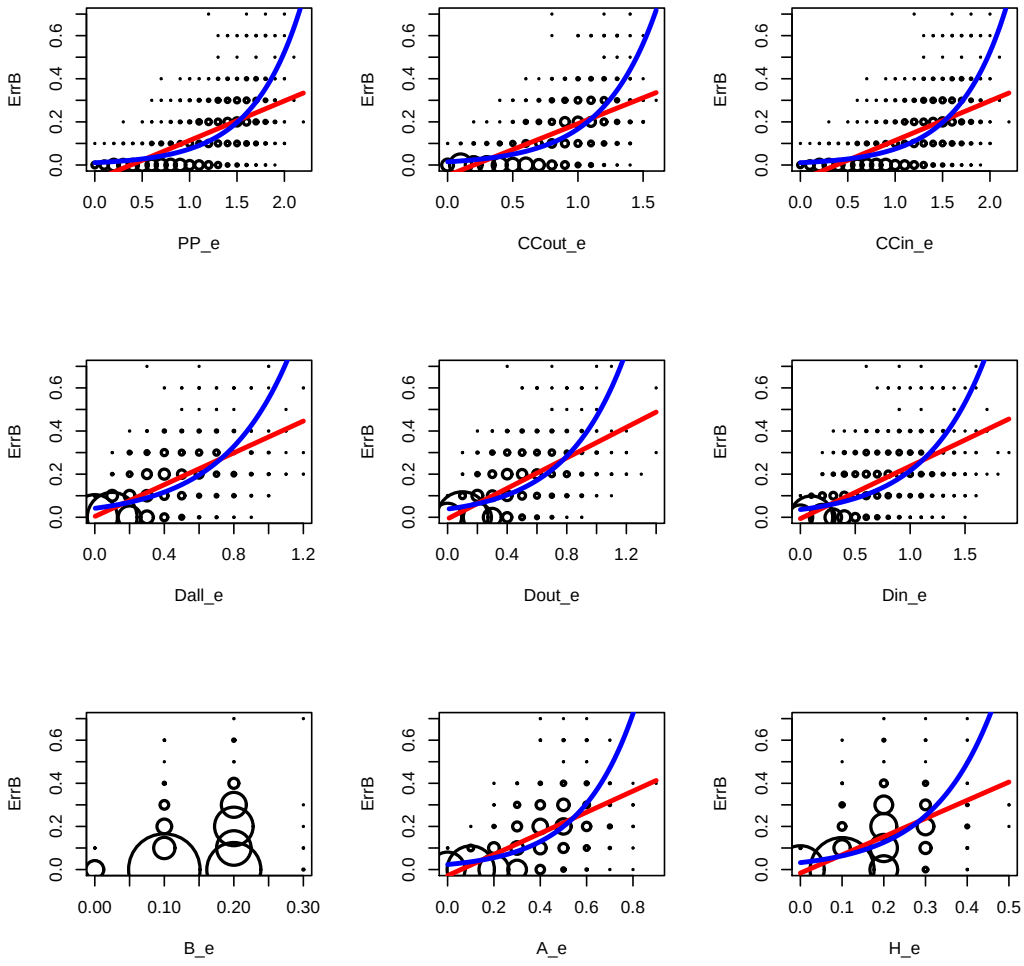


Figure 8.15: Impact of differences in network properties based on Euclidean distance to values of *ErrB* with note borrowing network

be explained with the use of relative difference in number of null dyads as a predictor (39.4%). Percent of explained variance in values of *ErrB* when indices based on Euclidean distance are used as a predictor is in range from 11.9% for *B_e* to 44.8% for *PP_e* and *CCin_e*. Among indices calculated with correlation between vectors of network properties the most variance can be explained with use of *A_cor*. The generalized linear models are presented in above figures (with blue) for those models, where at least 25% of variance in values of *ErrB* is explained.

Similarly as for the Adjusted Rand Index, the highest percent of explained variance (49%) is obtained with percent of changed ties as a predictor. Figure 8.17 shows data

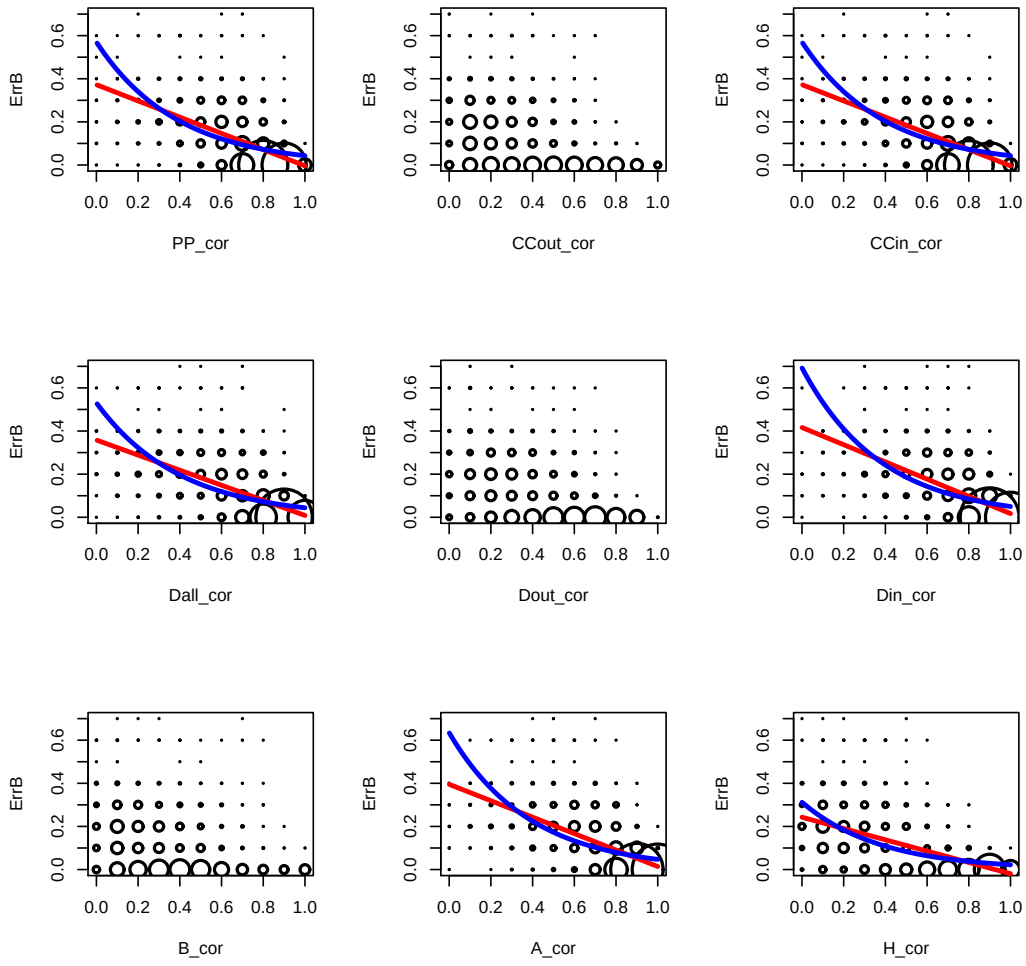


Figure 8.16: Impact of differences in network properties based oncorrelations to values of ARI with note borrowing network

from note borrowing network plotted against percents of changed ties.

The 'aggregated' scatterplot in Figure 8.17 indicates that instead of linear model, other models with functional dependency will be more suitable. The exponential model is able to explain 48.5% of variance in values of $ErrB$, and the quadratic model is able to explain 50.0% of variance in values of $ErrB$. Because the majority of $ErrB$ values between 1 and 15 percent of changed ties is 0, the two-piecewise model (drawn in green in Figure 8.17) was also fitted to the data. It is able to explain 50.8% of variance in values of $ErrB$ (Table 8.11).

Table 8.10: Results of fitted generalized linear models for $ErrB$ with data for the note borrowing network

index	a	b	Dispersion	Residual deviance	Scaled deviance	R^2
p.changed	-4.351	0.086	0.090	367.526	4062.532	0.485
Dens	-3.724	3.565	0.108	446.329	4130.240	0.375
Rec	-2.204	0.529	0.165	712.935	4312.481	0.001
D_Mut	-2.738	1.658	0.145	592.493	4085.404	0.170
D_Asym	-3.599	2.716	0.127	514.963	4044.137	0.279
D_Null	-3.996	5.587	0.106	433.005	4095.596	0.394
PP_e	-4.547	1.950	0.097	394.307	4072.665	0.448
CCout_e	-4.216	2.428	0.097	399.481	4113.363	0.440
CCin_e	-4.547	1.95	0.097	394.253	4072.548	0.448
Dall_e	-3.182	2.587	0.111	463.84	4195.574	0.350
Dout_e	-3.269	2.508	0.111	464.615	4181.825	0.349
Din_e	-3.332	1.800	0.106	442.522	4180.561	0.380
B_e	-3.738	10.272	0.145	629.227	4324.873	0.119
A_e	-3.755	4.288	0.100	415.222	4154.760	0.418
H_e	-3.439	6.835	0.127	510.608	4025.775	0.285
PP_cor	-0.562	-2.581	0.133	535.367	4033.008	0.250
CCout_cor	-1.475	-2.166	0.149	603.799	4054.447	0.154
CCin_cor	-0.561	-2.581	0.133	535.225	4032.904	0.250
Dall_cor	-0.633	-2.489	0.125	503.977	4019.441	0.294
Dout_cor	-1.366	-2.224	0.146	592.596	4050.055	0.170
Din_cor	-0.366	-2.629	0.123	505.552	4122.230	0.292
B_cor	-1.561	-1.862	0.156	632.857	4067.795	0.114
A_cor	-0.455	-2.601	0.120	494.476	4119.499	0.307
H_cor	-1.165	-2.643	0.123	486.664	3945.478	0.318

Degrees of freedom: 3949 for the null model and 3948 for the residual model

Null deviance of all models: 1152.917

All models are significant, p-value is 0.000

Legend:

R^2 - deviance explained

a, b - parameters in exponential glm $\hat{y}_{ErrB} = e^{a+b \cdot index}$

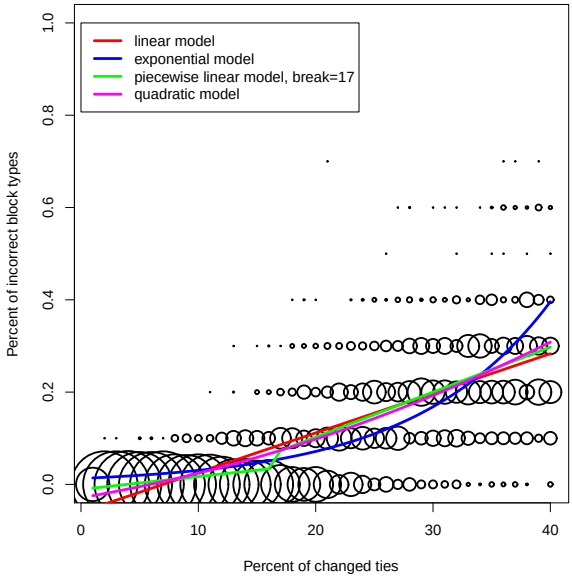


Figure 8.17: Impact of differences in network properties based on correlations to values of $ErrB$ with data for the note borrowing network

Table 8.11: Different fitted models for $ErrB$ with $p.changed$ ties as a predictor for data from the note borrowing network

Name of the model	Formula	R^2
Linear model	$\hat{y}_{ErrB} = -0.060 + 0.009 \cdot p.changed$	0.492
Exponential model	$\hat{y}_{ErrB} = e^{-4.351+0.086 \cdot p.changed}$	0.485
Piecewise linear models where		
break=17	$\hat{y}_{ErrB} = \begin{cases} -0.011 + 0.003 \cdot p.changed; & p.changed < 17 \\ -0.092 + 0.010 \cdot p.changed; & p.changed \geq 17 \end{cases}$	0.508
Quadratic model	$\hat{y}_{ErrB} = -0.029 - 0.004 \cdot p.changed - 0.0001 \cdot p.changed^2$	0.500

8.3 The impact of differences in network characteristic on the stability of blockmodeling in case of simulated networks

In this section the impact of differences in network characteristics and properties is investigated with data from three simulated blockmodel structures: completely symmetric blockmodel structure (Section 8.3.1) and two non-symmetric blockmodel structures (Sections 8.3.2 and 8.3.3).

8.3.1 The completely symmetric blockmodel structure

The blockmodel structure for the completely symmetric blockmodel structure is the same as for the boy-girl liking ties network. The construction of starting whole networks is described in Section 6.2.3.1.

8.3.1.1 Stability of partitions and block types

The stability of blockmodeling it terms of partitions and therefore in terms of ARI values is high for quite a large percent of changed ties *p.changed* as presented in Figure 7.49 in Section 7.5.3.1. If the percent of changed ties is lower than 20%, the mean value of Adjusted Rand Index is above 0.8, which indicates good agreement between partitions. The patterns in data on differences in network characteristics and their impact on indices of network stability observed with the boy-girl liking ties network (Section 8.2.1) are far more explicit with the simulated completely symmetric blockmodel structure.

Pearson correlation coefficient between percent of changed ties (*p.changed*) and ARI is -0.71 (Table B.3 in Appendix B). The linear model with *p.changed* ties as a predictor is therefore able to explain 50.4% of variation in ARI. This indicates that the model is good, but the 'aggregated' scatterplot can not confirm this conclusion (left part of Figure 8.18). The majority of ARI values is 1 for percent of changed ties in range from 1 to 30, and when percent of changed ties is higher than 30% there is absolutely no

agreement between whole starting and measured partition ($ARI = 0$). The variation in values of ARI increases with higher percent of changed ties as shown with boxplots on the right part of Figure 8.18 (the mean values of ARI are plotted with blue dots). This indicates that generalized linear model could be more appropriate. It explains 41.5% of variation in ARI, but visually it does not fit the data well.

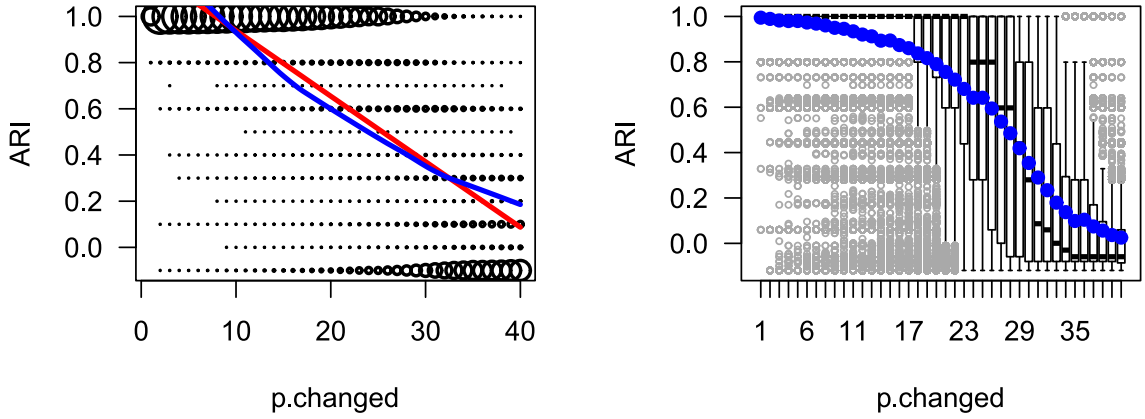


Figure 8.18: 'Aggregated' scatterplots with predictor *p.changed* to values of ARI with the completely symmetric blockmodel structure (left), and boxplots with mean values of ARI (right)

Other indices of changes in network characteristic and network properties reveal even less functional dependency with values of ARI. Figure 8.19 presents two 'aggregated' scatterplots with Euclidean distance (D_e) and correlation between two vectors of network indegree from the whole and measured networks as predictors. The linear regression model with Din_e as a predictor is able to explain 27.2% of variation in ARI (red line on the left part of Figure 8.19). The exponential model is drawn in blue dashed curve, because it explains only 24.6%, and usually the models are drawn if they explain more than 25% of variation. Other indices of differences in network characteristics (except $Dout\_e$) perform even worse. The absolute values of correlation coefficients between ARI and other indices are in range from 0.05 and 0.335 (Table B.3 in Appendix B). As presented on the right part of Figure 8.19 for all range of values of index Din_cor from 0 to 1, the Adjusted Rand Index can take the value 1, which indicates excellent agreement between two partitions, or value -0.1 which indicates that partitions from the whole and measured networks are obtained at random. There is clearly no pattern in data which could have predictive power. Similar results (not presented here) are

obtained with other indices of differences in network characteristics and properties.

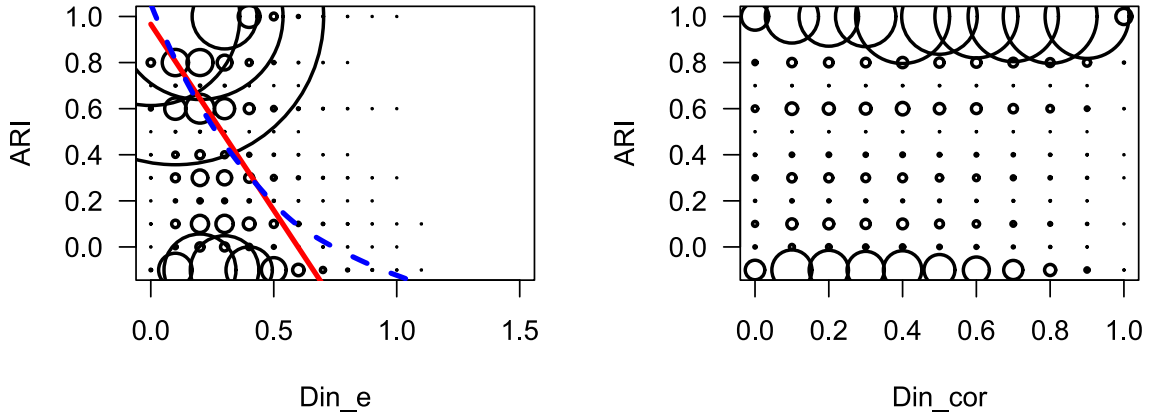


Figure 8.19: 'Aggregated' scatterplots of predictors *Din_e* and *Din_cor* to values of *ARI* with data for the completely symmetric blockmodel structure

The above results show that in the completely symmetric blockmodel structure indices of differences in network characteristics have practically no predictive power to values of *ARI* and similarly to percent of incorrect block types (*ErrB*). The explanation for these results is visible from the figure below. Figure 8.20 shows that both predictors, calculated with Euclidean distance and with correlation, are sensitive to percent of changed ties in a measured network. Those changes reflect in values of *Din_e* and/or *Din_cor* and their increasing variation practically from one or two changed ties onwards.

On the other hand, the blockmodel is stable for almost 20% of changed ties, which is shown on boxplots where 75% of *ARI* values are equal to 1 (Figure 8.19) or in Figure 7.49 where mean values are above 0.8. According to the above findings, we can conclude that index which is sensitive to changes in percent of changed ties (e.g. *Din_cor*) can not successfully predict another index which is extremely insensitive (*ARI*), at any rate not only with simple models. Results of low prediction power are even more extreme with second index of blockmodeling stability, the percent of incorrect block types (*ErrB*) and therefore they are not presented here.

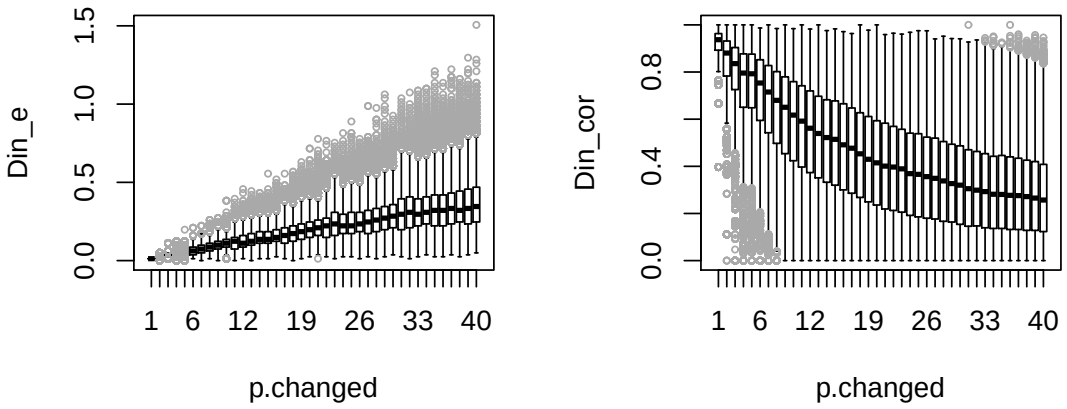


Figure 8.20: Boxplots for *Din_e* and *Din_cor* according to percent of changed ties (*p.changed*) with data for the completely symmetric blockmodel structure

8.3.2 The first non-symmetric blockmodel structure

The structure of a network which was used in simulation study is presented in Equation (6.3) in Section 6.2.3.2. The real whole starting networks have 15 actors and the real blockmodel structure based on structural equivalence has three-cluster partition with complete blocks on diagonal and one off-diagonal complete block.

First, the predictive power of differences in network characteristics and properties to the Adjusted Rand Index is examined in Section 8.3.2.1. The impact of differences in network characteristic on the stability of blockmodeling in terms of correctly identified block types is presented in Section 8.3.2.2.

8.3.2.1 Stability of partitions

The correlations between ARI and indices of network characteristics reveal (Table 8.12) that the highest linear relationship is between relative differences in number of mutual dyads D_{Mut} and values of ARI ($r = -0.411$). All correlations between indices of network characteristics and ARI are in range -0.288 (D_Asym) and -0.411, which indicates medium linear effect according to Cohen (1988). Among all indices of network properties calculated with Euclidean distances, there is the highest correlation between

all three indices of degree centrality and ARI (e.g. correlation between degree centrality based on indegree (Din_e) and ARI is $r = -0.701$). There is no linear effect between values of ARI and betweenness centrality based on Euclidean distance ($r = 0.085$). The correlations between indices of network properties calculated with correlations between two vectors show higher linear relationship with ARI than corresponding indices calculated with Euclidean distance. The highest positive correlation is between relative differences in hub weights and ARI ($r = 0.764$) and relative differences in closeness centrality based on all-degree and ARI ($r = 0.763$). The medium linear effect is also between differences in betweenness centrality B_cor and ARI ($r = 0.319$).

The linear regression models with the Adjusted Rand Index as dependent variables were fitted to data where correlations are higher than 0.5. Figure 8.21 presents 'aggregated' scatterplots where values of ARI are plotted against difference in network characteristics. There is no large linear effect (Table 8.12), therefore no linear model is fitted to the data.

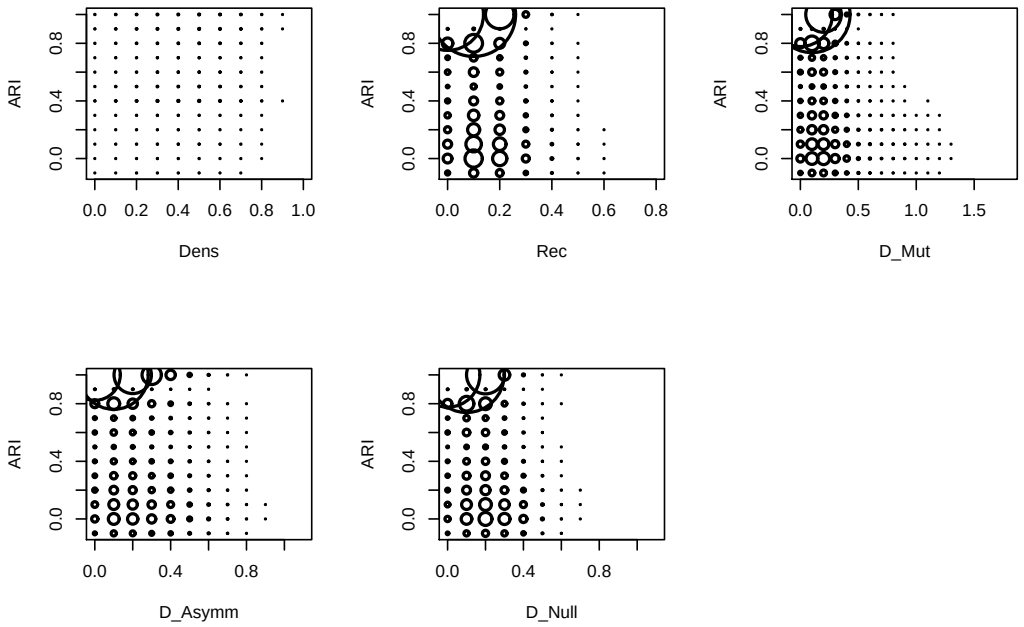


Figure 8.21: Impact of differences in network characteristics to values of ARI with data for the first non-symmetric blockmodel structure

Figure 8.22 shows the ARI values plotted against the indices calculated with Euclidean

Table 8.12: Correlations and results of fitted linear models for ARI with data for the first non-symmetric blockmodel structure

index	r	R^2	b_0	b_1
p.changed	-0.832	0.692	1.2	-0.03
Dens	-0.374	0.14	0.731	-1.671
Rec	-0.3	0.09	0.74	-1.317
D_Mut	-0.343	0.118	0.732	-0.902
D_Asym	-0.288	0.083	0.727	-0.752
D_Null	-0.411	0.169	0.803	-1.389
PP_e	-0.229	0.053	0.668	-0.218
CCout_e	-0.251	0.063	0.677	-0.281
CCin_e	-0.274	0.075	0.69	-0.314
Dall_e	-0.693	0.48	0.936	-1.939
Dout_e	-0.694	0.482	0.949	-1.082
Din_e	-0.701	0.491	0.935	-0.998
B_e	0.085	0.007	0.545	0.876
A_e	-0.638	0.407	0.853	-2.608
H_e	-0.697	0.486	0.927	-3.749
PP_cor	0.693	0.48	-0.28	1.257
CCout_cor	0.653	0.427	-0.132	1.075
CCin_cor	0.694	0.481	-0.282	1.258
Dall_cor	0.763	0.582	-0.427	1.375
Dout_cor	0.706	0.498	-0.229	1.158
Din_cor	0.738	0.545	-0.38	1.315
B_cor	0.319	0.102	0.386	0.482
A_cor	0.759	0.576	-0.392	1.319
H_cor	0.764	0.584	-0.321	1.254

Legend:

r - Pearson correlation coefficient

R^2 - variance explained

b_0 - the intercept parameter in regression model

b_1 - the slope parameter in regression model

distances between network properties. Linear models (red lines) are fitted for all three indices of degree centrality and correlations between vectors of hub and authority weights. The highest percent of explained variance (49.1%) in values of ARI is obtained with closeness centrality based on indegree (Din_e) as predictor. All linear models (except for betweenness centrality) show positive relationship between predictor and values of ARI. 'Aggregated' scatterplots also show that if values of predictor are 1 (perfect correlation between two vectors of network properties between real and mea-

sured network), there is almost perfect agreement between partitions (the majority of ARI is above 0.8).

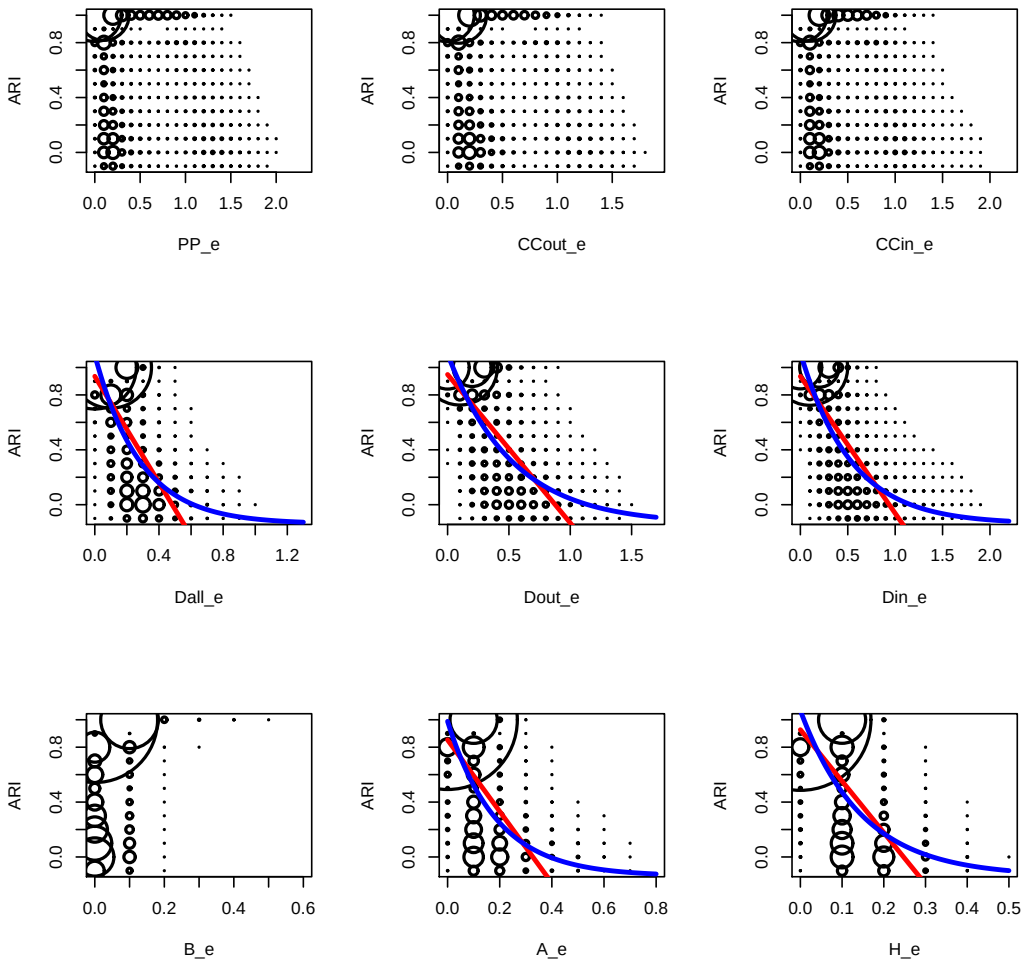


Figure 8.22: Impact of differences in network properties based on Euclidean distance to values of ARI with first non-symmetric blockmodel structure

The regression linear models are plotted for all indices of network properties calculated as correlations between two vectors except for betweenness centrality (Figure 8.22). The highest percent of variance in ARI is explained with use of hub weights (58.4%) and with use of closeness centrality based on all-degree (58.2%).

Different regression models were also fitted to the percent of changed ties as a predictor *p.changed*. Table 8.12 shows that linear regression model (red line in Figure 8.24) can explain 69.2% of variance in ARI.

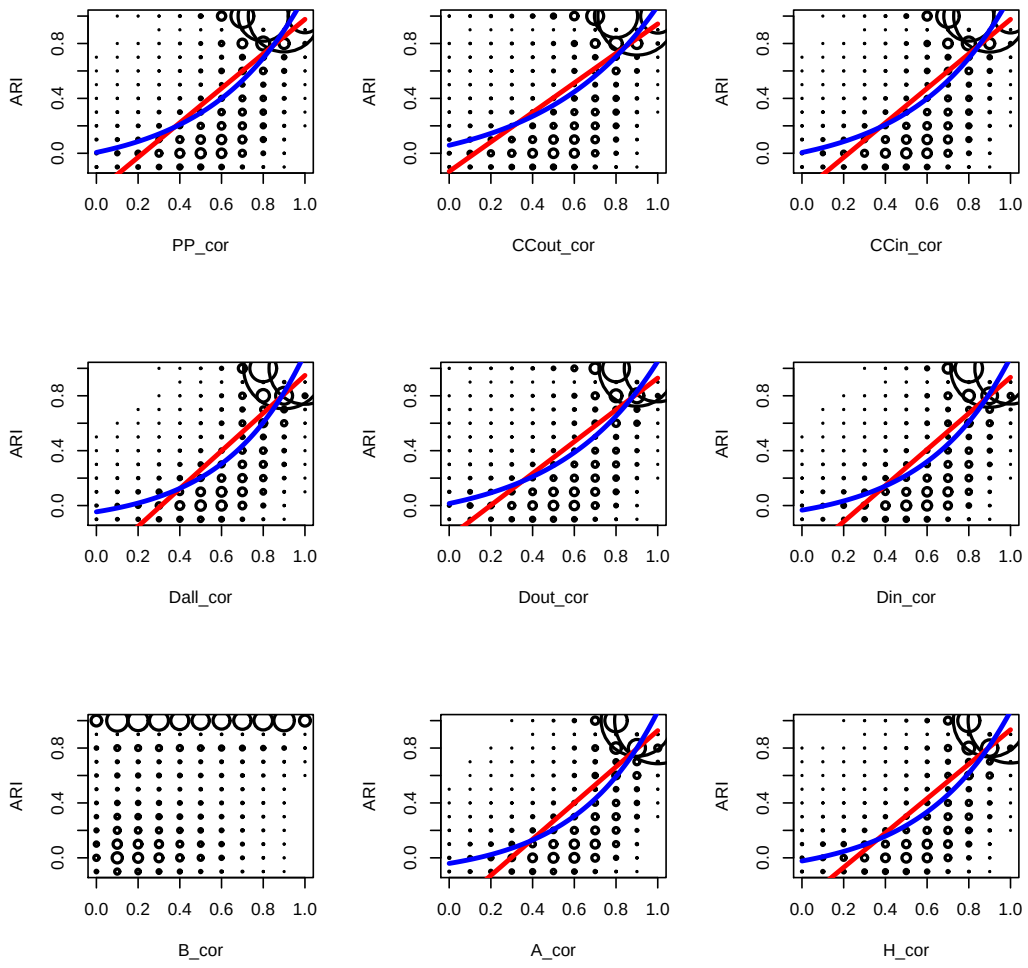


Figure 8.23: Impact of differences in network properties based on correlations to values of ARI with the first non-symmetric blockmodel structure

For small percent of changed ties the majority of ARI values is 1, which suggests that two-piecewise regression model will probably be able to explain more variance in ARI. Figure C.3(a) in Appendix C suggests that break at 14% of changed ties will lead to the best piecewise model. The model is presented in blue in Figure 8.24 and it is able to explain 71.4% of variance in ARI (Table 8.13).

Beside the linear regression models, also the quadratic model is fitted to the data. Because the piecewise model is better than the regular linear model, this indicates that other functional dependencies could be more suitable. The quadratic model is similar, but a little bit worser than piecewise model with break at 14% changed ties and can

Table 8.13: Different fitted models for ARI with $p.changed$ ties as a predictor for data from the first non-symmetric blockmodel structure

Name of the model	Formula	R^2
Linear model	$\hat{y}_{ARI} = 1.200 - 0.030 \cdot p.changed$	0.692
Exponential model	$\hat{y}_{ARI} = e^{0.453 - 0.043 \cdot p.changed}$	0.588
Piecewise linear models where		
break=14	$\hat{y}_{ARI} = \begin{cases} 1.009 - 0.007 \cdot p.changed; & p.changed < 14 \\ 1.348 - 0.035 \cdot p.changed; & p.changed \geq 14 \end{cases}$	0.714
Quadratic model	$\hat{y}_{ARI} = 1.074 - 0.015 \cdot p.changed - 0.0004 \cdot p.changed^2$	0.706

explain 70.6% of variance in ARI. It is drawn in magenta in Figure 8.24 where it can be seen that it fails to correctly predict values of ARI for small percentages of changed ties.

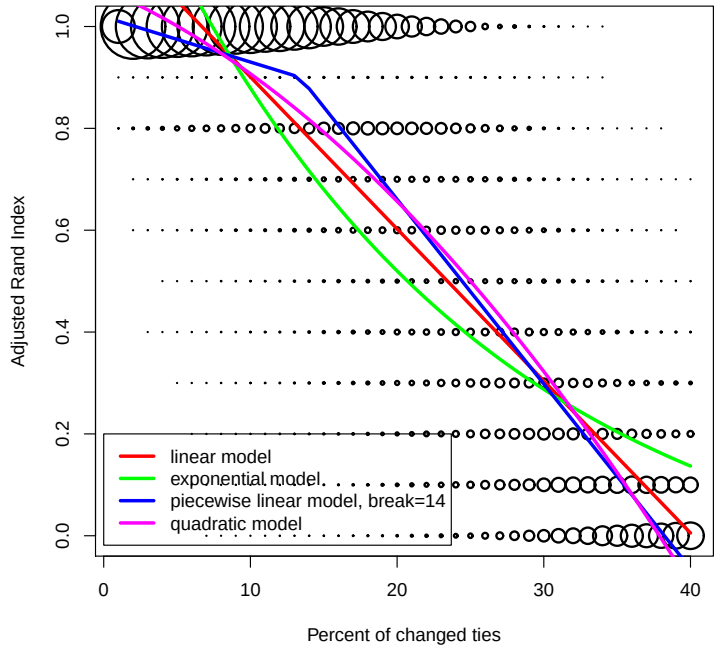


Figure 8.24: Impact of percent of changed ties on values of ARI with the first non-symmetric blockmodel structure

The last step of searching for the best fitting model for ARI was investigation of generalized linear models. Figures of ARI values plotted against percent of changed ties in Section 7.5.3.2 suggested that variance is proportional to the mean values of Adjusted

Rand Index. When values of ARI decrease, the variance increases, which is a sign that GLM could be appropriate selection. Table 8.14 presents results of fitted GLM with exponential dependency to 24 predictor variables.

Table 8.14: Results of fitted generalized linear models for ARI with data for the first non-symmetric blockmodel structure

index	a	b	Dispersion	Residual deviance	Scaled deviance	$R^2_{D,df}$
p.changed	0.453	-0.043	0.109	36404.076	334231.435	0.588
Dens	-0.095	-2.916	0.205	76543.524	373813.952	0.135
Rec	-0.102	-2.002	0.217	81426.956	374933.113	0.079
D_Mut	-0.093	-1.55	0.21	78482.373	374018.496	0.113
D_Asym	-0.122	-1.138	0.219	82011.193	374934.122	0.073
D_Null	-0.016	-2.142	0.203	75136.449	370327.275	0.15
PP_e	-0.205	-0.332	0.224	84336.807	377227.609	0.046
CCout_e	-0.189	-0.438	0.222	83460.588	375631.708	0.056
CCin_e	-0.171	-0.489	0.219	82512.873	376082.52	0.067
Dall_e	0.208	-3.533	0.137	46512.955	339309.011	0.474
Dout_e	0.214	-1.907	0.139	47298.684	341494.117	0.465
Din_e	0.222	-1.878	0.13	44615.052	342240.458	0.496
B_e	-0.374	1.135	0.232	87936.431	379676.093	0.006
A_e	0.121	-5.332	0.146	50226.405	344039.041	0.432
H_e	0.189	-6.792	0.137	46181.987	337058.585	0.478
PP_cor	-1.923	2.176	0.142	47907.74	336633.482	0.458
CCout_cor	-1.618	1.811	0.164	53079.125	322883.921	0.4
CCin_cor	-1.927	2.18	0.142	47799.345	336531.639	0.46
Dall_cor	-2.361	2.584	0.112	36872.618	328435.652	0.583
Dout_cor	-1.865	2.041	0.145	46000.849	316351.617	0.48
Din_cor	-2.24	2.439	0.123	40538.621	328724.248	0.542
B_cor	-0.605	0.649	0.222	81132.931	365894.06	0.083
A_cor	-2.309	2.497	0.115	37248.531	324659.695	0.579
H_cor	-2.145	2.335	0.115	37076.115	322052.062	0.581

Degrees of freedom: 315999 for the null model and 315998 for the residual model

Null deviance of all models: 88441.61

All models are significant, p-value is 0.000

Legend:

R^2 - deviance explained

$a,\ b$ - parameters in exponential glm $\hat{y}_{ARI} = e^{a+b \cdot index}$

Comparison of scaled deviance to residual degrees of freedom reveals the fit of the model. In order to compare generalized linear models to linear ones, the pseudo $R^2_{D,df}$ was used (Mittlbock, 2004). For example, the scale deviance for the model with *p.changed* as a predictor is 334231 which is close to residual degrees of freedom $df_R =$

315998. The pseudo $R^2_{D,df}$ shows that this model explains 58.8% of variance in ARI which is in fact a little bit worse than if linear model is fitted to the data ($R^2 = 69.2\%$ in Table 8.12). Percentages of explained variance are very similar in linear models and in GLMs. The GLMs perform slightly better when Din_e, $Dall\_cor$, and A_cor are predictors in the models. When models explain more than 25% of variance ($R^2_{D,df}$ is higher than 0.25), the exponential models are drawn (in blue) beside the linear ones in Figures 8.21, 8.22, and 8.23.

8.3.2.2 Stability of block types

In this section we tried to find out if differences in network characteristics and properties are able to predict stability of blockmodel structure or values of $ErrB$ in case of the first non-symmetric blockmodel structure. Table 8.15 shows correlations between predictors and values of $ErrB$ together with percent of explained variance in fitted linear regression model. Quick overview reveals that predictors have smaller power to predict values of $ErrB$ compared to corresponding linear model for ARI.

All indices of differences in network characteristics have small linear effect to stability of block types ($ErrB$) with correlation coefficients in range between 0.24 and 0.33. Among them, the highest percent of explained variance in $ErrB$ (11%) is obtained with use of differences in number of null dyads (D_Null). Among indices of network properties calculated with Euclidean distance the strong linear relationship with $ErrB$ is shown by all three indices of degree centrality ($Dall\_e$, $Dout\_e$, and Din_e), and differences in authority and hub weights (A_e, H_e). The review of correlation coefficients between indices of network properties obtained by correlations reveals that all indices, except $CCout\_cor$ and B_cor, have strong linear relationships with values of $ErrB$. Correlation coefficients among all indices are presented in Table B.4 in Appendix B.

As written above, the differences in network characteristics shows no clear linear relationship to values of $ErrB$. Therefore only 'aggregated' scatterplots without fitted linear models are presented in Figure 8.25.

Table 8.15: Correlations and results of fitted linear models for *ErrB* with data for the first non-symmetric blockmodel structure

index	r	R^2	b_0	b_1
p.changed	0.678	0.460	-0.085	0.011
Dens	0.254	0.064	0.095	0.509
Rec	0.240	0.058	0.084	0.474
D_Mut	0.248	0.061	0.092	0.292
D_Asym	0.246	0.061	0.085	0.289
D_Null	0.328	0.107	0.061	0.497
PP_e	0.151	0.023	0.115	0.064
CCout_e	0.195	0.038	0.107	0.098
CCin_e	0.180	0.032	0.109	0.092
Dall_e	0.557	0.310	0.013	0.700
Dout_e	0.591	0.350	0.000	0.414
Din_e	0.530	0.281	0.020	0.339
B_e	-0.092	0.008	0.158	-0.427
A_e	0.450	0.202	0.055	0.825
H_e	0.554	0.307	0.017	1.338
PP_cor	-0.584	0.341	0.467	-0.475
CCout_cor	-0.487	0.238	0.380	-0.360
CCin_cor	-0.584	0.342	0.467	-0.476
Dall_cor	-0.639	0.408	0.520	-0.517
Dout_cor	-0.535	0.287	0.417	-0.394
Din_cor	-0.632	0.399	0.510	-0.506
B_cor	-0.263	0.069	0.213	-0.179
A_cor	-0.656	0.431	0.519	-0.512
H_cor	-0.599	0.359	0.458	-0.441

Legend:

r - Pearson correlation coefficient

R^2 - variance explained

b_0 - the intercept parameter in regression model

b_1 - the slope parameter in regression model

Figure 8.26 shows 'aggregated' scatterplots for indices of network properties calculated with Euclidean distance. The linear models are fitted to those data, where models are able to explain at least 25% of variance in *ErrB* values and are represented as red lines on scatterplots. The most variance (35.0%) in values of *ErrB* can be explained by *Dout_e*, the Euclidean distance between two vectors of oudegree from whole and measured network.

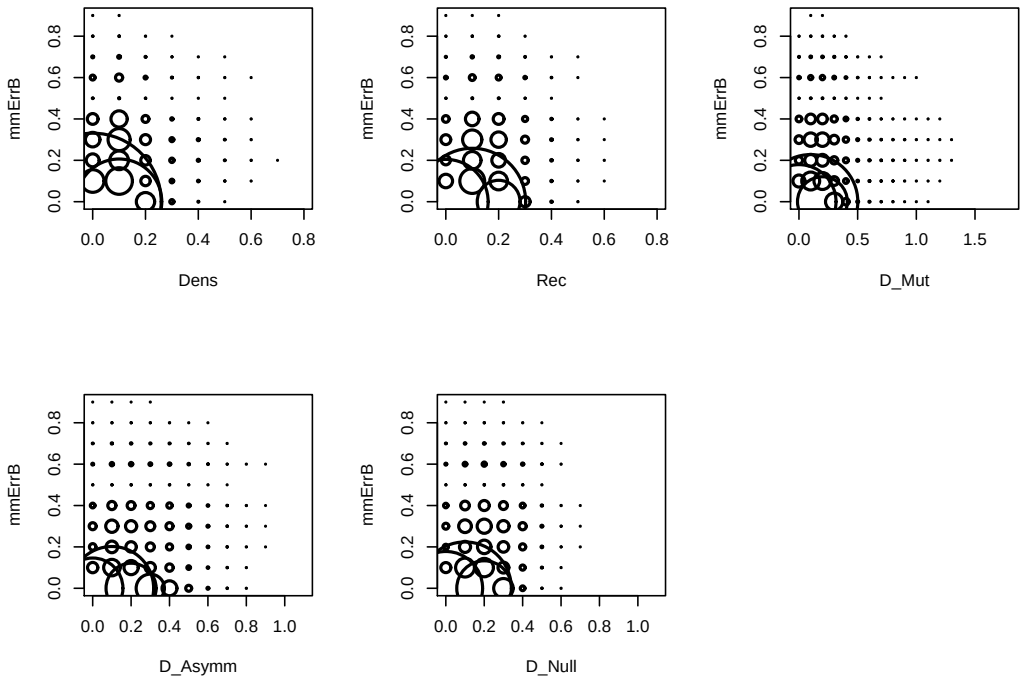


Figure 8.25: Impact of differences in network characteristics to values of *ErrB* with data for the first non-symmetric blockmodel structure

Figure 8.27 presents 'aggregated' scatterplots with linear models where predictors are differences in network properties calculated as correlation between two corresponding vectors. Scatterplot for differences in betweenness centrality (*B_cor*) reveals no clear predictive pattern. For all range of possible values $[0,\ 1]$ for *B_cor* there are networks with perfect stability of block types ($ErrB = 0$). The strongest linear relationship is obtained with use of *A_cor* as a predictor, where model explains 43.1% of variance in values of *ErrB*.

The strongest linear effect is shown when the percent of change ties (*p.changed*) is used as a predictor. The linear model presented in Figure 8.28 explains 46% of variance in *ErrB* (Table 8.16). The 'aggregated' scatterplot suggests that instead of linear model, piecewise model will be more appropriate. For 25% of changed ties, the majority of blocks were correctly identified (*ErrB*=0), and for larger percents of changed ties, the *ErrB* values start to increase.

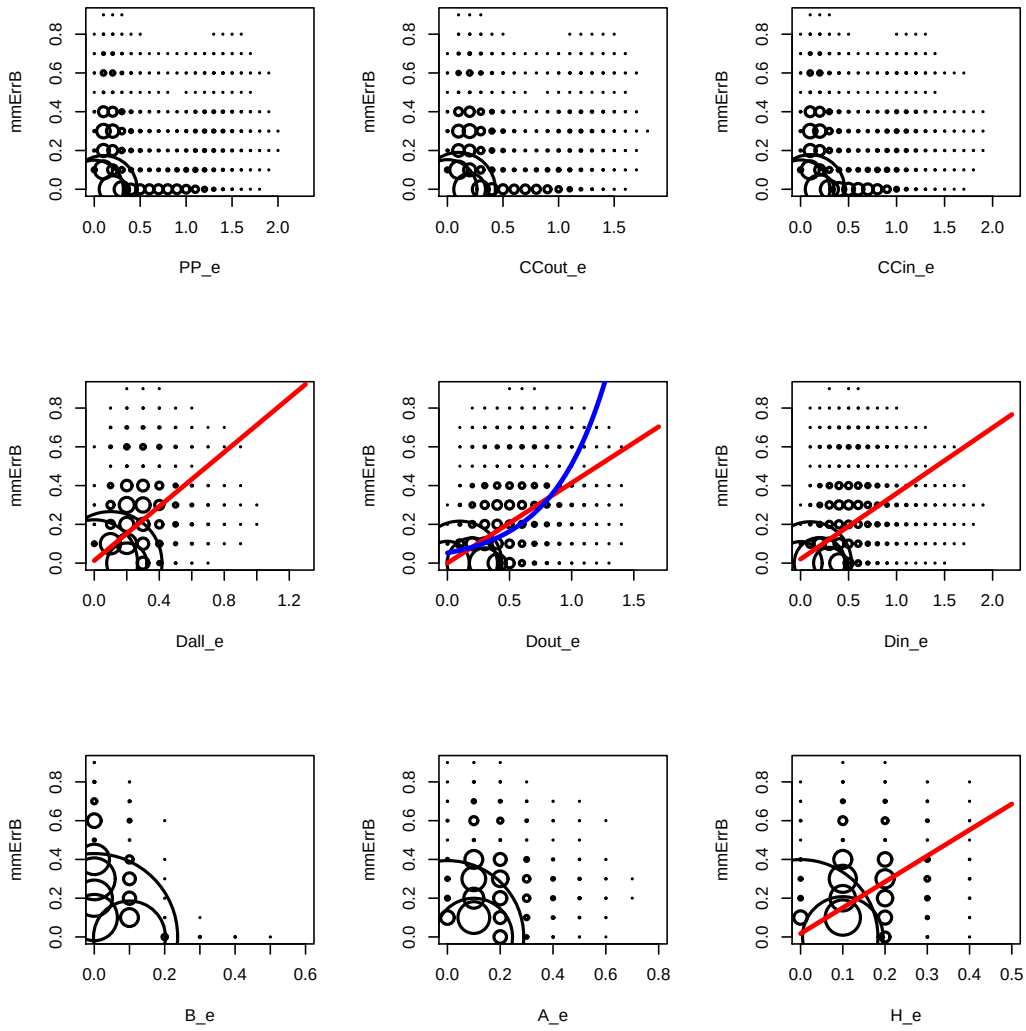


Figure 8.26: Impact of differences in network properties based on Euclidean distance to values of *ErrB* with the first non-symmetric blockmodel structure

Figure C.3 in Appendix C shows that break at 17% of changed ties would be the most appropriate for two-piecewise linear model. The model is drawn in green in Figure 8.28 and explains 48.8% of variance, which is for 2.8% higher than in case of linear model. Quadratic model is practically the same as piecewise one according to Figure 8.28 and also according to percent of explained variance 48.2% in values of *ErrB*.

The fitted generalized linear models for *ErrB* are presented In Table 8.17. The pseudo

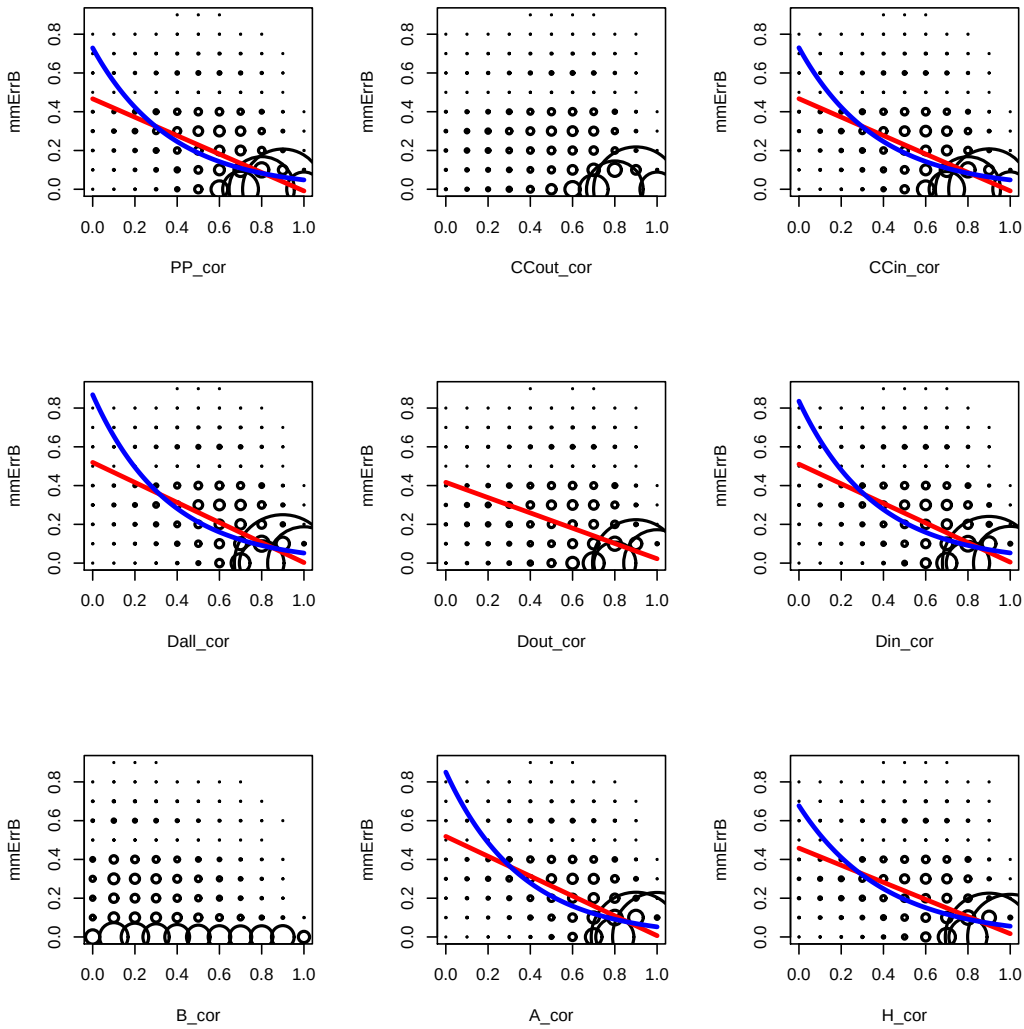


Figure 8.27: Impact of differences in network properties based on correlations to values of $ErrB$ with first non-symmetric blockmodel structure

R^2 shows that GLM for percent of changed ties (*p.changed*) explains 48.7% of variance, which is for 2.7% more than linear model. Exponential model with *p.changed* ties as a predictor is presented with blue curve in Figure 8.28. For other indices of network characteristics, the linear model performs better than generalized one with exponential dependency.

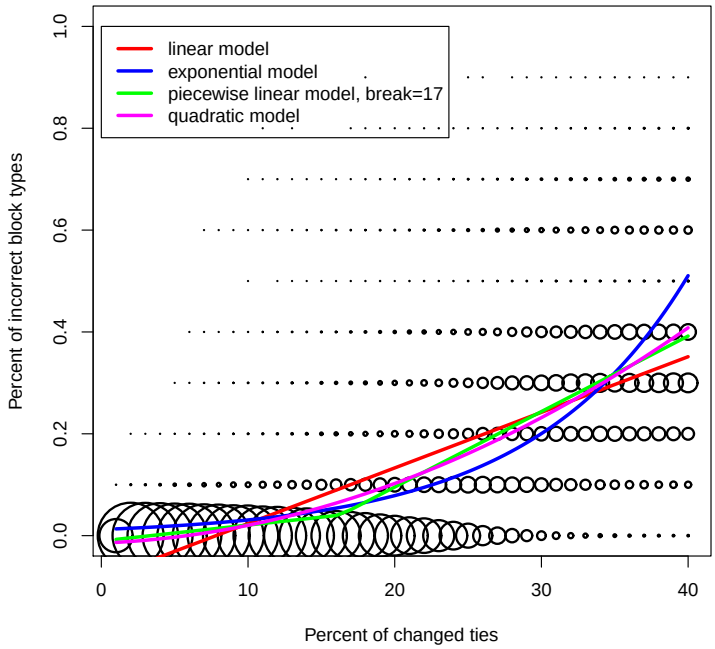


Figure 8.28: Impact of percent of changed ties on values of *ErrB* with the first non-symmetric blockmodel structure

Table 8.16: Different fitted models for *mErrB* with *p.changed* ties as a predictor for data from the first non-symmetric blockmodel structure

Name of the model	Formula	R^2
Linear model	$\hat{y}_{ErrB} = -0.085 + 0.011 \cdot p.changed$	0.460
Exponential model	$\hat{y}_{ErrB} = e^{-4.414+0.094 \cdot p.changed}$	0.487
Piecewise linear models where		
break=17	$\hat{y}_{ErrB} = \begin{cases} -0.010 + 0.003 \cdot p.changed; & p.changed < 17 \\ -0.201 + 0.015 \cdot p.changed; & p.changed \geq 17 \end{cases}$	0.488
Quadratic model	$\hat{y}_{ErrB} = -0.015 + 0.001 \cdot p.changed + 0.0002 \cdot p.changed^2$	0.482

8.3.3 The second non-symmetric blockmodel structure

The second non-symmetric structure is presented in Equation (6.4) in Section 6.2.3.3. The study in this section is an extension of the study for the note borrowing network in Section 8.2.2, because the starting blockmodel structure in simulation of whole starting networks is the same as for the note borrowing network.

Table 8.17: Results of fitted generalized linear models for $mErrB$ with data for the first non-symmetric blockmodel structure

index	a	b	Dispersion	Residual deviance	Scaled deviance	R^2
p.changed	-4.414	0.094	0.145	40081.459	276447.608	0.487
Dens	-2.25	2.796	0.237	74001.396	311595.6	0.052
Rec	-2.354	2.928	0.236	74104.448	313351.818	0.051
D_Mut	-2.252	1.544	0.235	74285.879	315516.602	0.049
D_Asym	-2.351	1.797	0.235	73896.339	313873.27	0.054
D_Null	-2.526	3.066	0.235	70710.996	300744.869	0.094
PP_e	-2.14	0.411	0.24	76475.546	318111.021	0.021
CCout_e	-2.191	0.598	0.238	75490.68	317406.064	0.033
CCin_e	-2.182	0.575	0.24	75854.898	315647.834	0.028
Dall_e	-2.794	3.62	0.193	59287.896	307517.669	0.241
Dout_e	-2.944	2.271	0.183	56100.428	306437.263	0.282
Din_e	-2.726	1.726	0.2	61220.446	306700.99	0.216
B_e	-1.824	-3.622	0.237	77356.758	326894.383	0.009
A_e	-2.486	4.042	0.217	66302.807	306151.569	0.151
H_e	-2.803	7.179	0.195	59026.866	302656.215	0.244
PP_cor	-0.316	-2.719	0.187	56088.661	300373.981	0.282
CCout_cor	-0.682	-2.173	0.205	62151.163	302639.471	0.204
CCin_cor	-0.315	-2.719	0.187	56057.206	300406.261	0.282
Dall_cor	-0.141	-2.82	0.173	52697.81	304469.518	0.325
Dout_cor	-0.534	-2.31	0.197	59284.885	301540.146	0.241
Din_cor	-0.18	-2.768	0.175	53156.396	303855.258	0.319
B_cor	-1.461	-1.401	0.227	72504.049	319598.486	0.071
A_cor	-0.162	-2.792	0.168	51336.29	305240.792	0.343
H_cor	-0.39	-2.503	0.185	55138.422	298822.702	0.294

Degrees of freedom: 315999 for the null model and 315998 for the residual model

Null deviance of all models: 878080.34

All models are significant, p-value is 0.000

Legend:

R^2 - deviance explained

a, b - parameters in exponential glm $\hat{y}_{ErrB} = e^{a+b \cdot index}$

8.3.3.1 Stability of partitions

The linear regression models for the second non-symmetric blockmodel structure are very similar to those for the first non-symmetric blockmodel structure in Section 8.3.2. The most variance (70.2%) in values of ARI is explained when pecent of changed ties (*p.changed*) ties is used as a predictor in simple linear regression model (Table 8.18). Among indices of differences in network characteristics, the most variation (30.9%) in

values of the Adjusted Rand Index can be explained with use of relative differences in number of null dyads (D_Null). Other indices of differences in network characteristics explain less than 17.0% of variance in ARI, therefore the linear models are not drawn on the 'aggregated' scatterplots in Figure 8.29.

Table 8.18: Correlations and results of fitted linear models for ARI with data for the second non-symmetric blockmodel structure

index	r	R^2	b_0	b_1
p.changed	-0.838	0.702	1.152	-0.03
Dens	-0.413	0.170	0.717	-0.793
Rec	-0.250	0.063	0.642	-0.720
D_Mut	-0.372	0.138	0.668	-0.379
D_Asym	-0.282	0.079	0.665	-0.616
D_Null	-0.556	0.309	0.874	-1.609
PP_e	-0.241	0.058	0.628	-0.103
CCout_e	-0.312	0.097	0.649	-0.253
CCin_e	-0.25	0.062	0.633	-0.113
Dall_e	-0.644	0.415	0.829	-1.144
Dout_e	-0.646	0.417	0.872	-1.129
Din_e	-0.730	0.533	0.897	-0.665
B_e	0.074	0.005	0.505	0.625
A_e	-0.695	0.483	0.865	-1.614
H_e	-0.641	0.410	0.868	-3.176
PP_cor	0.692	0.478	-0.518	1.450
CCout_cor	0.485	0.236	0.196	0.738
CCin_cor	0.694	0.482	-0.522	1.455
Dall_cor	0.769	0.591	-0.456	1.36
Dout_cor	0.595	0.354	0.076	0.894
Din_cor	0.781	0.610	-0.768	1.634
B_cor	0.400	0.160	0.246	0.607
A_cor	0.789	0.622	-0.726	1.589
H_cor	0.76	0.577	-0.102	1.067

Legend:

r - Pearson correlation coefficient

R^2 - variance explained

b_0 - the intercept parameter in regression model

b_1 - the slope parameter in regression model

Among predictors in linear models calculated with Euclidean distance, the most variance in values of ARI can be explained with use of Euclidean distance between vectors of indegree Din_e (53.3%) and with use of Euclidean distance between vectors of authority weights A_e (48.3). There is practically no linear relationship (R^2 is in all cases

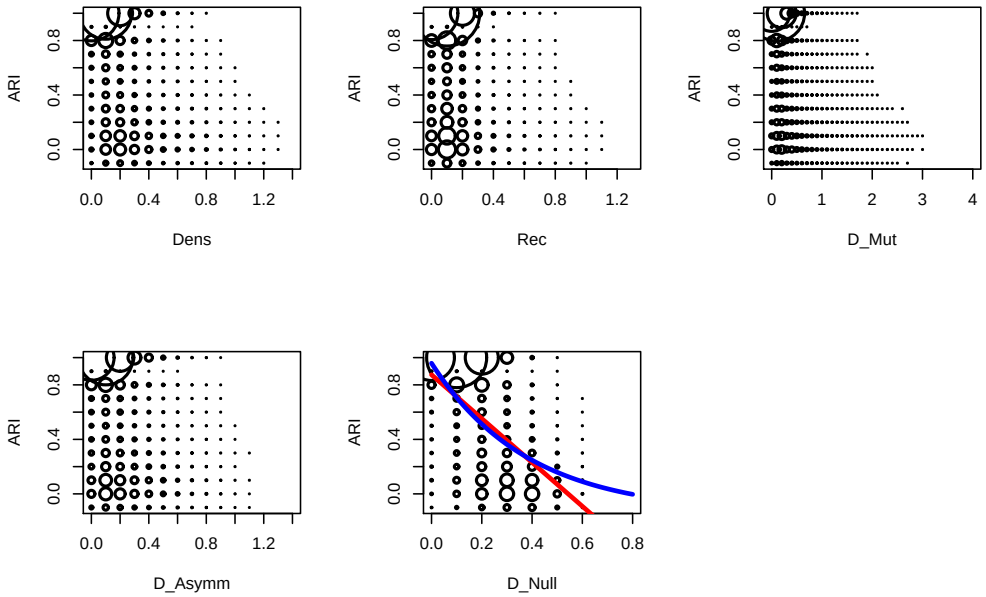


Figure 8.29: Impact of differences in network characteristics to values of ARI with data for the second non-symmetric blockmodel structure

lower or equal to 0.097) when PP_e, $CCout\_e$, $CCin\_e$, and B_e are used as predictors, therefore linear models are not added to 'aggregated' scatterplots in Figure 8.30.

Figure 8.31 presents 'aggregated' scatterplots for indices calculated with correlation. The most variance in values of ARI regression(62.2%) is explained when A_cor is used as a predictor. The linear models are also reasonable when PP_cor and $CCin\_cor$ are used as predictors, which was not true for the corresponding indices calculated with Euclidean distance.

The next step was to fit the generalized linear models with exponential dependency to the data from the second non-symmetric blockmodel structure. Among indices of relative differences in network characteristics the most variation in the Adjusted Rand Index can be explained with use of relative difference in number of null dyads D_Null (Table 8.19). Compared to the corresponding linear model, the exponential model performs a little bit worse, because it explains for 2.8% less variance in values of ARI.

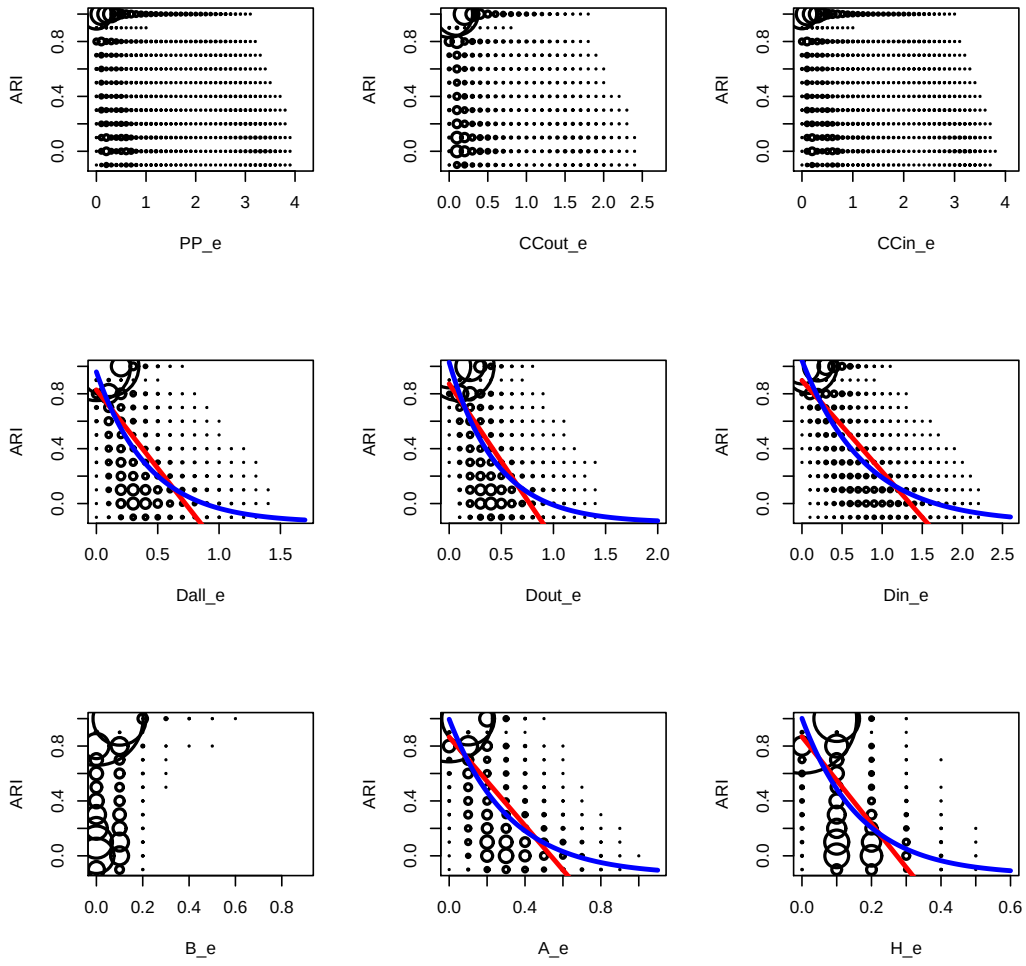


Figure 8.30: Impact of differences in network properties based on Euclidean distance to values of ARI with data for the second non-symmetric blockmodel structure

On the other hand, exponential models with $Dall\_e$ and $Dout\_e$ as predictor perform better than corresponding linear models. The best exponential model among indices calculated with correlation is A_cor, which is able to predict 70.0% of variation in ARI. This model is far better than corresponding linear model, because it is able to explain 7.8% more variation in values of ARI. The GLMs, which can explain at least 25% of variance in ARI, are on above figures drawn with blue curve.

Figure 8.32 presents different models fitted to data from the second non-symmetric blockmodel structure, where percent of changed ties $p.changed$ is a predictor. The linear model (drawn in red) explains 70.3% of variation in ARI. The exponential model,

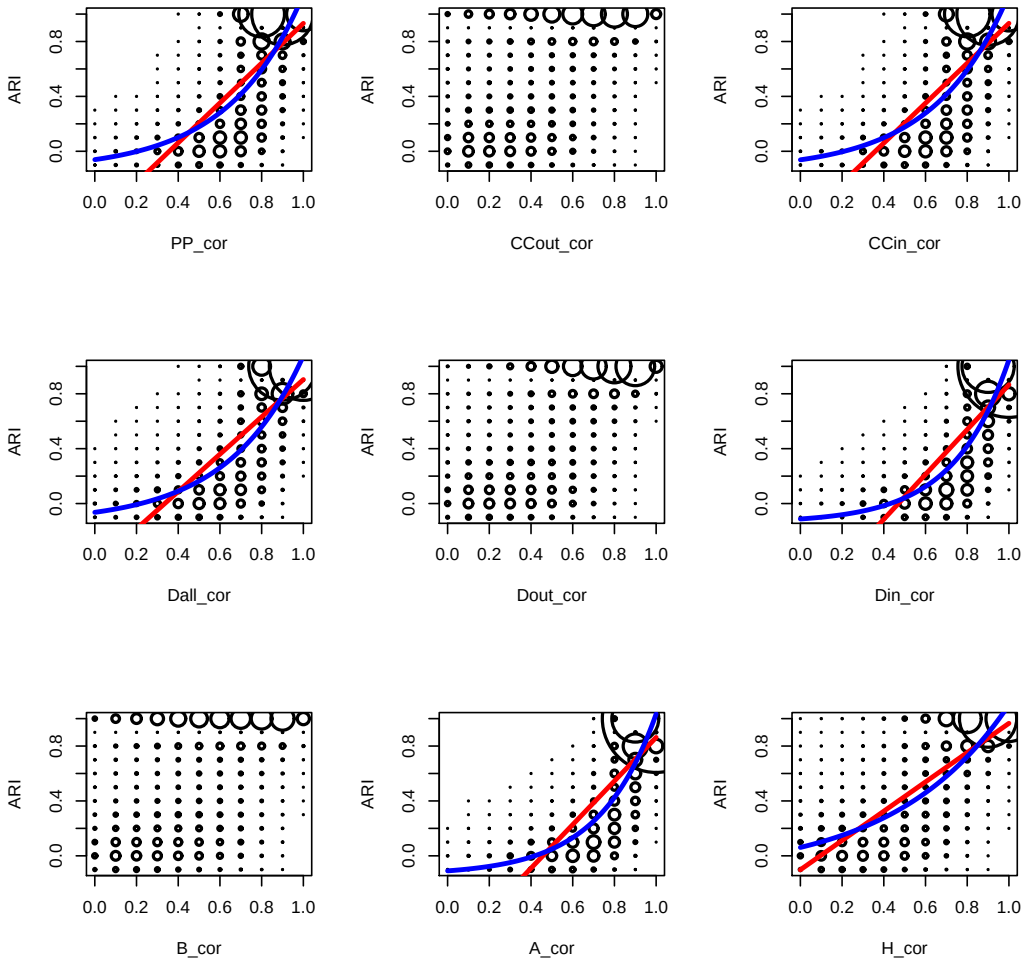


Figure 8.31: Impact of differences in network properties based on correlations to values of ARI with the second non-symmetric blockmodel structure

drawn with blue curve, performs worse, because it is able to explain just 61.8% of variation in values of ARI. The 'aggregated' scatterplot indicates that two-piecewise model should be more suitable, because for small percent of changed ties (less than 10 or 15 percent) the majority of ARI values is equal to 1. Figure C.4 in Appendix C indicates that break at 23 percent of changed leads to the lower residual standard error and therefore the highest percent of explained variance. The two piecewise model is drawn in green and it explains 71.4% of variation in ARI. The quadratic model, drawn in magenta, explains 70.6% of variance in ARI (Table 8.20).

Table 8.19: Results of fitted generalized linear models for ARI with data for the second non-symmetric blockmodel structure

index	a	b	Dispersion	Residual deviance	Scaled deviance	R^2
p.changed	0.431	-0.046	0.102	33641.068	330789.793	0.618
Dens	-0.103	-1.442	0.203	73291.057	360278.862	0.167
Rec	-0.221	-1.256	0.226	82727.901	365822.652	0.06
D_Mut	-0.166	-0.754	0.208	75372.544	361512.496	0.144
D_Asym	-0.194	-1.024	0.223	81543.819	364998.222	0.074
D_Null	0.094	-2.606	0.179	63312.349	353131.328	0.281
P_e	-0.252	-0.17	0.227	83316.287	366679.739	0.053
CCout_e	-0.216	-0.444	0.218	79805.426	366621.985	0.093
CCin_e	-0.245	-0.187	0.226	82932.939	366483.787	0.058
Dall_e	0.095	-2.346	0.149	49425.111	331609.956	0.438
Dout_e	0.156	-2.201	0.151	50358.018	333565.067	0.428
Din_e	0.169	-1.281	0.121	40565.001	334225.531	0.539
B_e	-0.435	0.871	0.238	87618.512	367643.702	0.005
A_e	0.127	-3.152	0.133	44598.366	335014.658	0.493
H_e	0.133	-5.982	0.154	51816.075	336459.751	0.411
PP_cor	-2.537	2.786	0.136	45289.935	333425.333	0.485
CCout_cor	-0.936	1.091	0.205	70249.071	342671.78	0.202
CCin_cor	-2.554	2.804	0.135	44859.279	333378.321	0.490
Dall_cor	-2.576	2.772	0.103	33569.583	324799.455	0.619
Dout_cor	-1.171	1.384	0.18	60561.46	336129.58	0.312
Din_cor	-3.585	3.757	0.087	27565.99	316775.146	0.687
B_cor	-0.85	0.9	0.218	75938.837	347645.202	0.137
A_cor	-3.489	3.649	0.084	26438.935	314674.165	0.700
H_cor	-1.603	1.819	0.124	40520.859	326331.45	0.540

Degrees of freedom: 315999 for the null model and 315998 for the residual model

Null deviance of all models: 88017.06

All models are significant, p-value is 0.000

Legend:

R^2 - deviance explained

a, b - parameters in exponential glm $\hat{y}_{ARI} = e^{a+b \cdot index}$

8.3.3.2 Stability of block types

The highest correlation coefficient between all indices of network characteristics and properties is between percent of changed ties (*p.changed*) and proportion of incorrect blocks (*ErrB*) and is equal to 0.664 (Table 8.21). In comparison to correlation coefficients between indices of network characteristic and indices of blockmodeling stability, we can notice that correlation coefficients between ARI values (Table 8.18) are higher

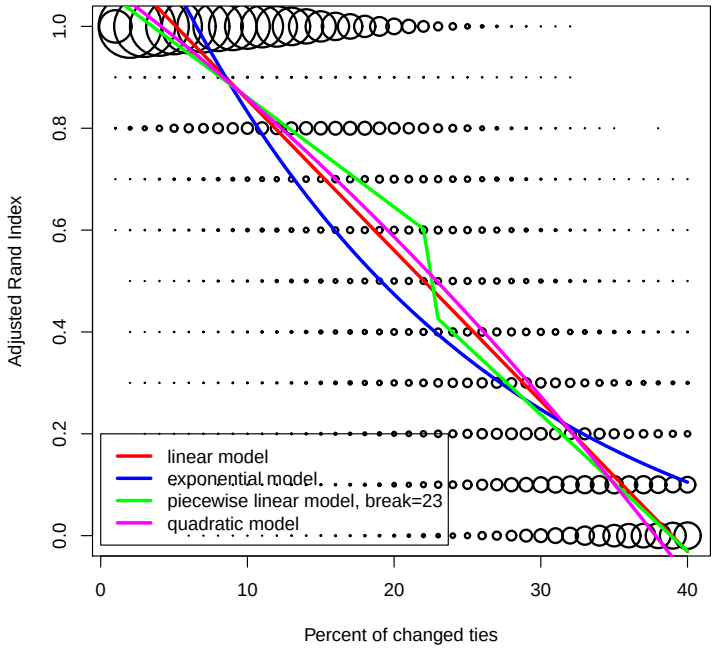


Figure 8.32: Impact of percent of changed ties on values of ARI with data for the second non-symmetric blockmodel structure

Table 8.20: Different fitted models for ARI with $p.changed$ ties as a predictor with data for the second non-symmetric blockmodel structure

Name of the model	Formula	R^2
Linear model	$\hat{y}_{ARI} = 1.152 - 0.030 \cdot p.changed$	0.702
Exponential model	$\hat{y}_{ARI} = e^{0.431 - 0.045 \cdot p.changed}$	0.618
Piecewise linear models where		
break=23	$\hat{y}_{ARI} = \begin{cases} 1.076 - 0.022 \cdot p.changed; & p.changed < 23 \\ 1.043 - 0.026 \cdot p.changed; & p.changed \geq 23 \end{cases}$	0.714
Quadratic model	$\hat{y}_{ARI} = 1.090 - 0.021 \cdot p.changed - 0.0002 \cdot p.changed^2$	0.706

than corresponding coefficient with $ErrB$ values.

Pearson correlation coefficients between indices of relative difference in network characteristics and the proportion of incorrect block types ($ErrB$) are presented in Table 8.21. The relative difference in number of null dyad (D_Null) is able to explain 21.8%

Table 8.21: Correlations and results of fitted linear models for *ErrB* with data for the second non-symmetric blockmodel structure

index	R	R^2	b_0	b_1
p.changed	0.664	0.442	-0.056	0.009
Dens	0.312	0.097	0.073	0.221
Rec	0.184	0.034	0.095	0.195
D_Mut	0.278	0.077	0.087	0.105
D_Asym	0.208	0.043	0.089	0.168
D_Null	0.467	0.218	0.019	0.499
PP_e	0.152	0.023	0.102	0.024
CCout_e	0.214	0.046	0.095	0.064
CCin_e	0.159	0.025	0.101	0.027
Dall_e	0.513	0.263	0.038	0.336
Dout_e	0.526	0.277	0.023	0.339
Din_e	0.566	0.320	0.021	0.19
B_e	-0.067	0.005	0.134	-0.21
A_e	0.512	0.262	0.034	0.439
H_e	0.472	0.223	0.034	0.863
PP_cor	-0.530	0.281	0.422	-0.410
CCout_cor	-0.367	0.134	0.219	-0.206
CCin_cor	-0.531	0.282	0.423	-0.411
Dall_cor	-0.606	0.367	0.413	-0.396
Dout_cor	-0.453	0.205	0.253	-0.251
Din_cor	-0.599	0.358	0.493	-0.462
B_cor	-0.314	0.098	0.208	-0.176
A_cor	-0.61	0.372	0.484	-0.453
H_cor	-0.599	0.359	0.310	-0.311

Legend:

R - Pearson correlation coefficient

R^2 - variance explained

b_0 - the intercept parameter in regression model

b_1 - the slope parameter in regression model

of variation in values of *ErrB*. All other indices perform worse and can explain at most 9.7% of variation. The 'aggregated' scatterplots, without fitted models, are presented in Figure 8.33.

There are four indices of network properties calculated with Euclidean distance which are able to explain more than 25% of variance in values of *ErrB* (Table 8.21). Three of them are calculated from vectors of network degree. The best linear predictor is Eu-

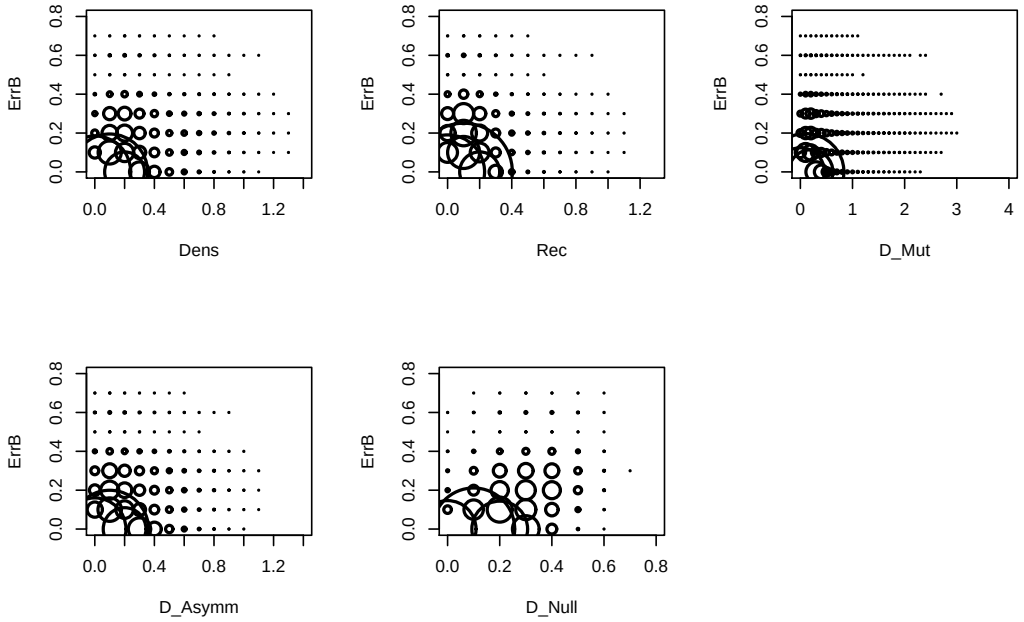


Figure 8.33: Impact of differences in network characteristics to values of *ErrB* with data for the second non-symmetric blockmodel structure

clidean distance between vectors of indegree Din_e (32.0%), $Dout\_e$ is able to explain 27.7% of variance, $Dall\_e$ can explain 26.3% of variance in values of *ErrB*. Linear model with Euclidean distance between vectors of authority weights as predictor can explain 26.2% of variation in *ErrB*. Linear models are drawn in blue on the 'aggregated' scatterplots in Figure 8.34.

The 'aggregated' scatterplots with indices calculated with correlation coefficient between two vectors of network properties are presented in Figure 8.35. Among indices of network properties calculated with correlation, A_cor has the highest percent of explained variance in values of *ErrB* (37.2%). The correlation between vectors of alldegree ($Dall\_cor$) can explain 36.7% of variance in *ErrB*. A little less predictive power has the correlation between vectors of indegree ($Dall\_cor$) which is able to explain 35.8% of variation in values of *ErrB*.

The next step was to examine the generalized linear models with exponential depen-

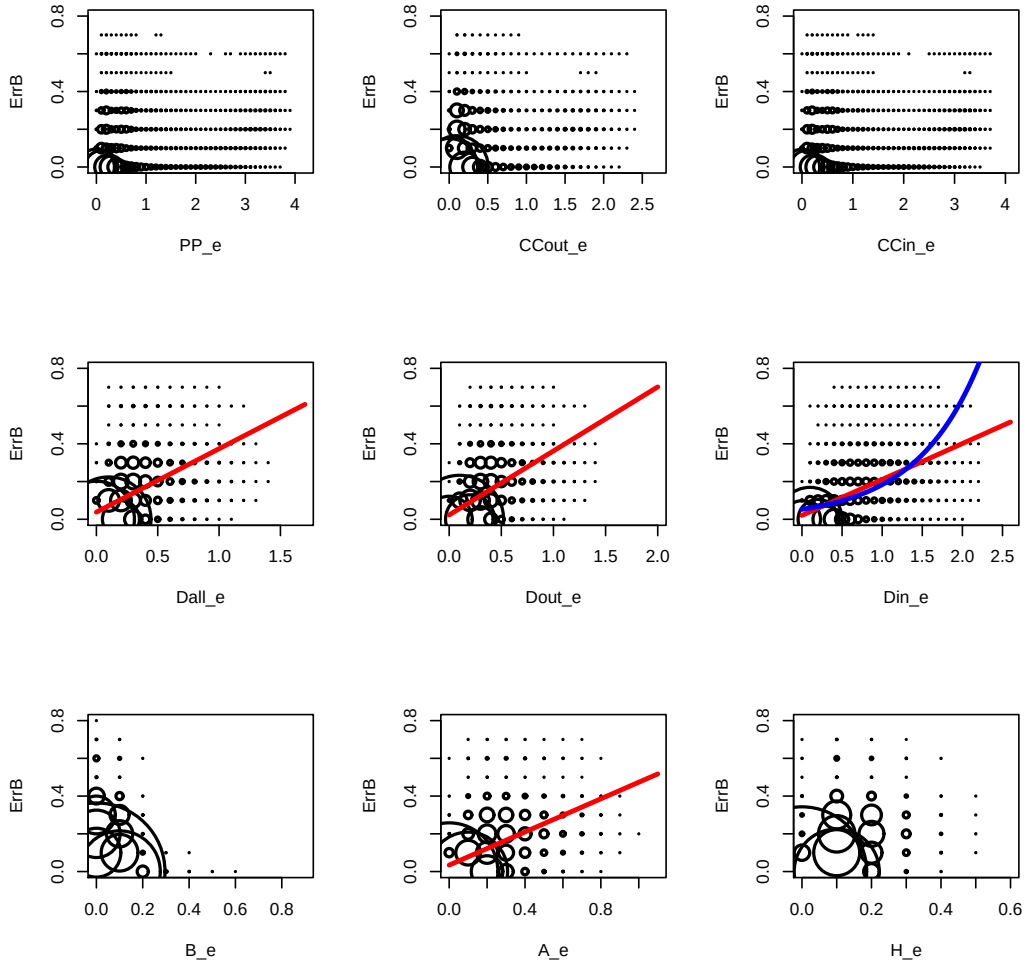


Figure 8.34: Impact of differences in network properties based on Euclidean distance to values of *ErrB* with data for the second non-symmetric blockmodel structure

dency. In comparison with corresponding linear models, they can explain less variance in values of *ErrB*. Among indices of network characteristic, the best predictor in exponential models is *D_Null* which can explain 19.1% of variation in *ErrB*. Among indices calculated with Euclidean distance the best predictor is *Din_e* ($R^2 = 0.255$). The index calculated as correlation between vectors of hub weights (*H_cor*) is the best predictor among indices calculated with correlation and it can explain 32.0% of variation in values of *ErrB* (Table 8.22). The exponential models, which are able to explain at least 25% of variance in *ErrB*, are in above figures drawn with blue curve.

The last step was to examine the predictive power of percent of changed ties *p.changed*

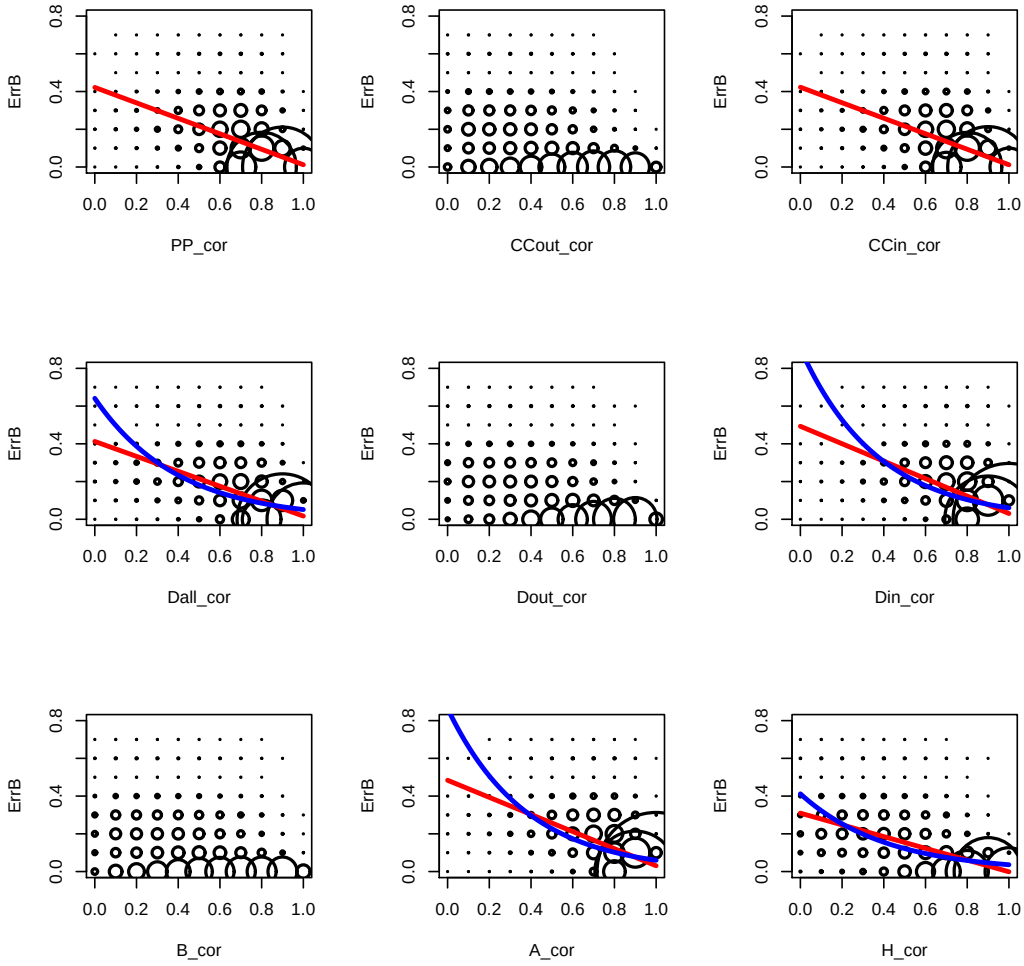


Figure 8.35: Impact of differences in network properties based on correlations to values of *ErrB* with data for the second non-symmetric blockmodel structure

for values of the proportion of incorrectly identified block types *ErrB*. Figure 8.36 presents 'aggregated' scatterplot with different fitted models. The simple linear model is able to explain 44.2% of variation in values of *ErrB*. The pattern in data indicates that two-piecewise model should be appropriate, because for lower percentages of changed ties the majority of *ErrB* values is equal to zero. Figure C.4 in Appendix C indicates that break at $p.changed = 14$ is the most appropriate. The suggested piecewise model, drawn in green, explains 1.5% more variation in *ErrB* than simple linear regression model. The generalized linear model with exponential dependency performs a little bit worse than the linear model, and the predictive power of the quadratic model is between the simple linear regression model and two-piecewise model. The differences

Table 8.22: Results of fitted generalized linear models for $mErrB$ with data for the second non-symmetric blockmodel structure

index	a	b	Dispersion	Residual deviance	Scaled deviance	R^2
p.changed	-4.186	0.082	0.116	34128.463	294854.677	0.44
Dens	-2.469	1.433	0.173	56263.215	324899.442	0.077
Rec	-2.298	1.288	0.178	59310.503	332908.169	0.027
D_Mut	-2.345	0.624	0.174	57430.748	329196.654	0.058
D_Asym	-2.362	1.182	0.178	58744.244	330348.207	0.036
D_Null	-3.025	3.768	0.161	49301.716	306689.279	0.191
PP_e	-2.26	0.174	0.18	59760.298	332771.191	0.02
CCout_e	-2.312	0.44	0.179	58658.779	328225.224	0.038
CCin_e	-2.27	0.192	0.179	59645.148	332292.805	0.022
Dall_e	-2.727	1.995	0.152	49041.874	323039.17	0.196
Dout_e	-2.837	2.056	0.15	48220.264	321968.299	0.209
Din_e	-2.943	1.247	0.144	45446.915	315443.106	0.255
B_e	-1.997	-1.898	0.178	60692.941	341414.942	0.005
A_e	-2.807	2.825	0.153	48400.606	315815.952	0.206
H_e	-2.785	5.522	0.158	50303.051	317738.841	0.175
PP_cor	-0.351	-2.593	0.151	47666.33	315526.246	0.218
CCout_cor	-1.382	-1.764	0.164	53131.595	323341.216	0.129
CCin_cor	-0.348	-2.597	0.151	47591.824	315409.288	0.219
Dall_cor	-0.445	-2.522	0.138	43563.678	315619.064	0.285
Dout_cor	-1.175	-2.071	0.157	49352.984	314741.158	0.191
Din_cor	-0.1	-2.706	0.141	44935.702	319654.887	0.263
B_cor	-1.459	-1.473	0.17	55309.867	325804.192	0.093
A_cor	-0.145	-2.672	0.138	44263.336	320076.739	0.274
H_cor	-0.887	-2.434	0.135	41467.64	306437.978	0.32

Degrees of freedom: 315999 for the null model and 315998 for the residual model

Null deviance of all models: 60969.16

All models are significant, p-value is 0.000

Legend:

R^2 - deviance explained

$a,\ b$ - parameters in exponential glm $\hat{y}_{mErrB} = e^{a+b \cdot index}$

in percent of explained variance in values of $ErrB$ are very small, therefore, according to principle of parsimony (Crawley, 2007), the linear regression model is preferred.

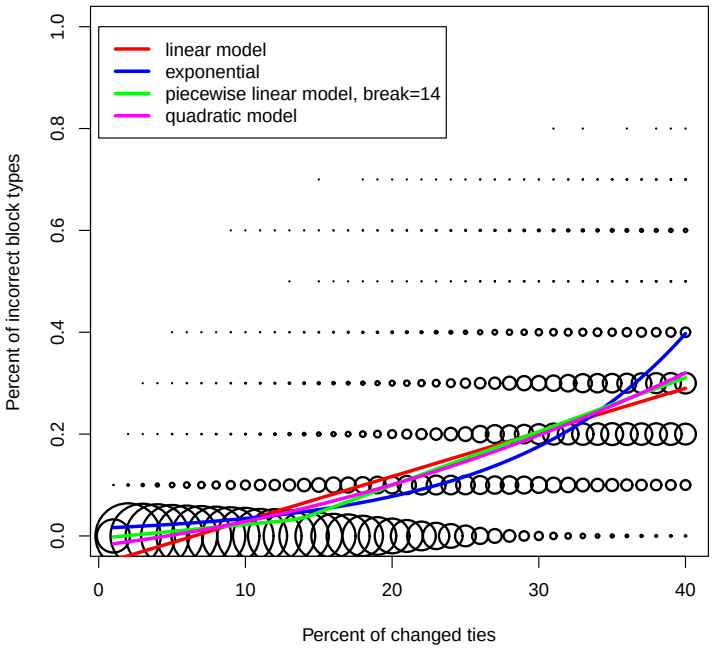


Figure 8.36: Impact of percent of changed ties on values of $ErrB$ with data for the second non-symmetric blockmodel structure

Table 8.23: Different fitted models for $mErrB$ with $p.changed$ ties as a predictor for data from the second non-symmetric blockmodel structure

Name of the model	Formula	R^2
Linear model	$\hat{y}_{ErrB} = -0.056 + 0.009 \cdot p.changed$	0.442
Exponential model	$\hat{y}_{ErrB} = e^{-4.186+0.082 \cdot p.changed}$	0.440
Piecewise linear models where		
break=14	$\hat{y}_{ErrB} = \begin{cases} -0.101 + 0.099 \cdot p.changed; & p.changed < 14 \\ -0.112 + 0.011 \cdot p.changed; & p.changed \geq 14 \end{cases}$	0.457
Quadratic model	$\hat{y}_{ErrB} = -0.019 + 0.004 \cdot p.changed + 0.0001 \cdot p.changed^2$	0.451

8.4 The impact of measured network characteristic on the stability of blockmodeling

In this section, before final conclusions, we expanded our studies of impact of relative differences in network characteristics on the stability of blockmodeling to investigation of predictive power of network characteristics of measured networks. When the network is sampled or measured, we in fact do not know what the hidden underlying structure is. According to Holland and Leinhardt (1973) the mathematical representation of a network as a graph or a sociomatrix is just approximation of true structure which is unobservable. Therefore, we decided to investigate the impact of network characteristics of measured network to the stability of blockmodeling.

Supposed that the whole network, before introduced random measurement error, presents the true underlying structure and the network with introduced errors is the measured network. In surveys the measured networks are collected. Therefore, the characteristic of the measured network will be used as predictors for two indices of blockmodeling stability, ARI and $ErrB$. In calculation of the Adjusted Rand index and proportion of incorrect blocks, the blockmodel or whole network (which presents the 'true' structure) is used as reference network for comparison. The simulation of random measurement errors is presented in Section 7.5.1.

The impact of characteristics of measured networks to stability of blockmodeling will be presented on two examples of real networks, the boy-girl liking ties network (8.4.1) and the note borrowing network (Section 8.4.2).

8.4.1 The impact of measured network characteristics on the stability of blockmodeling with data from the boy-girl liking ties network

Data were generated separately from the data of differences in network characteristics from previous sections. Therefore, the correlation coefficient between ARI, $ErrB$, and

p.changed differs a little from previously reported coefficients. For example, the Pearson coefficient between ARI and *p.changed* for the boy-girl liking ties network is in simulation of differences in network characteristics (Section 8.2.1) equal to $r = -0.773$. In new simulation, where measurement errors were again generated randomly, the Person correlation between these two indices is $r = -0.763$ (Table 8.24).

Table 8.24: Pearson correlation coefficients between indices of measured network characteristics and indices of stability of blockmodeling for the boy-girl liking ties network

		1	2	3	4	5	6	7	8
1	*ARI*		-0.825	-0.763	-0.448	0.535	0.253	-0.617	0.594
2	*ErrB*	-0.825		0.569	0.309	-0.421	-0.226	0.467	-0.434
3	*p.changed*	-0.763	0.569		0.713	-0.717	-0.252	0.857	-0.873
4	*Density*	-0.448	0.309	0.713		-0.292	0.339	0.631	-0.893
5	*Reciprocity*	0.535	-0.421	-0.717	-0.292		0.795	-0.921	0.688
6	*Mut*	0.253	-0.226	-0.252	0.339	0.795		-0.516	0.120
7	*Asymm*	-0.617	0.467	0.857	0.631	-0.921	-0.516		-0.912
8	*Null*	0.594	-0.434	-0.873	-0.893	0.688	0.120	-0.912	

*All correlation coefficients are significant at 0.001 significance level.

In this study, we investigated the impact of network density (*Density*), reciprocity of a network (*Reciprocity*), number of mutual (*Mut*), asymmetric (*Asymm*) and null dyads (*Null*) to the indices of network stability; the Adjusted Rand Index (ARI) and the percent of incorrectly identified blocktypes (*ErrB*).

The highest Pearson correlation coefficient is between ARI and the number of asymmetric dyads ($r = -0.617$), which indicates that higher number of asymmetric dyads leads to lower values of ARI and therefore lower stability of blockmodeling in terms of partitions. On the other hand, changes in number of mutual dyads have little positive effect ($r = 0.253$). The number of null dyads also have large effect, where higher number of null dyads indicates lower values of ARI ($r = 0.594$) and therefore lower network stability in terms of partitions. The correlation coefficient between ARI and reciprocity is equal to $r = 0.535$, which also indicates large effect (Table 8.24).

The next step was to establish multiple regression models with indices of network

characteristics, where number of changed ties (*p.changed*) is controlled. The summary of those models is presented in Table 8.25. The simple linear model with percent of changed ties (*p.changed*) as a predictor is significant (p-value=0.000) and is able to explain 58.2% of variation in values of ARI. If percent of changed ties increases for one changed tie, the value of ARI increases for 0.029. When *Density* is added to the model as a predictor (Model 2), the percent of explained variance increases for 1.9% ($R^2 = 60.1\%$). If the number of changed ties is held constant, the change in network density for 0.1 increases the ARI values for 0.1701. The overall model with reciprocity is significant (p-value in Model 3 is 0.000), but the coefficient for *reciprocity* is not significantly different from zero ($b = -0.030, p - value = 0.091$).

Table 8.25: Regression models for ARI with characteristics of measured networks as predictors with data for the boy-girl liking ties network

	Index added	Coefficients (p-value)			Model summary	
Model	in Model 1	(Intercept)	*p.changed*	index	R^2	p-value
1	/	1.291 (0.000)	-0.029 (0.000)	/	0.582	0.000
2	*Density*	0.703 (0.000)	-0.034 (0.000)	1.701 (0.000)	0.601	0.000
3	*Reciprocity*	1.359 (0.000)	-0.030 (0.000)	-0.096 (0.091)	0.583	0.000
4	*Mut*	1.135 (0.000)	-0.029 (0.000)	0.011 (0.000)	0.586	0.000
5	*Asymm*	1.185 (0.000)	-0.034 (0.000)	0.010 (0.000)	0.587	0.000
6	*Null*	2.071 (0.000)	-0.039 (0.000)	-0.025 (0.000)	0.604	0.000

The addition of number of mutual dyads or number of asymmetric dyads as predictor to the Model 1 increases the percent of explained variance just for 0.4% and 0.5%, respectively (Model 4 and Model 5 in Table 8.25). The reason for this are high correlations between both predictors, e.g. the Pearson correlation coefficient between *p.changed* and *Asymm* is 0.857, which indicates multicollinearity problem. A major change in percent of explained variance is obtained when number of null dyads is added in the model as a predictor (Model 6). If number of null dyads is held constant, than the change in percent of changed ties (*p.changed*) for 1, decreases the ARI values for 0.039. On the other hand, when the number of changed ties is held constant, the change in

number of null dyads for 1 decreases the ARI values for 0.025.

Pearson correlation coefficients between $ErrB$ and other indices are a little bit lower than corresponding indices with ARI (Table 8.24). This indicates that network characteristics have higher prediction power for stability of partitions (ARI) than number of incorrectly classified block types ($ErrB$). The highest correlation coefficient ($r = 0.467$) is between $ErrB$ and number of asymmetric dyads ($Asymm$), which indicates that higher number of asymmetric dyads leads to more unstable blockmodel in terms of block types (higher values of $ErrB$). Positive coefficient is also obtained with $Density$ ($r = 0.309$). Negative effect on values of $ErrB$ have reciprocity ($r = -0.421$), number of mutual dyads ($r = -0.226$), and number of null dyads ($r = -0.434$), which means that higher values of those indices lead to lower values of $ErrB$ and therefore more stable blockmodel in terms of correctly identified block types.

The simple linear regression model with percent of changed ties ($p.changed$) as a predictor is able to explain 32.4% of variation in $ErrB$ (Table 8.26). If the percent of changed ties increase for one, the values in $ErrB$ increases for 0.015.

Table 8.26: Regression models for $ErrB$ with characteristics of measured networks as predictors with data for the boy-girl liking ties network

	Index added	Coefficients (p-value)			Model summary	
Model	in Model 1	(Intercept)	*p.changed*	index	R^2	p-value
1	/	-0.151 (0.000)	0.015 (0.000)	/	0.324	0.000
2	*Density*	0.254 (0.000)	0.019 (0.000)	-1.175 (0.000)	0.343	0.000
3	*Reciprocity*	-0.103 (0.000)	0.014 (0.000)	-0.068 (0.170)	0.324	0.000
4	*Mut*	-0.006 (0.000)	0.014 (0.000)	-0.011 (0.000)	0.331	0.000
5	*Asymm*	-0.112 (0.000)	0.017 (0.000)	-0.004 (0.003)	0.325	0.000
6	*Null*	-0.616 (0.000)	0.021 (0.000)	0.015 (0.000)	0.340	0.000

When network density is added in the model as predictor (Model 2 in Table 8.26), the percent of explained variance in values of $ErrB$ increases for 1.9% ($R^2 = 34.3\%$). When

the number of changed ties in this model is held constant, the change in network density for 0.1 increases the values of *ErrB* for 0.1175. A higher density, when percent of changed ties is held constant, means that ties are more likely to be randomly added. This means that random addition of ties is more likely to preserve the blockmodel structure than random deletion of ties.

Similarly as in model for ARI, the coefficient in Model 3 for *Reciprocity* is not statistically different from zero.

Among indices of changes in dyad census, the highest change in percent of explained variance in values of *ErrB* is obtained with the use of number of null dyads as a predictor in multiple regression model. The obtained model (Model 6 in Table 8.26) is able to explain 34.0% of variation in *ErrB*, which is for 1.6% more than in simple model with *p.changed* ties as predictor. If the percent of changed ties is held constant, the increase of number of null dyads for one increases the value of *ErrB* for 0.015. This means that higher number of null dyads with the same number of changed ties leads to a little bit more unstable blockmodel in terms of block types (higher values of *ErrB*).

8.4.2 The impact of measured network characteristics on the stability of blockmodeling with data from the note borrowing network

The study of measured network characteristics on the stability of blockmodeling was also performed with data from the note borrowing network (Section 6.2.1.2).

The highest correlation coefficient is between ARI and percent of changed ties (r=-0.854). Among indices of network characteristics the highest absolute values of correlation coefficients are obtained with number of null dyads ($r = 0.768$), network density ($r = -0.750$), number of asymmetric dyads ($r = -0.684$), and number of mutual dyads ($r = -0.506$). All those indices indicate large linear effect to values of ARI. The correlation coefficient between *Reciprocity* and ARI values is not significant (Table 8.27).

The simple linear regression model with percent of changed ties (*p.changed*) as a pre-

Table 8.27: Pearson correlation coefficients between indices of measured network characteristics and indices of stability of blockmodels for the note borrowing network

		1	2	3	4	5	6	7	8
1	*ARI*		-0.729	-0.854	-0.750	0.001	-0.506	-0.684	0.768
2	*ErrB*	-0.729		0.705	0.635	0.009	0.434	0.572	-0.647
3	*p.changed*	-0.854	0.705		0.912	0.048	0.644	0.799	-0.921
4	*Density*	-0.750	0.635	0.912		0.219	0.807	0.768	-0.966
5	*Reciprocity*	0.001	0.009	0.048	0.219		0.746	-0.449	0.036
6	*Mut*	-0.506	0.434	0.644	0.807	0.746		0.241	-0.628
7	*Asymm*	-0.684	0.572	0.799	0.768	-0.449	0.241		-0.907
8	*Null*	0.768	-0.647	-0.921	-0.966	0.036	-0.628	-0.907	

*All correlation coefficients are significant at 0.001 significance level, except the correlation coefficient between *ARI* and *Reciprocity* (p-value=0.945) and between *ErrB* and *Reciprocity* (p-value=0.559).

dictor is able to explain 72.9% of variation in values of *ARI* (Table 8.28). If the network density is added as predictor in Model 1, the percent of explained variance in values of *ARI* increases just for 0.5%, but the obtained new model is significantly better from the simple linear one (p-value=0.000). If the percent of changed ties is held constant, then the increase of measured network density for 0.1 increases the values of *ARI* for 0.1098. This means that added ties (because *p.changed* is held constant) increase the stability of blockmodeling in terms of partitions (higher values of *ARI*).

Table 8.28: Regression models for *ARI* with characteristics of measured networks as predictors with data for the note borrowing network

	Index added	Coefficients (p-value)			Model summary	
Model	in Model 1	(Intercept)	*p.changed*	index	R^2	p-value
1	/	1.021 (0.000)	-0.029 (0.000)		0.729	0.000
2	*Density*	0.728 (0.000)	-0.034 (0.000)	1.098 (0.000)	0.734	0.000
3	*Reciprocity*	0.911 (0.000)	-0.029 (0.000)	0.261 (0.000)	0.731	0.000
4	*Mut*	0.938 (0.000)	-0.031 (0.000)	0.007 (0.000)	0.732	0.000
5	*Asymm*	1.027 (0.000)	-0.029 (0.000)	-0.0002 (0.812)	0.729	0.000
6	*Null*	1.319 (0.000)	-0.033 (0.000)	-0.005 (0.000)	0.731	0.000

The multiple regression model with added reciprocity to predictors (Model 3 in Table 8.28) is able to explain just 0.2% more variance in values of ARI than simple linear model. If the *p.changed* is held constant, then the change in reciprocity for 0.1 increases the ARI values for 0.026. This means that for selected percent of changed ties (*p.changed*) more symmetrical network leads to more stable blockmodel in terms of partitions.

In the multiple regression model with number of asymmetric dyads (*Asymm*) as a predictor, the coefficient at *Asymm* is not significantly different from zero (p-value=0.812).

Another two models with indices from dyad census, number of mutual dyads and number of null dyads in measured network are also able to explain just 0.3% and 0.2% more variation in values of ARI, respectively (Models 4 and 6 in Table 8.28). Reasons for that are high correlation coefficient between percent of changed ties and other indices of network characteristics presented in Table B.2. If the percent of changed ties in the Model 5 is held constant, then the increase of number of mutual dyads for 1, increases the ARI values for 0.007. This means that a higher number of mutual dyads, with selected percent of changed ties, leads to more stable blockmodel in values of ARI. On the other hand Model 6 shows that the increase of number of null dyads for 1 causes the drop of ARI values for 0.005, if the percent of changed ties is held constant. Therefore, for selected *p.changed* a higher number of null dyads leads to less stable blockmodel in terms of agreement between two partitions.

Among all indices, the highest correlation coefficient ($r = 0.705$) is between percent of incorrect block types in a blockmodel and percent of changed ties (Table 8.27). The simple linear regression model with percent of changed ties as a predictor (Model 1 in table 8.29) can explain 49.65% of variation in values of $ErrB$.

When density of a measured network is added in the model, the percent of explained variance increases for just 0.04%, but the coefficient at *Density* is not significant.

Table 8.29: Regression models for *ErrB* with characteristics of measured networks as predictors with data for the note borrowing network

	Index added	Coefficients (p-value)			Model summary	
Model	in Model 1	(Intercept)	*p.changed*	index	R^2	p-value
1	/	-0.056 (0.000)	0.008 (0.000)	/	0.4965	0.000
2	*Density*	-0.028 (0.108)	0.009 (0.000)	-0.104 (0.102)	0.4969	0.000
3	*Reciprocity*	-0.033 (0.002)	0.008 (0.028)	-0.054 (0.170)	0.4972	0.000
4	*Mut*	-0.043 (0.000)	0.009 (0.000)	-0.001 (0.024)	0.4972	0.000
5	*Asymm*	-0.070 (0.000)	0.0004 (0.000)	-0.004 (0.204)	0.4967	0.000
6	*Null*	-0.065 (0.012)	0.0001 (0.000)	0.015 (0.725)	0.4965	0.000

In regression models with added *Reciprocity*, number of asymmetric dyads (*Asymm*) or number of null dyads (*Null*), the coefficients for corresponding indices are not statistically significantly different from zero.

In multiple regression model with added number of mutual dyads (*Mut*) as a predictor, the percent of explained variance is higher for 0.07% compared to simple linear model (Model 1 in Table 8.29). If the number of changed ties is held constant, the increase of number of mutual dyads for one, causes the drop of *ErrB* values for 0.001. This means that higher number of mutual dyads in measured network for selected percent of changed ties decreases the *ErrB* values, and therefore the blockmodel is a little bit more stable.

If we compare the results for the boy-girl liking ties network (Section 8.4.1) and the note borrowing network, we can conclude that *Density* and number of mutual dyads (*Mut*) are the best predictors (when percent of changed ties is constant) for stability of blockmodeling. This conclusion is made just on two special examples of networks, therefore an extended simulation study with different blockmodel structures and larger networks should be performed.

8.5 Conclusions

In previous sections numerous linear regression models and their generalizations were established. We try to answer the question (the first research question on page 37), if the changes in network characteristics and properties are able to predict the stability of blockmodeling and to what extent. The results are far from simple.

First, we were tried to summarize the results from the Pearson correlation coefficients, and therefore also from linear regression models, for all starting networks used in simulations. Table 8.30 presents results for two real networks and three simulated blockmodel structures, which are in detailed presented in Sections 8.2 and 8.3. The Pearson correlation coefficients between indices of changes in network characteristics and properties and two indices of blockmodeling stability (ARI and $ErrB$) are presented with the following graphic signs. Correlation lower than 0.5 is marked with sign $-$, correlation between 0.5 and 0.7 are denoted with sign $\circ$, and correlation coefficients higher than 0.7 is presented with sign +.

Three obvious conclusions can be made. First, the best predictor of stability ob blockmodeling partitions (presented with ARI values) is percent of changed ties (*p.changed*). In the simple linear model the percent of changed ties can explain at least 50% of variation of ARI irrespective of blockmodel structure.

Second, all indices, which were used as predictors, have less power to predict percent of incorrectly identified block types ($ErrB$) compared to stability of partition (ARI) in two blockmodels.

Third, as explained in Section 8.3.1, indices of network characteristics have less power to predict stability of blockmodeling if the blockmodel structure is highly symmetric. In case of structural equivalence, the blockmodel is stable for relatively high percent of changed ties (Section 7.5.3.1). On the other hand, indices of network characteristics are able to perceive the small changes in network structure. One possible solution, which should be investigated further, is use of more universal models, which can take this

Table 8.30: Summary of predictive powers for linear regression models for values of *ARI* and *ErrB*

Blockmodel	Symmetric				Non-symmetric					
					Simulated					Simulated
	Real		Simulated		First		Real		Second	
Index	ARI	ErrB	ARI	ErrB	ARI	ErrB	ARI	ErrB	ARI	ErrB
p.changed	+	○	+	○	+	○	+	○	+	○
Dens	-	-	-	-	-	-	+	○	-	-
Rec	○	-	-	-	-	-	-	-	-	-
D_Mut	-	-	-	-	-	-	○	-	-	-
D_Asym	○	-	-		-	-	○	○	-	-
D_Null	○	-	-	-	-	-	○	○	○	-
PP_e	○	-	-	-	-	-	○	○	-	-
CCout_e	○	-	-	-	-	-	○	○	-	-
CCin_e	○	-	-	-	-	-	○	○	-	-
Dall_e	○	-	-	-	○	○	+	○	○	○
Dout_e	○	-	○	-	○	○	+	○	○	○
Din_e	○	-	○	-	○	○	+	○	+	○
B_e	-	-	-	-	-	-	-	-	-	-
A_e	-	-	-	-	○	-	+	○	○	○
H_e	-	-	-	-	○	○	+	○	○	-
PP_cor	-	-	-	-	○	○	○	○	○	○
CCout_cor	-	-	-	-	○	-	○	-	-	-
CCin_cor	-	-	-	-	○	○	○	○	○	○
Dall_cor	-	-	-	-	+	○	+	○	+	○
Dout_cor	-	-	-	-	○	○	○	-	○	○
Din_cor	-	-	-	-	+	○	+	○	+	○
B_cor	-	-	-	-	-	-	-	-	-	-
A_cor	-	-	-	-	+	○	+	○	+	○
H_cor	-	-	-	-	+	○	+	○	+	○

attribute into account.

Two types of models or solution were used with *p.changed* ties as a predictor; the two- (or more) piecewise regression model and generalized linear models with exponential dependency. In future work, these two types of models can be fitted to the data with indices of network characteristics as a predictor.

If we look closer to indices of network characteristics and their meaning, we can conclude that the best predictors are those indices which are calculated based on correlation between two vectors of network properties from the whole and the measured

network. One possible interpretation can be that the linear relationship between two vectors of network properties (which is measured by correlation coefficient) has higher impact to stability of blockmodeling than the magnitude (measured by distance) between two vectors. Another possible explanation why indices calculated with Euclidean distance perform worse as predictors is that unstandardized vectors were used in calculation of Euclidean distance between them. Faust and Romney (1985, 101) argued that "Distance as a measure of similarity applied to nonstandardized variables confounds information about the similarity in the patterns of values with information about the differences in the mean and variance of each variable".[26]

The best predictors of indices of blockmodeling stability are correlation coefficient between vectors of degree centrality based on all-degree and indegree, and correlation between vectors of authority and hub weights.

Probably better question than the first research question on page 37 is to investigate the impact of properties of measured network to the blockmodeling stability, because in real research studies are the whole networks which represent the unobservable underlying structure unknown. Therefore the differences between characteristics of the whole and measured networks can not be measured or calculated. The results presented in Section 8.4 suggest that network density and number of mutual dyads are the best predictors (when percent of changed ties is constant) for stability of blockmodeling. The conclusions unfortunately can not be generalized because of the small number of starting network in the simulations. Therefore, further work on characteristics of measured network and their impact on (indices of) network stability should be done with larger range of starting blockmodel structures.

[26]Faust and Romney (1985) use the square of Euclidean distance as a measure of structural equivalence, which was calculated for actors i and j as distance between row i and row j in sociomatrix. In spite of the fact that Euclidean distance in our study was calculated on different vectors, their suggestion should be taken into account.

9 Conclusions

After a short overview of the dissertation, an evaluation of blockmodeling stability on errors in research design is presented together with some recommendations. The chapter concludes with some ideas for further research.

9.1 A short overview

The generalized blockmodeling is a popular and widely used technique inside the social network analysis. Another fact is that networks are measured with errors and there was no adequate study about the impact of design errors on results of blockmodeling. The original contribution of my dissertation is therefore a systematic research of different types of errors on the stability of blockmodeling.

First, networks and their main characteristics, relations and generalized blockmodeling together with different types of equivalence are presented in Chapters 2 and 3.

The discussion about evaluation of network stability is presented in Chapter 4. The whole starting blockmodel and the measured blockmodel from network with introduced errors have to be compared. Because the result of using a blockmodeling procedure is a partition (of actors) determining positions and image matrix with selected block types, two indices for measuring the stability are needed. The first index, the Adjusted Rand Index, measures the agreement between both partitions, and the second index compares block types in image matrices and their positions and is calculated as the percent of incorrectly identified block types. The described indices can reach both levels of the blockmodel; stability of macro structure is estimated with proportion of

incorrectly identified block types, and the changes in the micro level of an actor are evaluated with the Adjusted Rand Index.

The extensive part of the dissertation is the review of the literature on the errors in the research design and their classification in Chapter 4. The boundary specification problem, errors caused by design and errors caused by actors are three main categories of design errors. A questionnaire can be a large source of errors, especially with specification of number of choices and recall method. The impact on the established blockmodel also has the direction of question where the perceptions of giving or receiving of social support can be gathered. An important source of errors could be also actors themselves. They could refuse to respond to the entire questionnaire or only on the particular tie(s). For actor (and tie) non-response different possible treatment are examined, such as the complete-case approach, reconstruction procedure and imputations. The measurement errors occur where there is a discrepancy between the true value of a concept and the observed (or measured) value of that concept. The definition of measurement error in the social network analysis is presented together with main sources.

The design of simulation studies together with networks used in the studies is presented in Chapter 6, while the results of evaluation of blockmodling stability are presented in Chapters 7 and 8. The main conclusions from both chapters are presented below.

9.2 Evaluation of stability of blockmodeling on design errors

The evaluation of blockmodeling stability is quite different regarding the type of an introduced error and type of selected equivalence in the generalized blockmodeling process.

The consequences of introduced design errors and their strength depend on the em-

ployed type of equivalence. First, we ascertained that the structural equivalence is more stable compared to other types of equivalence. This is an expected result according to the remark by Batagelj et al. (1992b, 67), who emphasized that "although the definition of structural equivalence is 'local' it has 'global' implications - structurally equivalent units behave in the same way also to all other units. A position is defined in terms of all other units in a network". If the actors are structurally equivalent, a locally changed tie has small impact on overall structure of the network. On the other hand, regularly equivalent actors have the same or similar connection patterns to the different neighbours. In that case a small amount of changed ties destroys the local structure of the network and therefore the clusters of equivalent units. Despite these theoretical expectancies, the level of instability of blockmodeling based on regular equivalence was quite surprising. The smallest change in the composition of network ties (one changed tie) completely destroys the established blockmodel on both, micro and macro level of the network. The destroyed position membership of an actor (completely changed partition) affects the micro level of the network, while the destroyed blockmodel structure with changed and reorganized block types destroys the macro level.

Different types of errors from the research design have different implications on the resulting blockmodeling. First, we emphasized the findings about limitation of number of choices instead of free choice design. The limitation of number of choices may destroy the blockmodel structure if the restriction is unrealistic or too far from the true number of desired nominations. As pointed out by Newman (2010), the fixed choice design is often selected only for practical purposes to reduce the work of the researcher. We would like to emphasize that this is not the right reason for selection of fixed choice questionnaire format which has high ability to destroy the underlying true structure of the network. From blockmodeling point of view the fixed number of choices should not be enforced. If there is a reasonable argumentation for use of fixed choice design, the limitations should not be set too strict. For established blockmodel it is better that questionnaire format forces the respondents to nominate more friends (than is the real number) than make them impossible to list all their friends. Therefore, the blockmodel-

ing is more instable to lower number of choices than to higher number of nominations according to the real underlying structure.

The impact of direction of question on the established blockmodel structure was studied with real networks. The main conclusion is that both results of the blockmodeling procedure, the position membership and the image matrix, depend on the method used for gathering social network data. Therefore, further research should establish if there is a common pattern in the blockmodels obtained with different questions formats in data collection process. The confirmation of ties from the 'original' network from the ties from the 'reversed' one could probably be used to find the most dense, stable and cohesive subgroups of a network.

The most extensive studies with blockmodeling based on structural equivalence were performed with actor non-response. The main conclusion is that the performance of the non-response data treatments in social networks depends on the symmetry of the networks. The symmetry of the network refers to reciprocity value and also to symmetry of the blockmodel structure. The best treatments for the symmetric networks are reconstruction and combination of reconstruction and mode imputations. For the non-symmetric network the best treatments are the imputations based on mode and the complete-case approach. However, the use of complete-case approach is not advisable, because we lose information about the location of actor(s) in a position, because non-respondents are deleted from the network. We also do not advise using the null tie imputation, because its performance is always the worst. Therefore, the simple recording of zeros instead of absent ties is the worst solution, although it is frequently used in network data collection process.

The tie non-response study is the continuation of the actor non-response simulations. The above conclusions that the selection of the best non-response treatment depends on the symmetry of the network, also hold true in that case. Additional conclusion is that imputations based on modes fares badly for core-periphery structures, while reconstruction works well for them.

The main conclusion from the simulation of random measurment error is that the blockmodeling based on structural equivalence is highly stable. The blockmodeling procedure is a little bit more stable in terms of blockmodel structure than in terms of position membership of actors. As described above, the blockmodels based on regular or generalized equivalence are extremely sensitive to the minor changes in network ties. One randomly changed tie could completely destroy both, position membership and the blockmodel structure. We should emphasize that with our results the guidelines made by Doreian et al. (2005) are even more important. In the generalized blockmodeling prior knowledge of the researcher should be incorporated in prespecifed blockmodels prior to blockmodeling analysis.

Another important question we try to answer in the dissertation is whether the relative changes in network characteristics and properties are able to predict the stability of blockmodeling and to what extent. The results are far from simple, but the following conclusions can still be drawn. The best predictor of stability of blockmodeling partitions is percent of changed ties. All indices, which were used as predictors, have more power to predict the position membership of the actors than percent of incorrectly identified block types. In real studies the real underlying structure (presented with whole networks in our simulations) is unknown, which makes the comparison between whole and measured network impossible. Therefore, the impact of properties of measured network (alone) to the blockmodeling stability was investigated with two real networks. The results (presented in Section 8.4) suggest that network density and number of mutual dyads are the best predictors (when percent of changed ties is held constant) for stability of blockmodeling. Due to the small number of starting network in the simulations the conclusions unfortunately can not be generalized and therefore further simulations are needed.

9.3 Guidelines for researchers

Part of guidelines about actor non-response and preferable non-response data treatments have already been published in Žnidaršič 2012. Based on the study and results given in the previous sections the following recommendations for the researchers can be given:

- **During research design**
  - **Correctly define network boundaries.**
    Exclusion of actors could changed the obtained blockmodel, but if the number of missing actors is low, the blockmodel could be correctly identified in terms of positions and image matrix (results on the complete-case approach in Section 7.3).
  - **Define the research question according to providing or receiving social support.**
    Networks and consecutively established blockmodels could be quite different because different concepts are measured (Section 7.2).
  - **Select free choice design instead of fixed choice design.**
    If limitation of number of actors is necessary, do not set the limit to strict (Section 7).
- **During data collection**
  - **Report the missing ties by coding them as such, for example by NA, in the matrix representation of the network.**
    Missing ties are too often recoded as 0 which is the worst solution when analyzing blockmodels (results on the null tie imputations in Section 7.3).
  - **Never replace absent ties with 0s,**
    because the null tie imputation treatment was the worst treatment regarding both micro (position membership) and macro level (block structure) of the network (results on the null tie imputations in Section 7.3).
  - **Identify the item and/or actor non-response. Report the percentage of actor and/or item non-response** together with the size of the network.

- **When choosing the type of blockmodel**
  - **Structural equivalence is very stable** up to 50% of non-respondents or 15% of random measurement errors (Sections 7.3 and 7.5).
  - **Requlars and generalized types of equivalence are extremely unstable**, because one changed tie could completely destroy the blockmodel structure (Sections 7.3.5, 7.5.4 and 7.5.5).
- **During data analysis (blockmodeling)**
  - **Estimate the reciprocity of the fully observed network** in order to decide about the best non-response treatment.
  - **If the reciprocity is lower than 0.5 the complete-case approach or imputations based on mode should be used.**
  - **If the reciprocity is higher than 0.5 the complete-case approach or one of the reconstruction treatment is suggested.**
  - **Do not use the complete-case approach if the aim of the study is to investigate the position of non-respondents of the network.**

9.4 Ideas for future research

During the work on this dissertation several existing questions were left unanswered or are just partially answered, and at the same time several new questions arose. The majority of them is pointed out in the conclusions of the individual sections. The most important open questions with outline of further research are presented here.

All simulations in this dissertation should be extended to larger networks. At the moment there is no (cmpletely) suitable solution in programs used in the simulations. The simulations were performed in R package called `blockmodeling` developed by Žiberna (2008) which is appropriate for generation of whole networks and introduction or errors but it is a little bit slow in running the blockmodeling procedure. Its deficiency is slow running of the blockmodeling algorithm with larger networks (and/or larger

number of clusters). One of the solutions available now is the testing version of package adopted by Žiberna called `testBlockmodelingTestC`.

Another option would be to link together Pajek (Batagelj and Mrvar, 2010a,b) and R in such a way that Pajek could be called from R for blockmodeling procedures, while the generation of whole networks, introduction of errors, and the storage of data would still run in R.

A broader set of different types of block patterns for structural equivalence should be used. We could start from well constructed artificial networks of different numbers and sizes of clusters with precisely determined block structure or patterns of block types in the image matrices. Similarly as for networks based on regular equivalence (presented in Section 6.2.4), the well known models such as the cohesive subgroups model and core-periphery model, could be used for initial simulations.

The existent simulations together with recommendations presented above should be extended to other equivalence types and other blockmodel structures. The generalized type of equivalence with its broad collection of suitable block types offers a broad set of possible combination of block types in the image matrix. Therefore, the impact of starting blockmodel structure in terms of generalized equivalence should be studied in more systematic way.

The extreme instability of blockmodeling based on regular equivalence (Section 7.5.4) demands special attention. First, extensive research on literature about theoretical fundamentals, definitions and properties should be examined. In addition, all usages of regular equivalence should be studied where probably also the re-establishment of the models will be necessary. Based on those findings the interpretation of instable results of blockmodeling together with some recommendations will hopefully be possible.

The treatments of the actor (and tie) non-response should be extended to more complex treatments. In ordinary surveys with missing data a great success is achieved with

the use of EM algorithm or multiple imputations. Both approaches should be adopted to the social networks data.

In case of tie non-response (short overview presented in Section 7.4) other than random missing mechanisms should be used. In both types of non-response in social networks, actor and tie non-response, characteristics of actors should be used in construction of missing data. The patterns in non-response should reflect the real situation from social data collection process.

Another important extension of present work could be study of errors with sign and valued networks. The definition of error in binary social network that an error is a missing or extra tie (Holland and Leinhardt, 1973), can be extended to valued networks, where error occurs when a wrong value (of strength tie) is recorded. Therefore, beside the amount of introduced errors also the expanse or magnitude of the change together with direction of the change should be controlled.

Bibliography

Adar, Eytan, and Christopher Ré. 2007. Managing uncertainty in social networks. *IEEE Data Engineering Bulletin* 30:23–31.

Batagelj, Vladimir. 1993. *Centrality in social networks*, vol. 9 of *Metodološki zvezki, Developments in statistics and methodology*, 129–138. FDV, Ljubljana.

———. 1997. Notes on blockmodeling. *Social Networks* 19:143–155.

Batagelj, Vladimir, Patrick Doreian, and Anuška Ferligoj. 1992a. An optimizational approach to regular equivalence. *Social Networks* 14(1-2):121 – 135.

Batagelj, Vladimir, Anuška Ferligoj, and Patrick Doreian. 1992b. Direct and indirect methods for structural equivalence. *Social Networks* 14(1-2):63 – 90.

Batagelj, Vladimir, and Andrej Mrvar. 2010a. *Pajek 1.28*. Available at http://pajek.imfm.si/doku.php?id=download (July 25, 2010).

———. 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 (July 25, 2010).

Batagelj, Vladimir, Andrej Mrvar, Anuška Ferligoj, and Patrick Doreian. 2004. Generalized blockmodeling with pajek. *Metodološki zvezki* 1:455–467.

Bell, David C., Benedetta Belli-McQueen, and Ali Haider. 2007. Partner naming and forgetting: Recall of network members. *Social Networks* 29(2):279 – 299.

Bernard, H. R., P. Killworth, D. B. Kronenfeld, and L. Sailer. 1984. The Problem of Informant Accuracy: The Validity of Retrospective Data. *Annual Review of Anthropology* 13:495–517.

Bernard, H. R., and Peter D. Killworth. 1977. Informant Accuracy in Social Network Data II. *Human Communication Research* 4(1):3–18.

Bernard, Russell H., Peter D. Killworth, and Lee Sailer. 1982. Informant accuracy in social-network data V. An experimental attempt to predict actual communication from recall data. *Social Science Research* 11(1):30–66.

Biemer, Paul P. 2010. *Overview of Design Issues: Total Survey Error*, 27–57. Handbook of Survey Research, Howard Hous, UK: Emerald Group Publishing Limited.

Biemer, Paul P., and Lars E. Lyberg. 2003. *Introduction to survey quality*. New Jersey, USA: John Willey & Sons.

Borgatti, Stephen P., Kathleen M. Carley, and David Krackhardt. 2006. On the robustness of centrality measures under conditions of imperfect data. *Social Networks* 28(2): 124 – 136.

Borgatti, Stephen P., Martin G. Everett, and Linton C. Freeman. 2002. *Ucinet for windows. version 6: Software for social network analysis*. Harvard, MA: Analytic Technologies. Available at http://www.analytictech.com/ucinet/ (October 10, 2010).

Brewer, Devon D. 2000. Forgetting in the recall-based elicitation of personal and social networks. *Social Networks* 22(1):29–43.

Brewer, Devon D., and Cynthia M. Webster. 2000. Forgetting of friends and its effects on measuring friendship networks. *Social Networks* 21(4):361 – 373.

Burt, Ronald S. 1987. A note on missing network data in the general social survey. *Social Networks* 9(1):63–73.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. 2nd ed. New Jersey: Lawrence Erlbaum Associates.

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

Crawley, Michael J. 2007. *The R Book*. John Wiley & Sons Ltd.

Dixon, John, and Clyde Tucker. 2010. *Survey Nonresponse*, 593–630. Handbook of Survey Research, Howard Hous, UK: Emerald Group Publishing Limited.

Doreian, Patrick. 2008. *Positional analysis and blockmodeling*. Encyclopedia of Complexity and Systems Science, New York, USA: Cambridge University Press.

Doreian, Patrick, Vladimir Batagelj, and Anuška Ferligoj. 1994. Partitioning networks based on generalized concepts of equivalence1. *Journal of Mathematical Sociology* 19(1):1–27.

———. 2005. *Generalized blockmodeling*. New York, NY, USA: Cambridge University Press.

Doreian, Patrick, and Katherine L. Woodard. 1994. Defining and locating cores and boundaries of social networks. *Social Networks* 16(4):267 – 293.

Erickson, Bonnie H., and T. A. Nosanchuk. 1983. Applied network sampling. *Social Networks* 5(4):367 – 382.

Erickson, Bonnie H., T. A. Nosanchuk, and Edward Lee. 1981. Network sampling in practice: Some second steps. *Social Networks* 3(2):127 – 136.

Faust, Katherine. 1988. Comparison of methods for positional analysis: Structural and general equivalences. *Social Networks* 10(4):313 – 341.

———. 2007. Very local structure in social networks. *Sociological Methodology* 1:209–256.

Faust, Katherine, and A. Kimball Romney. 1985. Does structure find structure?: A critique of burt's use of distance as a measure of structural equivalence. *Social Networks* 7(1):77 – 103.

Faust, Katherine, and Stanely Wasserman. 1992. Blockmodels: Interpretation and evaluation. *Social Networks* 14(1):5–61.

Feld, Scott L. 1991. Why your friends have more friends than you do. *The American Journal of Sociology* 96(6):464–1477.

Feld, Scott L., and William C. Carter. 2002. Detecting measurement bias in respondent reports of personal networks. *Social Networks* 24(4):365 – 383.

Ferligoj, Anuška. 1989. Razvrščanje v skupine, teorija in uporaba v družboslovju. *Metodološki zvezki* 4. In Slovene.

Ferligoj, Anuška, and Valentina Hlebec. 1998. Socialna opora dijakov Gimnazije Bežigrad [datoteka podatkov]. Ljubljana: Fakulteta za družbene vede, Center za metodologijo in informatiko [izdelava], 1998. Ljubljana: Arhiv družboslovnih podatkov [distribucija], 2007. In Slovene.

———. 1999. Evaluation of social network measurement instruments. *Social Networks* 21(2):111 – 130.

Ferligoj, Anuška, Karmen Leskošek, and Tina Kogoj. 1995. Zanesljivost in veljavnost merjenja. *Metodološki zvezki* 11. In Slovene.

Field, Andy. 2009. *Discovering Statistics Using SPSS*. SAGE Publications. 3rd edition.

Fine, Gary Alan. 1987. *With the Boys: Little League Baseball and Preadolescent Culture*. Chicago, USA: University of Chicago Press.

Freeman, Linton C. 1978-1979. Centrality in social networks: Conceptual clarification. *Social Networks* 1(3):215 – 239.

Freeman, Linton C., A. Kimball Romney, and Sue C. Freeman. 1987. Cognitive Structure and Informant Accuracy. *American Anthropologist* 89(2):310–325.

Friedl, Herwig. 2010. Generalized linear models and extensions - theory and examples. Slides.

Granovetter, Mark. 1976. Network Sampling: Some First Steps. *The American Journal of Sociology* 81(6):1287–1303.

Groves, Robert M. 2004. *Survey errors and survey costs*. New Jersey, USA: John Willey & Sons.

Hammer, Muriel. 1985. Implications of behavioral and cognitive reciprocity in social network data. *Social Networks* 7(2):189 – 201.

Handcock, Mark S., and Krista Gile. 2007. Modeling social networks with sampled or missing data. working paper no. 75. Tech. Rep., Center for Statistics and the Social Sciences, University of Washington, Seattle. Available at http://www.csss.washington.edu/Papers/wp75.pdf (April 27, 2011).

Hlebec, Valentina. 1992. Merjenje v analizi omrežij. Diplomska naloga. FDV, Ljubljana. In Slovene.

———. 1993. *Recall versus recognition: Comparison of the two alternative procedures for collecting social network data*, vol. 9 of *Metodološki zvezki, Developments in statistics and methodology*, 121–128. Faculty of Social Sciences, Ljubljana.

———. 1999. Evaluation of survey measurement instruments for measuring social networks. Ph.D. thesis, Faculty of social sciences, University of Ljubljana.

———. 2001. Meta-analiza zanesljivosti anketnega merjenja socialne opore v popolnih omrežjih. *Teorija in praksa* 38:63–76. In Slovene.

Hlebec, Valentina, and Anuška Ferligoj. 2001. Respondent mood and the instability of survey network measurements. *Social Networks* 23(2):125 – 140.

———. 2002. Reliability of social network measurement instruments. *Field method* 14: 288–306.

Hlebec, Valentina, and Tina Kogovšek. 2006. *Merjenje socialnih omrežij*. Študentska založba, Ljubljana. In Slovene.

Holland, Paul W., and Samuel Leinhardt. 1970. A method for detecting structure in sociometric data. *The American Journal of Sociology* 76(3):492–513.

———. 1973. The structural implications of measurement error in sociometry. *The Journal of Mathematical Sociology* 3(1):85–11.

Hubert, Lawrence, and Phipps Arabie. 1985. Comparing Partitions. *Journal of Classification* 2:193–218.

Huisman, Mark. 2009. Imputation of missing network data: Some simple procedures. *Journal of Social Structure* 10(1). Available at:

http://www.cmu.edu/joss/content/articles/volume10/ huisman.pdf (April 10, 2010).

Huisman, Mark, and Christian Steglich. 2008. Treatment of non-response in longitudinal network studies. *Social Networks* 30(4):297 – 308.

Kang, Soong Moon. 2007. Equicentrality and network centralization: A micro-macro linkage. *Social Networks* 29(4):585 – 601.

Killworth, Peter D., and Russell H. Bernard. 1979 - 1980. Informant accuracy in social network data III: A comparison of triadic structure in behavioral and cognitive data,. *Social Networks* 2(1):19–46.

Kleinberg, Jon M. 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, ed. Howard Karloff. SIAM/ACM-SIGACT. Available at: http://www.cs.cornell.edu/home/kleinber/auth.pdf (January 14, 2011).

Knoke, David, and James H. Kuklinski. 1982. *Networks analysis*. Los Angeles, USA: Sage Publications.

Knoke, David, and Song Yang. 2008. *Social networks analysis*. 2nd ed. Los Angeles, USA: Sage Publications.

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

Košmelj, Katarina, and Damijana Kastelec. 2003. Analiza variance in regresija. Delovno gradivo 2003/04 za podiplomski študij. Available at http://www.bf.uni-lj.si/fileadmin/groups/2721/MSc_študiji/UPORABNA_BIOSTATISTIKA/ regresija_prvi_del.pdf (December 15, 2010). In Slovene.

Krackhardt, David. 1987. Cognitive social structures. *Social Networks* 9(2):109 – 134.

Krosnick, Jon A., and Stanley Presser. 2010. *Question and Guestionnaire Design*, 263–313. Handbook of Survey Research, Howard Hous, UK: Emerald Group Publishing Limited.

Laumann, E. O., P. V. Marsden, and D. Prensky. 1989. The boundary specification problem in network analysis. In *Research Methods in Social Network Analysis*, ed. L. C. Freeman, D. White, and A. K. Romney, 62–87. Fairfax, VA: George Mason University Press. Reprinted from Burt, R. S. and Minor, M. J. 1983. Applied Network Analysis: A Methodological Introduction, pg 18-34.

Lee, Sang Hoon, Pan-Jun Kim, and Hawoong Jeong. 2006. Statistical properties of sampled networks. *Physical Review E* 73(1):016102.

Lin, Nan. 1976. *Foundations in Social Research*. McGraw-Hill.

Lorrain, Francoise, and Harrison C. White. 1971. Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology* 1:49–80.

Manfreda, Katja Lozar, Vasja Vehovar, and Valentina Hlebec. 2004. Collecting ego-centred network data via the web. *Metodološki zvezki* 1:295–321.

Marsden, Peter V. 1990. Network data and measurement. *Annual Review of Sociology* 26:435–463.

———. 2005. *Recent developments in network measurement*, 8–30. Models and Methods In Social Network Analysis, New York, USA: Cambridge University Press.

———. 2011. *Survey methods for network data*, 370–388. The SAGE Handbook of Social Network Analysis, Thousand Oaks: Cambridge Sage Publications.

McCullagh, Peter, and John Nelder. 1989. *Generalized linear models*. 2nd ed. Boca Raton: Chapman and Hall/CRC.

Mittlbock, Martina andand Harald Heinzl. 2004. Pseudo R-squared measures for generalized linear models. 1st European Workshop on the Assessment of Diagnostic Performance, Milan, Italy. Availabe at: http://www.tech.plym.ac.uk/spmc/pdf/EWADP2004/EWADP2004.PAPER06.P71-80.MITTLBOCK.pdf (January 10, 2011).

Newman, M. E. J. 2003. Ego-centered networks and the ripple effect. *Social Networks* 25(1):83 – 95.

———. 2010. *Networks. An Introduction*. New York, NY, USA: Oxford University Press Inc.

de Nooy, Wouter, Andrej Mrvar, and Vladimir Batagelj. 2005. *Exploratory Social Network Analysis with Pajek (Structural Analysis in the Social Sciences)*. Cambridge University Press.

Piazza, Thomas. 2010. *Fundamentals of Applied Sampling*, 140–168. Handbook of Survey Research, Howard Hous, UK: Emerald Group Publishing Limited.

Robins, Garry, Philippa Pattison, and Jodie Woolcock. 2004. Missing data in networks: exponential random graph (p*) models for networks with non-respondents. *Social Networks* 26(3):257 – 283.

Rubin, Donald B. 1976. Inference and missing data. *Biometrika* 63(3):581–592. `http://biomet.oxfordjournals.org/content/63/3/581.full.pdf+html`.

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

Santos, Jorge M., and Mark Embrechts. 2009. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. In *ICANN '09: Proceedings of the 19th International Conference on Artificial Neural Networks*, 175–184. Berlin, Heidelberg: Springer-Verlag.

Saporta, Gilbert, and Genane Youness. 2002. Comparing two partitions: Some proposals and experiments. In *Proceedings in computational statistics*, ed. W. Hardle and B. Ronz, 243–248. Physica Verlag Berlin.

Schafer, Joseph L., and John W. Graham. 2002. Missing Data: Our View of the State of the Art. *Psychological Methods* 7(2):147 – 177.

Scott, John. 2000. *Social Network Analysis. A Handbook.* 2nd ed. SAGE Publicationd Ltd.

Steinley, Douglas. 2004. Properties of the Hubert-Arabie Adjusted Rand Index. *Psychological Methods* 9(3):386 – 396.

Stork, Diana, and William D. Richards. 1992. Nonrespondents in Communication Network Studies: Problems and Possibilities. *Group and Organization Management* 17: 193–209.

Tourangeau, Roger, and Norman M. Bradburn. 2010. *The Psychology of Survey Response*, 315–346. Handbook of Survey Research, Howard Hous, UK: Emerald Group Publishing Limited.

Vehovar, Vasja, Katja Lozar Manfreda, Gašper Koren, and Valentina Hlebec. 2008. Measuring ego-centered social networks on the web: Questionnaire design issues. *Social Networks* 30(3):213 – 222.

Vinh, Nguyen Xuan, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Icml '09: Proceedings of the 26th annual international conference on machine learning*, 1073–1080. New York, NY, USA: ACM.

Viswanathan, Madhu. 2005. *Measurement error and research design*. Thousand Oaks, California, USA: Sage publications, Inc.

Warrens, Matthijs. 2008. On the Equivalence of Cohen's Kappa and the Hubert-Arabie Adjusted Rand Index. *Journal of Classification* 25:177–183.

Wasserman, Stanley, and Katherine Faust. 1998. *Social Network Analysis: Methods and Applications*. 2nd ed. Cambridge, USA: Cambridge University Press.

White, Douglas R., and Karl P. Reitz. 1983. Graph and semigroup homomorphisms on networks of relations. *Social Networks* 5(2):193 – 234.

Wright, Eric R., and Bernice A. Pescosolido. 2002. "Sorry, I Forgot": : The Role of Recall Error in Longitudinal Personal Network Studies. *Social Networks and Health* 8: 113–129.

Yeung, Ka Yee, and Walter L. Ruzzo. 2001. An empirical study on principal component analysis for clustering gene expression data. *Bioinformatics* 17:763–774.

Zemljič, Barbara, and Valentina Hlebec. 2001. Zanesljivost mer središčnosti in pomembnosti v socialnih omrežjih. *Družboslovne razprave* 17(37-38):73 – 88.

———. 2005. Reliability of measures of centrality and prominence. *Social Networks* 27(1):73 – 88.

Žiberna, Aleš. 2007. Generalized blockmodeling of valued networks. Ph.D. thesis, Faculty of social sciences, University of Ljubljana.

———. 2008. *Blockmodeling 0.1.7: An r package for generalized and classical blockmodeling of valued networks*. Available at http://www2.arnes.si/ aziber4/ (January 5, 2010).

Žnidaršič, Anja, Patrick Doreian, and Anuška Ferligoj. 2011a. Tie non-response in social networks, treatments of tie non-response, and blockmodeling outcomes. *Submitted*.

Žnidaršič, Anja, Anuška Ferligoj, and Patrick Doreian. 2011b. The Impact of Non-Response Treatments on the Stability of Blockmodels. *58th Congres of International Statistical Institute*.

———. 2012. Non-response in social networks: The impact of different non-response treatments on the stability of blockmodels. *Social Networks*. Doi: 10.1016/j.socnet.2012.02.002.

Index of Authors

Subject Index

Terms *blockmodeling*, *blockmodel*, *image matrix*, partition, *Adjusted Rand Index*, and *proportion of incorrect block types* are used through the whole dissertation and are marked only in their first definition and few other remarkable places.

10 Stabilnost bločnega modeliranja (razširjen povzetek)

Namen doktorske disertacije je raziskati, kako stabilni so postavljeni bločni modeli oziroma bločno modeliranje na različne tipe napak v zasnovi raziskave. Različni avtorji so preučevali vpliv posameznih napak na karakteristike omrežij (npr. gostota omrežja, tranzitivnost ...), vendar pa do sedaj še ni bil raziskan njihov vpliv na posplošeno bločno modeliranje (Doreian in drugi, 2005).

V disertaciji so tako predstavljeni osnovni pojmi analize omrežij in posplošenega bločnega modeliranja. Posebej sta predstavljena kazalnika, s katerima lahko primerjamo dva bločna modela. Na podlagi obširnega pregleda literature s področja napak v zasnovi raziskave smo sestavili shemo napak ter predstavili najpomembnejše izsledke predhodnih raziskav. Načrt raziskave oziroma simulacij smo dopolnili s pregledom uporabljenih omrežij. Stabilnost bločnih modelov je ocenjena na podlagi dveh vrst omrežij, in sicer (realnih) omrežij iz literature in simuliranih omrežij. Na koncu smo poskušali še oceniti, v kolikšni meri lahko spremembe v karakteristikah omrežja napovedo rezultat bločnega modeliranja.

10.1 Omrežja

Socialno omrežje je sestavljeno iz končne množice enot (ali akterjev) in relacije (ali relacij) med njimi (Wasserman in Faust, 1998). Akterji v omrežju lahko predstavljajo posameznike ali pa skupine enot, kot so formalne ali neformalne organizacije. Relacija je v splošnem definirana kot posebna vrsta kontakta oziroma povezave med parom

akterjev (Knoke in Yang, 2008, 7).

Omrežja lahko predstavimo z grafom, kjer točke predstavljajo akterje, povezave pa rišemo s puščicami oziroma daljicami ter tako ločimo med usmerjenimi in neusmerjenimi povezavami. Poleg grafa lahko omrežje predstavimo tudi z matriko z n vrsticami in stolpci, kjer element r_{ij} predstavlja obstoj in/ali moč povezave med akterjema i in j.

Znanih je več vrst omrežij. Glede na število različnih množic enot v omrežju delimo le-ta na enovrstna, dvovrstna in večvrstna omrežja. Druga delitev se nanaša na uporabljeno mersko lestvico za merjenje relacij, kjer ločimo binarna omrežja (povezava obstaja ali ne), predznačena omrežja (povezava med enotami ima pozitivno ali negativno vrednost) in omrežja z vrednostmi na povezavah (vrednosti na povezavah so vsaj intervalnega tipa). Omrežja delimo še na popolna, kjer opazujemo relacije vsake enote z vsemi ostalimi enotami v omrežju, in egocentrična oziroma osebna omrežja, kjer opazujemo izbrane enote (ege) in njegove povezave do drugih članov omrežja (alterjev).

V disertaciji smo se osredotočili na popolna enovrstna binarna omrežja. Vsa v disertaciji uporabljena omrežja lahko označimo kot majhna (nekaj 10 enot). Razlog za to je računska zahtevnost implementiranih algoritmov bločnega modeliranja v programski paket `blockmodeling` (Žiberna, 2008) v R-u.

Cilj analize socialnih omrežij je iz surovih relacijskih podatkov dobiti uporaben in enostaven opis sistema razmerij (Stork in Richards, 1992). Uporabna in popularna tehnika za iskanje strukturnih vzorcev je posplošeno bločno modeliranje.

10.1.1 Lastnosti omrežij

V disertaciji smo na kratko predstavili lastnosti oz. karakteristike omrežij, ki smo jih uporabili za preučevanje oziroma napovedovanje stabilnosti bločnega modela. Karakteristike omrežja so lahko podane za omrežje kot celoto (npr. gostota), lahko pa so izračunane za vsakega akterja posebej (npr. mere središčnosti).

Gostota omrežja je definirana kot vsota povezav v omrežju deljeno s številom vseh možnih povezav v omrežju. Gostota tako opisuje splošni nivo povezanosti med akterji v omrežju (Scott, 2000, 93).

Naslednji pristop proučevanja povezanosti je raziskovanje parov akterjev oz. diad (Holland in Leinhardt, 1970; Wasserman in Faust, 1998). Ločimo asimetrične diade, kjer obstaja ena izmed povezav med akterjema i in j ($i \rightarrow j$ ali $j \rightarrow i$), ne pa obe hkrati. Vzajemna diada obstaja, kadar obstajata tako povezava med akterjema i in j kot tudi obratna povezava (v takem primeru rečemo, da je med akterjema neusmerjena povezava). Prazno diado imamo v primeru, ko akterja nista povezana.

S pomočjo diad se definira recipročnost (Huisman, 2009), ki meri simetričnost omrežja ter se izračuna kot dvakratnik vzajemnih diad deljeno z vsoto dvakratnika vzajemnih ter asimetičnih diad.

Naslednja skupina mer so mere središčnosti (oz. centralnosti) in pomembnosti, s katerimi iščemo najbolj 'središčne' akterje v omrežju. Najpomembnejša delitev mer središčnosti se nanaša glede na tip relacije v omrežjih (Batagelj, 1993), in sicer govorimo v primeru usmerjenih povezav o merah pomembnosti, pri neusmerjenih omrežjih pa o merah središčnosti.

Prva skupina so mere središčnosti in pomembnosti glede na stopnjo, kjer je akter najbolj središčen v omrežju, če ima največ povezav do ostalih akterjev (Wasserman in Faust, 1998). Relativna mera središčnosti glede na stopnjo je tako za posameznega akterja definirana kot število povezav do ostalih akterjev v omrežju deljeno z $n-$, kjer je n velikost omrežja. V primeru usmerjenih omrežij lahko koncept središčnosti glede na stopnjo razširimo, in sicer merimo vpliv akterja, če upoštevamo le izhodne povezave. V primeru upoštevanja vhodnih povezav pa merimo podporo akterju.

Pri izračunu mere središčnosti glede na dostopnost se upoštevajo razdalje posame-

znega akterja do vseh ostalih akterjev. Tako so najbolj središčni tisti akterji v omrežju, ki so blizu vsem ostalim akterjem (Freeman 197989; Wasserman in Faust, 1998). Prednost mere je, da upošteva tudi posredne sosede akterja, hkrati pa to pomeni, da jo lahko izračunamo le za krepko povezana omrežja. V primeru usmerjenih omrežij lahko izračunamo dostopnost glede na izhodne povezave, torej, kako blizu so vsi ostali akterji izbranemu akterju. V primeru vhodnih povezav pa lahko dostopnost interpretiramo kot bližino izbranega akterja do vseh ostalih.

Ideja mere središčnosti glede na vmesnost je, da je akter središčen, če leži na veliko najkrajših poteh med ostalimi akterji v omrežju (Wasserman in Faust, 1998).

Bližino izbranega akterja izračunamo kot delež akterjev v območju vpliva posameznega akterja deljeno s povprečno oddaljenostjo tega akterja od vseh drugih akterjev v omrežju. Pri tem je območje vpliva posameznega akterja enako številu ali deležu vseh akterjev, ki so s potjo povezani z izbranim akterjem (Freeman 197989; Wasserman in Faust, 1998; de Nooy, 2005).

Zadnji dve uporabljeni meri sta opisi in kazala. Akter je dobro kazalo, če kaže na veliko dobrih opisov ter je dober opis, če nanj kaže veliko dobrih kazal.

10.2 Bločno modeliranje

Namen bločnega modeliranja je razvrstitev akterjev omrežja v skupine (pozicije) in hkrati razvrstitev povezav v bloke, ki so določeni s povezavami med enotami v skupini (Wasserman in Faust, 1998; Doreian in drugi, 2005). Rezultat bločnega modeliranja je kompaktna predstavitev omrežja, torej model, ki predstavlja bistveno (poenostavljeno) strukturo omrežja, ki jo lahko tudi enostavneje interpretiramo. Bločni model lahko predstavimo na dva načina: z bločnim grafom ali z bločno matriko. Enote v tej poenostavljeni strukturi so skupine oziroma pozicije enakovrednih akterjev, medtem ko povezave predstavljajo odnose med skupinami.

Akterji znotraj skupine (in med skupinami) imajo enake oziroma zelo podobne vzorce povezav glede na izbrano enakovrednost. Najbolj znani sta strukturna in regularna enakovrednost:

i) Akterja sta strukturno enakovredna, če sta povezana do ostalih akterjev v omrežju na enak način (Lorrain in White, 1971).

ii) Akterja sta regularno enakovredna, če sta povezana z ostalimi enakovrednimi akterji na enak način (White in Reitz, 1983; Doreian in drugi, 2005).

Batagelj in drugi (1992b) so dokazali, da so za strukturno enakovrednost možni le štirje idealni bloki: prazni, prazni blok s povezavami na diagonali, polni ter polni blok z ničlami na diagonali. Izven diagonale sta možna le polni in prazni blok.

Za regularno enakovrednost obstajata le dva idealna bloka, in sicer prazni in regularni (vsaj ena 1 je v vsaki vrstici in vsakem stolpcu bloka) blok (Batagelj in drugi, 1992a).

Koncept posplošene enakovrednosti so prvič predstavili Doreian in drugi leta 1994. Posplošena enakovrednost tako dovoli tudi druge vzorce povezav in jo v bistvu lahko definiramo z množico dovoljenih blokov. Poleg praznega, polnega in regularnega bloka so tako dovoljeni še vrstično-dominantni, stolpično dominantni, vrstično-regularni, stolpično-regularni, vrstično-funkcijski in stolpično funkcijski blok (predstavljeni v Tabeli 3.2 na strani 31).

V bločnem modeliranju obstajata dva glavna pristopa: posredni in neposredni pristop (Batagelj in drugi, 1992b). V posrednem pristopu se izračuna matrika različnosti med akterji in tako problem modeliranja prevedemo na problem standardne analize podatkov (npr. razvrščanja). Pri neposrednem pristopu z optimizacijskim algoritmom iščemo najboljšo razvrstitev z najmanjšo vrednostjo kriterijske funkcije, ki jo določimo na podlagi izbrane enakovrednosti.

Posplošeno bločno modeliranje so obširno predstavili Doreian in drugi (2005) v knjigi Generalized blockmodeling. Njegove tri glavne karakteristike in prednosti pred posrednim pristopom so (Doreian in drugi 2005, 25−26):

i) v direktnem pristopu so uporabljeni osnovni omrežni podatki;

ii) uporabljen je širši nabor dovoljenih blokov;

iii) možna je tudi opredelitev položaja blokov v modelu, kar omogoča vključitev raziskovalčevega znanja v model pred samim bločnim modeliranjem.

Kriterijska funkcija je definirana kot seštevek odstopanj med idealnimi in empiričnimi bloki v bločnem modelu. Za vsak empirični blok izračunamo odstopanja od vseh dovoljenih idealnih blokov. V bločnem modelu je tako empirični blok predstavljen z idealnim blokom, ki najmanj odstopa od empiričnega. Vrednosti odstopanj posameznih blokov se seštejejo in dobljeni seštevek predstavlja odstopanje omrežja od bločnega modela. Kriterijsko funkcijo uporabimo v optimizacijskem algoritmu, kjer za več začetnih razvrstitev iščemo najboljšo rešitev, torej tisto razvrstitev z minimalno vrednostjo kriterijske funkcije.

Posplošeno bločno modeliranje je vključeno v program Pajek (Batagelj in Mrvar, 2010a, b) ter v R-paket `blockmodeling` (Žiberna, 2008), ki smo ju uporabljali pri naših simulacijah, ter v nekatere druge pakete (npr. UCINET (Borgatti in drugi, 2002)).

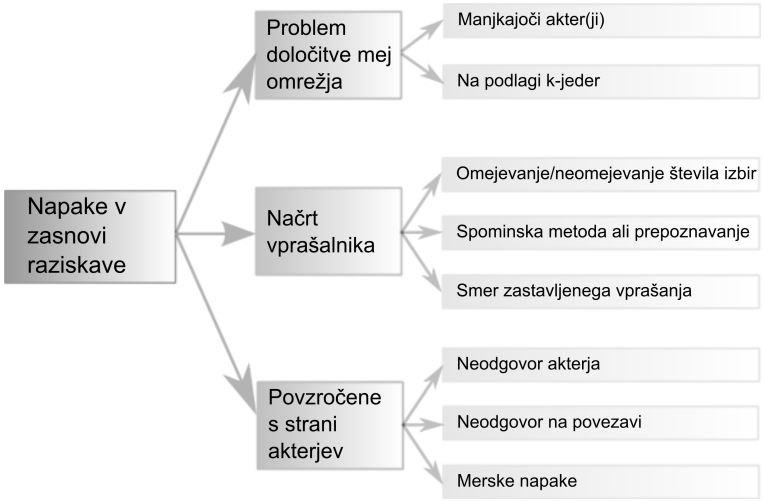
V skladu s predstavljenimi enakovrednostmi, kjer je glavna prednost posplošene enakovrednosti njena prilagodljivost, in opombo Batagelja in drugih (1992b), da čeprav je definicija strukturne ekviavalentnosti lokalna, ima le-ta globalne učinke, smo postavili prvo tezo: *Strukturna enakovrednost daje bolj stabilne rezultate kot regularna (ali drugi posplošeni tipi) enakovrednosti.*

Poleg ocene stabilnosti različnih enakovrednosti nas je zanimala še napovedna moč različnih karakteristik omrežja na stabilnost dobljenih bločnih modelov. Tako smo postavili prvo raziskovalno vprašanje: *V kakšnem obsegu so (relativne) razlike v karakteristikah omrežja (npr. gostota omrežja, recipročnost, število različnih tipov diad) ter korelacije in/ali Evklidske razdalje med vektorjema z lastnostmi akterjev (npr. mere centralnosti) sposobne napovedati rezultat bločnega modeliranja (stabilnost, razvrstitev in tip bločnega modela).*

10.3 Napake v zasnovi raziskave

Najpogostejše tehnike zbiranja podatkov (če izvzamemo arhivske podatke) so ankete in vprašalniki (Marsden, 2005; Wasserman in Faust, 1998). Vsaka metoda lahko vpliva na prisotnost napak. Njihov vpliv je potrebno preiskati na dva načina: potrebno je ugotoviti, kako zmanjšati prisotnost posameznih napak ter določiti vpliv napak na rezultate, ki jih dobimo z analizo omrežij. V disertaciji smo se osredotočili na drug problem, in sicer smo poskušali določiti vpliv različnih napak v zasnovi raziskave na stabilnost bločnih modelov. Tako smo postavili drugo raziskovalno vprašanje: *Kako stabilno je bločno modeliranje na različne količine in tipe napak?*

Na podlagi pregleda literature smo napake v zasnovi raziskave najprej razdelili v tri skupine (Slika 10.1): problem določitve mej omrežja, napake iz načrta vprašalnika ter na napake, povzročene s strani akterjev.



Slika 10.1: Shema napak v zasnovi raziskave

10.3.1 Problem določitve mej omrežja

Problem določitve mej omrežja se nanaša na pravila za vključevanje akterjev v preučevano omrežje (Laumann in drugi, 1989). Najbolj znana sta dva pristopa: realističen pristop, kjer akterji v omrežju sami določijo meje le-tega oziroma njihovo skupno pripadnost omrežju, in nominalističen pristop, kjer meje omrežja določijo raziskovalci

glede na nek kriterij, ki se lahko nanaša na akterja, relacijo ali aktivnost (oziroma kombinacijo teh treh faktorjev).

Doreian in Woodard (1994) sta ugotovila, da je tveganje za napačno določene meje še posebej veliko pri analizah, ki upoštevajo oziroma definirajo položaj akterja glede na vzorce povezav do vseh ostalih akterjev v omrežju. Predlagala sta uporabo *k*-jeder za določitev mej omrežja.

Napačno določene meje omrežja se lahko pokažejo v eni izmed treh oblik: vključitev akterjev, ki ne spadajo v omrežje, izključitev akterjev, ki spadajo v omrežje, ter kombinacija neprave vključitve in izključitve akterjev.

10.3.2 Napake v zasnovi vprašalnika

Druga skupina napak se nanaša na zasnovo vprašalnika, razdelili pa smo jo na tri podskupine: omejevanje oziroma neomejevanje števila izbir, spominska metoda ali prepoznavanje ter smer zastavljenega vprašanja.

Vprašalnik za zbiranje omrežnih podatkov lahko vključuje navodila o zahtevanem številu izbir, torej številu akterjev, ki jih je potrebno imenovati. Pri omejevanju števila izbir se lahko zgodi ena izmed treh možnosti (Holland in Leinhardt, 1973):

i) prava struktura je enaka zbranim omrežnim podatkom v sociogramu;

ii) v sociogramu je predstavljena le podmnožica prave strukture;

iii) prava struktura je podmnožica povezav v sociogramu.

Pomembno je poudariti, da neomejevanje števila izbir v vprašalniku sicer zagotavlja bogatejše podatke, ne pomeni pa, da ni prisotnih nobenih napak. Le-te se lahko nanašajo na različno razumevanje v vprašalniku uporabljenih pojmov, grafično oblikovanje vprašalnika in tako naprej.

Druga podskupina napak iz zasnove vprašalnika zajema napake, ki nastanejo pri priklicu akterjev (iz spomina) oziroma prepoznavanju (s seznama). Število prepoznanih

akterjev je navadno večje kot število priklicanih po spominu, hkrati pa uporaba seznama poenostavi poročanje akterjev.

Pri velikem številu relacij, ki se uporabljajo pri zbiranju omrežnih podatkov, je pomembna smer le-te (Stork in Richards, 1992; Ferligoj in Hlebec, 1999). Tako lahko akterje sprašujemo o prejemanju oziroma nudenju socialne opore.

10.3.3 Napake, povzročene s strani akterjev

Tretjo skupino napak v zasnovi raziskave sestavljajo napake, povzročene s strani akterjev: neodgovor akterja, neodgovor na povezavi in merske napake.

V primeru, da imamo v omrežju z n akterji m akterjev, ki zavrnejo sodelovanje, je stopnja odgovorov akterjev (in stopnja odgovorov na relaciji) enaka $1 - \frac{m}{n}$ (Knoke and Yang, 2008). Podatki med respondenti so popolni, medtem ko so podatki med respondenti in nerespondenti le delni in jih lahko uporabimo za nadomeščanje manjkajočih vrednosti. V matrični predstavitvi omrežja se nerespondenti kažejo kot vrstice manjkajočih podatkov. Postopki za delo z manjkajočimi podatki zajemajo tri glavne pristope: pristop popolnih podatkov, rekonstrukcija in imputacije.

Pristop popolnih podatkov upošteva le podatke med respondenti in čeprav imamo o nerespondentih zbrane vhodne povezave, le-te zavrže. Poleg vrstice z nerespondenti se zbrišejo tudi ustrezni stolpci in rezultat je manjše omrežje.

Pri rekonstrukciji upoštevamo delno zbrane podatke med respondenti in nerespondenti tako, da manjkajočo vrstico podatkov zamenjamo z ustreznim stolpcem oziroma manjkajočo izhodno povezavo r_{ij} zamenjamo z vhodno povezavo r_{ji} (Stork in Richards, 1992; Huisman, 2009). Slabost rekonstrukcije je, da le-ta ni možna med dvema nerespondentoma. V takem primeru so potrebne dodatne imputacije, v najenostavnejšem primeru se namesto povezav med nerespondenti vstavi 0.

Imputacije povezav v omrežjih nadomestijo manjkajoče povezave z ocenami in tako

ustvarijo navidezno popolno omrežje. Manjkajoče povezave se lahko ocenijo s povprečnim številom povezav v omrežju, kar pomeni, da se za redka omrežja z gostoto manjšo ali enako 0,5, vstavi 0, za gosta omrežja pa 1. Druga možnost je ocena manjkajočih vrednosti na podlagi povprečja vhodnih povezav. V primeru binarnih omrežij to pomeni, da se vstavi 1 (torej povezava) za akterje, ki so popularni glede na vhodne povezave. Potrebna je izbira meje, in sicer: če je le-ta 0,5, se vstavi 1 (če je akterja zbralo vsaj pol respondentov) in 0 v nasprotnem primeru. Opisane imputacije imenujemo imputacije na podlagi modusa (vhodnih povezav). Imputacije na podlagi modusa se lahko uporabijo tudi kot dodatne imputacije pri rekonstrukciji povezav med nerespondenti.

V analizah omrežij z nerespondenti se manjkajoče vrednosti vse prevečkrat zanemari in kodira z 0, kot da gre za neobstoječe povezave. Tako smo tudi v naših simulacijah zaradi primerjave uporabili tudi to možnost, in sicer smo namesto manjkajočih povezav vstavili 0 (imputacije praznih povezav).

Problem nerespondentov je lahko rešen oziroma vsaj zmanjšan z uporabo primernih tretmajev. Zato smo postavili drugo tezo: *Stabilnost bločnega modeliranja omrežja z neodgovori (glede na bločni model popolnega omrežja) je večja, če je uporabljena rekonstrukcija kot pa vstavljanje brezpogojnih povprečij (na podlagi števila vhodnih povezav).*

Neodgovori so lahko prisotni tudi le na posamezni povezavi oziroma povezavah. V tem primeru akter sodeluje v raziskavi, vendar ne poda odgovora o povezavah do vseh ostalih akterjev. Tudi v tem primeru lahko uporabimo podobne postopke, kot smo jih opisali pri neodgovorih akterjev.

Tretja podskupina napak povzročenih s strani akterjev so merske napake. Prvo definicijo merskih napak v omrežjih sta leta 1973 podala Holland in Leinhardt, in sicer se merska napaka pojavi (ne glede na vzrok) kadar se zabeležen odgovor akterja ne ujema s pravo prikrito strukturo. Natančneje; merska napaka se pojavi, če povezava ni zabeležena v sociogramu, obstaja pa v pravi strukturi in obratno, če je povezava

zabeležena v sociogramu, vendar v resnici ne obstaja.

10.4 Stabilnost bločnega modeliranja

Rezultat bločnega modeliranja sta razvrstitev akterjev v skupine in bločna matrika. Stabilnost bločnega modeliranja na napake lahko tako definiramo oziroma merimo z dvema kazalnikoma, kjer primerjamo originalni bločni model ('popolni bločni model') in bločni model dobljen iz omrežja z napakami ('izmerjeni bločni model').

Prvi kazalnik, posplošeni Randov kazalnik (ARI, angl: *Adjusted Rand Index*), meri ujemanje med obema razvrstitvama akterjev. Vrednost 1 pomeni, da sta razvrstitvi popolnoma enaki, medtem ko vrednost 0 pomeni, da sta razvrstitvi dobljeni naključno.

Na podlagi obširnih simulacij, ki jih je predstavil Steinley (2004) smo se odločili, da bomo bločni model označili kot stabilen glede na razvrstitev, če bodo vrednosti posplošenega Randovega kazalnika nad 0,8.

Drugi kazalnik, odstotek napačnih blokov ($ErrB$, angl: *proportion of incorrect block types*), primerja originalno in izmerjeno bločno matriko oziroma natančneje: primerja tipe blokov in njihov položaj. Vrednost 0 pomeni, da so vsi bloki v obeh bločnih matrikah enaki in enako razvrščeni, medtem ko največja vrednost 1 pomeni, da se nobena dva bloka (iz originalne in izmerjene bločne matrike) ne ujemata. Za sam bločni model (oz. bločno matriko) pa bomo rekli, da je stabilen, če bo povprečen odstotek napačnih blokov pod 0,2.

Primernost obeh kazalnikov potrjujeta obe osrednji ideji analize socialnih omrežij, kot jih je podal Doreian (2008). Prva ideja pravi, da je struktura socialnega omrežja kot celota pomembna pri skupinskem izidu na nivoju omrežja. Druga ideja pa pravi, da je pozicija v omrežju pomemben izid na nivoju akterja. Če primerjamo to z rezultati bločnega modeliranja, je torej bločna matrika pomembna na nivoju omrežja, razvrstitev pa na nivoju akterja.

10.5 Zasnova simulacij za oceno stabilnosti bločnega modeliranja

10.5.1 Osnovna shema simulacij

Vse simulacije v disertaciji sledijo osnovni shemi s štirimi koraki:

1. Izbira popolnega omrežja iz literature oziroma generiranje omrežja glede na znani model.
2. Postavitev bločnega modela popolnega omrežja[27].
3. Naj $nGen$ predstavlja število simulacij za dano kombinacijo tipa napake, količino le-teh ter v nekaterih primerih še uporabo tretmajev za manjkajoče podatke. Za $i = 1$ do $i = nGen$ naredimo naslednje:
   a) Konstrukcija omrežja z napakami - izmerjeno omrežje.
   b) Postavitev bločnega modela izmerjenega omrežja.
   c) Primerjamo rezultate bločnega modeliranja popolnega in izmerjenega omrežja z obema kazalnikoma: posplošeni Randov kazalnik (ARI) in odstotek napačnih blokov ($ErrB$).
4. Raziščemo vpliv napak v zasnovi raziskave glede na povprečne vrednosti obeh kazalnikov.

10.5.2 Omrežja, uporabljena v simulacijah

V simulacijah smo uporabili dve vrsti omrežij: realna omrežja, znana iz literature, in simulirana omrežja glede na znan začetni model. Pri simuliranih omrežjih smo tako kot vhodne parametre potrebovali velikost omrežja s številom skupin akterjev oziroma natančneje začetno razvrstitev akterjev, bločno matriko ter verjetnosti povezav v posameznih blokih.

[27] *Popolno omrežje* je začetno znano omrežje (ali začetno omrežje v simulacijah).

10.6 Ocena stabilnosti bločnega modeliranja glede na napake v zasnovi raziskave

S simulacijo različnih tipov napak in količine napak smo poskušali odgovoriti na drugo raziskovalno vprašanje o stabilnosti bločnega modeliranja.

10.6.1 Napake zaradi omejevanja števila izbir

Najprej smo poskušali oceniti stabilnost bločnega modeliranja v primeru, ko je namesto neomejenega števila izbir v zasnovi vprašalnika postavljena neka omejitev nominacij. Rezultati, dobljeni na podlagi simulacij z dvema začetnima realnima omrežjema, kažejo, da lahko omejevanje števila izbir uniči bločno strukturo, če je omejitev postavljena prestrogo oziroma predaleč od resničnega števila želenih nominacij posameznih akterjev. Kot je poudaril Newman (2010) so omejitve pogosto postavljene zaradi praktičnih razlogov, da zmanjšajo delo raziskovalca. Radi bi poudarili, da to ni pravi razlog za postavljanje omejitev števila izbir, prav tako pred njihovo uporabo svari več drugih avtorjev (Holland in Leinhardt, 1973; Kossinets, 2006).

Glede na dobljene bločne modele s strukturno enakovrednostjo lahko zaključimo, da omejevanje števila izbir ni priporočljivo. Če so omejitve potrebne še iz kakšnega drugega razloga, le-teh ne smemo postaviti preveč strogo. Bolje je namreč, da akterje prisilimo, da nominirajo več prijateljev kot jih imajo v resnici, kot pa da jim onemogočimo, da naštejejo vse svoje prijatelje.

10.6.2 Napake zaradi smeri vprašanja

Stabilnost bločnega modeliranja s strukturno enakovrednostjo glede na smer zastavljenega vprašanja smo preverjali z dvema paroma realnih omrežij. V prvem primeru so bili akterji člani študentske vlade, ki smo jih spraševali po relacijah 'vprašati za mnenje' in 'biti vprašan za mnenje'. V drugem primeru so bili člani omrežja dijaki, ki so odgovarjali na vprašanji o dajanju in prejemanju emocionalne opore.

Ugotovili smo, da ima smer zastavljenega vprašanja velik vpliv na dobljeni bločni model. Tako razvrstitev akterjev kot bločna matrika sta odvisni od zastavljenega vprašanja. Z nadaljnjimi raziskavami bi morali ugotoviti, ali obstaja kakšen skupen vzorec povezav v bločnih modelih, dobljenimi iz omrežij zbranimi z originalnim in obrnjenim vprašanjem. Omrežje, dobljeno s potrditvijo povezav iz obeh omrežij, bi bilo morda lahko uporabljeno za iskanje najbolj stabilnih oziroma povezanih skupin akterjev.

10.6.3 Napake zaradi neodgovora akterja

Ocenjevanje stabilnosti bločnega modeliranja s strukturno enakovrednostjo na napake zaradi neodgovorov akterjev smo izvedli na podlagi dveh realnih omrežij in treh obsežnih simulacij. V osnovno shemo simulacij smo vključili različne načine določanja manjkajočih akterjev oziroma akterjev z neodgovori ter različne načine nadomeščanja manjkajočih povezav.
Tako smo v osnovno shemo simulacij vključili tri različne načine generiranja manjkajočih podatkov, in sicer:

i) popolnoma naključno;

ii) na podlagi vhodne stopnje;

iii) na podlagi izhodne stopnje.

V primeru določanja nerespondentov glede na vhodno oziroma izhodno stopnjo, je manjša vhodna oziroma izhodna stopnja pomenila večjo verjetnost, da je akter izbran kot nerespondent. Huisman in Steglich (2008) pravita, da tak izbor respondentov odseva karakteristike realnih zbranih omrežij, kjer so popularni akterji (z visoko vhodno stopnjo) bolj pripravljeni sodelovati v raziskavah, kot neaktivni akterji z nizko izhodno stopnjo. Costenbader in Valente (2003) tako ugotavljata, da akterji, ki odklonijo sodelovanje oziroma so manjkajoči, prihajajo z obrobja omrežja.

Prvi način generiranja nerespondentov se ujema z Rubinovo klasifikacijo (1976) manjkajočih podatkov *popolnoma nakljucno* (*angl.* MCAR - missing completely at random), saj akterji z neodgovori niso povezani z lastnostmi omrežja ali samih akterjev. Določanje

nerespondentov na podlagi vhodne ali izhodne stopnje je *nenakljucno* (*angl.* NMAR - not missing at random), saj je generiranje nerespondentov odvisno od omrežja oz. lastnosti akterjev v omrežju. *Naključno* generiranje manjkajočih podatkov (*angl.* MAR - missing at random) ni bilo vključeno v naše simulacije. V tem primeru je določanje nerespondentov odvisno od njihovih lastnosti, in ne od manjkajočih povezav. Huisman in Steglich (2008) sta na primer v svojih simulacijah uporabila podatke o porabi alkohola.

Za generiranjem manjkajočih akterjev smo v simulacijah uporabili še pet različnih postopkov oziroma tretmajev za zmanjšanje vpliva neodgovorov: pristop popolnih podatkov, rekonstrukcijo, imputacije na podlagi modusa, kombinacijo rekonstrukcije in imputacij na podlagi modusa ter imputacije prazne povezave. Za vsako omrežje je bil postavljen bločni model, ki smo ga primerjali z začetnim znanim bločnim modelom.

Stabilnost bločnih modelov glede na različno število manjkajočih akterjev in postopke za zmanjšanje vpliva neodgovorov smo ocenjevali s pomočjo povprečnih vrednosti obeh predstavljenih kazalnikov. Kot smo zapisali pri definiciji obeh kazalnikov, pravimo, da je bločni model stabilen glede na razvrstitev akterjev, če so povprečne vrednosti prilagojenega Randovega kazalnika ARI nad 0,8. V primeru drugega kazalnika, odstotka napačnih blokov ($ErrB$), pravimo, da je bločni model stabilen, če so vrednosti kazalnika pod 0,2.

Poleg tega smo postavili še multiple regresijske modele, kjer smo vrednosti kazalnikov napovedovali s številom manjkajočih akterjev, postopkom za manjkajoče podatke in načinom generiranja manjkajočih podatkov. Splošna ugotovitev je, da imajo odvisne spremenljivke večjo napovedno moč v primeru kazalnika ARI, torej pri napovedovanju ujemanja razvrstitev, kot pri napovedovanju bločne strukture (kazalnik $ErrB$).

Glavna ugotovitev je, da je uspešnost posameznih postopkov za manjkajoče podatke odvisna od simetrije omrežja. Simetrijo omrežja smo merili z recipročnostjo[28] in s sime-

[28]Recipročnost (Huisman, 2009) je definirana kot: recipročnost $= \frac{2 \cdot M}{2 \cdot M + A}$, kjer je M število medseboj-

trijo bločnega modela. Tretmaji, ki so uspešni v primeru simetričnih omrežij, se slabo obnašajo v primeru nesimetričnih omrežij in obratno. Za simetrična omrežja sta tako najboljša tretmaja rekonstrukcija in kombinacija rekonstrukcije z imputacijami na podlagi modusa za povezave med nerespondenti. Za nesimetrična omrežja sta najboljša tretmaja imputacije na podlagi modusa in pristop popolnih podatkov.

Tako lahko našo tezo 2 le delno potrdimo. Stabilnost bločnega modeliranja je višja, če uporabimo rekonstrukcijo v primerjavi z imputacijami na podlagi modusa le za simetrična omrežja. V nasprotnem primeru, torej pri nesimetričnih omrežjih, so imputacije na podlagi modusa boljši tretma kot rekonstrukcija.

Imputacije praznih povezav so v vseh primerih najslabši možni tretma in zato odsvetujemo njihovo uporabo. V praksi to pomeni, da je najslabše, če namesto manjkajočih podatkov v omrežje vstavimo kar 0. Čeprav se je pristop popolnih podatkov v simulacijah izkazal kot uspešen tretma v primeru nesimetričnih omrežij, njegovo uporabo odsvetujemo. Z odstranitvijo nerespondentov iz omrežja izgubimo namreč informacijo o njihovem položaju glede na druge akterje.

Simulacija neodgovorov akterjev na podlagi vhodne in izhodne stopne ni pokazala bistvenih razlik z naključno generiranimi nerespondenti. Razlog za to so najverjetneje majhna začetna omrežja, kjer razlike med akterji niso bili zelo izrazite.

10.6.4 Napake zaradi neodgovora na povezavi

V nadaljevanju raziskovanja stabilnosti bločnega modeliranja z neodgovori smo se osredotočili na neodgovore na povezavah. Podobno kot pri manjkajočih akterjih so tudi manjkajoče povezave v raziskavah in analizah vse prevečkrat kodirane kot 0, kar pomeni, da povezava ne obstaja. S simulacijami smo poskušali preveriti, ali je ta preprost tretma manjkajočih povezav sprejemljiv ter kateri tretma najbolj zmanjša vpliv manjkajočih povezav.

nih diad, A pa število asimetričnih diad.

Stabilnost bločnega modeliranja na podlagi strukturne enakovrednosti v omrežjih z manjkajočimi povezavami smo ocenjevali s simulacijami na štirih realnih omrežjih. Imputacije na podlagi modusa so najbolj primerne za omrežja z nizko recipročnostjo ter niso primerne za strukture oblike jedro-periferija. Za omenjeno strukturo se je kot najboljši tretma izkazala rekonstrukcija. Imputacije praznih povezav so, podobno kot pri neodgovorih akterjev, najslabši možni tretma. To pomeni, da kodiranje manjkajočih vrednosti z 0 ni sprejemljivo. Na splošno je stabilnost bločnega modeliranja z omrežji z manjkajočimi povezavami višja pri odkrivanju strukture modela kot pri odkrivanju položaja akterjev.

10.6.5 Slučajne merske napake

Kot smo že povedali, so merske napake v omrežju prisotne, če je zabeležena povezava, ki v resnični strukturi ne obstaja in obratno, če povezava v omrežju ni zabeležena, čeprav v resnici obstaja. Merske napake smo generirali naključno, kontrolirali smo le količino simuliranih napak. Merske napake smo simulirali tako, da smo naključno izbrane povezave spremenili, kar pomeni, da smo obstoječe povezave spremenili v prazne (1 smo spremenili v 0) in obratno, iz neobstoječih povezav smo ustvarili povezave (0 smo spremenili v 1).

Najprej smo ocenjevali stabilnost bločnega modeliranja s strukturno enakovrednostjo. Glede na rezultate, dobljene z realnimi in simuliranimi omrežji, lahko rečemo, da je bločno modeliranje s strukturno enakovrednostjo zelo stabilno. Če primerjamo stabilnost bločne strukture in razvrstitev akterjev, lahko zaključimo, da je bločna struktura nekoliko bolj stabilna. V prikazanih primerih omrežij je tako do 20% merskih napak še zagotavljalo sprejemljivo razvrstitev akterjev, medtem ko smo dobili sprejemljivo strukturo modela v povprečju tudi pri omrežjih, ki so imela med 25% in 30% spremenjenih povezav. Torej, bločno modeliranje s strukturno enakovrednostjo (ne glede na število skupin in simetričnost omrežja) je bolj stabilno na makro nivoju kot na mikro nivoju omrežja.

Simulacije s strukturno enakovrednostjo smo nato razširili še na regularno enakovrednost. Omrežja smo generirali na podlagi znane razvrstitve akterjev in bločne matrike. Verjetnost povezav v regularnih blokih smo najprej računali s pomočjo velikosti bloka. Na ta način smo dobili regularne bloke, ki ravno zadoščajo temu kriteriju. Povprečne gostote regularnih blokov so bile v različnih modelih tako med 0,2 in 0,4.

Ne glede na izbrani model (jedro-periferija ali povezane skupine) in število skupin so se bločni modeli z regularno enakovrednostjo izkazali za ekstremno nestabilne. Dobljena razvrstitev izmerjenega bločnega modela je bila v vseh primerih nesprejemljiva že za 2% spremenjenih povezav, v polovici predstavljenih začetnih omrežij pa že celo pri samo 1% spremenjenih povezav. Struktura bločnega modela je bila v povprečju sprejemljiva za največ 3% spremenjenih povezav.

Poskušali smo ugotoviti, zakaj je regularna enakovrednost tako zelo nestabilna in kaj se zgodi z bločnim modelom, če zamenjamo npr. samo eno povezavo. Ugotovili smo, da večkrat dobimo več enakovrednih razvrstitev (glede na vrednost kriterijske funkcije), med katerimi brez dodatnega znanja ne moremo izbrati najprimernejše. Med enakovrednimi razvrstitvami je lahko katera izmed rešitev enaka originalni razvrstitvi, druge pa se navadno povsem razlikujejo. Ker trenutno nimamo objektivnih kriterijev za izbiro najprimernejše razvrstitve v takih primerih, moramo bločne modele z regularno enakovrednostjo obravnavati še posebej previdno, obvezno pa moramo vključiti dodatna znanja raziskovalcev.

Možno razlago nestabilnosti regularne enakovrednosti smo iskali tudi v nizki povprečni gostoti regularnih blokov. Tako smo namesto izračunane verjetnosti povezav v regularnih blokih to verjetnost povečali na 0,6 oziroma 0,8. V obeh primerih se regularna enakovrednost izkaže za še bolj nestabilno, saj samo 1% spremenjenih povezav poruši tako razvrstitev kot bločni model. Ena od možnih rešitev je uporaba strukturne enakovrednosti, tudi če omrežje vsebuje vzorce povezav, ki ustrezajo definiciji regularne enakovrednosti.

Kljub nestabilnosti regularne enakovrednosti smo preverili, kako občutljiva je na merske napake posplošena enakovrednost, ki jo lahko definiramo z izborom blokov. Poskusili smo več različnih kombinacij dovoljenih idealnih blokov, vendar smo v vseh primerih že z 1% spremenjenih povezav dobili povsem nesprejemljiv bločni model glede na originalnega.

Tako smo našo tezo 1, da je strukturna enakovrednost bolj stabilna kot regularna oziroma posplošena enakovrednost, potrdili. Nestabilni rezultati regularne enakovrednosti zahtevajo natančnejši pregled definicije regularne enakovrednosti ter primerov njene uporabe v literaturi in postavitev smernic za njeno uporabo in interpretacijo.

10.7 Vpliv razlik v karakteristikah omrežij na stabilnost bločnega modeliranja

V nadaljevanju smo poskušali najti odgovor na prvo raziskovalno vprašanje, pri katerem nas je zanimalo, v kakšnem obsegu so razlike v karakteristikah popolnega in izmerjenega omrežja ter korelacije in/ali Evklidske razdalje med vektorjema z lastnostmi akterjev v obeh omrežjih sposobne napovedati rezultat bločnega modeliranja.

Glede na nestabilnost regularne in posplošene enakovrednosti smo v analizah vpliva sprememb lastnosti omrežij uporabili le strukturno enakovrednost. Vsako izmerjeno omrežje z naključno spremenjenimi povezavami smo primerjali s pravim, in sicer smo primerjali tako bločna modela (s prej predstavljenima kazalnikoma ARI in $ErrB$) kot karakteristike obeh omrežij. Tako smo izračunali relativno razliko v gostoti obeh omrežij, recipročnosti, številu medsebojnih diad, številu asimetričnih diad in številu praznih diad.

Precej karakteristik omrežja opišemo z vektorji, saj so mere podane na nivoju akterjev, torej za vsakega člana omrežja posebej. V takih primerih smo izračunali Pearsonov koeficient korelacije ter Evklidsko razdaljo med vektorjema popolnega in izmerjenega omrežja. Vektorje smo primerjali na dva načina, saj obe izbrani meri različnosti raz-

krivata različne vzorce med vektorjema (Ferligoj, 1989). Tako je npr. Pearsonov koeficient korelacije med dvema vzporednima vektorjema 1. Lahko bi rekli, da imata taka dva vektorja enak 'profil', le da je vzporedno premaknjen. Razlike v vrednostih obeh vektorjev pa lahko tudi v primeru vzporednih vektorjev zazna Evklidska razdalja. Tako v popolnem kot v izmerjenem omrežju smo izračunali bližino,središčnost glede na vhodno dostopnost, središčnost glede na izhodno dostopnost, središčnost glede na izhodno stopnjo, središčnost glede na vhodno stopnjo, središčnost glede na stopnjo, središčnost glede na vmesnost ter uteži kazal in vsebin.

V analizi smo tako raziskali multiple linearne regresijske modele ter, kjer so podatki nakazovali eksponentno odvisnost, še posplošene linearne modele (GLM). Simulacije smo, tako kot pri preučevanju vpliva merskih napak, izvedli na dveh realnih omrežjih ter treh simuliranih modelih.

Najprej smo tako preučili Pearsonove koeficiente korelacije oziroma enostavne linearne regresijske modele med vsemi napovednimi spremenljivkami (odstotek spremenjenih povezav, relativne razlike med karakteristikami omrežja in korelacija oz. Evklidska razdalja med vektorji z lastnostmi akterjev) in odvisnima kazalnikoma ujemanja obeh bločnih modelov (ARI in $ErrB$). Daleč največjo napovedno moč ima odstotek spremenjenih povezav, saj lahko napove vsaj 50% variacije kazalnika ARI ne glede na začetno strukturo modela. Vse napovedne spremenljivke imajo manjšo napovedno moč pri napovedovanju bločne strukture (kazalnik $ErrB$) kot razvrstitve akterjev (kazalnik ARI).

Relativne razlike v karakteristikah omrežja imajo manjšo napovedno moč, če je struktura omrežja simetrična. Kot smo povedali že pri analizi merskih napak, je strukturna enakovrednost zelo stabilna tudi za relativno visok odstotek spremenjenih povezav. Na drugi strani so karakteristike omrežja in lastnosti akterjev bolj občutljive na majhne spremembe v omrežju. Tako bi morali v nadaljnje raziskave vključiti bolj univerzalne modele, ki bi lahko upoštevali to različno občutljivost neodvisnih in odvisnih spremenljivk.

Če primerjamo napovedne spremenljivke izračunane s korelacijo in Evklidsko razdaljo med ustreznima vektorjema lastnosti akterjev, lahko zaključimo, da imajo večjo napovedno moč stabilnosti bločnega modela spremenljivke izračunane s korelacijo. Najboljše spremenljivke so tako korelacijski koeficient med vektorji središčnosti glede na vhodno in skupno stopnjo (vhodne in izhodne povezave) ter korelacije med vektorjema uteži kazal in vsebin. Prva možna razlaga je, da ima linearnost, ki jo izmerimo s korelacijo med dvema vektorjema, večji vpliv na stabilnost bločnega modela. Odgovor, zakaj spremenljivke, izračunane z Evklidsko razdaljo, pojasnijo manj variance, se skriva najbrž v komentarju Fausta in Romneya (1985), ki pravita, da se z uporabo razdalje kot mere podobnosti na nestandardiziranih spremenljivkah pomešajo informacije o podobnosti v vzorcih z informacijami o razlikah povprečja in variance posamezne spremenljivke.

Bolj podrobno smo raziskali odvisnost med odstotkom spremenjenih povezav in obema kazalnikoma stabilnosti bločnega modeliranja. Linearnemu regresijskemu modelu smo tako dodali še sestavljen regresijski model (angl. *piecewise linear regression model*) ter posplošen linearen model z eksponentno odvisnostjo. V večini primerov se je za najboljšega izkazal dvodelni regresijski model. Seveda bi dobili še boljše modele, če bi glede na podatke izbrali tri- ali večdelni model, vendar je vprašanje, če so glede na naravo podatkov takšni modeli smiselni.

V prvem raziskovalnem vprašanju smo se osredotočili na razlike med omrežjema in njihovo zmožnostjo napovedati stabilnost bločnega modela. Najbrž bi bilo bolj smiselno vprašanje, ali lahko same karakteristike izmerjenega omrežja napovedo stabilnost bločnega modela, saj pravega začetnega omrežja (in tako tudi ne razlik med pravim in izmerjenim omrežjem), ki prestavlja skrito temeljno strukturo, v resničnih raziskavah ne poznamo. V tem primeru se kot najboljše napovedne spremenljivke za stabilnost bločnega modeliranja izkažejo gostota omrežja in število medsebojnih diad (če je odstotek spremenjenih povezav konstanten). Teh zaključkov zaradi majhnega števila začetnih omrežij žal trenutno ne moremo posplošiti. Potrebne bi bile nadaljnje

simulacije z večjim številom začetnih omrežij ter različno velikostjo in strukturo le-teh. V disertaciji so sistematično predstavljene napake v zasnovi raziskav iz socialnih omrežij. Glavni doprinos disertacije k razvoju znanosti je obširna raziskava o vplivu posameznih tipov napak na rezultate posplošenega bločnega modeliranja. Za namen primerjanja bločnih modelov smo poiskali oziroma razvili dva kazalnika, ki primerjata bločna modela tako na individualnem mikro nivoju akterja, kot na makro nivoju omrežja oziroma modela.

Najbolj podrobno je raziskana strukturna enakovrednost, ki se je izkazala tudi za najstabilnejšo. Regularna enakovrednost in posplošena enakovrednost so se izkazale za izjemno nestabilne, tako glede na razvrstitev kot strukturo bločnega modela.

Najbolj raziskan tip napak so neodgovori akterjev, ki so tudi zelo pogost vir napak v dejanskih raziskavah. Na podlagi obsežnih simulacij smo podali priporočila za uporabo različnih tretmajev za manjkajoče podatke zaradi neodgovorov. Izbira tretmaja je odvisna od simetričnosti omrežja, in sicer tako od recipročnosti kot simetričnosti samega modela. Pomembna ugotovitev je, da je (v praksi vse prevečkrat uporabljeno) ignoriranje manjkajočih povezav, ki se kodirajo kar kot prazne, najslabša možna rešitev za stabilnost bločnega modela oziroma za postavitev pravega modela. Podobne rezultate in priporočila kot v primeru neodgovora akterja smo dobili tudi v primeru neodgovorov na povezavi.

10.8 Ideje za nadaljnje raziskovanje

V disertaciji predstavljene simulacije bo v prihodnosti potrebno izvesti za večja omrežja. Trenutno žal ne obstaja (povsem) zadovoljiva rešitev v programih, ki smo jih uporabljali pri pripravi disertacije. Simulacije so bile narejene v programu R s paketom `blockmodeling` (Žiberna, 2008), ki je primeren za generiranje začetnih omrežij in simuliranje napak, vendar je nekoliko počasen pri izvedbi bločnega modeliranja v primeru večjih omrežij ter večjega števila skupin. Trenutno je ena izmed možnih rešitev delo s testnim paketom `testBlockmodelingTestC`, ki ga je priredil Žiberna.

Druga možnost bi bila povezava programa Pajek (Batagelj in Mrvar, 2010a,b) z R-om na način, da bi v R-u lahko zagnali Pajka in bločno modeliranje, podatki oz. rezultati pa bi se shranjevali v R-u.

V primeru strukturne enakovrednosti bi morali uporabiti širši nabor vzorcev postavitve blokov. Poleg različnih tipov bločnih matrik bi morali obravnavati še različne velikosti omrežij ter različno število skupin. Podobno kot pri regularni enakovrednosti bi lahko kot začetne modele vzeli modele s povezanimi skupinami oziroma modele, ki predstavljajo jedro in periferijo.

Simulacije skupaj z napotki raziskovalcem bi bilo potrebno razširiti tudi na druge tipe enakovrednosti. V primeru posplošene enakovrednosti imamo širok nabor blokov, ki jih lahko kombiniramo na ogromno načinov, zato bi morale biti simulacije pripravljene še posebej skrbno.

Možne postopke v primeru neodgovorov akterja ali neodgovora na povezavah bi bilo potrebno razširiti na kompleksnejše tretmaje. V običajnih družboslovnih raziskavah se pogosto uporablja EM algoritem in večkratne (multiple) imputacije, zato bi bilo potrebno preučiti možnost njihove implementacije v analizo omrežij.

V primeru neodgovorov na povezavah bi bilo potrebno uporabiti nenaključno generiranje manjkajočih odgovorov. Prav tako bi morali uporabiti lastnosti akterjev pri simuliranju manjkajočih podatkov.

Poleg večjih omrežij bo potrebno analizo napak v zasnovi raziskave razširiti tudi na predznačena omrežja ter na omrežja z vrednostmi na povezavah. V tem primeru bo potrebno poleg količine napak v omrežju nadzirati tudi moč oziroma razsežnost napak skupaj s smerjo spremembe.

10.9 Napotki raziskovalcem

Na podlagi rezultatov smo oblikovali osnovne smernice raziskovalcem, ki se nanašajo tako na načrtovanje raziskave kot na analizo zbranih podatkov z bločnim modeliranjem.

- **Med pripravo raziskave**
  - **Pravilno določite meje omrežij.**
    Izključitev akterjev iz omrežja lahko spremeni dobljeni bločni model. V primeru, da je število manjkajočih akterjev majhno, lahko dobimo pravilen bločni model glede na tip blokov in njihov položaj v bločni matriki (rezultati dobljeni s pristopom popolnih podatkov v poglavju 7.3).
  - **Definicija raziskovalnega vprašanja glede na zagotavljanje ali prejemanje socialne opore.**
    Omrežja ter posledično dobljeni bločni modeli se lahko precej razlikujejo, ker merimo različne koncepte (poglavje 7.2).
  - **Izberite neomejeno število izbir namesto omejevanja izbir.**
    če je omejevanje števila akterjev nujno, naj meja ne bo postavljena prestrogo (poglavje 7).
- **Med zbiranjem podatkov**
  - **Manjkajoče povezave naj bodo jasno zabeležene, na primer z NA, v matričnem prikazu omrežja.**
    Manjkajoče povezave so prepogosto označene z 0, kar je najslabša možnost pri analizi omrežij z bločnim modeliranjem (rezultati o imputacijah (vstavljanju) prazne povezave v poglavju 7.3).
  - **Nikoli ne nadomestite manjkajočih povezav z 0,**
    ker so imputacije praznih povezav najslabši možni postopek tako glede na mikro nivo akterja (pozicija v omrežju) kot tudi glede na skupinski nivo omrežja, podan z bločno strukturo (rezultati o imputacijah prazne povezave v poglavju 7.3).

  - **Ugotovite, ali v omrežju obstaja neodgovor aktera in/ali neodgovor na povezavi. Poročajte o odstotku neodgovorov akterjev in/ali neodgovorov na povezavah** skupaj z velikostjo omrežja-

- **Pri izbiri tipa bločnega modela**

  - **Strukturna enakovrednost je zelo stabilna** do 50% nerespondentov ali 15% slučajnih merskih napak (poglavji 7.3 in 7.5).
  - **Regularna in posplošeni tipi enakovrednosti so izjemno nestabilni**, ker lahko ena spremenjena povezava popolnoma spremeni strukturo modela (poglavja 7.3.5, 7.5.4 in 7.5.5).

- **Med analizo podatkov (bločno modeliranje)**

  - **Ocenite recipročnost popolnega omrežja** z namenom izbire najboljšega postopka za nadomeščanje manjkajočih podatkov zaradi neodgovorov.
  - **Če je recipročnost nižja kot 0,5, naj bo uporabljen pristop popolnih podatkov ali imputacije na osnovi modusa.**
  - **Če je recipročnost višja kot 0,5, naj bo uporabljen pristop popolnih podatkov ali eden izmed obeh postopkov z rekonstrukcijo.**
  - **Ne uporabite pristopa popolnih podatkov, če je namen raziskave ugotoviti pozicijo nerespondentov v omrežju.**

Appendices

A All simple linear regressions with data for actor non-response

Table A.1: Linear regression models for all combinations of whole networks, missing data mechanism and treatments for ARI

Net	Missing mechanism	Treatment	Estimate $\hat{\beta}$	Std. Error	Confid. interval 2.5%	97.5%	t-test β_0	t value	Pr(<|t|)
boy-girl liking ties network	random	NTI	-0.117	0.002840	-0.122	-0.111	-0.040	-27.009	0.000
		RE	-0.014	0.001659	-0.017	-0.011	-0.040	15.662	1.000
		MO	-0.030	0.002540	-0.034	-0.025	-0.040	4.134	1.000
		REMO	-0.009	0.001393	-0.012	-0.007	-0.040	22.026	1.000
		CC	-0.008	0.001451	-0.011	-0.005	-0.040	22.226	1.000
	based on outdegree	NTI	-0.112	0.002571	-0.117	-0.107	-0.040	-27.956	0.000
		RE	-0.004	0.000777	-0.005	-0.002	-0.040	46.638	1.000
		MO	-0.129	0.003753	-0.137	-0.122	-0.040	-23.831	0.000
		REMO	-0.002	0.000445	-0.003	-0.001	-0.040	85.952	1.000
		CC	-0.008	0.001480	-0.011	-0.005	-0.040	21.420	1.000
	based on indegree	NTI	-0.112	0.002571	-0.117	-0.107	-0.040	-27.956	0.000
		RE	-0.004	0.000777	-0.005	-0.002	-0.040	46.638	1.000
		MO	-0.131	0.003061	-0.137	-0.125	-0.040	-29.594	0.000
		REMO	-0.002	0.000445	-0.003	-0.001	-0.040	85.952	1.000
		CC	-0.003	0.000795	-0.005	-0.002	-0.040	46.241	1.000
note borrowing network	random	NTI	-0.131	0.001859	-0.134	-0.127	-0.033	-52.354	0.000
		RE	-0.109	0.001884	-0.113	-0.105	-0.033	-40.134	0.000
		MO	-0.084	0.001711	-0.088	-0.081	-0.033	-29.901	0.000
		REMO	-0.077	0.001934	-0.081	-0.074	-0.033	-22.794	0.000
		CC	-0.050	0.002377	-0.055	-0.045	-0.033	-6.945	0.000
	based on outdegree	NTI	-0.133	0.001583	-0.136	-0.130	-0.033	-63.204	0.000
		RE	-0.106	0.001904	-0.110	-0.103	-0.033	-38.358	0.000
		MO	-0.072	0.001861	-0.076	-0.069	-0.033	-20.904	0.000
		REMO	-0.063	0.002228	-0.067	-0.058	-0.033	-13.153	0.000
		CC	-0.042	0.002214	-0.046	-0.037	-0.033	-3.818	0.000
	based on indegree	NTI	-0.133	0.001583	-0.136	-0.130	-0.033	-63.204	0.000
		RE	-0.097	0.001996	-0.101	-0.093	-0.033	-31.725	0.000
		MO	-0.072	0.001861	-0.076	-0.069	-0.033	-20.904	0.000
		REMO	-0.079	0.002532	-0.084	-0.074	-0.033	-18.173	0.000
		CC	-0.083	0.001806	-0.086	-0.079	-0.033	-27.260	0.000
completely symmetric network	random	NTI	-0.117	0.000371	-0.117	-0.116	-0.040	-206.382	0.000
		RE	-0.024	0.000245	-0.025	-0.024	-0.040	65.264	1.000
		MO	-0.141	0.000436	-0.141	-0.140	-0.040	-230.605	0.000
		REMO	-0.033	0.000287	-0.034	-0.032	-0.040	24.548	1.000
		CC	-0.045	0.000343	-0.045	-0.044	-0.040	-13.515	0.000
	based on outdegree	NTI	-0.117	0.000384	-0.118	-0.116	-0.040	-200.176	0.000
		RE	-0.021	0.000244	-0.021	-0.020	-0.040	78.671	1.000
		MO	-0.139	0.000439	-0.139	-0.138	-0.040	-224.567	0.000
		REMO	-0.033	0.000302	-0.034	-0.033	-0.040	22.233	1.000

continued on next page

continued from previous page

Net	Missing mechanism	Treatment	Estimate $\hat{\beta}$	Std. Error	Confid. interval 2.5%	97.5%	t-test β_0	t value	Pr(<|t|)
		CC	-0.044	0.000338	-0.045	-0.044	-0.040	-12.598	0.000
	based on indegree	NTI	-0.114	0.000355	-0.115	-0.113	-0.040	-208.364	0.000
		RE	-0.026	0.000245	-0.026	-0.025	-0.040	58.645	1.000
		MO	-0.135	0.000431	-0.135	-0.134	-0.040	-219.660	0.000
		REMO	-0.033	0.000277	-0.033	-0.032	-0.040	27.026	1.000
		CC	-0.097	0.000457	-0.098	-0.096	-0.040	-125.367	0.000
second non-symmetric blockmodel structure	random	NTI	-0.099	0.000111	-0.100	-0.099	-0.033	-597.382	0.000
		RE	-0.097	0.000154	-0.098	-0.097	-0.033	-416.152	0.000
		MO	-0.101	0.000147	-0.101	-0.101	-0.033	-460.084	0.000
		REMO	-0.092	0.000159	-0.092	-0.092	-0.033	-368.923	0.000
		CC	-0.015	0.000120	-0.016	-0.015	-0.033	150.012	1.000
	based on outdegree	NTI	-0.057	0.000136	-0.058	-0.057	-0.033	-177.420	0.000
		RE	-0.083	0.000132	-0.083	-0.082	-0.033	-374.175	0.000
		MO	-0.114	0.000126	-0.114	-0.113	-0.033	-638.696	0.000
		REMO	-0.091	0.000133	-0.091	-0.091	-0.033	-433.321	0.000
		CC	-0.037	0.000151	-0.037	-0.036	-0.033	-22.648	0.000
	based on indegree	NTI	-0.090	0.000096	-0.090	-0.089	-0.033	-587.444	0.000
		RE	-0.069	0.000158	-0.069	-0.068	-0.033	-222.891	0.000
		MO	-0.097	0.000125	-0.097	-0.096	-0.033	-505.026	0.000
		REMO	-0.063	0.000152	-0.064	-0.063	-0.033	-196.877	0.000
		CC	-0.028	0.000120	-0.028	-0.028	-0.033	45.593	1.000
second non-symmetric blockmodel structure	random	NTI	-0.092	0.000113	-0.093	-0.092	-0.033	-521.772	0.000
		RE	-0.096	0.000115	-0.096	-0.095	-0.033	-540.395	0.000
		MO	-0.049	0.000107	-0.049	-0.048	-0.033	-142.646	0.000
		REMO	-0.071	0.000160	-0.071	-0.071	-0.033	-236.210	0.000
		CC	-0.027	0.000146	-0.027	-0.027	-0.033	43.080	1.000
	based on outdegree	NTI	-0.080	0.000119	-0.080	-0.080	-0.033	-390.861	0.000
		RE	-0.085	0.000125	-0.086	-0.085	-0.033	-415.715	0.000
		MO	-0.064	0.000101	-0.064	-0.064	-0.033	-302.589	0.000
		REMO	-0.073	0.000151	-0.074	-0.073	-0.033	-264.612	0.000
		CC	-0.038	0.000155	-0.038	-0.038	-0.033	-30.004	0.000
	based on indegree	NTI	-0.094	0.000113	-0.094	-0.093	-0.033	-532.565	0.000
		RE	-0.091	0.000121	-0.091	-0.091	-0.033	-476.149	0.000
		MO	-0.042	0.000112	-0.043	-0.042	-0.033	-81.454	0.000
		REMO	-0.090	0.000125	-0.090	-0.089	-0.033	-449.112	0.000
		CC	-0.019	0.000123	-0.019	-0.019	-0.033	116.039	1.000

Table A.2: Linear regression models for all combinations of whole networks, missing data mechanism and treatments for *ErrB*

Net	Missing mechanism	Treatment	Estimate $\hat{\beta}$	Std. Error	Confid. interval 2.5%	97.5%	t-test β_0	t value	Pr(>|t|)
boy-girl liking ties network	random	NTI	0.058	0.000859	0.057	0.060	0.040	21.493	0.000
		RE	0.015	0.001207	0.012	0.017	0.040	-21.122	1.000
		MO	0.029	0.001660	0.025	0.032	0.040	-6.853	1.000
		REMO	0.012	0.001144	0.010	0.014	0.040	-24.495	1.000
		CC	0.006	0.000917	0.004	0.008	0.040	-36.984	1.000
	based on outdegree	NTI	0.057	0.000869	0.055	0.059	0.040	19.802	0.000
		RE	0.009	0.001026	0.007	0.011	0.040	-30.426	1.000
		MO	0.087	0.001926	0.083	0.091	0.040	24.443	0.000
		REMO	0.008	0.000981	0.006	0.010	0.040	-33.028	1.000
		CC	0.007	0.001022	0.005	0.009	0.040	-31.874	1.000
	based on indegree	NTI	0.057	0.000869	0.055	0.059	0.040	19.802	0.000
		RE	0.009	0.001026	0.007	0.011	0.040	-30.426	1.000
		MO	0.091	0.001483	0.088	0.094	0.040	34.263	0.000
		REMO	0.008	0.000981	0.006	0.010	0.040	-33.028	1.000
		CC	0.004	0.000705	0.003	0.006	0.040	-50.779	1.000
note borrowing network	random	NTI	0.045	0.001177	0.043	0.047	0.033	9.961	0.000
		RE	0.038	0.001053	0.036	0.040	0.033	4.519	0.000
		MO	0.030	0.001293	0.027	0.032	0.033	-2.885	0.998
		REMO	0.030	0.001114	0.028	0.033	0.033	-2.584	0.995
		CC	0.031	0.001268	0.028	0.033	0.033	-1.949	0.974
	based on outdegree	NTI	0.045	0.001209	0.043	0.047	0.033	9.573	0.000
		RE	0.038	0.001088	0.036	0.040	0.033	4.182	0.000
		MO	0.028	0.001255	0.026	0.030	0.033	-4.256	1.000
		REMO	0.028	0.001178	0.026	0.031	0.033	-4.215	1.000
		CC	0.029	0.001232	0.026	0.031	0.033	-3.918	1.000
	based on indegree	NTI	0.045	0.001209	0.043	0.047	0.033	9.573	0.000
		RE	0.030	0.000992	0.028	0.032	0.033	-3.172	0.999
		MO	0.028	0.001255	0.026	0.030	0.033	-4.256	1.000
		REMO	0.030	0.001145	0.028	0.032	0.033	-2.786	0.997
		CC	0.029	0.001208	0.027	0.031	0.033	-3.606	1.000
completely symmetric network	random	NTI	0.067	0.000196	0.066	0.067	0.040	135.647	0.000
		RE	0.017	0.000155	0.016	0.017	0.040	-150.167	1.000
		MO	0.091	0.000208	0.090	0.091	0.040	243.624	0.000
		REMO	0.018	0.000164	0.017	0.018	0.040	-137.074	1.000
		CC	0.063	0.000282	0.063	0.064	0.040	82.453	0.000
	based on outdegree	NTI	0.068	0.000201	0.068	0.069	0.040	140.169	0.000
		RE	0.015	0.000166	0.015	0.015	0.040	-149.593	1.000
		MO	0.089	0.000204	0.089	0.089	0.040	239.882	0.000
		REMO	0.017	0.000177	0.017	0.018	0.040	-127.601	1.000
		CC	0.063	0.000278	0.063	0.064	0.040	83.301	0.000
	based on indegree	NTI	0.063	0.000174	0.063	0.064	0.040	133.586	0.000
		RE	0.019	0.000152	0.019	0.019	0.040	-138.948	1.000
		MO	0.092	0.000204	0.091	0.092	0.040	253.571	0.000
		REMO	0.020	0.000166	0.019	0.020	0.040	-122.564	1.000
		CC	0.071	0.000252	0.070	0.071	0.040	121.168	0.000
first non-symmetric blockmodel structure	random	NTI	0.029	0.000038	0.029	0.029	0.033	-111.914	1.000
		RE	0.034	0.000067	0.034	0.034	0.033	7.721	0.000
		MO	0.034	0.000084	0.034	0.034	0.033	8.480	0.000
		REMO	0.025	0.000077	0.024	0.025	0.033	-113.008	1.000
		CC	0.015	0.000076	0.015	0.015	0.033	-240.053	1.000
	based on outdegree	NTI	0.033	0.000035	0.033	0.033	0.033	-5.061	1.000
		RE	0.030	0.000065	0.030	0.030	0.033	-53.889	1.000
		MO	0.055	0.000071	0.054	0.055	0.033	297.916	0.000
		REMO	0.030	0.000082	0.030	0.030	0.033	-37.399	1.000
		CC	0.045	0.000100	0.044	0.045	0.033	113.154	0.000
	based on indegree	NTI	0.031	0.000037	0.031	0.031	0.033	-58.096	1.000
		RE	0.028	0.000067	0.028	0.029	0.033	-73.423	1.000
		MO	0.051	0.000073	0.051	0.051	0.033	238.261	0.000

continued on next page

continued from previous page

Net	Missing mechanism	Treatment	Estimate $\hat{\beta}$	Std. Error	Confid. interval 2.5%	97.5%	β_0	t-test t value	Pr(>|t|)
		REMO	0.024	0.000066	0.024	0.024	0.033	-142.352	1.000
		CC	0.040	0.000099	0.040	0.040	0.033	68.653	0.000
second non-symmetric blockmodel structure	random	NTI	0.040	0.000068	0.040	0.040	0.033	92.506	0.000
		RE	0.042	0.000072	0.042	0.042	0.033	118.016	0.000
		MO	0.030	0.000084	0.030	0.030	0.033	-35.824	1.000
		REMO	0.035	0.000081	0.035	0.035	0.033	17.734	0.000
		CC	0.031	0.000084	0.031	0.031	0.033	-24.514	1.000
	based on outdegree	NTI	0.041	0.000070	0.041	0.041	0.033	106.320	0.000
		RE	0.042	0.000072	0.042	0.042	0.033	121.001	0.000
		MO	0.039	0.000092	0.039	0.039	0.033	60.678	0.000
		REMO	0.039	0.000084	0.039	0.039	0.033	67.437	0.000
		CC	0.040	0.000090	0.040	0.040	0.033	75.730	0.000
	based on indegree	NTI	0.039	0.000078	0.039	0.039	0.033	75.873	0.000
		RE	0.037	0.000077	0.037	0.037	0.033	48.146	0.000
		MO	0.028	0.000083	0.028	0.028	0.033	-63.880	1.000
		REMO	0.037	0.000079	0.037	0.037	0.033	47.743	0.000
		CC	0.029	0.000079	0.028	0.029	0.033	-59.738	1.000

B Pearson correlation coefficients for network indices and indices of stability of blockmodels

Table B.1: Pearson correlation coefficients between indices of network properties and indices of stability of blockmodels for the boy-girl liking ties network

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	ARI	1.00	-0.82	-0.77	-0.48	-0.56	-0.33	-0.64	-0.62	-0.56	-0.57	-0.57	-0.57	-0.59	-0.59	0.37	-0.46	-0.45	0.21	0.18	0.21	0.48	0.41	0.45	0.09	0.45	0.42
2	ErrB	-0.82	1.00	0.58	0.35	0.44	0.27	0.49	0.47	0.42	0.42	0.42	0.41	0.44	0.44	-0.27	0.34	0.32	-0.14	-0.12	-0.15	-0.33	-0.28	-0.33	-0.06	-0.33	-0.30
3	p.changed	-0.77	0.58	1.00	0.72	0.74	0.37	0.87	0.88	0.85	0.85	0.85	0.78	0.79	0.80	-0.57	0.67	0.65	-0.31	-0.30	-0.33	-0.70	-0.63	-0.70	-0.14	-0.66	-0.64
4	Dens	-0.48	0.35	0.72	1.00	0.32	-0.15	0.64	0.89	0.91	0.92	0.92	0.91	0.82	0.82	-0.58	0.47	0.45	-0.21	-0.21	-0.22	-0.48	-0.44	-0.48	-0.09	-0.49	-0.48
5	Rec	-0.56	0.44	0.74	0.32	1.00	0.79	0.92	0.70	0.53	0.52	0.52	0.37	0.47	0.48	-0.37	0.61	0.58	-0.24	-0.22	-0.26	-0.56	-0.50	-0.54	-0.10	-0.58	-0.56
6	D_Mut	-0.33	0.27	0.37	-0.15	0.79	1.00	0.55	0.23	0.07	0.06	0.06	-0.02	0.09	0.11	-0.07	0.35	0.33	-0.12	-0.09	-0.14	-0.31	-0.28	-0.29	-0.05	-0.31	-0.29
7	D_Asymm	-0.64	0.49	0.87	0.64	0.92	0.55	1.00	0.92	0.78	0.77	0.77	0.65	0.70	0.70	-0.53	0.66	0.63	-0.27	-0.25	-0.29	-0.62	-0.56	-0.61	-0.12	-0.65	-0.62
8	D_Null	-0.62	0.47	0.88	0.89	0.70	0.23	0.92	1.00	0.94	0.93	0.93	0.85	0.83	0.83	-0.61	0.63	0.60	-0.27	-0.26	-0.29	-0.61	-0.55	-0.60	-0.12	-0.63	-0.61
9	PP_e	-0.56	0.42	0.85	0.91	0.53	0.07	0.78	0.94	1.00	0.94	1.00	0.88	0.79	0.88	-0.62	0.66	0.60	-0.33	-0.28	-0.35	-0.62	-0.56	-0.64	-0.12	-0.65	-0.62
10	CCout_e	-0.57	0.42	0.85	0.92	0.52	0.06	0.77	0.93	0.94	1.00	0.93	0.90	0.91	0.80	-0.61	0.60	0.63	-0.26	-0.34	-0.29	-0.62	-0.61	-0.59	-0.13	-0.61	-0.62
11	CCin_e	-0.57	0.42	0.85	0.92	0.52	0.06	0.77	0.93	1.00	0.93	1.00	0.90	0.80	0.90	-0.62	0.65	0.58	-0.34	-0.27	-0.36	-0.62	-0.55	-0.65	-0.13	-0.63	-0.61
12	Dall_e	-0.57	0.41	0.78	0.91	0.37	-0.02	0.65	0.85	0.88	0.90	0.90	1.00	0.89	0.89	-0.52	0.54	0.53	-0.27	-0.26	-0.28	-0.62	-0.51	-0.56	-0.15	-0.52	-0.51
13	Dout_e	-0.59	0.44	0.79	0.82	0.47	0.09	0.70	0.83	0.79	0.91	0.80	0.89	1.00	0.76	-0.49	0.52	0.60	-0.22	-0.35	-0.24	-0.60	-0.62	-0.53	-0.15	-0.53	-0.56
14	Din_e	-0.59	0.44	0.80	0.82	0.48	0.11	0.70	0.83	0.88	0.80	0.90	0.89	0.76	1.00	-0.50	0.62	0.50	-0.38	-0.22	-0.40	-0.63	-0.48	-0.71	-0.16	-0.57	-0.51
15	B_e	0.37	-0.27	-0.57	-0.58	-0.37	-0.07	-0.53	-0.61	-0.62	-0.61	-0.62	-0.52	-0.49	-0.50	1.00	-0.39	-0.38	0.06	0.05	0.06	0.39	0.37	0.40	-0.07	0.41	0.41
16	A_e	-0.46	0.34	0.67	0.47	0.61	0.35	0.66	0.63	0.66	0.60	0.65	0.54	0.52	0.62	-0.39	1.00	0.79	-0.41	-0.20	-0.41	-0.71	-0.53	-0.66	-0.07	-0.86	-0.73
17	H_e	-0.45	0.32	0.65	0.45	0.58	0.33	0.63	0.60	0.60	0.63	0.58	0.53	0.60	0.50	-0.38	0.79	1.00	-0.19	-0.38	-0.21	-0.70	-0.66	-0.52	-0.06	-0.74	-0.84
18	PP_cor	0.21	-0.14	-0.31	-0.21	-0.24	-0.12	-0.27	-0.27	-0.33	-0.26	-0.34	-0.27	-0.22	-0.38	0.06	-0.41	-0.19	1.00	-0.04	0.98	0.40	0.14	0.68	0.22	0.37	0.18
19	CCout_cor	0.18	-0.12	-0.30	-0.21	-0.22	-0.09	-0.25	-0.26	-0.28	-0.34	-0.27	-0.26	-0.35	-0.22	0.05	-0.20	-0.38	-0.04	1.00	0.01	0.34	0.71	0.19	0.25	0.20	0.34
20	CCin_cor	0.21	-0.15	-0.33	-0.22	-0.26	-0.14	-0.29	-0.29	-0.35	-0.29	-0.36	-0.28	-0.24	-0.40	0.06	-0.41	-0.21	0.98	0.01	1.00	0.43	0.18	0.73	0.25	0.36	0.20
21	Dall_cor	0.48	-0.33	-0.70	-0.48	-0.56	-0.31	-0.62	-0.61	-0.62	-0.62	-0.62	-0.62	-0.60	-0.63	0.39	-0.71	-0.70	0.40	0.34	0.43	1.00	0.66	0.72	0.33	0.66	0.65
22	Dout_cor	0.41	-0.28	-0.63	-0.44	-0.50	-0.28	-0.56	-0.55	-0.56	-0.61	-0.55	-0.51	-0.62	-0.48	0.37	-0.53	-0.66	0.14	0.71	0.18	0.66	1.00	0.47	0.23	0.51	0.62
23	Din_cor	0.45	-0.33	-0.70	-0.48	-0.54	-0.29	-0.61	-0.60	-0.64	-0.59	-0.65	-0.56	-0.53	-0.71	0.40	-0.66	-0.52	0.68	0.19	0.73	0.72	0.47	1.00	0.22	0.61	0.51
24	B_cor	0.09	-0.06	-0.14	-0.09	-0.10	-0.05	-0.12	-0.12	-0.12	-0.13	-0.13	-0.15	-0.15	-0.16	-0.07	-0.07	-0.06	0.22	0.25	0.25	0.33	0.23	0.22	1.00	0.05	0.04
25	A_cor	0.45	-0.33	-0.66	-0.49	-0.58	-0.31	-0.65	-0.63	-0.65	-0.61	-0.63	-0.52	-0.53	-0.57	0.41	-0.86	-0.74	0.37	0.20	0.36	0.66	0.51	0.61	0.05	1.00	0.74
26	H_cor	0.42	-0.30	-0.64	-0.48	-0.56	-0.29	-0.62	-0.61	-0.62	-0.62	-0.61	-0.51	-0.56	-0.51	0.41	-0.73	-0.84	0.18	0.34	0.20	0.65	0.62	0.51	0.04	0.74	1.00

Table B.2: Pearson correlation coefficients between indices of network properties and indices of stability of blockmodeling for the note borrowing network

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	ARI	1.00	-0.72	-0.85	-0.75	-0.10	-0.51	-0.67	-0.76	-0.79	-0.79	-0.79	-0.74	-0.74	-0.79	-0.46	-0.83	-0.73	0.67	0.54	0.67	0.74	0.57	0.74	0.46	0.76	0.74
2	ErrB	-0.72	1.00	0.70	0.64	0.04	0.47	0.55	0.64	0.66	0.66	0.66	0.67	0.66	0.68	0.34	0.68	0.57	-0.57	-0.39	-0.57	-0.61	-0.42	-0.62	-0.33	-0.63	-0.56
3	p.changed	-0.85	0.70	1.00	0.91	0.07	0.65	0.80	0.92	0.95	0.94	0.95	0.91	0.90	0.93	0.55	0.93	0.79	-0.80	-0.60	-0.80	-0.85	-0.68	-0.86	-0.54	-0.86	-0.80
4	Dens	-0.75	0.64	0.91	1.00	-0.05	0.78	0.78	0.97	0.94	0.96	0.94	0.96	0.95	0.92	0.50	0.85	0.69	-0.70	-0.55	-0.70	-0.75	-0.60	-0.75	-0.50	-0.76	-0.72
5	Rec	-0.10	0.04	0.07	-0.05	1.00	-0.22	0.37	0.11	0.03	-0.00	0.03	-0.05	-0.03	-0.01	0.15	0.06	0.12	-0.06	-0.11	-0.06	-0.06	-0.15	-0.05	-0.14	-0.05	-0.11
6	D_Mut	-0.51	0.47	0.65	0.78	-0.22	1.00	0.28	0.62	0.69	0.72	0.69	0.79	0.77	0.73	0.32	0.60	0.45	-0.50	-0.37	-0.51	-0.53	-0.39	-0.55	-0.31	-0.55	-0.48
7	D_Asymm	-0.67	0.55	0.80	0.78	0.37	0.28	1.00	0.91	0.78	0.78	0.78	0.73	0.73	0.73	0.50	0.74	0.65	-0.60	-0.50	-0.60	-0.67	-0.58	-0.64	-0.50	-0.65	-0.67
8	D_Null	-0.76	0.64	0.92	0.97	0.11	0.62	0.91	1.00	0.93	0.95	0.93	0.92	0.92	0.90	0.53	0.85	0.71	-0.70	-0.56	-0.70	-0.76	-0.63	-0.75	-0.53	-0.76	-0.74
9	PP_e	-0.79	0.66	0.95	0.94	0.03	0.69	0.78	0.93	1.00	0.97	1.00	0.91	0.88	0.93	0.56	0.91	0.73	-0.83	-0.63	-0.83	-0.81	-0.65	-0.84	-0.56	-0.82	-0.76
10	CCout_e	-0.79	0.66	0.94	0.96	-0.00	0.72	0.78	0.95	0.97	1.00	0.97	0.92	0.93	0.91	0.57	0.88	0.72	-0.72	-0.66	-0.72	-0.77	-0.66	-0.77	-0.57	-0.78	-0.76
11	CCin_e	-0.79	0.66	0.95	0.94	0.03	0.69	0.78	0.93	1.00	0.97	1.00	0.91	0.88	0.93	0.56	0.91	0.73	-0.83	-0.63	-0.83	-0.81	-0.65	-0.84	-0.56	-0.82	-0.76
12	Dall_e	-0.74	0.67	0.91	0.96	-0.05	0.79	0.73	0.92	0.91	0.92	0.91	1.00	0.96	0.96	0.48	0.86	0.68	-0.75	-0.49	-0.75	-0.85	-0.58	-0.82	-0.47	-0.81	-0.69
13	Dout_e	-0.74	0.66	0.90	0.95	-0.03	0.77	0.73	0.92	0.88	0.93	0.88	0.96	1.00	0.91	0.48	0.83	0.72	-0.69	-0.52	-0.69	-0.78	-0.63	-0.75	-0.47	-0.76	-0.71
14	Din_e	-0.79	0.68	0.93	0.92	-0.01	0.73	0.73	0.90	0.93	0.91	0.93	0.96	0.91	1.00	0.48	0.94	0.70	-0.86	-0.52	-0.86	-0.86	-0.58	-0.93	-0.48	-0.91	-0.71
15	B_e	-0.46	0.34	0.55	0.50	0.15	0.32	0.50	0.53	0.56	0.57	0.56	0.48	0.48	0.48	1.00	0.51	0.46	-0.45	-0.63	-0.45	-0.50	-0.48	-0.44	-0.92	-0.44	-0.48
16	A_e	-0.83	0.68	0.93	0.85	0.06	0.60	0.74	0.85	0.91	0.88	0.91	0.86	0.83	0.94	0.51	1.00	0.80	-0.85	-0.57	-0.85	-0.87	-0.62	-0.93	-0.51	-0.96	-0.80
17	H_e	-0.73	0.57	0.79	0.69	0.12	0.45	0.65	0.71	0.73	0.72	0.73	0.68	0.72	0.70	0.46	0.80	1.00	-0.61	-0.48	-0.61	-0.72	-0.64	-0.66	-0.45	-0.73	-0.90
18	PP_cor	0.67	-0.57	-0.80	-0.70	-0.06	-0.50	-0.60	-0.70	-0.83	-0.72	-0.83	-0.75	-0.69	-0.86	-0.45	-0.85	-0.61	1.00	0.47	1.00	0.80	0.50	0.92	0.44	0.86	0.60
19	CCout_cor	0.54	-0.39	-0.60	-0.55	-0.11	-0.37	-0.50	-0.56	-0.63	-0.66	-0.63	-0.49	-0.52	-0.52	-0.63	-0.57	-0.48	0.47	1.00	0.47	0.46	0.63	0.48	0.62	0.48	0.52
20	CCin_cor	0.67	-0.57	-0.80	-0.70	-0.06	-0.51	-0.60	-0.70	-0.83	-0.72	-0.83	-0.75	-0.69	-0.86	-0.45	-0.85	-0.61	1.00	0.47	1.00	0.80	0.50	0.92	0.44	0.86	0.60
21	Dall_cor	0.74	-0.61	-0.85	-0.75	-0.06	-0.53	-0.67	-0.76	-0.81	-0.77	-0.81	-0.85	-0.78	-0.86	-0.50	-0.87	-0.72	0.80	0.46	0.80	1.00	0.61	0.87	0.49	0.85	0.69
22	Dout_cor	0.57	-0.42	-0.68	-0.60	-0.15	-0.39	-0.58	-0.63	-0.65	-0.66	-0.65	-0.58	-0.63	-0.58	-0.48	-0.62	-0.64	0.50	0.63	0.50	0.61	1.00	0.53	0.48	0.54	0.66
23	Din_cor	0.74	-0.62	-0.86	-0.75	-0.05	-0.55	-0.64	-0.75	-0.84	-0.77	-0.84	-0.82	-0.75	-0.93	-0.44	-0.93	-0.66	0.92	0.48	0.92	0.87	0.53	1.00	0.44	0.96	0.65
24	B_cor	0.46	-0.33	-0.54	-0.50	-0.14	-0.31	-0.50	-0.53	-0.56	-0.57	-0.56	-0.47	-0.47	-0.48	-0.92	-0.51	-0.45	0.44	0.62	0.44	0.49	0.48	0.44	1.00	0.43	0.48
25	A_cor	0.76	-0.63	-0.86	-0.76	-0.05	-0.55	-0.65	-0.76	-0.82	-0.78	-0.82	-0.81	-0.76	-0.91	-0.44	-0.96	-0.73	0.86	0.48	0.86	0.85	0.54	0.96	0.43	1.00	0.70
26	H_cor	0.74	-0.56	-0.80	-0.72	-0.11	-0.48	-0.67	-0.74	-0.76	-0.76	-0.76	-0.69	-0.71	-0.71	-0.48	-0.80	-0.90	0.60	0.52	0.60	0.69	0.66	0.65	0.48	0.70	1.00

Table B.3: Pearson correlation coefficients between indices of network properties and indices of stability of blockmodeling for the completely symmetric blockmodel structure

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	ARI	1.00	-0.80	-0.71	-0.27	-0.34	-0.29	-0.03	-0.38	-0.18	-0.25	-0.26	-0.48	-0.53	-0.52	0.09	-0.05	-0.05	0.23	0.24	0.24	0.29	0.29	0.28	0.23	0.30	0.30
2	ErrB	-0.80	1.00	0.52	0.20	0.27	0.23	0.02	0.28	0.14	0.20	0.20	0.35	0.39	0.39	-0.05	0.05	0.05	-0.18	-0.18	-0.18	-0.22	-0.21	-0.21	-0.17	-0.23	-0.22
3	p.changed	-0.71	0.52	1.00	0.40	0.62	0.52	0.27	0.66	0.41	0.47	0.48	0.68	0.73	0.72	-0.04	0.21	0.21	-0.49	-0.51	-0.50	-0.59	-0.58	-0.57	-0.44	-0.57	-0.58
4	Dens	-0.27	0.20	0.40	1.00	0.18	0.01	0.08	0.54	0.73	0.78	0.78	0.77	0.58	0.59	0.04	0.37	0.37	-0.22	-0.22	-0.22	-0.23	-0.22	-0.22	-0.22	-0.26	-0.27
5	Rec	-0.34	0.27	0.62	0.18	1.00	0.81	0.51	0.58	0.41	0.39	0.40	0.29	0.37	0.38	0.14	0.37	0.37	-0.38	-0.40	-0.39	-0.46	-0.44	-0.43	-0.35	-0.44	-0.46
6	D_Mut	-0.29	0.23	0.52	0.01	0.81	1.00	0.36	0.27	0.10	0.10	0.11	0.20	0.29	0.29	0.07	0.15	0.15	-0.28	-0.29	-0.28	-0.36	-0.35	-0.34	-0.25	-0.33	-0.33
7	D_Asym	-0.03	0.02	0.27	0.08	0.51	0.36	1.00	0.42	0.39	0.29	0.31	0.10	0.12	0.12	0.16	0.40	0.40	-0.25	-0.26	-0.25	-0.29	-0.29	-0.27	-0.22	-0.25	-0.27
8	D_Null	-0.38	0.28	0.66	0.54	0.58	0.27	0.42	1.00	0.59	0.60	0.62	0.55	0.54	0.55	0.01	0.31	0.31	-0.37	-0.39	-0.38	-0.44	-0.42	-0.41	-0.35	-0.44	-0.45
9	PP_e	-0.18	0.14	0.41	0.73	0.41	0.10	0.39	0.59	1.00	0.89	0.98	0.59	0.43	0.55	0.22	0.72	0.69	-0.39	-0.35	-0.39	-0.34	-0.32	-0.35	-0.34	-0.38	-0.37
10	CCout_e	-0.25	0.20	0.47	0.78	0.39	0.10	0.29	0.60	0.89	1.00	0.88	0.67	0.64	0.50	0.21	0.59	0.63	-0.34	-0.43	-0.34	-0.36	-0.40	-0.31	-0.35	-0.36	-0.42
11	CCin_e	-0.26	0.20	0.48	0.78	0.40	0.11	0.31	0.62	0.98	0.88	1.00	0.67	0.50	0.66	0.21	0.64	0.60	-0.42	-0.36	-0.43	-0.37	-0.34	-0.40	-0.36	-0.41	-0.39
12	Dall_e	-0.48	0.35	0.68	0.77	0.29	0.20	0.10	0.55	0.59	0.67	0.67	1.00	0.81	0.81	0.00	0.26	0.26	-0.36	-0.38	-0.37	-0.51	-0.43	-0.42	-0.36	-0.41	-0.42
13	Dout_e	-0.53	0.39	0.73	0.58	0.37	0.29	0.12	0.54	0.43	0.64	0.50	0.81	1.00	0.64	-0.01	0.19	0.27	-0.35	-0.50	-0.35	-0.49	-0.59	-0.40	-0.36	-0.42	-0.54
14	Din_e	-0.52	0.39	0.72	0.59	0.38	0.29	0.12	0.55	0.55	0.50	0.66	0.81	0.64	1.00	-0.01	0.28	0.19	-0.48	-0.35	-0.49	-0.49	-0.40	-0.57	-0.35	-0.53	-0.42
15	B_e	0.09	-0.05	-0.04	0.04	0.14	0.07	0.16	0.01	0.22	0.21	0.21	0.00	-0.01	-0.01	1.00	0.35	0.35	-0.25	-0.26	-0.25	-0.16	-0.12	-0.11	-0.37	-0.09	-0.11
16	A_e	-0.05	0.05	0.21	0.37	0.37	0.15	0.40	0.31	0.72	0.59	0.64	0.26	0.19	0.28	0.35	1.00	0.93	-0.35	-0.29	-0.35	-0.30	-0.25	-0.30	-0.31	-0.35	-0.32
17	H_e	-0.05	0.05	0.21	0.37	0.37	0.15	0.40	0.31	0.69	0.63	0.60	0.26	0.27	0.19	0.35	0.93	1.00	-0.30	-0.35	-0.30	-0.30	-0.31	-0.25	-0.31	-0.30	-0.38
18	PP_cor	0.23	-0.18	-0.49	-0.22	-0.38	-0.28	-0.25	-0.37	-0.39	-0.34	-0.42	-0.36	-0.35	-0.48	-0.25	-0.35	-0.30	1.00	0.45	0.99	0.52	0.41	0.82	0.50	0.62	0.41
19	CCout_cor	0.24	-0.18	-0.51	-0.22	-0.40	-0.29	-0.26	-0.39	-0.35	-0.43	-0.36	-0.38	-0.50	-0.35	-0.26	-0.29	-0.35	0.45	1.00	0.45	0.56	0.84	0.41	0.51	0.40	0.64
20	CCin_cor	0.24	-0.18	-0.50	-0.22	-0.39	-0.28	-0.25	-0.38	-0.39	-0.34	-0.43	-0.37	-0.35	-0.49	-0.25	-0.35	-0.30	0.99	0.45	1.00	0.53	0.41	0.83	0.50	0.63	0.41
21	Dall_cor	0.29	-0.22	-0.59	-0.23	-0.46	-0.36	-0.29	-0.44	-0.34	-0.36	-0.37	-0.51	-0.49	-0.49	-0.16	-0.30	-0.30	0.52	0.56	0.53	1.00	0.64	0.62	0.59	0.53	0.56
22	Dout_cor	0.29	-0.21	-0.58	-0.22	-0.44	-0.35	-0.29	-0.42	-0.32	-0.40	-0.34	-0.43	-0.59	-0.40	-0.12	-0.25	-0.31	0.41	0.84	0.41	0.64	1.00	0.45	0.44	0.45	0.74
23	Din_cor	0.28	-0.21	-0.57	-0.22	-0.43	-0.34	-0.27	-0.41	-0.35	-0.31	-0.40	-0.42	-0.40	-0.57	-0.11	-0.30	-0.25	0.82	0.41	0.83	0.62	0.45	1.00	0.43	0.73	0.45
24	B_cor	0.23	-0.17	-0.44	-0.22	-0.35	-0.25	-0.22	-0.35	-0.34	-0.35	-0.36	-0.36	-0.36	-0.35	-0.37	-0.31	-0.31	0.50	0.51	0.50	0.59	0.44	0.43	1.00	0.35	0.38
25	A_cor	0.30	-0.23	-0.57	-0.26	-0.44	-0.33	-0.25	-0.44	-0.38	-0.36	-0.41	-0.41	-0.42	-0.53	-0.09	-0.35	-0.30	0.62	0.40	0.63	0.53	0.45	0.73	0.35	1.00	0.60
26	H_cor	0.30	-0.22	-0.58	-0.27	-0.46	-0.33	-0.27	-0.45	-0.37	-0.42	-0.39	-0.42	-0.54	-0.42	-0.11	-0.32	-0.38	0.41	0.64	0.41	0.56	0.74	0.45	0.38	0.60	1.00

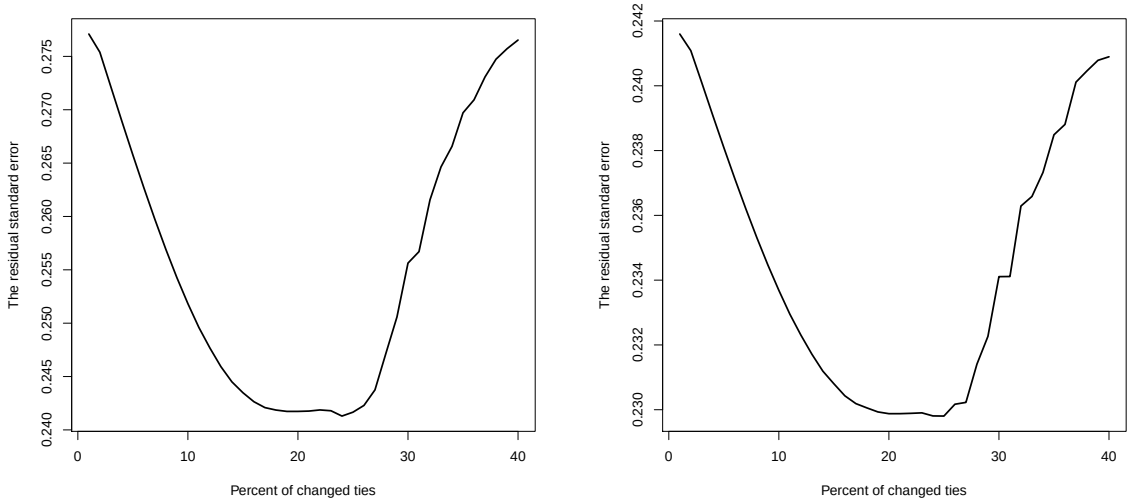
Table B.4: Pearson correlation coefficients between indices of network properties and indices of stability of blockmodels for the first non-symmetric blockmodel structure

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	ARI	1.00	-0.80	-0.83	-0.37	-0.30	-0.34	-0.29	-0.41	-0.23	-0.25	-0.27	-0.69	-0.69	-0.70	0.08	-0.64	-0.70	0.69	0.65	0.69	0.76	0.71	0.74	0.32	0.76	0.76
2	ErrB	-0.80	1.00	0.68	0.25	0.24	0.25	0.25	0.33	0.15	0.20	0.18	0.56	0.59	0.53	-0.09	0.45	0.55	-0.58	-0.49	-0.58	-0.64	-0.54	-0.63	-0.26	-0.66	-0.60
3	p.changed	-0.83	0.68	1.00	0.46	0.43	0.43	0.49	0.55	0.40	0.43	0.44	0.84	0.85	0.82	-0.03	0.72	0.82	-0.79	-0.71	-0.79	-0.84	-0.76	-0.82	-0.49	-0.83	-0.81
4	Dens	-0.37	0.25	0.46	1.00	0.18	0.63	0.24	0.68	0.54	0.43	0.61	0.71	0.41	0.67	0.08	0.69	0.45	-0.29	-0.50	-0.29	-0.35	-0.50	-0.26	-0.18	-0.27	-0.48
5	Rec	-0.30	0.24	0.43	0.18	1.00	0.67	0.65	0.24	0.16	0.18	0.17	0.39	0.40	0.35	0.04	0.28	0.33	-0.32	-0.28	-0.32	-0.33	-0.29	-0.32	-0.26	-0.34	-0.30
6	D_Mut	-0.34	0.25	0.43	0.63	0.67	1.00	0.26	0.31	0.24	0.17	0.30	0.57	0.40	0.54	0.05	0.48	0.32	-0.29	-0.39	-0.29	-0.33	-0.40	-0.28	-0.19	-0.29	-0.40
7	D_Asym	-0.29	0.25	0.49	0.24	0.65	0.26	1.00	0.55	0.36	0.40	0.36	0.48	0.49	0.39	0.06	0.34	0.46	-0.36	-0.26	-0.36	-0.35	-0.27	-0.37	-0.32	-0.38	-0.30
8	D_Null	-0.41	0.33	0.55	0.68	0.24	0.31	0.55	1.00	0.53	0.50	0.57	0.65	0.50	0.59	0.04	0.58	0.54	-0.37	-0.43	-0.37	-0.42	-0.45	-0.37	-0.28	-0.39	-0.46
9	PP_e	-0.23	0.15	0.40	0.54	0.16	0.24	0.36	0.53	1.00	0.93	0.99	0.50	0.41	0.50	0.19	0.56	0.56	-0.33	-0.28	-0.33	-0.27	-0.28	-0.28	-0.33	-0.26	-0.27
10	CCout_e	-0.25	0.20	0.43	0.43	0.18	0.17	0.40	0.50	0.93	1.00	0.88	0.49	0.54	0.40	0.15	0.42	0.63	-0.31	-0.27	-0.31	-0.31	-0.26	-0.30	-0.37	-0.29	-0.27
11	CCin_e	-0.27	0.18	0.44	0.61	0.17	0.30	0.36	0.57	0.99	0.88	1.00	0.56	0.43	0.57	0.19	0.63	0.58	-0.36	-0.35	-0.36	-0.31	-0.34	-0.30	-0.33	-0.29	-0.33
12	Dall_e	-0.69	0.56	0.84	0.71	0.39	0.57	0.48	0.65	0.50	0.49	0.56	1.00	0.82	0.88	0.04	0.80	0.78	-0.67	-0.68	-0.67	-0.83	-0.72	-0.70	-0.42	-0.72	-0.75
13	Dout_e	-0.69	0.59	0.85	0.41	0.40	0.40	0.49	0.50	0.41	0.54	0.43	0.82	1.00	0.65	-0.03	0.52	0.87	-0.66	-0.67	-0.66	-0.76	-0.72	-0.70	-0.45	-0.71	-0.74
14	Din_e	-0.70	0.53	0.82	0.67	0.35	0.54	0.39	0.59	0.50	0.40	0.57	0.88	0.65	1.00	0.06	0.92	0.68	-0.75	-0.67	-0.75	-0.74	-0.70	-0.77	-0.37	-0.75	-0.73
15	B_e	0.08	-0.09	-0.03	0.08	0.04	0.05	0.06	0.04	0.19	0.15	0.19	0.04	-0.03	0.06	1.00	0.12	0.04	-0.02	-0.06	-0.02	0.04	0.00	0.07	-0.39	0.08	0.02
16	A_e	-0.64	0.45	0.72	0.69	0.28	0.48	0.34	0.58	0.56	0.42	0.63	0.80	0.52	0.92	0.12	1.00	0.69	-0.62	-0.67	-0.62	-0.67	-0.70	-0.62	-0.30	-0.65	-0.72
17	H_e	-0.70	0.55	0.82	0.45	0.33	0.32	0.46	0.54	0.56	0.63	0.58	0.78	0.87	0.68	0.04	0.69	1.00	-0.62	-0.71	-0.62	-0.75	-0.77	-0.65	-0.40	-0.67	-0.83
18	PP_cor	0.69	-0.58	-0.79	-0.29	-0.32	-0.29	-0.36	-0.37	-0.33	-0.31	-0.36	-0.67	-0.66	-0.75	-0.02	-0.62	-0.62	1.00	0.50	1.00	0.75	0.54	0.95	0.46	0.89	0.59
19	CCout_cor	0.65	-0.49	-0.71	-0.50	-0.28	-0.39	-0.26	-0.43	-0.28	-0.27	-0.35	-0.68	-0.67	-0.67	-0.06	-0.67	-0.71	0.50	1.00	0.50	0.68	0.95	0.49	0.32	0.52	0.87
20	CCin_cor	0.69	-0.58	-0.79	-0.29	-0.32	-0.29	-0.36	-0.37	-0.33	-0.31	-0.36	-0.67	-0.66	-0.75	-0.02	-0.62	-0.62	1.00	0.50	1.00	0.76	0.54	0.95	0.46	0.89	0.59
21	Dall_cor	0.76	-0.64	-0.84	-0.35	-0.33	-0.33	-0.35	-0.42	-0.27	-0.31	-0.31	-0.83	-0.76	-0.74	0.04	-0.67	-0.75	0.75	0.68	0.76	1.00	0.73	0.81	0.44	0.84	0.79
22	Dout_cor	0.71	-0.54	-0.76	-0.50	-0.29	-0.40	-0.27	-0.45	-0.28	-0.26	-0.34	-0.72	-0.72	-0.70	0.00	-0.70	-0.77	0.54	0.95	0.54	0.73	1.00	0.54	0.30	0.57	0.93
23	Din_cor	0.74	-0.63	-0.82	-0.26	-0.32	-0.28	-0.37	-0.37	-0.28	-0.30	-0.30	-0.70	-0.70	-0.77	0.07	-0.62	-0.65	0.95	0.49	0.95	0.81	0.54	1.00	0.42	0.96	0.62
24	B_cor	0.32	-0.26	-0.49	-0.18	-0.26	-0.19	-0.32	-0.28	-0.33	-0.37	-0.33	-0.42	-0.45	-0.37	-0.39	-0.30	-0.40	0.46	0.32	0.46	0.44	0.30	0.42	1.00	0.39	0.30
25	A_cor	0.76	-0.66	-0.83	-0.27	-0.34	-0.29	-0.38	-0.39	-0.26	-0.29	-0.29	-0.72	-0.71	-0.75	0.08	-0.65	-0.67	0.89	0.52	0.89	0.84	0.57	0.96	0.39	1.00	0.66
26	H_cor	0.76	-0.60	-0.81	-0.48	-0.30	-0.40	-0.30	-0.46	-0.27	-0.27	-0.33	-0.75	-0.74	-0.73	0.02	-0.72	-0.83	0.59	0.87	0.59	0.79	0.93	0.62	0.30	0.66	1.00

Table B.5: Pearson correlation coefficients between indices of network properties and indices of stability of blockmodels for the second non-symmetric blockmodel structure

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	ARI	1.00	-0.76	-0.84	-0.41	-0.25	-0.37	-0.28	-0.56	-0.24	-0.31	-0.25	-0.64	-0.65	-0.73	0.07	-0.69	-0.64	0.69	0.49	0.69	0.77	0.59	0.78	0.40	0.79	0.76
2	ErrB	-0.76	1.00	0.66	0.31	0.18	0.28	0.21	0.47	0.15	0.21	0.16	0.51	0.53	0.57	-0.07	0.51	0.47	-0.53	-0.37	-0.53	-0.61	-0.45	-0.60	-0.31	-0.61	-0.60
3	p.changed	-0.84	0.66	1.00	0.57	0.29	0.46	0.44	0.73	0.43	0.49	0.44	0.81	0.80	0.89	0.03	0.84	0.76	-0.77	-0.65	-0.77	-0.84	-0.76	-0.83	-0.54	-0.83	-0.84
4	Dens	-0.41	0.31	0.57	1.00	0.41	0.82	0.72	0.84	0.86	0.93	0.87	0.87	0.84	0.73	0.18	0.79	0.68	-0.39	-0.42	-0.39	-0.44	-0.47	-0.42	-0.34	-0.42	-0.44
5	Rec	-0.25	0.18	0.29	0.41	1.00	0.74	0.11	0.25	0.31	0.37	0.31	0.37	0.41	0.31	0.09	0.31	0.31	-0.23	-0.19	-0.23	-0.24	-0.21	-0.24	-0.19	-0.24	-0.21
6	D_Mut	-0.37	0.28	0.46	0.82	0.74	1.00	0.29	0.55	0.65	0.74	0.65	0.72	0.73	0.58	0.09	0.61	0.53	-0.33	-0.29	-0.33	-0.37	-0.34	-0.37	-0.25	-0.37	-0.35
7	D_Asym	-0.28	0.21	0.44	0.72	0.11	0.29	1.00	0.76	0.70	0.71	0.71	0.64	0.58	0.55	0.19	0.63	0.56	-0.28	-0.38	-0.28	-0.33	-0.41	-0.29	-0.29	-0.30	-0.36
8	D_Null	-0.56	0.47	0.73	0.84	0.25	0.55	0.76	1.00	0.64	0.70	0.64	0.86	0.83	0.78	0.12	0.77	0.67	-0.51	-0.51	-0.51	-0.57	-0.58	-0.54	-0.42	-0.55	-0.59
9	PP_e	-0.24	0.15	0.43	0.86	0.31	0.65	0.70	0.64	1.00	0.96	1.00	0.72	0.66	0.64	0.27	0.73	0.60	-0.29	-0.36	-0.30	-0.31	-0.39	-0.28	-0.31	-0.27	-0.30
10	CCout_e	-0.31	0.21	0.49	0.93	0.37	0.74	0.71	0.70	0.96	1.00	0.96	0.79	0.76	0.68	0.21	0.76	0.67	-0.29	-0.41	-0.29	-0.36	-0.45	-0.31	-0.32	-0.32	-0.37
11	CCin_e	-0.25	0.16	0.44	0.87	0.31	0.65	0.71	0.64	1.00	0.96	1.00	0.73	0.66	0.65	0.27	0.73	0.61	-0.30	-0.37	-0.30	-0.31	-0.40	-0.29	-0.32	-0.28	-0.31
12	Dall_e	-0.64	0.51	0.81	0.87	0.37	0.72	0.64	0.86	0.72	0.79	0.73	1.00	0.93	0.92	0.07	0.91	0.77	-0.62	-0.53	-0.62	-0.75	-0.62	-0.68	-0.46	-0.68	-0.66
13	Dout_e	-0.65	0.53	0.80	0.84	0.41	0.73	0.58	0.83	0.66	0.76	0.66	0.93	1.00	0.84	0.06	0.83	0.83	-0.57	-0.61	-0.57	-0.69	-0.72	-0.63	-0.44	-0.64	-0.73
14	Din_e	-0.73	0.57	0.89	0.73	0.31	0.58	0.55	0.78	0.64	0.68	0.65	0.92	0.84	1.00	0.06	0.95	0.74	-0.75	-0.56	-0.75	-0.79	-0.65	-0.82	-0.49	-0.81	-0.72
15	B_e	0.07	-0.07	0.03	0.18	0.09	0.09	0.19	0.12	0.27	0.21	0.27	0.07	0.06	0.06	1.00	0.10	0.09	-0.06	-0.27	-0.06	-0.01	-0.15	0.03	-0.44	0.04	0.01
16	A_e	-0.69	0.51	0.84	0.79	0.31	0.61	0.63	0.77	0.73	0.76	0.73	0.91	0.83	0.95	0.10	1.00	0.83	-0.71	-0.55	-0.71	-0.75	-0.64	-0.78	-0.46	-0.79	-0.70
17	H_e	-0.64	0.47	0.76	0.68	0.31	0.53	0.56	0.67	0.60	0.67	0.61	0.77	0.83	0.74	0.09	0.83	1.00	-0.57	-0.63	-0.57	-0.67	-0.76	-0.61	-0.40	-0.64	-0.83
18	PP_cor	0.69	-0.53	-0.77	-0.39	-0.23	-0.33	-0.28	-0.51	-0.29	-0.29	-0.30	-0.62	-0.57	-0.75	-0.06	-0.71	-0.57	1.00	0.48	1.00	0.78	0.55	0.90	0.49	0.88	0.63
19	CCout_cor	0.49	-0.37	-0.65	-0.42	-0.19	-0.29	-0.38	-0.51	-0.36	-0.41	-0.37	-0.53	-0.61	-0.56	-0.27	-0.55	-0.63	0.48	1.00	0.48	0.54	0.87	0.47	0.54	0.48	0.68
20	CCin_cor	0.69	-0.53	-0.77	-0.39	-0.23	-0.33	-0.28	-0.51	-0.30	-0.29	-0.30	-0.62	-0.57	-0.75	-0.06	-0.71	-0.57	1.00	0.48	1.00	0.79	0.55	0.91	0.49	0.88	0.63
21	Dall_cor	0.77	-0.61	-0.84	-0.44	-0.24	-0.37	-0.33	-0.57	-0.31	-0.36	-0.31	-0.75	-0.69	-0.79	-0.01	-0.75	-0.67	0.78	0.54	0.79	1.00	0.64	0.87	0.55	0.86	0.74
22	Dout_cor	0.59	-0.45	-0.76	-0.47	-0.21	-0.34	-0.41	-0.58	-0.39	-0.45	-0.40	-0.62	-0.72	-0.65	-0.15	-0.64	-0.76	0.55	0.87	0.55	0.64	1.00	0.57	0.49	0.58	0.84
23	Din_cor	0.78	-0.60	-0.83	-0.42	-0.24	-0.37	-0.29	-0.54	-0.28	-0.31	-0.29	-0.68	-0.63	-0.82	0.03	-0.78	-0.61	0.90	0.47	0.91	0.87	0.57	1.00	0.46	0.98	0.69
24	B_cor	0.40	-0.31	-0.54	-0.34	-0.19	-0.25	-0.29	-0.42	-0.31	-0.32	-0.32	-0.46	-0.44	-0.49	-0.44	-0.46	-0.40	0.49	0.54	0.49	0.55	0.49	0.46	1.00	0.43	0.41
25	A_cor	0.79	-0.61	-0.83	-0.42	-0.24	-0.37	-0.30	-0.55	-0.27	-0.32	-0.28	-0.68	-0.64	-0.81	0.04	-0.79	-0.64	0.88	0.48	0.88	0.86	0.58	0.98	0.43	1.00	0.72
26	H_cor	0.76	-0.60	-0.84	-0.44	-0.21	-0.35	-0.36	-0.59	-0.30	-0.37	-0.31	-0.66	-0.73	-0.72	0.01	-0.70	-0.83	0.63	0.68	0.63	0.74	0.84	0.69	0.41	0.72	1.00

C Piecewise regression models with *p.changed* ties as a predictor



(a) Mean of the Adjusted Rand Index, $mARI$ (b) Mean of Incorrect block types, $ErrB$

Figure C.1: The residual errors plots for determining the break in piecewice regression models for boy-girl liking ties network

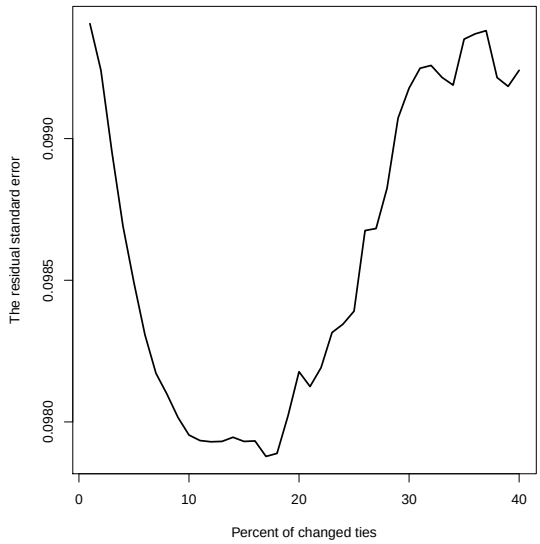
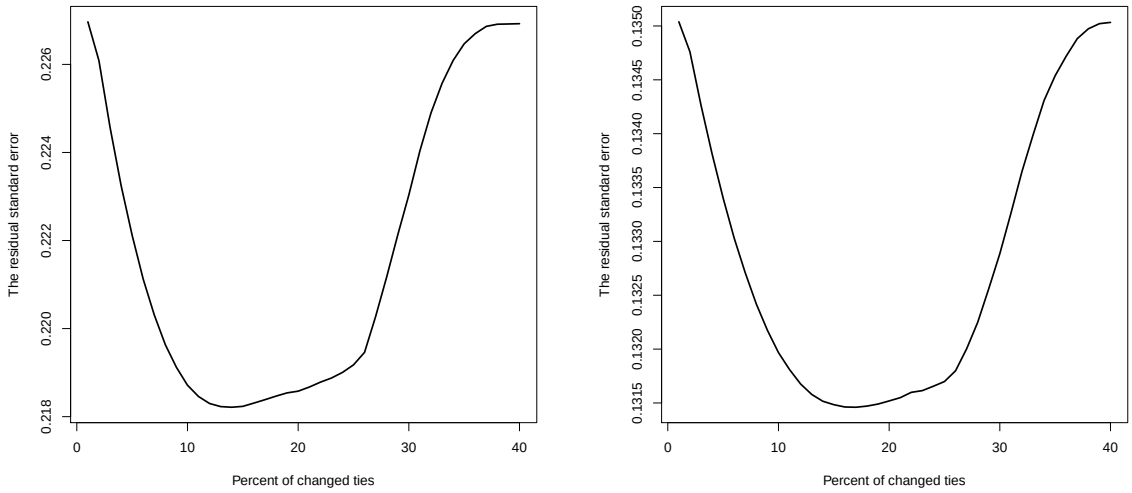
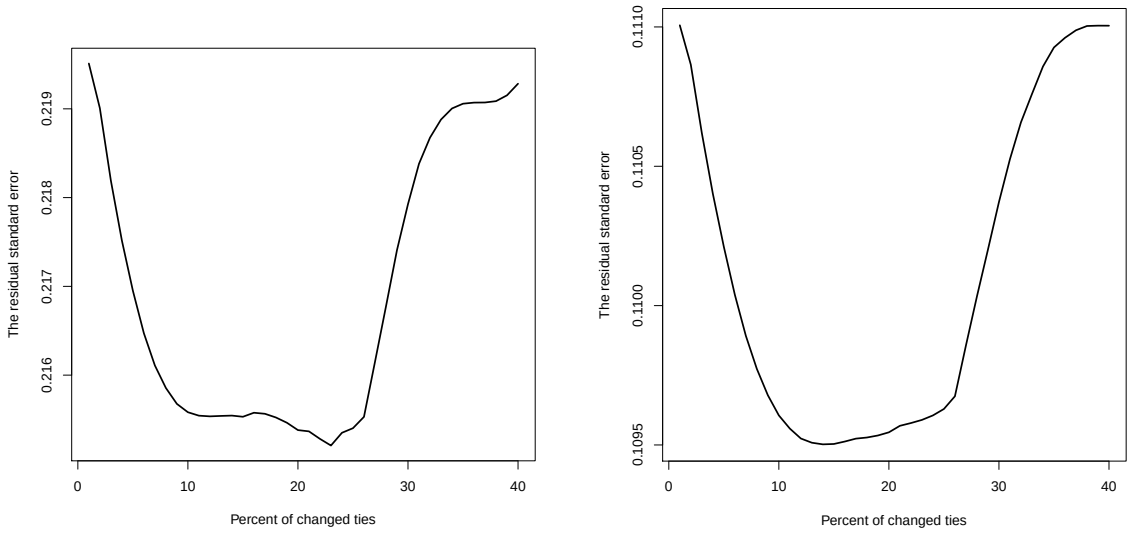


Figure C.2: The residual errors plots for determining the break in piecewice regression models for $ErrB$ for the note borrowing network



(a) Mean of the Adjusted Rand Index, ARI (b) Mean of Incorrect block types, $ErrB$

Figure C.3: The residual errors plots for determining the break in piecewice regression models for the first non-symmetric blockmodel structure



(a) Mean of the Adjusted Rand Index, ARI (b) Mean of Incorrect block types, $ErrB$

Figure C.4: The residual errors plots for determining the break in piecewice regression models for the second non-symmetric blockmodel structure