REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY) 29-11-2016	2. REPORT TYPE Final Report	3. DATES COVERED (From - To) 16-May-2012 - 15-May-2015

4. TITLE AND SUBTITLE
Systematic Investigation of the Effects of Missing Data on Statistical Models for Networks

5a. CONTRACT NUMBER
W911NF-12-1-0176

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHORS
Harold Green, Stanley Wasserman

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Indiana University at Bloomington
509 East 3rd ST

Bloomington, IN 47401 -3654

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
60734-NS.3

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

There is little published research on the impact of network sampling on network structures and structural measures. There is even less research that investigates how network sampling impacts models that link network structure with behaviors and attitudes. We need not only to study how sampling designs for social network studies impact network measures, but also investigate how these sampling designs impact models that include social influence effects. The old studies of measurement error in network analysis (circa 1975) are woefully old. This study, thus, contributes much needed rarely discussed but important information to a rapidly growing field.

15. SUBJECT TERMS
networks, statistical models

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Stanley Wasserman
UU	UU	UU	UU		19b. TELEPHONE NUMBER 812-856-7828

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

RPPR Final Report
as of 13-Oct-2022

Agency Code: 21XD

Proposal Number: 60734NS **Agreement Number: W911NF-12-1-0176**

INVESTIGATOR(S):

Name: Stanley Wasserman Phd
Email: stanwass@indiana.edu
Phone Number: 8128567828
Principal: Y

Organization: **Indiana University at Bloomington**
Address: 509 East 3rd ST, Bloomington, IN 474013654
Country: USA
DUNS Number: 006046700 EIN: 356001673
Report Date: 15-Feb-2013 Date Received: 29-Nov-2016
Final Report for Period Beginning 16-May-2012 and Ending 15-May-2015
Title: Network Analysis: Effects of Sampling and Measurement Error on Models and Statistics
Begin Performance Period: 16-May-2012 **End Performance Period:** 15-Aug-2015
Report Term: 0-Other
Submitted By: Email:
Phone:
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: **STEM Participants:**

Major Goals:

Accomplishments: This research was done with several Rand Corp team members:

Kayla de la Haye, Josh EmBree, Matthew Hoover, and Marc PunKay

Our scientific progress and accomplishments are discussed at length in the attachment.

Training Opportunities:

Results Dissemination:

Honors and Awards: none

Protocol Activity Status:

Technology Transfer:

RPPR Final Report
as of 13-Oct-2022

Partners

,

I certify that the information in the report is complete and accurate:
Signature:
Signature Date:

Systematic Investigation of the Effects of Missing Data on Statistical Models for Networks

Final Report for the US Army Research Office, for award W911NF-12-1-0176, WASSERMAN

Edward T. Palazzolo, Program Officer, Network Science Program, Army Research Office

Principal Investigators:

Harold D. Green, Jr.
Rand Corporation
hgreen@rand.org

Stanley Wasserman
Indiana University
stanwass@indiana.edu

Co-Investigators

Kayla de la Haye
University of Southern California
delahaye@usc.edu

Josh EmBree,
RAND Corporation
jembree@rand.org

Matthew Hoover,
Gallup, Inc.
matthew.a.hoover@gmail.com

Marc PunKay
University of Chicago Urban Labs
mopunkay@gmail.com

Introduction and Research Questions

Many have argued that theoretical perspectives and promising findings from social network analysis are an important influence on contemporary social and behavioral sciences; while at the same time the current theoretical and empirical convergence across the social and behavioral sciences have strengthened network science. Network studies facilitate direct behavioral intervention by pinpointing how human relationships encourage or discourage attitudes, actions, and behaviors. Social network analysis (SNA) provides important tools for identifying and understanding the social and contextual factors relevant to engagement in particular behaviors.[1-4] By quantifying relational information and linking it to human behavior, many important quantitative methods, such as matrix algebra, graph theory and statistical analysis, can be applied to identify structural patterns in social networks and measure the association of those patterns with various behavioral outcomes.[5]

Sophisticated techniques have been created to analyze many kinds of network data.[6-9] However, the impact of *network-based study designs* on methodologies that apply social network analysis to understand specific substantive questions is largely unknown. From empirical research, social network theory, and methodology, we know quite a bit about the statistical properties of networks measured where data are gathered on every social actor about every other social actor on all relations. Yet such knowledge is lacking for social network studies where network data sets are incomplete; that is, when sociocentric network information is missing (by design or by chance) or when it is studied at the local or personal level. There is a need for basic methodological research on the effect of missing data on substantive conclusions drawn from analysis. That was the goal of this study.

Network sampling design and measurement strategies tend to correspond with study size: Larger studies use social support scales, smaller studies use network batteries, and the smallest studies use roster approaches. However, study size should not be the sole factor influencing the choice of network study design. Clearly, one major concern in network research is optimal sampling of individual actors while, at the same time, gathering relevant and adequate information on relational ties.

Random sampling (taking many possible forms) is a standard approach in cross-sectional and longitudinal social and behavioral research. Unfortunately, it is simply not appropriate for a network-based study. Sampled individuals are unlikely to have relational ties to each other.[10-13] Consequently, network researchers have developed a variety of sampling techniques specific to studies containing network measurements. The 'complete' network approach involves collecting data from every member of a target population. Although very common for small, closed populations, it is not feasible for large-scale public studies largely due to resource limitations. Snowball sampling (more generally respondent driven sampling) and link-tracing—are more applicable to large-scale network studies but less frequently implemented, requiring fewer resources but having largely unknown effects on the measurement of network properties and model fits.[11-14] Personal and ego-centric network designs,[15,16] in which respondents are asked to report on their relational ties and on their perceptions of social structure among network alters, are most often implemented in large-scale studies, but their relationship to actual social structures is largely unknown and generate a great deal of what is, essentially, missing data.[17,18]

Modeling sampled data can be complex, since much of network analysis is based on the assumption that data are from complete networks with no missing values.[14,19-23] Any other data collection strategy may introduce bias into parameter estimates or yield inaccurate models. Currently, it is unknown which error is introduced or how it may affect a study's outcomes.[13] Few systematic investigations have examined how sampling approaches and the associated data loss affect models that link behaviors to the social context. Those studies primarily use simulations based on hypothetical or randomly generated networks and behavioral data to compare competing strategies of network data collection. Clearly, more studies that examine the reliability of network indices and of cross-sectional and longitudinal analyses that incorporate them are needed to contribute to our understanding of this problem, which is likely to become more significant as interest in network studies continues to grow and the studies themselves become larger and more ambitious. Even more important is the use of empirical data collected in real-world research contexts so that we understand what can happen when data are missing from actual social network and behavioral data sets. Researchers need guidance on the feasibility and advantages or disadvantages of different sampling designs, similar to previous work completed by Eudey, Johnson, and Schade (1994) that rated and ranked network nomination tasks to provide guidance to researchers on the pros and cons of various network data collection approaches.[24]

Analytic techniques for complete networks are more extensive than techniques for sampled or local networks.[6-8] The traditional complete network tradition argues that the structural properties of networks combined with attributes of the network members (including behavioral measures) map the entire social environment (or the social system, which might be complex). The standards of data collection are quite different from and more difficult to achieve than traditional social and behavioral science studies.[10-12] Little is known, if anything, about how data missing at random in complete network studies affects estimates of individual and network-wide indices. Even the simplest indicators may be strongly biased and lead to erroneous findings. Relationships between network position and behaviors can be badly misinterpreted.[19] The robustness and resistance of standard network indicators to sample design and measurement error are mostly unaddressed in the methodological literature. Exactly how study design impacts estimates of network structure and composition (and subsequent models that include those estimates) is still unknown.

Many studies use complicated network data sets. Take, for example, the Framingham Heart Study, which was never intended for use in network analyses, but whose family and friend contact information can be considered relational data and thus can be massaged into a social network.[25] Such data sets are often sampled in various ways, but the effects of the inherent sampling design on the findings of the study are unknown. Nevertheless, the sampling design will certainly influence the results in some way. There is little published research on the impact of network sampling on network structures and structural measures. There is even less research that investigates how network sampling impacts models that link network structure with behaviors and attitudes. We need not only to study how sampling designs for social network studies impact network measures, but also investigate how these sampling designs impact models that include social influence effects. The old studies of measurement error in network analysis (circa 1975) are woefully old.[19] This study, thus, contributes much-needed rarely discussed but important information to a rapidly growing field. Specifically, we proposed to answer the following methodological questions:

1. How does eliminating links bias cross sectional and longitudinal parameter estimates and statistical models? Missing links approximate both fixed choice designs where only a portion of links are provided and situations where respondents do not report all relationships that exist.
2. How does eliminating alters and their links bias cross sectional and longitudinal parameter estimates and statistical models? Missing alters represent panel-type loss of entire groups from a study as well as intermittent joiners and leavers in studies (alters that appear at multiple, but not sequential time points, as if a student were enrolled in a study and appeared at baseline but then not again until the fourth and fifth waves of a five wave study).
3. How does the method of sampling bias estimates and models? That is, do missing edges lead to greater problems than missing nodes? Does “churn” in these models lead to greater problems than the loss of alters at baseline due to sampling issues? Churn here refers to the presence of intermittent joiners and leavers.
4. How do systematic sampling approaches such as snowball sampling or link-tracing bias estimates and models? These approaches approximate respondent driven sampling approaches and other ‘trace’ based designs.

5. How does the information that is lost when one has personal, rather than complete, network information bias estimates and models?
6. Does missingness in alter attribute variables have the same effect as missingness in structural variables?
7. How much does measurement error (caused, for example, by forced fixed-choice designs) affect statistical findings from these new models, particularly for social influence parameters?

Our analytic strategy consisted of three phases.

Phase 1. The first phase of our research focused on models of networks and behavior: co-evolutionary and cross-sectional. In particular, we focused on stochastic actor-oriented models (based on certain assumptions about co-evolutionary processes) fit using R-SIENA, and ERGM/p* models. These models allow estimation of parameters that represent influence and selection mechanisms in the case of co-evolutionary models and homophily in the case of cross-sectional models. With longitudinal data and dynamic statistical models, one can answer questions such as: How do those connected to an actor influence that actor's behavior, and how do the actor's attributes affect choice of friends? With cross-sectional data and statistical models, a researcher is restricted to questions concerning the association of structure in networks with homophily, such as: How similar is an ego/actor to his alters? If two actors have alters in common, are they more similar than any two actors chosen at random? If an actor is popular, is the actor more similar to the actor's alters? Such questions are at the heart of the study of network processes and individual behavior.

The substantive motivation for models of *co-evolution* of behavior and structure is presented in the work of Snijders and colleagues,[9,26] and, for example, in the applied work of Christakis and Fowler.[27] Co-evolutionary models incorporate parameters that assess the amount of selection operating in the social system, the degree to which individual actors tend to change their friendships based on their behavioral or attitudinal attributes, their alters' behavioral or attitudinal attributes, or some combination of the two. When combined with effects for social influence and structural evolution, we gain an understanding of how all of these mechanisms work in concert to drive social evolution.

Co-evolutionary models (here we focus on stochastic actor-based models) assume that an actor will change his or her behavior at a subsequent time point based on the behaviors or attitudes of those actors he or she is connected to (and the relative positions of those actors) at a previous time point. These are called social influence processes. These models also assume that the behaviors and attitudes of an actor (and his or her relative position in a network) will affect a decision to create, maintain, or dissolve a relational tie at a subsequent time point. These are called selection processes. Both endogenous structural and behavioral factors (and combination of those factors) can be included as 'explanatory' or predictor variables in these models. These variables may include the overall number of connections an actor reports, as reflected by an outdegree parameter; the number of reciprocal connections an actor has, as reflected by a 'reciprocity' parameter; the number of transitive triangles an actor is included in (how likely it is that ties exist because of tendencies toward triadic closure or structural balance) as reflected or quantified by a transitivity parameter; and the tendency for an actor to have relational ties to similar others ('birds of a feather flock together'), as quantified by a homophily parameter.

The motivation for cross-sectional models of the underlying selection and influence processes that link behaviors and attitudes is that homophily represents the result of these processes in action. These models are based on *p**, an exponential family of random graph distributions. The parameters associated with specific network statistics are designed to reflect the structure and composition of the network. The beauty of this family of models comes from the fact that its members:

- are statistical and have probabilistic interpretation;

- apply to the entire network; and
- are designed to reflect how network structure may affect behavior.

*p** models assume that specific characteristics of social structure affect the probability that two actors interact with each other. Both endogenous (structural) and exogenous (composition) factors can be included as ‘explanatory’ or predictor variables in these models. These variables may include the overall level of ‘connectivity’ in the network, as reflected by an ‘expansiveness’ outdegree parameter; the overall level of ‘mutuality’ in the network (how likely a relational tie is ‘reciprocated’ or returned) as reflected by a ‘reciprocity’ parameter; the overall transitivity of the network (how likely it is that ties exist because of tendencies toward triadic closure or structural balance) as reflected or quantified by a transitivity parameter; and the overall tendency for network members to have relational ties to similar others (‘birds of a feather flock together’), as quantified by a homophily parameter.

Phase 2. The goal of the second phase is the study of the effect of missing data and network “sampling” on relevant statistical models. Such questions are critically important, since most studies confront missing data problems but have no clear strategy for addressing the issues and no understanding of how the missing data may affect their findings. Furthermore, a large percentage of network studies can gather only sampled network data and thus must address this problem. In this phase, we first simulate data that have missing edges and missing nodes. Then, we will simulate two network sampling designs on the “population”: snowball sampling, and ego-centered network sampling, taking repeated samples as described previously. Respondent driven sampling is a general network sampling scheme, and includes snowball sampling as a special case in which all alters named are contacted. Personal network sampling generates the egonets of a sample of individuals in a population, allowing a researcher to focus on the network neighborhoods that surround the sampled individuals. Sampling from cross-sectional data is easily achieved for each approach. For longitudinal data, missing nodes are chosen in two ways: once at baseline to represent complete loss to the study and also at each wave, to represent turnover or “churn” present in longitudinal studies. Further details on the sampling approaches are provided below. Our plan for the second phase of this research involves fitting statistical models to these simulated sampled network data. The fitted models will be identical to those estimated in Phase 1. However, in Phase 2, we will test hypotheses using statistical models fit to just the sampled network data. Specifically, once we have created multiple simulated sampled network data sets, we will complete the same modeling as described for Aim 1 above.

Phase 3. In this last phase, the assumed models based on the complete network data from Phase 1 become the target. We then compare the target models to models fit to datasets sampled from the complete network in Phase 2. Specifically, we used the models previously estimated as a benchmark to study how sampling might affect network parameters and the models that incorporate them. We will study where errors happen when the same models are fit to partial networks sampled from the benchmark data.

Analytic Strategy

Data and Modeling

The original data were constructed from three waves of surveys designed to collect friendship nominations and individual covariates from a group of 190 students. At each wave, each student was asked questions about his or her recent behavior and current friendships. The responses to these questions were used to construct a social network with three relations and several individual covariates. Note that there were some missing values in the responses. A few students did not complete all three waves of the survey. We focused on data from the University of Illinois Bullying and Sexual Violence Study[28] and Context of Adolescent Risk Behaviors in Networks (CARBIN) Study.[29]

Entire cohorts of middle school students (6th through 8th graders) were recruited in 3 Illinois public middle schools. In-school surveys were administered biannually/annually through high school

(2008 through 2015). The surveys included measures of substance use, bullying, violence, friendship nominations, school performance, demographics. For the SABMs, the analytic sample for this study includes 3 waves of data from 1 middle school. For the ERGM/p* models we focus only on the first wave of data. At wave one, the sample size was 190. The sample was 47% male, an average of 12.1 years old, with students in grades 6-8. The sample was 58% Black, 31% Non-Hispanic White, and 6% Hispanic.

Respondents were asked to name up to 8 friends in their school. These friendship nominations form the basis for our student networks. The key outcome variables were bullying perpetration and homophobic bullying perpetration. Bullying perpetration (BP) was measured with a 9 item scale that assessed past month behaviors with responses ranging from 1 (never) to 5 (seven or more times), including items such as: I upset others for the fun of it. Homophobic bullying perpetration (HBP) was measured with a 5 item scale that also assessed past month behaviors with responses again ranging from 1 (never) to 5 (seven or more times), including items such as: 'I called a friend a homo, gay, lesbo, fag, or dyke'. Covariates included measures of impulsivity; control and dominance; traditional masculinity; femininity; attitudes toward bullying; and neighborhood violence. We also collected demographic information: gender, race/ethnicity, grade level at school, and parents' highest level of education. These variables were used to build well-fitting SABMs and ERGM/p* models that became the targets models for those built in Phase 2.

First, stochastic actor-oriented models (*SABMs*) and Exponential Random Graph Models (ERGM/p*) were fit to this complete set of data. SABMs were estimated using the *RSIENA* software library in R. As the exact nature of the parameters is not particularly germane to the analytic strategy of this study, the complete set of parameters in the final models are only briefly described in further detail in the results section of this report. The *RSIENA* software estimates parameters using complex simulation routines based on the observed network statistics. It is common for particular SABM models to fail to reach convergence which causes the specified parameters to be estimated inaccurately.

Moreover, it is often unclear exactly how missing values are handled in the estimation process as they are sometimes imputed and included in estimation but not in simulation. A primary aim of this work was to explore the empirical limitations of the estimation procedures present in the RSIENA software with various types and levels of missing data without the assistance of imputation, thus we forced missing data when necessary to assess the impact on the models without any amelioration via imputation. For each type and level of missing data detailed below, 10 datasets were randomly generated and used to estimate the same set of parameters specified in the original benchmark model. The results from each of these runs are summarized in the relevant results section.

Next, *ERGM/p**models were fit to the complete data set using the *statnet* software package. Missing data were forced for the ERGM models as well, though missing data in these cross-sectional models is more easily understood and controlled. The complete set of parameters in this final model is presented in further detail in the results section of this report. *Statnet* uses count statistics and similarly complex simulation routines to estimate parameters in a set of conditional models. Convergence may be an issue in models with too many parameters. For each type and level of missing data below (with one exception), 10 datasets were randomly generated and used to estimate the same set of parameters specified in the original model. The results of these runs are summarized in the relevant results section.

1. Missing edges (90%, 70%, 50%, 30% complete)

To simulate networks with missing edges at various levels, we randomly select edges in the social network at each wave and remove them. For example, if there were 500, 600, and 400 edges (or friendship nominations) in each wave respectively and we simulate at the 90% complete level, then 10% of the edges in each wave will be randomly selected and removed from the data. That is, we remove 50, 60, and 40 edges from the edge list in each wave respectively.

It is important to note that there are two possible methods for recording missing edges in social network data. The first option is to change the randomly chosen edges from one to missing in the sociomatrix. Alternatively, the randomly chosen edges can be changed from one to zero in the sociomatrix, which is also equivalent to removing said edges from the edge list. Note that in the first option, the sociomatrix will be a mixture of ones, zeros, and missing values whereas the second option yields a traditional sociomatrix with only one and zero values. In practice, it is unreasonable to distinguish between missing edges and edges that truly do not exist. As a result, we simulate missing edges as non-existent rather than as missing data.

2. Missing nodes (90%, 70%, 50%, 30% complete)

To simulate networks with missing nodes at various levels, we randomly select a particular number of actors from the complete actor set and then remove them completely. For example, if there were 100 actors and we simulate at the 90% complete level, then 10 actors will be randomly chosen and removed along with all associated edges at every wave. Furthermore, relevant covariates associated with the nodes are removed from the data set.

Again, there are two possible methods for simulating missing nodes in social network data. The first method is to set all edges to and from the missing actors to missing in the sociomatrix. That is if actor j is missing, then row j and column j in the sociomatrix is comprised of only missing values. Alternatively, the missing actors can be completely removed from the data and disregarded. The latter appears to be much more common in practice. Imposing the missing values in the former method requires knowledge of the complete population. For example, if social network data is collected in a small class of students and some of them opt out, then those students may be included in the data with missing values. However, in cases when the number of missing actors is unclear, it is extremely difficult to justify this method. Additionally, imposing large amounts of missing data in the sociomatrix considerably complicates model estimation. Therefore, we chose the latter method for simulating missing nodes, removing the nodes and their associated edges completely rather than coding them as missing.

3. Missing nodes with churn (90%, 70%, 50%, 30% complete)

To simulate networks with missing nodes with churn at various levels, we randomly select nodes at each wave to remove. That is, a different set of nodes is removed from each wave. If a node is removed from every wave, it is treated as if it were never observed. However, if a node is removed from one or more waves but not every wave, then the ties associated with that node are set to missing when the node is removed for that particular wave. Furthermore, non-varying covariates for these actors are preserved across all waves while varying covariates are set to missing in waves when the actors are missing. For example, if there were 100 actors and we simulate at the 90% level, then 10 difference actors will be randomly selected and removed at each wave. If two actors were randomly chosen at all three waves, then those two are completely removed while the others will have their associated ties and varying covariates set to missing only for the waves they were removed from.

This scenario is very common empirically since it is difficult to isolate static group of people and follow them over multiple time periods. It is far more likely to observe some degree of churn (actors entering/exiting the social group of interest) as the number of waves increases. Additionally, there is an important distinction to make when documenting churn in network data: actors can either be missing due to non-response or missing due to churn. For example, a student who opts out of taking a survey but still attends the school is a non-respondent while a student who moved to another school can be considered churn. We have assumed that when an actor is missing due to non-response, he/she is still a member of the social group and therefore can receive nominations for social ties while his/her nominations should be regarded as missing. These students are still members of the school population.

Conversely, an actor who is missing due to churn should have all of his/her in/out ties regarded as missing or structurally non-existent. This student is not considered a member of the school population for that wave (because they may have moved away, etc.). These were our assumptions for modeling this type of turnover and making it as similar to the realities of collecting this type of data as we could.

4. Personal networks (90%, 70%, 50%, and 30% complete)

To simulate networks constructed through personal network sampling at various levels, we randomly select a set of actors and then use the personal networks from those actors at every wave to construct the overall network. That is, if there were 100 actors and we simulate at the 90% level, then 90 actors are randomly chosen. At each wave, any other actor nominated by at least one of the 90 selected actors will be included in the sample. To fill out the sampled network, first all nominations from the 90 selected actors are included at each wave. Second, any existing nomination between two actors in a personal network will also be included. That is, if actor A is in the sample, A nominated B and C, and B nominated C, then the tie from B to C will be included in the sample. However, if actor B nominated actor D but none of the sampled actors nominated both B and D, then the tie from B to D will not be included in the sample.

The personal networks sampling strategy is popular because it is less demanding than a complete network approach where every actor must be surveyed. In practice, the sampled actors are asked to nominate other actors as social ties and to identify which of their nominations have social ties between them as well. For example, a participant might be asked who all of his/her friends are and who amongst their friends are friends with one another. We simulate this process with the strategy described above. Note that in this strategy we assume that actors can correctly identify which of their nominations would nominate one another.

5. Link tracing/snowball sampling (5% seeds, 10% seeds and 90%, 70%, 50%, 30% complete edges)

To simulate networks constructed through link tracing (snowball sampling), we specify either 5% or 10% of the actors as seeds and then perform link tracing style sampling until the specified proportion of complete edges is reached in the first wave and then use the set of actors obtained through this strategy in each subsequent wave. For example, suppose that there were 100 actors, we use a 10% seed, and want 90% complete edges. Then 10 actors are randomly chosen from the first wave and we include all of their nominations as edges. If that number of edges is less than 90% of the overall number of edges in the complete first wave, then we include all of the nominated actors from the original set of 10 along with all of their nominations as edges. We continue incorporating actors and edges in this manner until the desired level is met. Once the desired level is met in the first wave, the unique set of actors obtained is carried through in each subsequent wave and any actor not included is completely removed as if not observed at all. That is, there are no tie variables or covariates set to missing in this framework. Note that if there are large, isolated components of the network, then it may not be possible to obtain higher proportions of complete edges through link tracing depending on the initial seed.

Link tracing is commonly used when the population of actors is not well defined or is hard to reach for surveying. In such cases, it is often more practical to recruit an initial set of actors (randomly chosen or otherwise) and then use their social connections to recruit more actors into the study. This can introduce bias depending on the initial seed and their connections. We simulate using this strategy to recruit a single set of actors and then observe them at two subsequent waves of data collection. While link tracing is more common for cross-sectional studies, this provides an interesting examination of the effect it may have on longitudinal social dynamics.

Comparisons

Phase three consisted of comparisons of simulated models against the ‘target’ models. It was our assumption that visual analytics provide the best and most informative approach to assessing parameter and model fit; so, we present figures that summarize much of our simulation data. First, we plotted estimates graphically such that target estimates represent the origin of an axis, with bars indicating 2 standard errors from the estimates (That is, we center each parameter at zero). Then, we plotted each of the ten estimates from a particular sampling type at a particular level of missing data. These points were sized by their significance level (or, how statistically large the estimates are). It is simple to visually assess whether a certain level of missing data was associated with more significant parameters located in areas where they shouldn’t be considered significant (similar to a Type 1 error) or more non-significant parameters within bounds that should be considered significant (similar to a Type 2 error). Furthermore, visual analytics allowed us to assess the range of parameter estimates for a particular sampling type and level of missing data. Taking these assessments together, we now discuss the overall goodness of fit of the model associated with the type/missing data combination.

Thus, including baseline models and ten randomly sampled data sets with missing data levels of 90%, 70%, 50%, and 30% across six missing-data states (randomly missing edges, randomly missing nodes, randomly missing nodes with churn, personal network type, snowball sampling with 5% of the population chosen as seeds, and snowball sampling with 10% of the population chosen as seeds), we estimated 264 coevolutionary models using RSIENA. Within the same framework, we also calculated 220 ERGM/p* models (the churn-based sampling strategy is irrelevant for cross-sectional data).

As before, we also created summary tables that enable comparison across sampling strategies. The sheer number of parameters made direct summary comparisons far too complicated and were not really relevant to the main goal of this study --- to assess, in general, how missing data affected these models and parameters across each sampling type. Thus, we decided to group parameters together into 'classes' for these summary tables. Longitudinal model parameters yield six classes: network rate parameters, structural parameters (e.g., reciprocity, triadic closure), selection parameters (e.g., covariate alter, covariate ego, and covariate same), behavioral rate parameters, behavioral parameters (e.g., linear shape, quadratic shape), influence parameters (e.g., average similarity), and covariate parameters (covariate effect on behavior). ERGM/p* parameters are grouped into three classes: structural parameters (e.g., mutuality, three-cycles, edgewise shared partners), homophily parameters (e.g., behavior match), and covariate parameters (e.g., gender match).

The tables shown below present the proportion of times that a sampled data set was considered significant though outside the confidence bounds of the original model and the number of times a parameter was not significant though inside those bounds. Because these tables were difficult to interpret, we also created table-like figures that shade each cell based on the magnitude of this proportion. We envision two users of these reports: researchers who want to know how a particular type of sampling will affect their models (across all classes) relative to the levels of missing data and those who are more statistically inclined, who want to know how a particular class of parameters (across all sampling types) fares relative to levels of missing data.

We also expect that there are researchers who will want figures that present gradually increasing levels of missingess across both types of error and researchers who will want to group the error types together. Thus we provide four summary figures that present these data.

Results

In brief, for ERGM models, even just 10% missing edges can lead to significant underestimation of network structural parameters (edges, mutual ties, edgewise shared partners, triangles) and of parameters that incorporate structural and behavioral information (same school grade, for example).

This small amount of missing data also substantially increases the standard deviation of the parameters. Missing *nodes* increase these problems for structural and structural/behavioral parameters, leading to a skew of parameter estimates across the range (both over- and under-estimation) and further increasing standard deviations. Missing nodes can also lead to poor estimation of behavioral parameters that are not affected by missing edges, reinforcing arguments for the need for complete participation from all members of a population. Figures and tables representing these modeling runs and assessing divergence from parameter estimates for ERGM models are included in the appendix.

SIENA models, of course reveal similar but more extreme patterns. Again, just 10% missing edges significantly impact parameter estimates (overestimating rates of friendship formation and underestimating the outdegree distribution, for example), though behavioral parameters remain fairly stable. When there are 10% missing edges, we see significant impact on behavioral parameter estimates as well. When we reach 70% we are, interestingly, underestimating rate parameters, doing very poorly on structural parameters (outdegree, reciprocity, triangles, etc.), and not very well in estimating influence or selection parameters. Once we reach 70% missing nodes, parameter estimates vary so widely across the range that it is difficult to assess if any systematic patterns are present. Figures and tables representing these modeling runs and assessing divergence from parameter estimates for SABM models are included in the appendix.

In these figures, circles are plotted relative to the parameter estimate for a particular draw. Placement on the horizontal axis represents actual deviations from the original parameter estimate (e.g., 0 is the original estimate), which is represented by the vertical line through the plot. The light gray bars underneath each row of parameters indicate plus/minus one standard error (estimated using the original models). The colors and horizontal white lines allow us to group similar parameters together (structural, behavioral, etc.). Circle size corresponds to p-values (in the legend) associated with tests of significance for a parameter estimate from a particular draw. The p-values are calculated using estimated standard errors (in the models calculated with sampled data) and assume large-sample normal distributions.

Baseline Model: SIENA

The baseline SIENA model included two behavioral outcome variables: bullying perpetration and homophobic bullying perpetration. Structural variables included outdegree, reciprocity, transitive triplets, 3-cycles, indegree-activity (sqrt), and outdegree activity (sqrt). Covariates included gender and ethnicity (alter, ego, and same parameters) as well as impulsivity. Linear and quadratic shape parameters were estimated for each behavior, as well as similarity on each behavior and the impact of impulsivity on each behavior. Results from these models have been presented previously[30] but Table 10 in this document provides final estimates, standard errors, and p-values for each parameter. We refer the reader to the table for individual details and to the forthcoming manuscript for detailed assessment and interpretation of the models, as this is outside the scope of this particular study. We will, however, summarize the findings herein.

Structural and rate parameters for network evolution were statistically large and in the expected directions with expected parameter values. Parameters for gender ego and gender similarity were statistically large as were race and grade similarity. Behavioral rate parameters were significant for both bullying and homophobic bullying and increased over each period (tested independently for heterogeneity reasons). Bullying behavior decreased in our sample over time in a linear fashion. Homophobic bullying decreased in a linear fashion but also had a significant quadratic shape parameter, suggesting greater change at the extremes of the range of homophobic bullying scores. With respect to co-evolution of friendships and bullying, the friendship dynamics associated with bullying perpetration and homophobic bullying perpetration differ. *Friendship choices* were driven by a preference for friends with above average homophobic bullying perpetration as well as similarity in homophobic bullying perpetration. Homophobic bullying perpetration was also influenced by friends' behavior, though bullying was not. Impulsivity was the only covariate of many explored (including depression, delinquency, etc.) that predicted both bullying and homophobic bullying perpetration. Overall, these models showed that friendships may be an especially important context for the emergence of homophobic bullying perpetration and related beliefs and norms.

*Baseline Model: ERGM/p**

The baseline ERGM/p* model included one behavioral outcome variable: homophobic bullying perpetration. Structural variables included edges, mutual ties, 2-paths, geometrically weighted indegree, outdegree parameters, geometrically weighted outdegree, geometrically weighted edgewise shared partners and 3-cycles. Covariates included matches on gender, ethnicity, age and parental education. Similarity on homophobic bullying perpetration as well as the impact of traditional masculinity and impulsivity were assessed as well. Table 11 in this document provides final estimates, standard errors, and p-values for each parameter. We refer the reader to the table for individual details. We will, however, summarize the findings herein.

Structural parameters were statistically large and in the expected directions with expected parameter values. Gender, grade and race similarity parameters were statistically large as was similarity in traditional masculinity attitudes. We found homophily with respect to homophobic bullying perpetration, but, in this model, impulsivity and traditional masculinity attitudes did not affect this behavior. Overall, this model demonstrated that friendships may be an especially important context for the emergence of homophobic bullying perpetration and related beliefs and norms.

Results: SIENA Model Comparisons

Results from our comparative simulation studies are presented both in summary tables and figures. Here, we assess the impact of missing data at 90, 70, 50, and 30% complete on the classes of parameters we included in the models (network rate, structure, behavior rate, behavior, influence, selection, covariate). We believe that the assessment of the parameters' behavior under missing data will be of greatest interest to those fitting models and estimating parameters, particularly those initially developing the models. We also assess the impact of missing data on the types of random and non-random sampling we investigated (missing edges, missing nodes, missing nodes with churn, personal network sampling, link tracing with 5% seed nodes, link tracing with 10% seed nodes). We believe that the assessment of the type of missing data will be of interest to those practitioners who make decisions about which research design may be most effective given limited resources. We will assess these parameter classes with respect to the likelihood of finding a parameter significant when it should not be (a "Type 1-like error", which we will refer to as simply a Type 1 error) and with respect to the likelihood of finding a parameter non-significant when it should be (a "Type 2-like error", which we will refer to as simply a Type 2 error).

Parameter Classes

Network rate parameters were most susceptible to Type 1 errors, with greater levels of missing data associated with greater proportions of parameter estimates being found significant though they were outside the range provided by the baseline model. This was the case across all random and non-random sampling approaches, though personal network approaches and edges missing at random were the most problematic in this regard. Missing nodes at random and link tracing with 10% of nodes as seeds performed best.

Structural parameters were equally susceptible to both Type 1 and Type 2 errors, with level of missing data having much less impact on the proportion of erroneous parameter estimates than would be expected. Overall it appears that, again, missing edges and personal network approaches were most problematic. Nodes missing with churn over time and link tracing with 10% of seeds performed best across both Type 1 and Type 2 errors.

Behavior Rate parameters were most susceptible to Type 2 errors, with level of missing data seeming unrelated to the proportion of errors recorded. Not surprisingly, missing nodes and missing nodes with churn were most problematic, followed by link tracing approaches with 5% of nodes used as seeds. Missing edges and personal network approaches performed best across Type 2 errors.

Behavior parameters (linear/quadratic shape) were also most susceptible to Type 2 errors, with level of missing data seeming unrelated to the proportion of errors recorded. Missing nodes were most problematic, followed by link tracing approaches with 5% of nodes used as seeds. Missing nodes with churn and link tracing approaches with 10% of nodes used as seeds performed (minimally) better than other approaches.

Influence parameters were also most susceptible to Type 2 errors, with greater levels of missing data associated with greater proportions of errors in parameter estimates recorded. Missing nodes and personal network approaches were most problematic across all levels of missing data, while nodes missing with churn and link tracing with 10% of nodes used as seeds performed best.

Selection parameters were susceptible to both Type 1 and Type 2 errors, with greater levels of missing data associated with greater proportions of errors in parameter estimates recorded. With respect to Type 1 errors, missing edges and link tracing with 5% of nodes used as seeds performed worst, while nodes missing with churn and personal network sampling approaches performed best. With respect to Type 2 errors, missing edges and link tracing with 10% of nodes used as seeds performed worst at higher levels of missing data, while nodes missing with churn and personal network approaches performed (marginally) better at higher levels of missing data.

Covariate parameters were most susceptible to Type 2 errors, with greater levels of missing data minimally related to proportion of missing data. Missing nodes and personal network sampling approaches performed marginally worse than other approaches. Nodes missing with churn performed marginally better than other approaches, as did link tracing with 5% of nodes used as seeds.

These findings are visually summarized in Figures 25 and 26. Overall it seems that Type 1 errors (false positives) most likely occur for structural rate parameters and for approaches where edges are more likely to be missing or to shift frequently (missing edges/nodes with churn). Not surprisingly, errors are greater with more missing data. Type 2 errors (false negatives) most likely occur for selection and influence parameters with errors again greater with more missing data. The preponderance of Type 2 errors is not surprising since, with more missing data, there is less information to use to calculate parameter estimates. However, the level of Type 1 errors reported for network, structural, and selection parameters bears further investigation, as finding significant parameters when they should not be significant can greatly influence model interpretation. While influence processes are currently of most interest to researchers, these other features of selection processes should not be overlooked. Behavioral parameters themselves are generally resistant to errors because these parameters are primarily calculated using means and thus are less problematic. Covariate parameters appear problematic with respect to Type 2 errors at greater rates than other behavior and influence parameters, which may be due to the inability to accurately assess the relationships between two variables in the presence of missing data.

Effects of Missing Data on Errors

Missing edges certainly cause both Type 1 and Type 2 errors, though level of missing data has differing impact on the various classes of parameters. For Type 1 errors, network rate and structural parameters are most affected, while behavior rate, behavior and covariate parameters least affected.

For Type 2 errors, selection and influence parameters are most affected, while network rate and structural parameters are least affected.

Missing nodes have a greater impact on Type 2 errors. Again, level of missing data has differing impact on the various classes of parameters. Selection and Influence parameters are most affected, while network rate and structural parameters are least affected.

Missing nodes with churn leads to both Type 1 and Type 2 errors. Level of missing data varies in its impact on the various classes of parameters, but structure-related parameters are most likely to have Type 1 errors. The effects on other parameter types are similar, and minimally impacted by missing data. With respect to Type 2 errors, we see network influence and covariate parameters most affected, as are selection parameters. Least affected are the network rate and structural parameters.

Personal network sampling approaches lead to both Type 1 and Type 2 errors, though the various classes of parameters seem unaffected by the level of missing data. With respect to Type 1 errors, personal network sampling approaches are associated with high proportions of problems with network rate parameters and, less so, with structural parameters. Other parameter types are minimally affected by missing data. With respect to Type 2 errors, it appears that network influence and covariate parameters are most affected, as are selection parameters. Least affected are behavioral parameters.

Link tracing with 5% of nodes as seeds leads primarily to Type 2 errors. While there are some Type 1 errors reported in network rate and structural parameters, the proportions are fairly small. With respect to Type 2 errors, most affected are, again, network influence and covariate parameters, with moderate levels of error also shown for selection parameters. Least affected are network rate and structural parameters.

Link tracing with 10% of nodes as seeds leads to both Type 1 and Type 2 errors, with level of missing data differentially associated with the various parameter classes. With respect to Type 1 errors, most impacted are network rate and structural parameters, although we see errors appearing across most parameter classes and levels of missing data. Least impacted here are the network influence parameters. With respect to Type 2 errors, most affected are covariate parameters, followed by network influence and network selection parameters. Least affected are network rate and structural parameters.

These findings are visually summarized in Figures 27 and 28. There is no single best or worst performer in this set of analyses. Missing edges cause errors, and personal network approaches consistently display errors, but other approaches are more problematic for the types of parameters researchers are most interested in: influence, selection, and covariate parameters. Nodes missing with churn and missing edges lead to large numbers of Type 1 errors, as does link tracing with 10% of nodes as seeds. With respect to Type 2 errors, missing edges are most problematic and show errors for selection, influence and covariate parameters. Missing nodes show errors for influence and covariate parameters. Missing nodes with churn shows errors primarily for influence, but less so than the overall missing nodes class. Personal network approaches show errors for influence and for covariates, as does link tracing with 5% of nodes as seeds. Link tracing with 10% of nodes as seeds also shows errors for influence and covariate parameters but less so than link tracing with 5% of nodes as seeds.

Results: ERGM/p Model Comparisons*

Results from our comparative simulation studies are presented both in summary tables and figures. Here, we assess the impact of missing data at 90, 70, 50, and 30% on the classes of parameters included in the models (structure, homophily). We believe that the assessment of the parameters' behavior under missing data will be of greatest interest to those interested in the modeling and estimation processes, particularly those developing the models. We also assess the impact of missing data on the types of random and non-random sampling we investigated (missing edges, missing nodes, missing nodes with churn, personal network sampling, link tracing with 5% seed nodes, link tracing with 10% seed nodes). We believe that the assessment of the type of missing data will be of interest to those practitioners who seek to make decisions about which sort of research design may be most effective given the resources on hand. We will assess these parameter classes and with respect to the likelihood of finding a parameter significant when it should not be (a "Type 1-like error") and with respect to the likelihood of finding a parameter non-significant when it should be (a "Type 2-like error").

Parameter Classes

Structural parameters show both Type 1 and Type 2 errors. For Type 1 errors, less missing data increases the chance that a parameter will be found significant that shouldn't be, while for Type 2 errors, the association seems less related to level of missing data (for those models where we could actually do comparisons). In some cases models would not converge and thus no comparisons were possible. Missing edges had the least impact on Type 1 errors, while personal network approaches had the greatest impact. Missing edges had the greatest impact on Type 2 errors while personal network approaches had the least impact.

Homophily parameters also show both Type 1 and Type 2 errors. For Type 1 errors, as before, less missing data increases the chance that a parameter will be found statistically large that should not be, while for Type 2 errors, the effect seems less related to level of missing data. Again, missing edges had the least impact on Type 1 errors, while personal network approaches had the greatest impact. With respect to Type 2 errors, it appears that most sampling approaches had equivalent impacts on parameter estimates, with missing edges producing marginally worse parameter estimates.

The results of these analyses are visually presented in Figures 49 and 50. Of interest is that, overall, structural parameters were more prone to Type 1 errors, while homophily parameters were more prone to Type 2 errors. Personal network approaches performed consistently worse with respect to Type1 errors across both parameter types while missing edges were associated with greater Type 2 errors across both parameter types.

Effects of Missing Data on Errors

Missing edges had greatest impact on homophily measures overall, with most of them causing Type 2 errors. Type 1 errors consistently occured across structure and homophily parameters, while Type 2 errors were more common for homophily parameters.

Missing nodes had the greatest impact on homophily measures overall, with most of them causing Type 2 errors. Type 1 errors were greater for structural parameters while Type 2 errors were greater for homophily parameters.

Personal network sampling approaches showed greater Type 1 errors for structure parameters and greater Type 2 errors for homophily parameters.

Snowball sampling with 30% of the sample had the greatest impact on homophily parameters. This approach showed roughly equivalent levels of Type 1 errors across structure and homophily parameters, with higher levels of Type 2 errors for homophily parameters relative to structural parameters and Type 1 errors.

Snowball sampling with 10% of the sample had greatest impact on homophily parameters overall. This approach showed higher levels of Type 1 errors for structure parameters, with higher levels of Type 2 errors for homophily parameters relative to structural parameters and Type 1 errors.

The results of these analyses are visually represented in Figures 51 and 52. With respect to Type 1 errors, the proportion of errors made was fairly consistent across sampling types and level of missing data.   Most non-random sampling approaches were consistently problematic. With respect to Type 2 errors, homophily parameters are more consistently problematic than structural parameters. While the proportion of errors made was consistently high, missing edges showed the highest level of errors.

Discussion

Our analyses created a great deal of information, and required a large synthesis to achieve our goal of understanding the effects of sampling designs. We have tried to assess the key findings for model developers as well as practitioners. Overall, we believe that the “take home message” is that we are conservative modelers. We are more likely to overlook significance where it should be, than to find significance where it isn’t. This is particularly true for SIENA models. p* models/ERGMs, however, are prone to higher levels of error across both Type 1 and Type 2 errors. Not surprisingly, more missing data leads to greater likelihood of Type 2 errors. Interestingly, in some cases, less missing data leads to greater likelihood of Type 1 errors.

With respect to SIENA models, for the methodologists, our most important findings are that structural and selection parameters are susceptible to both Type 1 and Type 2 errors, while influence parameters seem most susceptible to Type 2 errors. Strategies to reduce Type 1 errors for structural and selection parameters should be explored. For the practitioners, it seems that missing edges are likely to lead to consistent Type 1 and Type 2 errors, when data are missing at random, while nodes missing with churn, the most likely scenario in an applied setting, lead to higher overall levels of Type 1 errors. Such errors should be avoided in practice as they can be problematic when we seek to understand social systems. Link tracing, however, may be a promising approach for sampling in networks if the set of seeds chosen is large. In reality, however, it may be that researchers will have to present their findings and then discuss the limitations that their data present based on the design they have chosen.

With respect to the ERGM/p* models, it appears that any level of missing data leads to both Type 1 and Type 2 errors. Type 1 and Type 2 errors are less frequent for structural parameters than for homophily parameters, though homophily parameters display more frequent Type 2 errors than do structural parameters. For practitioners, it is important to note that missing edges in cross-sectional studies are more impactful than missing nodes and thus should be avoided if at all possible in research designs.

This study has a number of limitations that should be noted. First, the data we used for our test bed were collected in the field and clearly are prone to error. While this should not affect the results of the simulation and comparison study, it does impact the results of the baseline models. This is one reason why our first assessment was designed to be broad and shallow. We chose to assess a number of sampling approaches across a range of levels of missing data to help determine where to target subsequent, more focused simulation studies. Time and resource limitations forced us to generate just 10 data sets for each type of sampling. This is a small number for a simulation study. We intend to continue this work with more targeted studies using a much larger number. This will allow for more powerful (and actually quantitative, rather than qualitative) comparisons.

Given these limitations, we do believe that our study has revealed some interesting patterns associated with missing data that deserve exploration. It has also suggested that it may be useful to find ways to adjust such network models to improve them. We believe that these changes can be made on both the data side and on the modeling side. With respect to the data, are there methods for data imputation that would account for missing data more effectively than standard mean-based or regression-based imputations? With respect to modeling, are there ways to adjust the models so that missing data or sampling approach more accurately represents the observed data? Can we adjust the parameter estimation procedures to minimize both types of errors so that we can more carefully report interesting but less apparent relationships? Such relationships may reveal important associations that could generate effective interventions or lead to better insight into how networks affect behavior (and vice versa). Alternatively, are there models we could apply that are less impacted by missing data? Latent space models, relational events models and others in development may be of interest in this context. These questions bear further research. More work needs to be done.

FIGURES

Figure 1: Deviation and significance level of parameters based on SIENA models with 10% missing edges.

Figure 2: Deviation and significance level of parameters based on SIENA models with 30% missing edges.

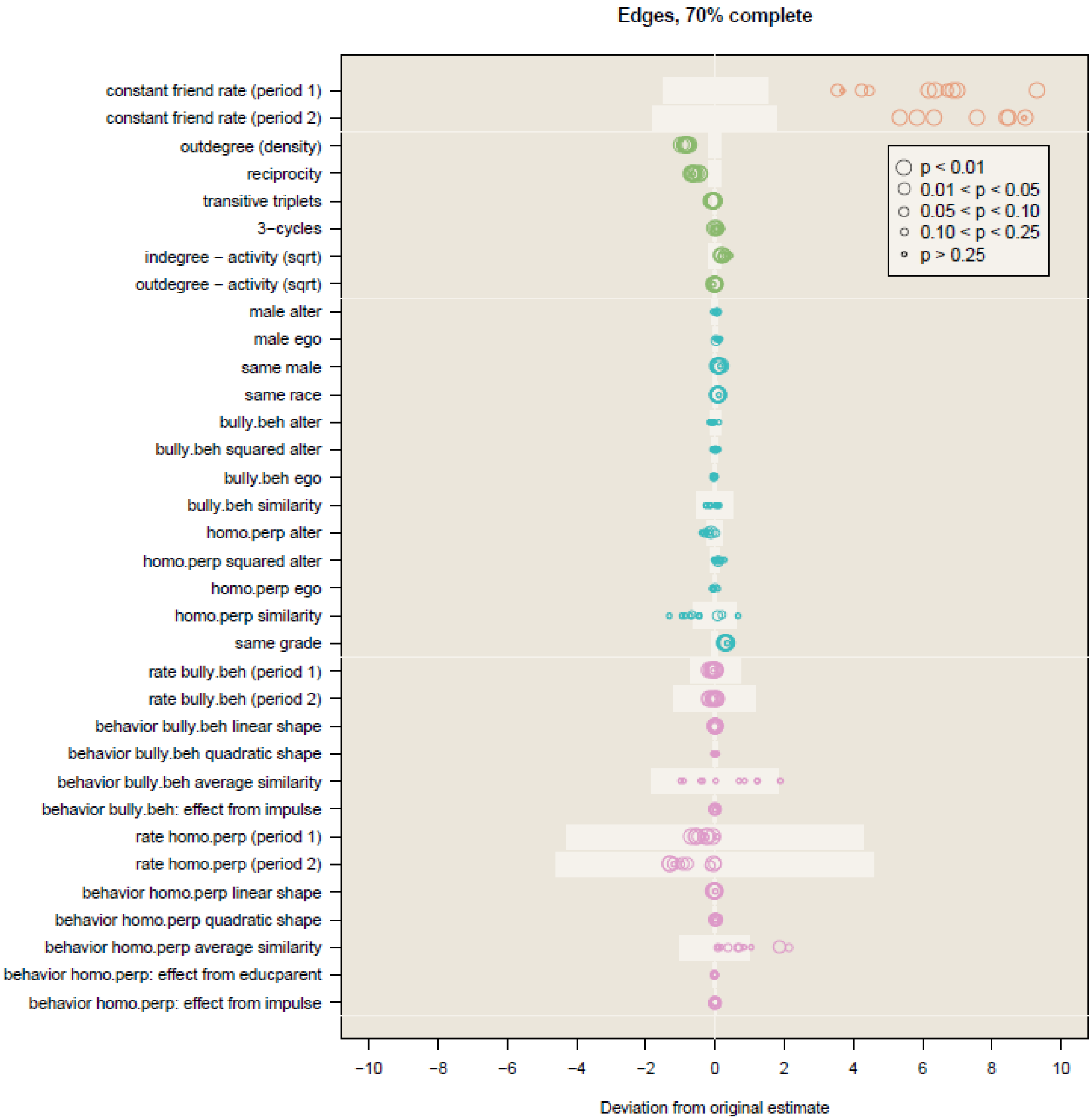


Figure 3: Deviation and significance level of parameters based on SIENA models with 50% missing edges

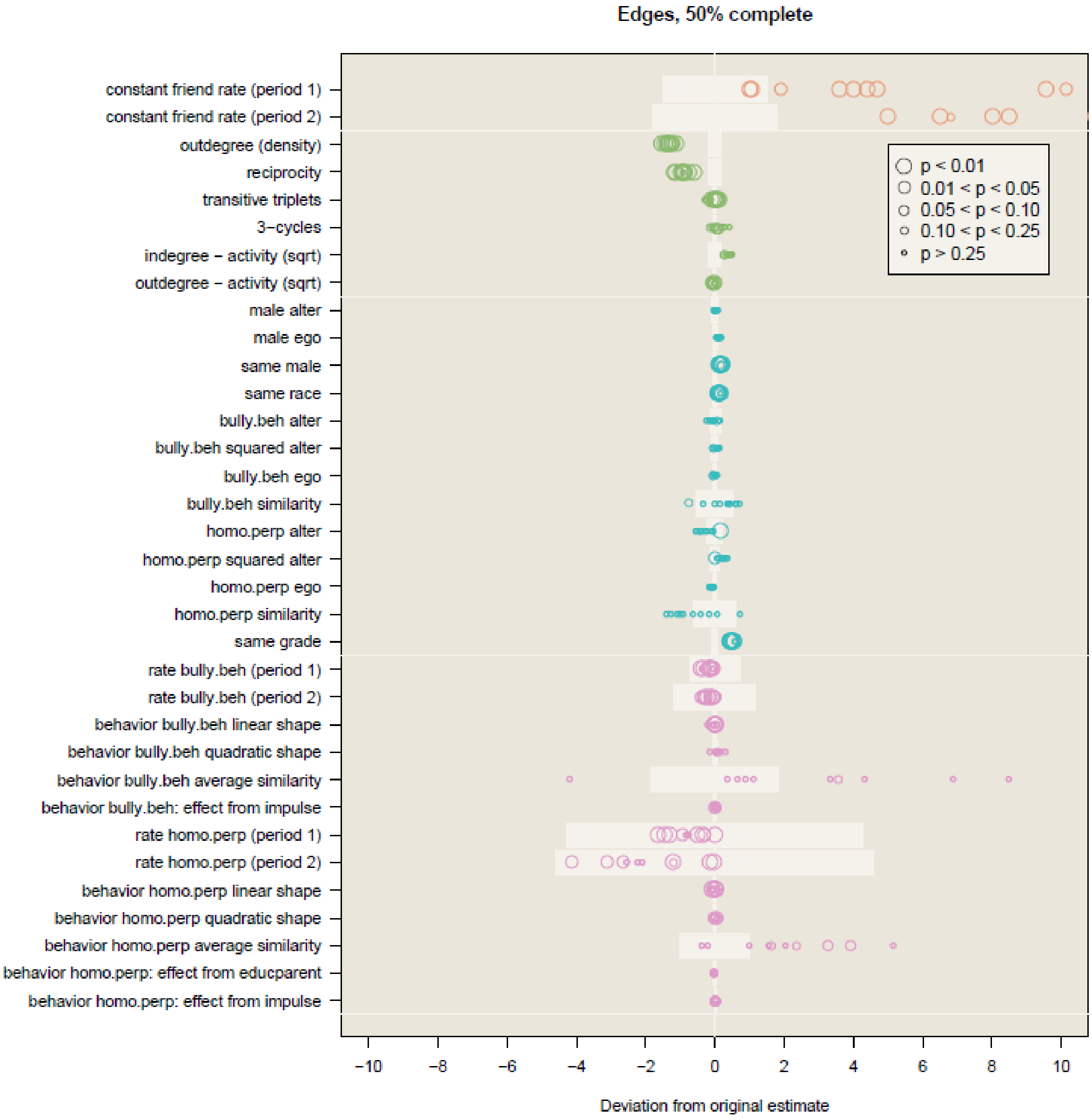


Figure 4: Deviation and significance level of parameters based on SIENA models with 70% missing edges

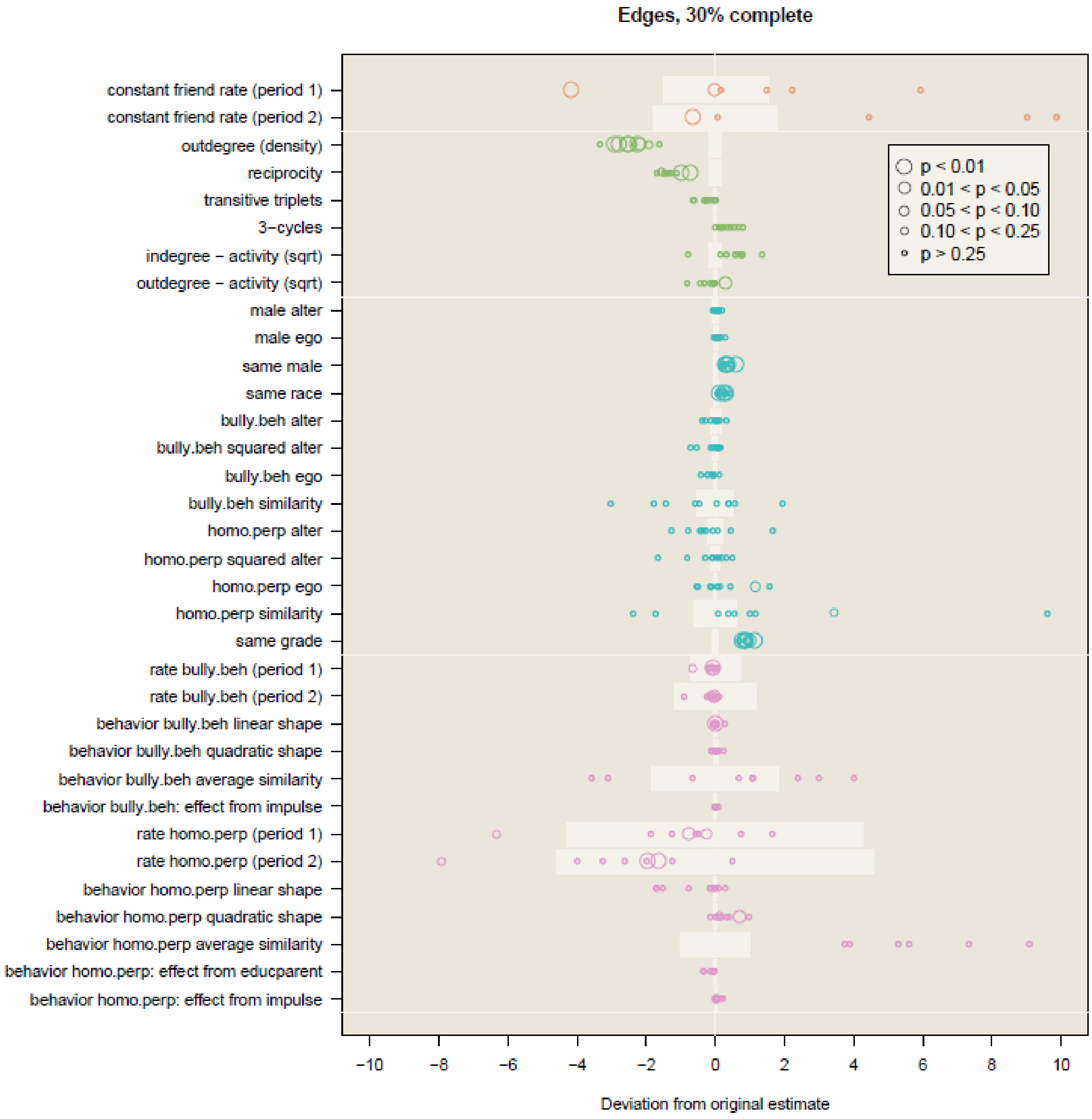


Figure 5: Deviation and significance level of parameters based on SIENA models with 10% missing nodes.

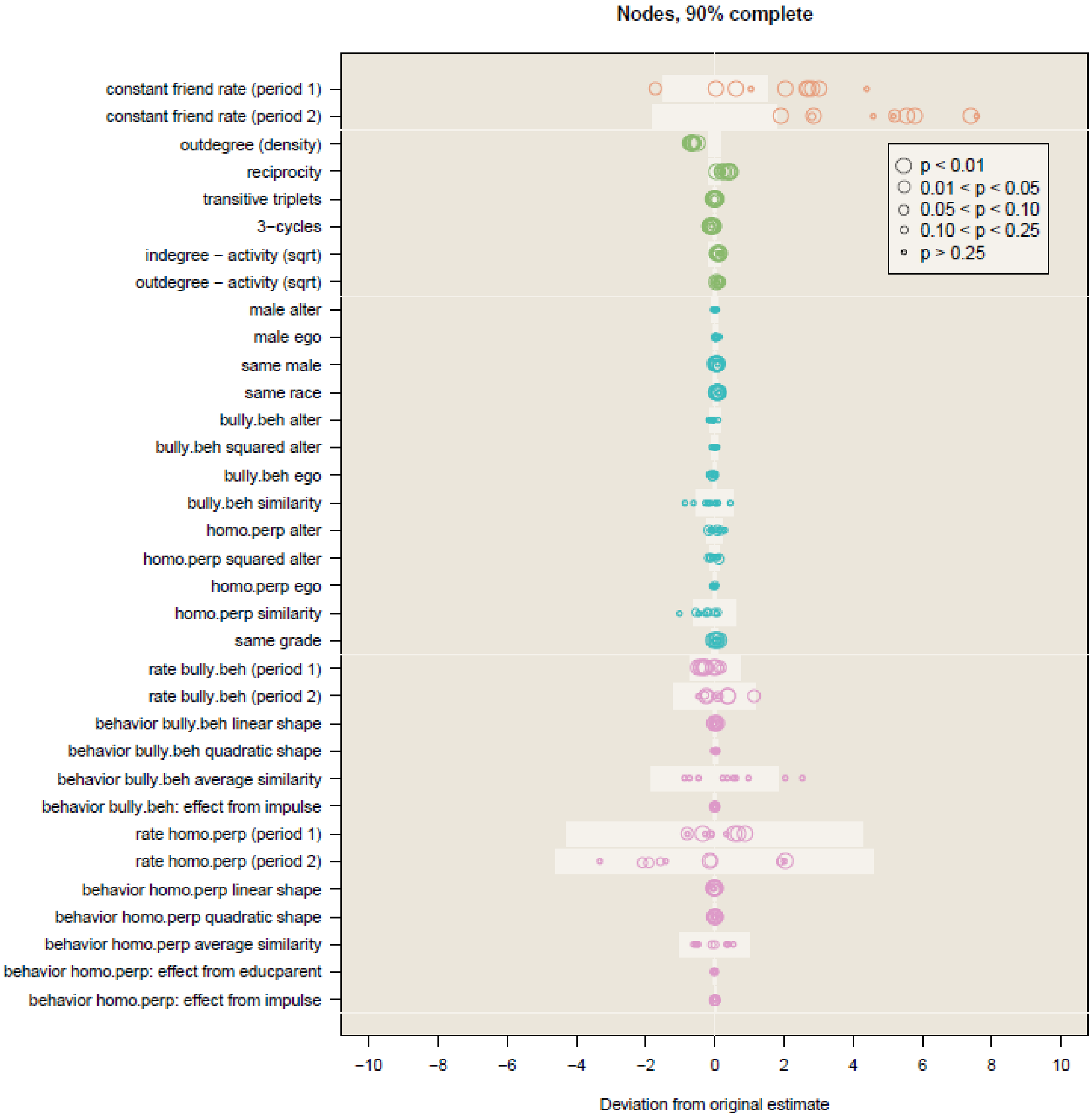


Figure 6: Deviation and significance level of parameters based on SIENA models with 30% missing nodes.

Figure 7: Deviation and significance level of parameters based on SIENA models with 50% missing nodes.

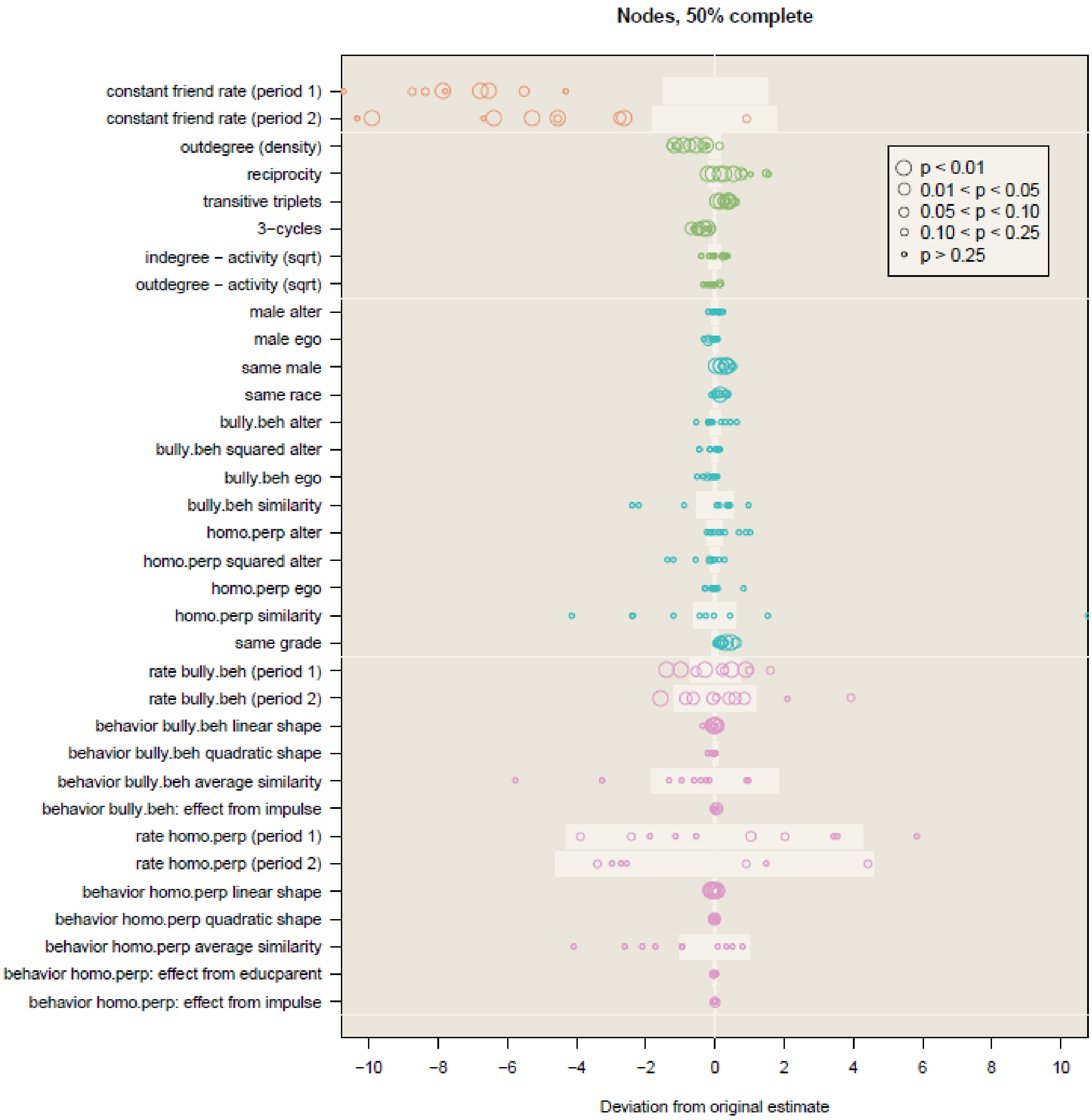


Figure 8: Deviation and significance level of parameters based on SIENA models with 70% missing nodes.

Figure 9: Deviation and significance level of parameters based on SIENA models with 10% missing nodes, varying from wave to wave.

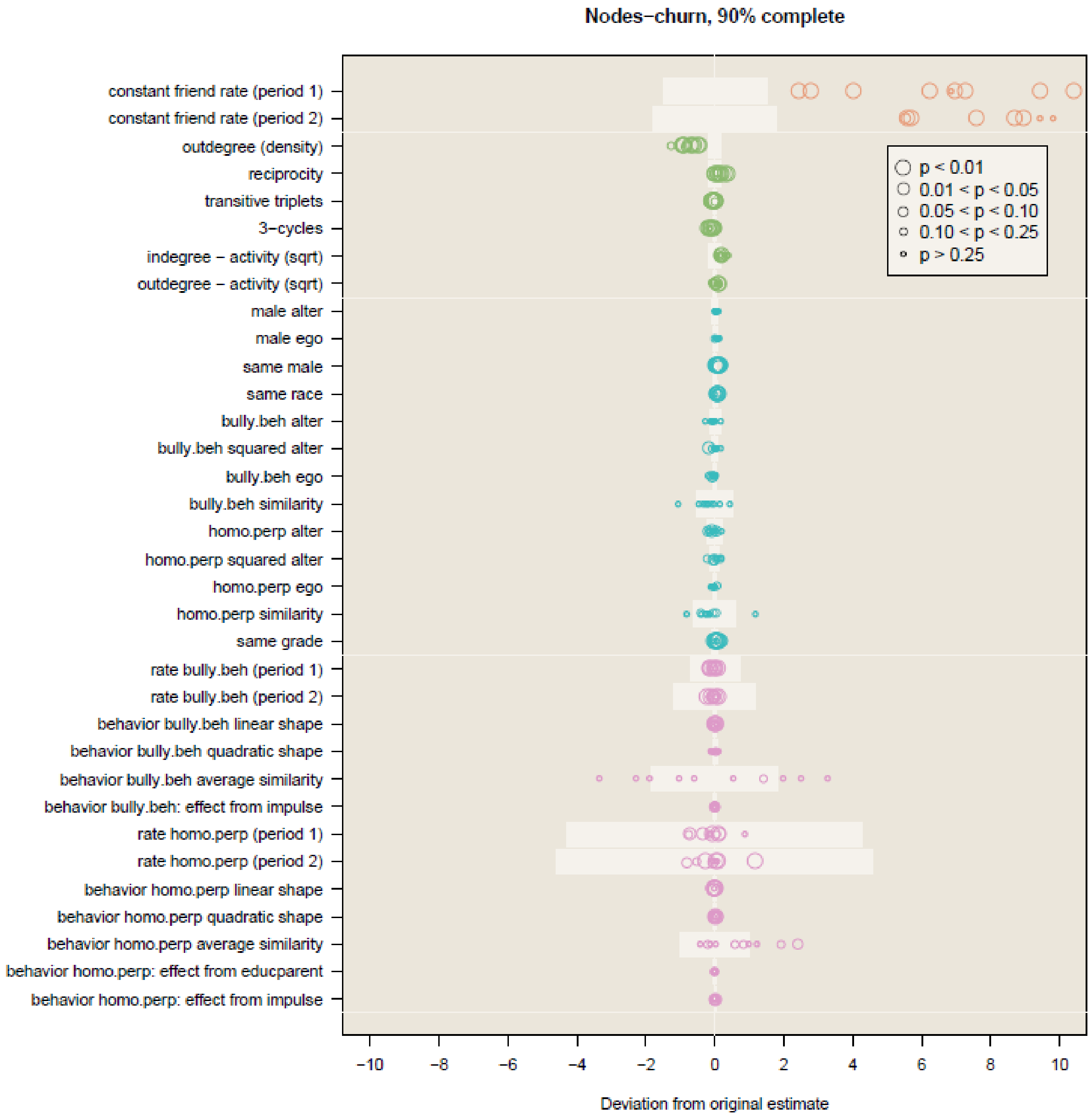


Figure 10: Deviation and significance level of parameters based on SIENA models with 30% missing nodes, varying from wave to wave.

Figure 11: Deviation and significance level of parameters based on SIENA models with 50% missing nodes, varying from wave to wave.

Figure 12: Deviation and significance level of parameters based on SIENA models with 70% missing nodes, varying from wave to wave.

Figure 13: Deviation and significance level of parameters based on SIENA models with 10% missing nodes, built from link tracing from randomly chosen seeds that account for 5% of the population.

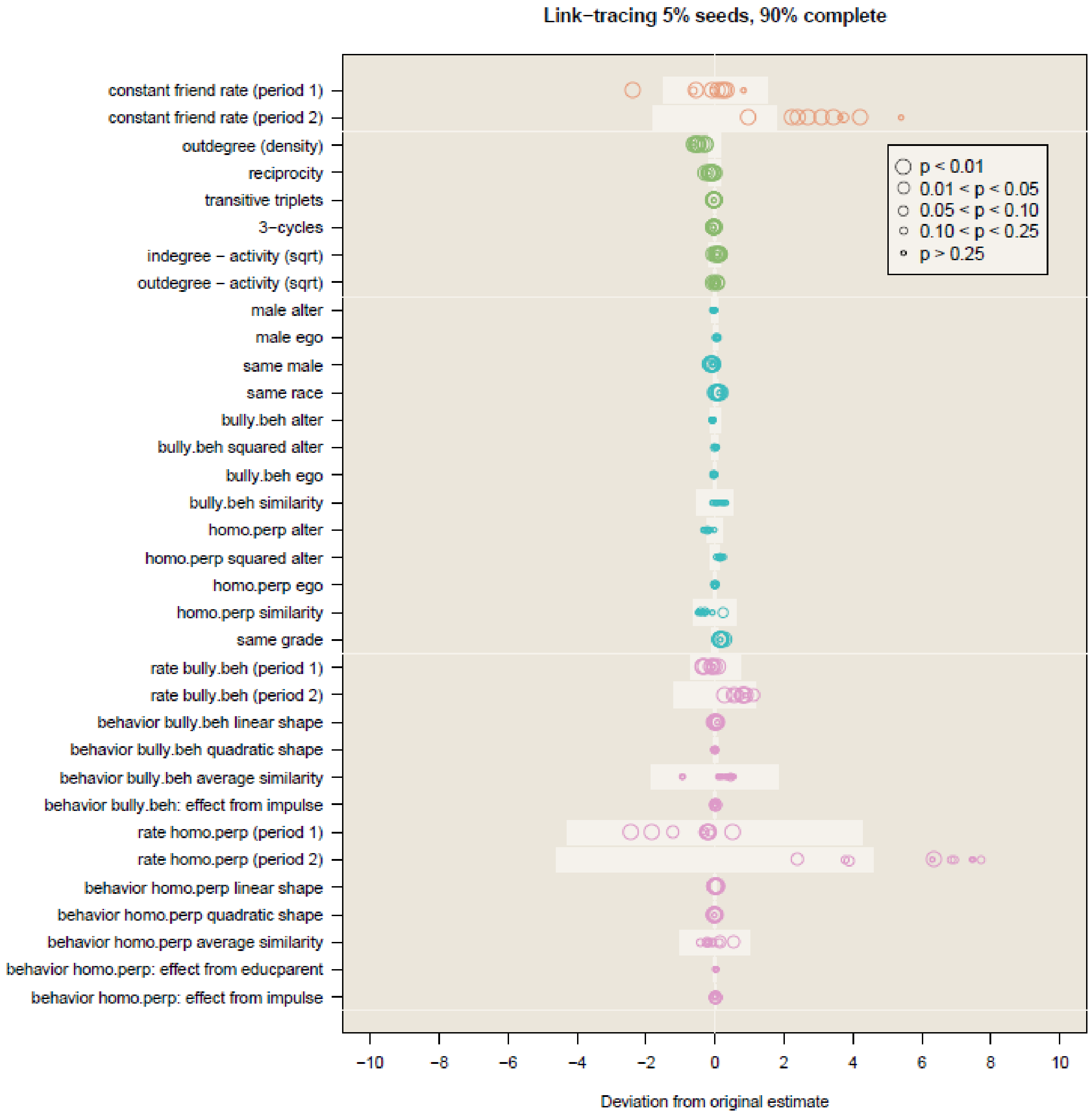


Figure 14: Deviation and significance level of parameters based on SIENA models with 30% missing nodes, built from link tracing from randomly chosen seeds that account for 5% of the population.

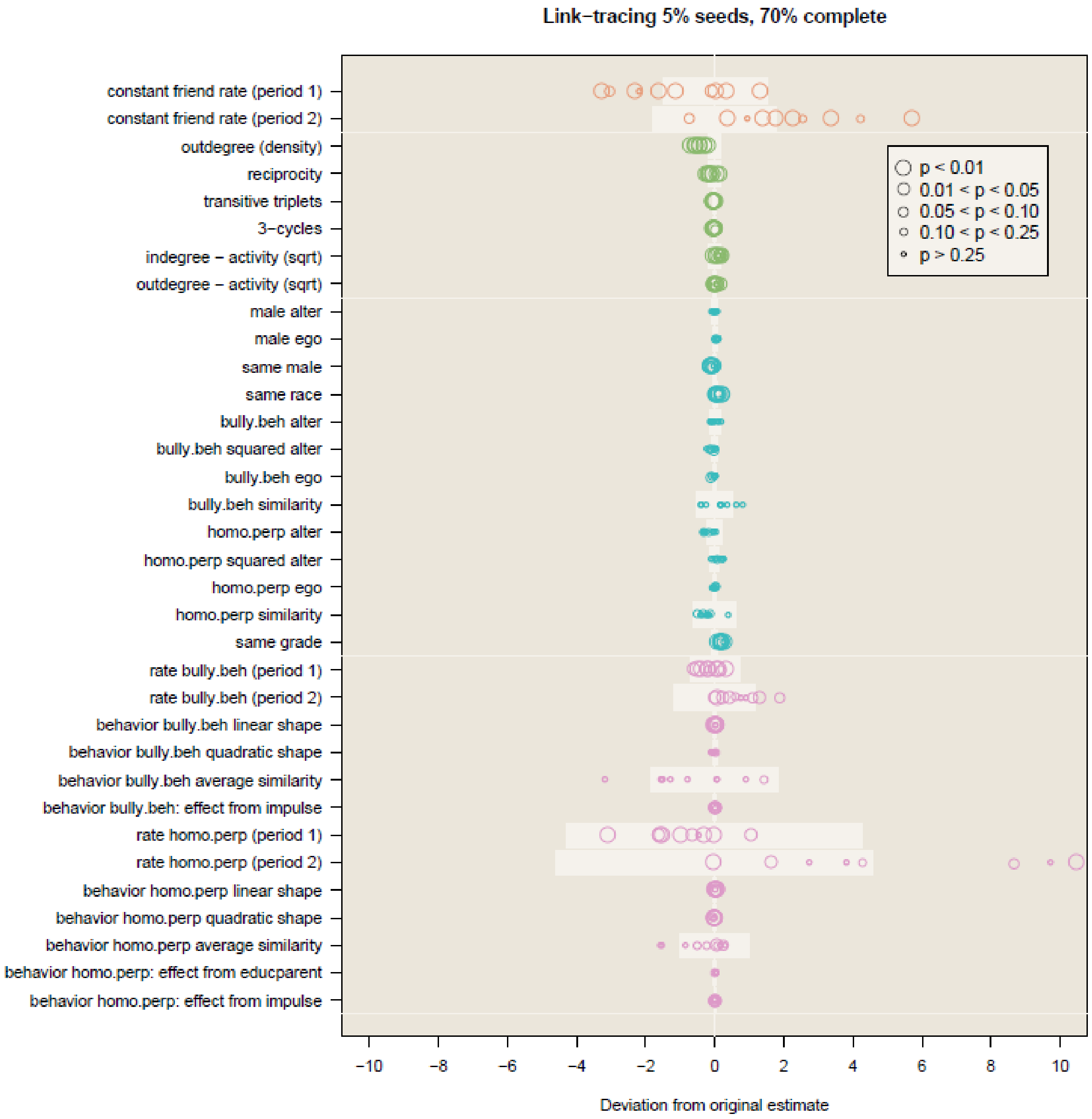


Figure 15: Deviation and significance level of parameters based on SIENA models with 50% missing nodes, built from link tracing from randomly chosen seeds that account for 5% of the population.

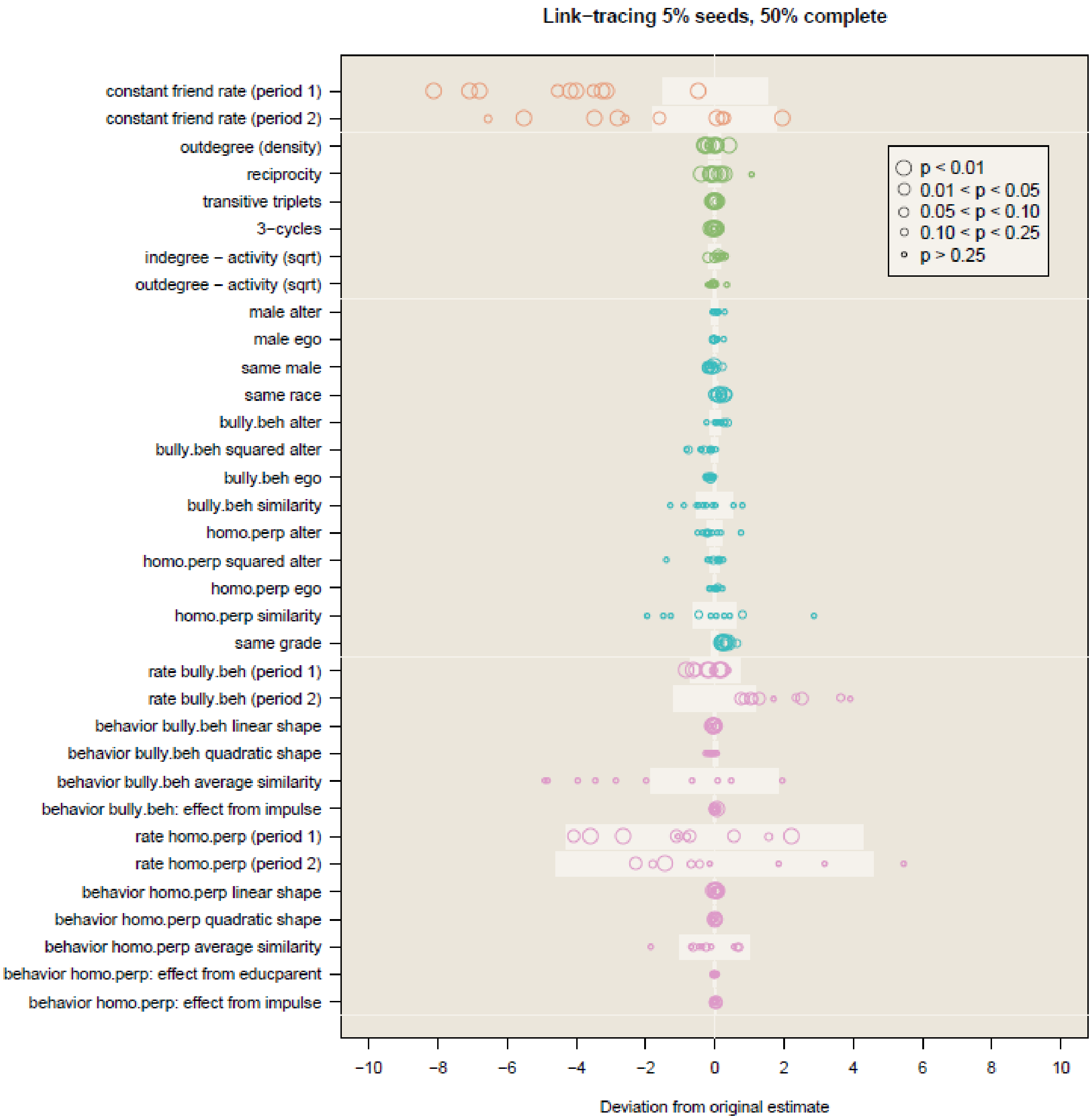


Figure 16: Deviation and significance level of parameters based on SIENA models with 70% missing nodes, built from link tracing from randomly chosen seeds that account for 5% of the population.

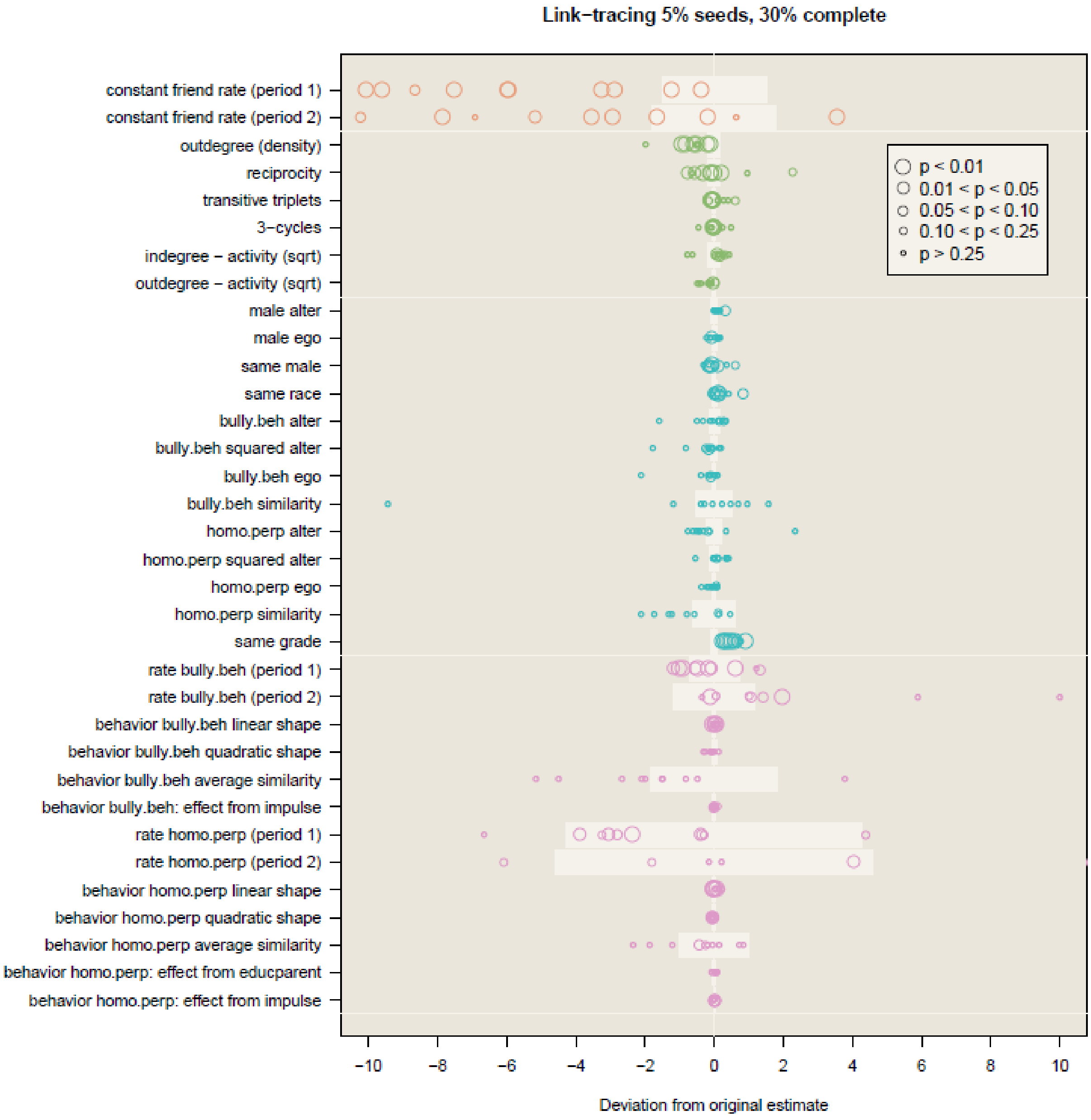


Figure 17: Deviation and significance level of parameters based on SIENA models with 10% missing nodes, built from link tracing from randomly chosen seeds that account for 10% of the population.

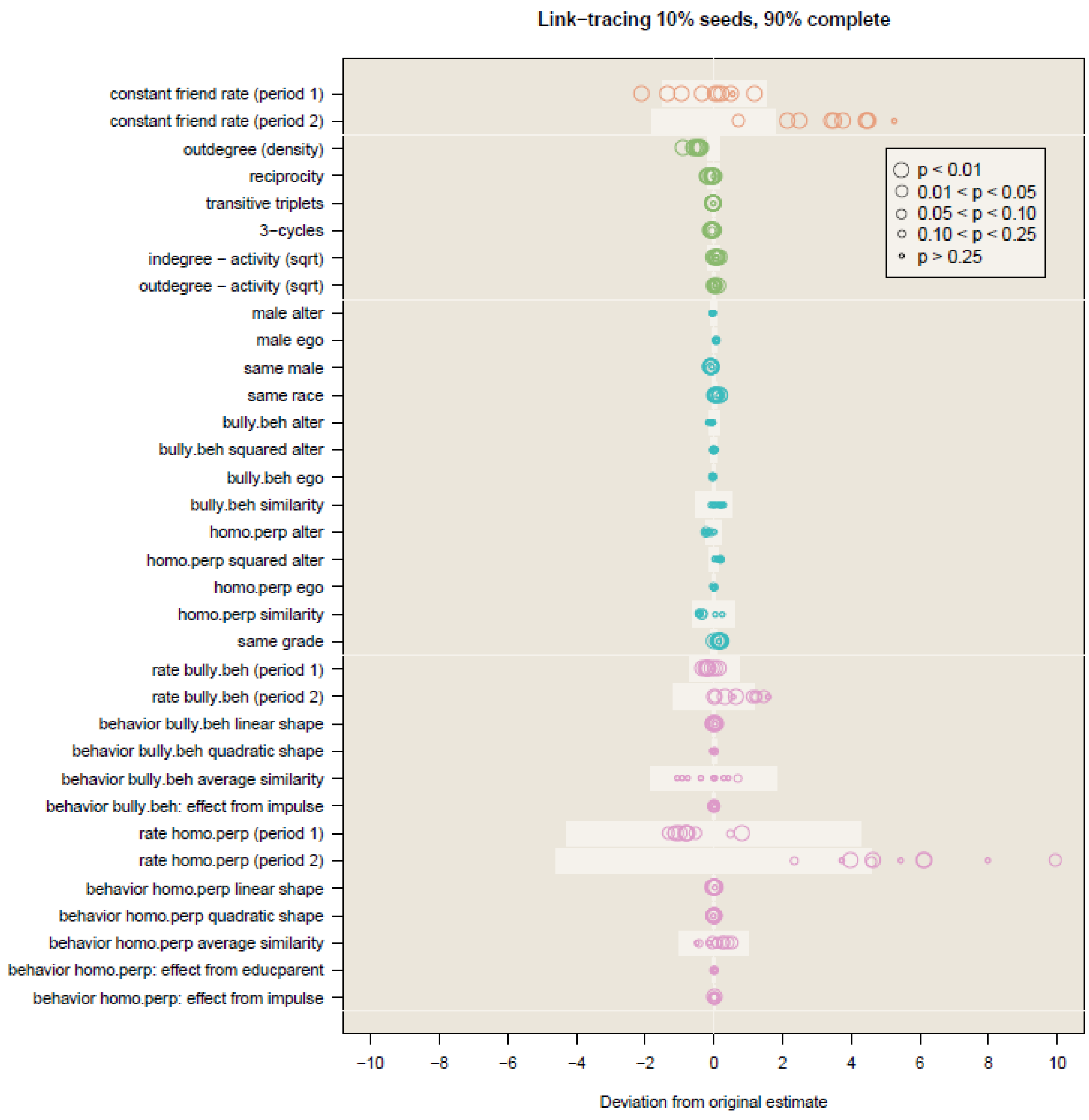


Figure 18: Deviation and significance level of parameters based on SIENA models with 30% missing nodes, built from link tracing from randomly chosen seeds that account for 10% of the population.

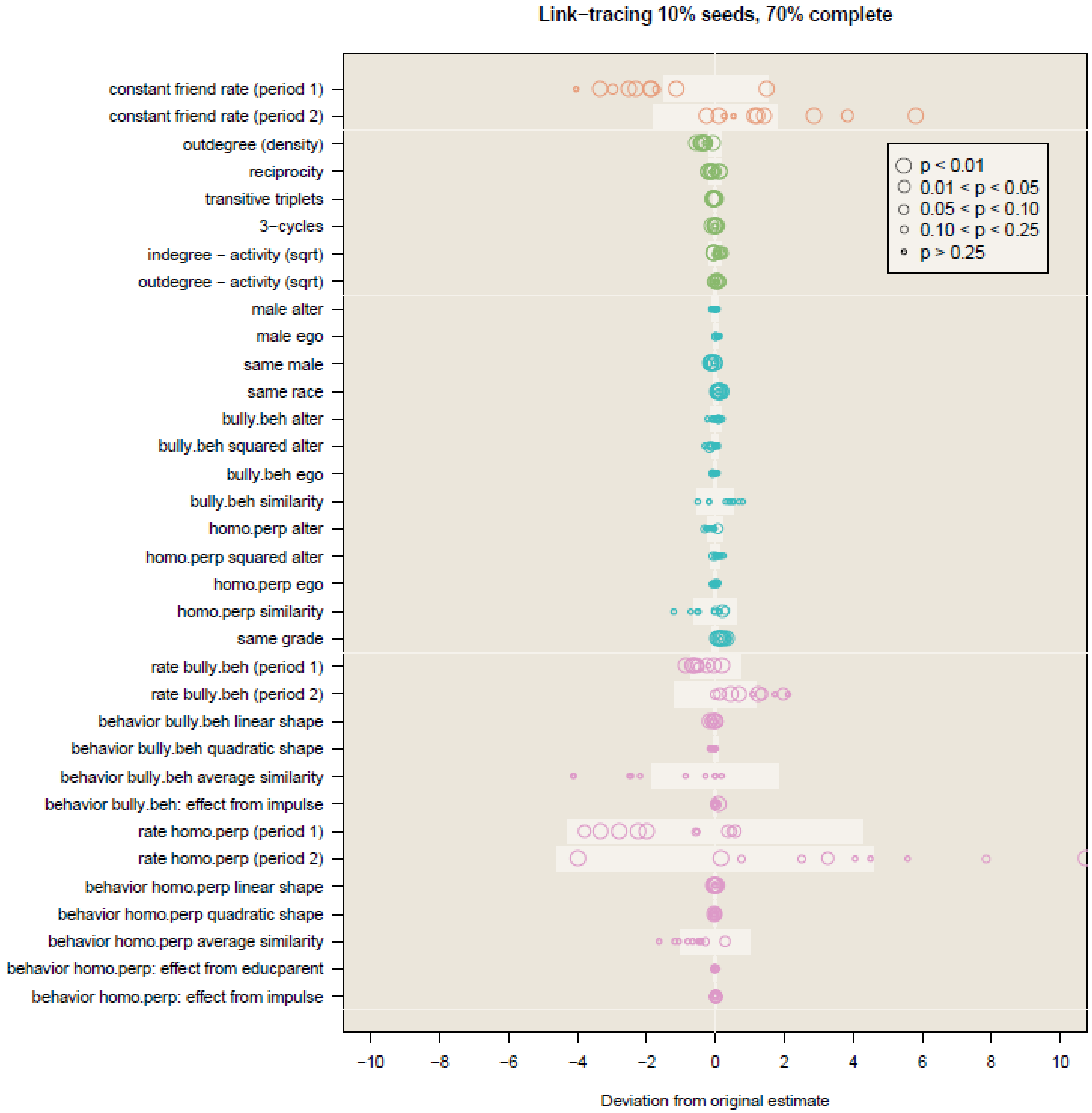


Figure 19: Deviation and significance level of parameters based on SIENA models with 50% missing nodes, built from link tracing from randomly chosen seeds that account for 10% of the population.

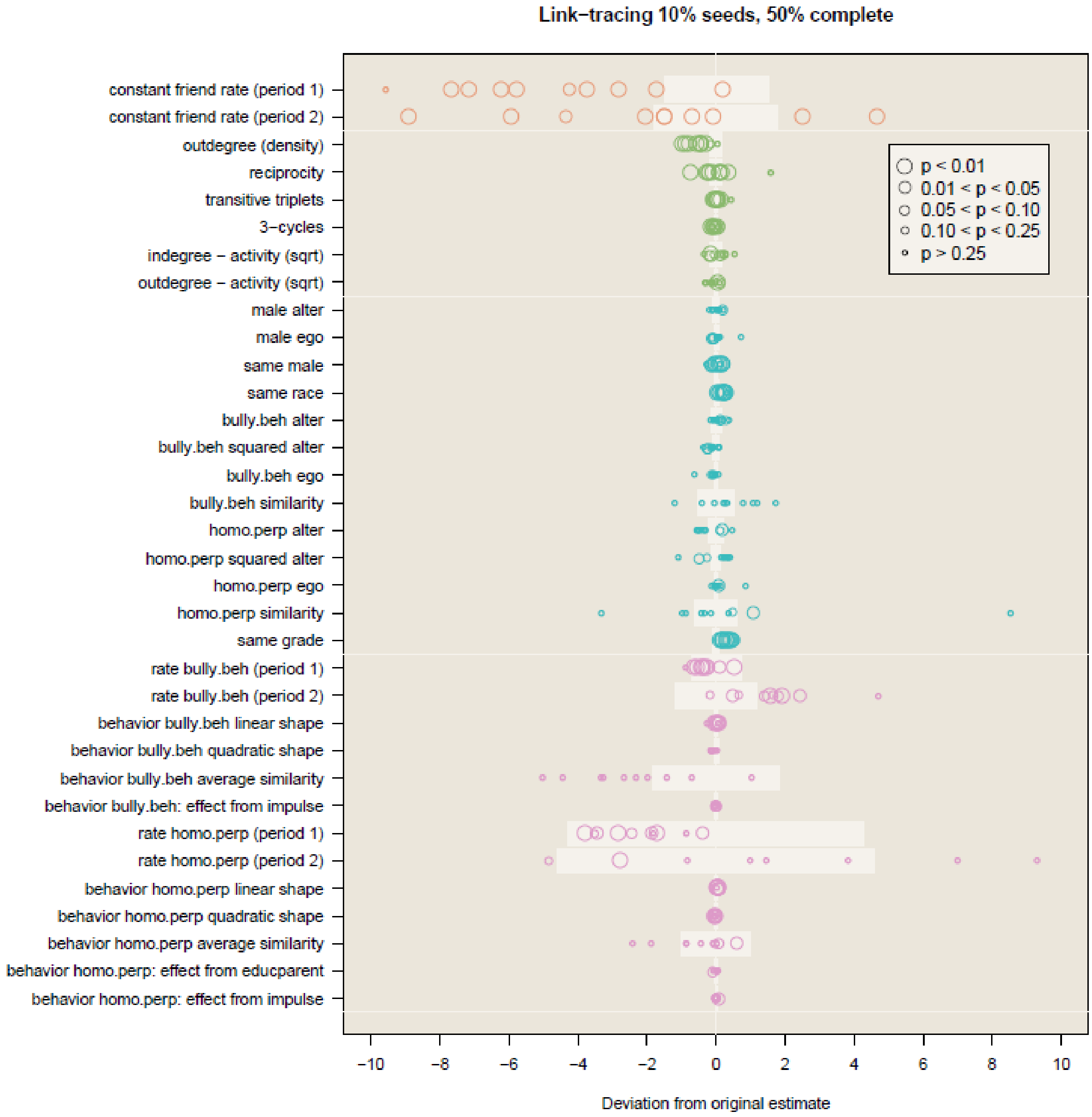


Figure 20: Deviation and significance level of parameters based on SIENA models with 70% missing nodes, built from link tracing from randomly chosen seeds that account for 10% of the population.

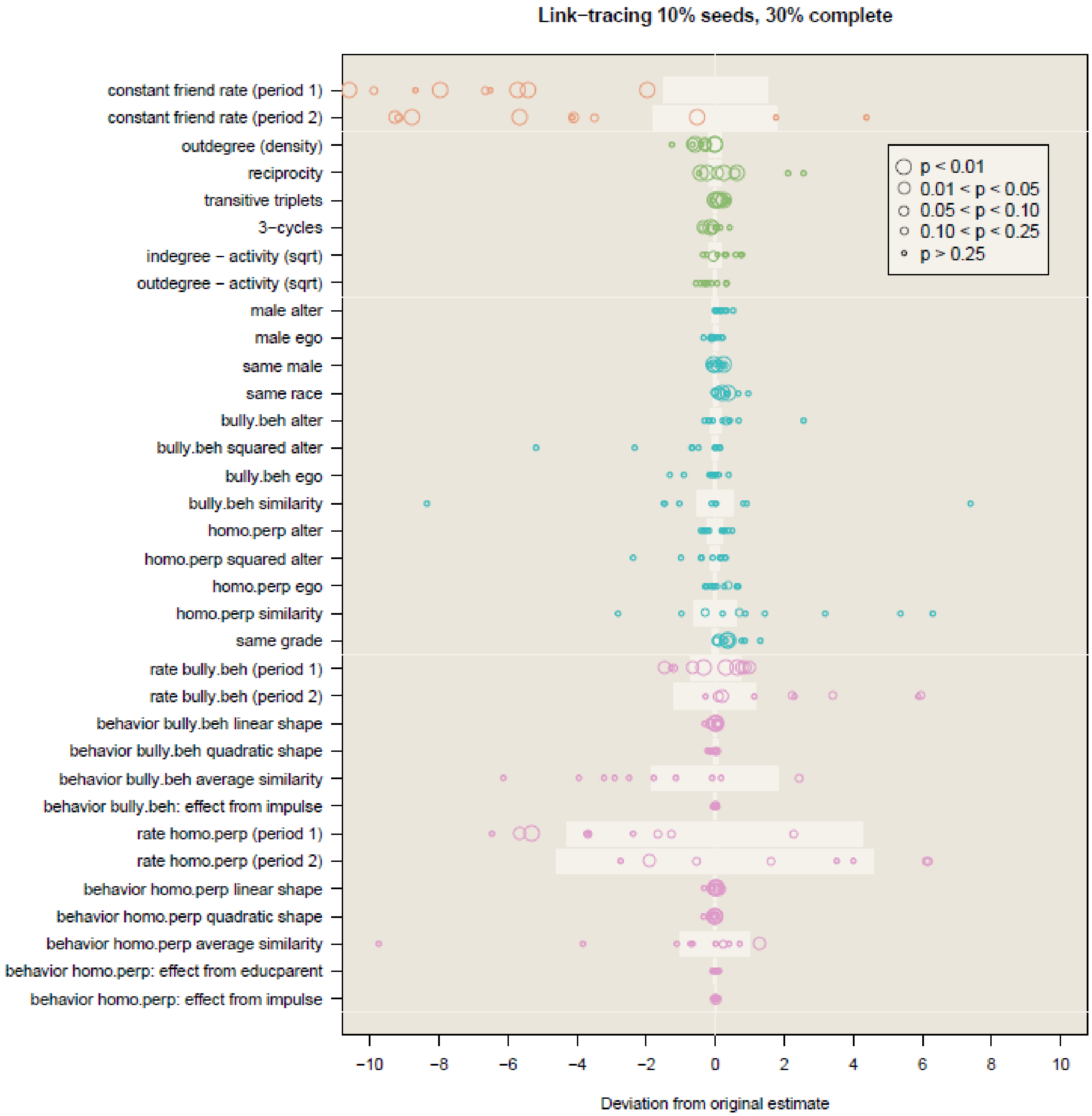


Figure 21: Deviation and significance level of parameters based on SIENA models with 10% missing nodes, built from personal network sampling of the population.

Figure 22: Deviation and significance level of parameters based on SIENA models with 30% missing nodes, built from personal network sampling of the population.

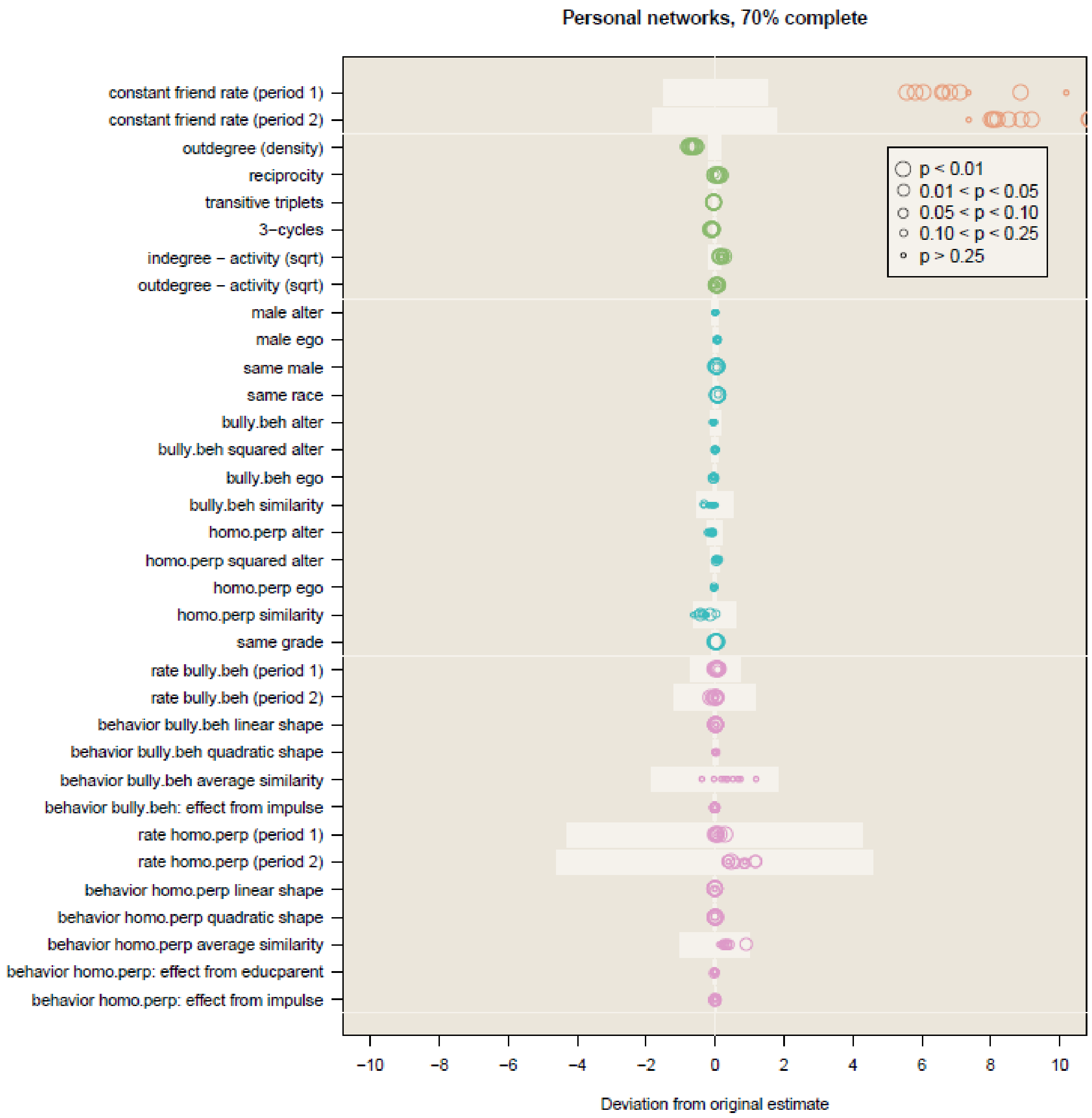


Figure 23: Deviation and significance level of parameters based on SIENA models with 50% missing nodes, built from personal network sampling of the population.

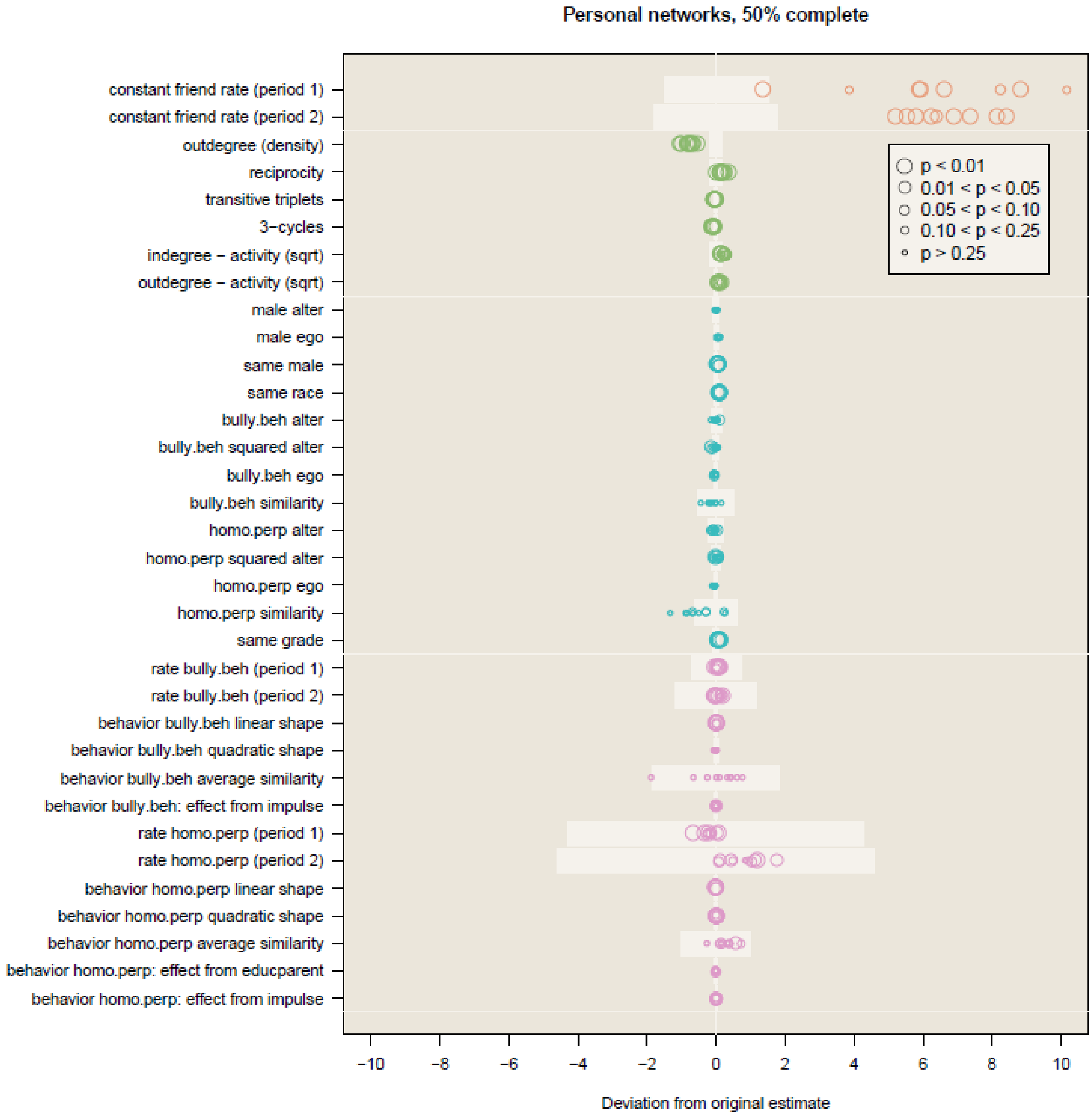


Figure 24: Deviation and significance level of parameters based on SIENA models with 70% missing nodes, built from personal network sampling of the population.

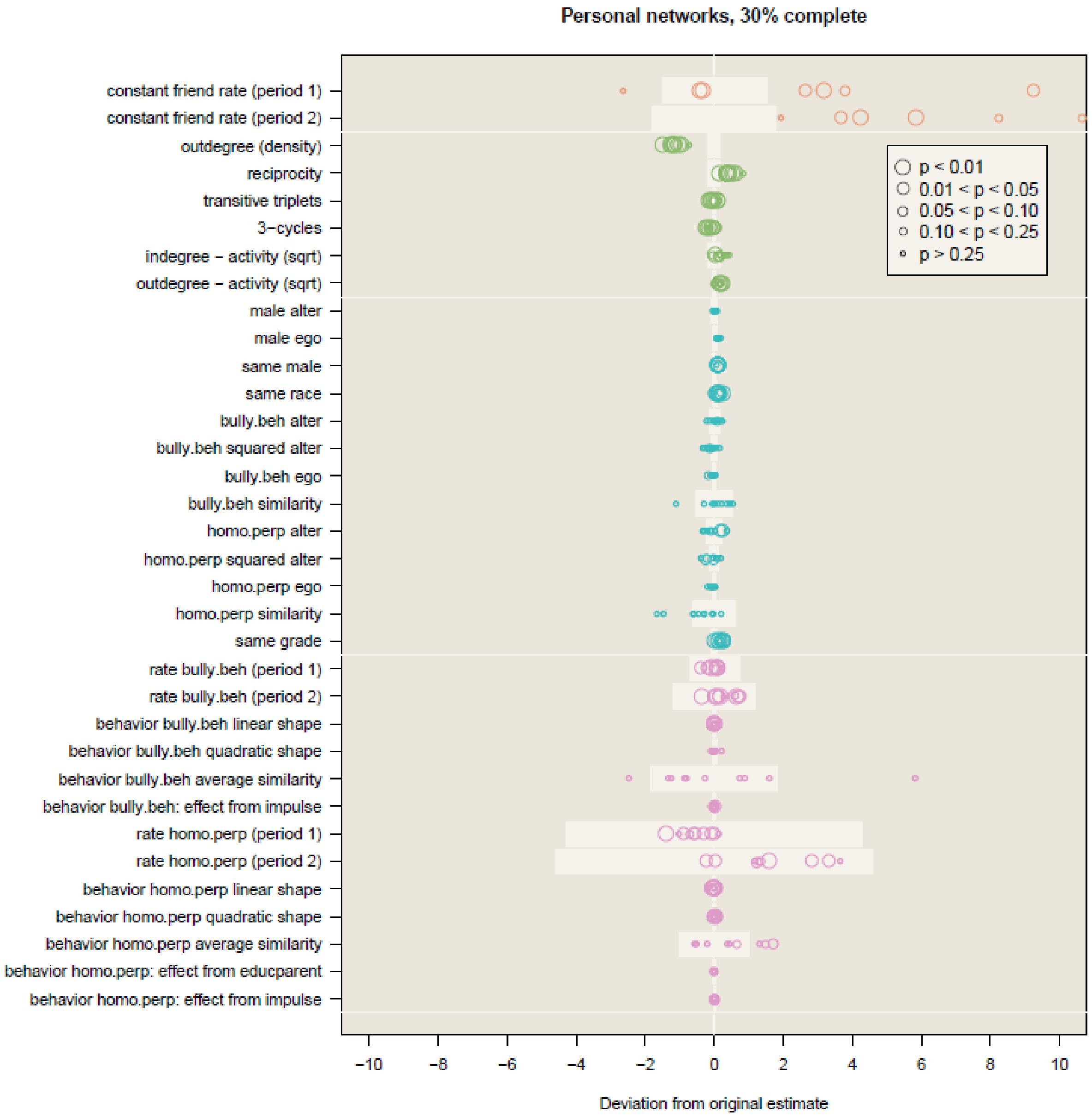


Figure 25: Summary across SIENA model parameter classes of the proportion of errors of each type for each sampling approach, across level of missingness; errors grouped by level.

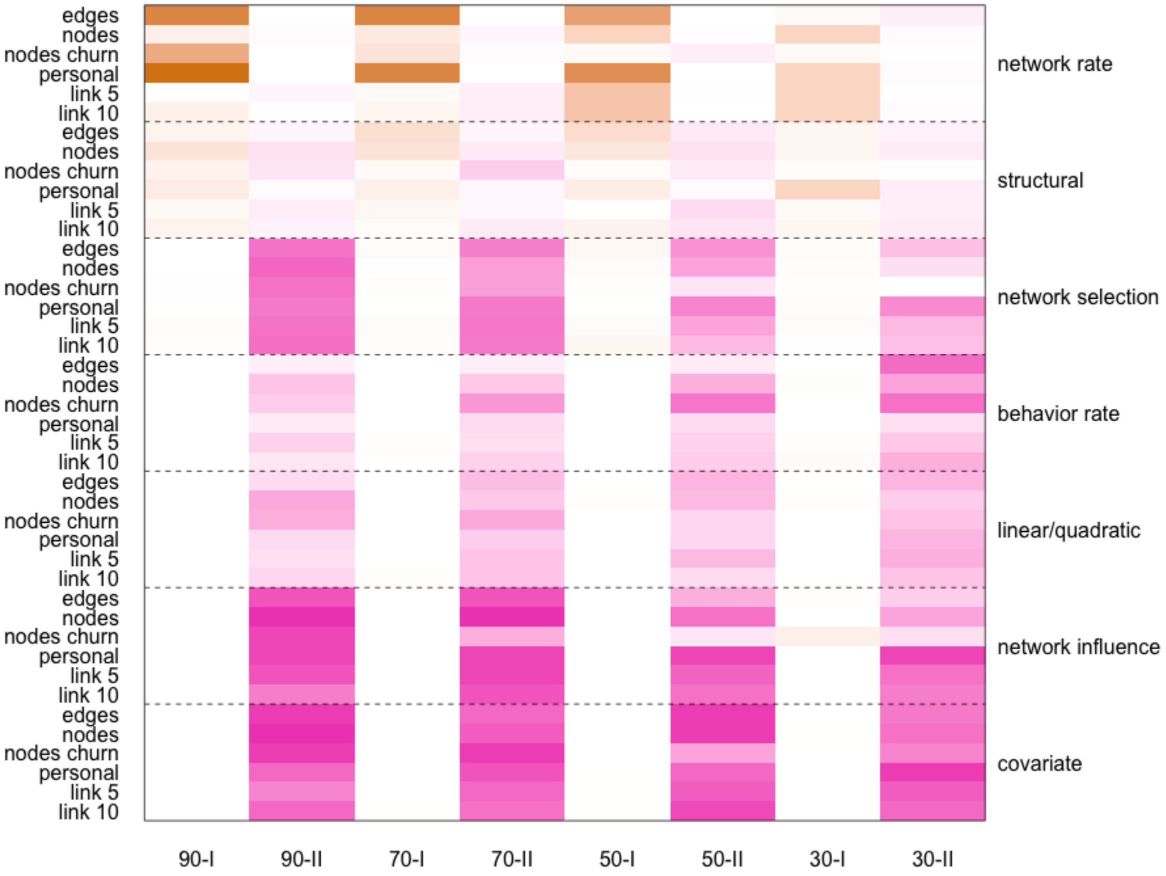


Figure 26: Summary across SIENA model parameter classes of the proportion of errors of each type for each sampling approach, across level of missingness; errors grouped by type.

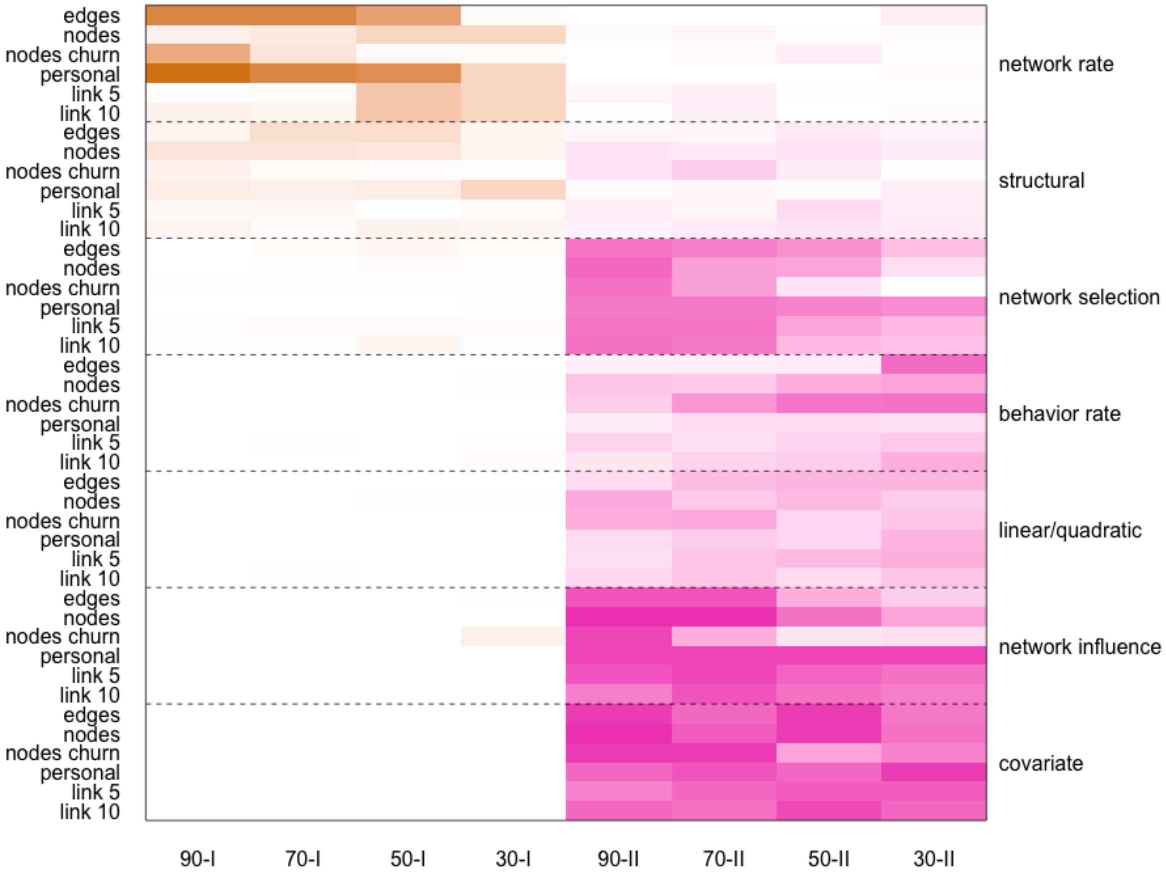


Figure 27: Summary across sampling approach of the proportion of errors of each type for each SIENA model parameter class, across level of missingness; errors grouped by level.

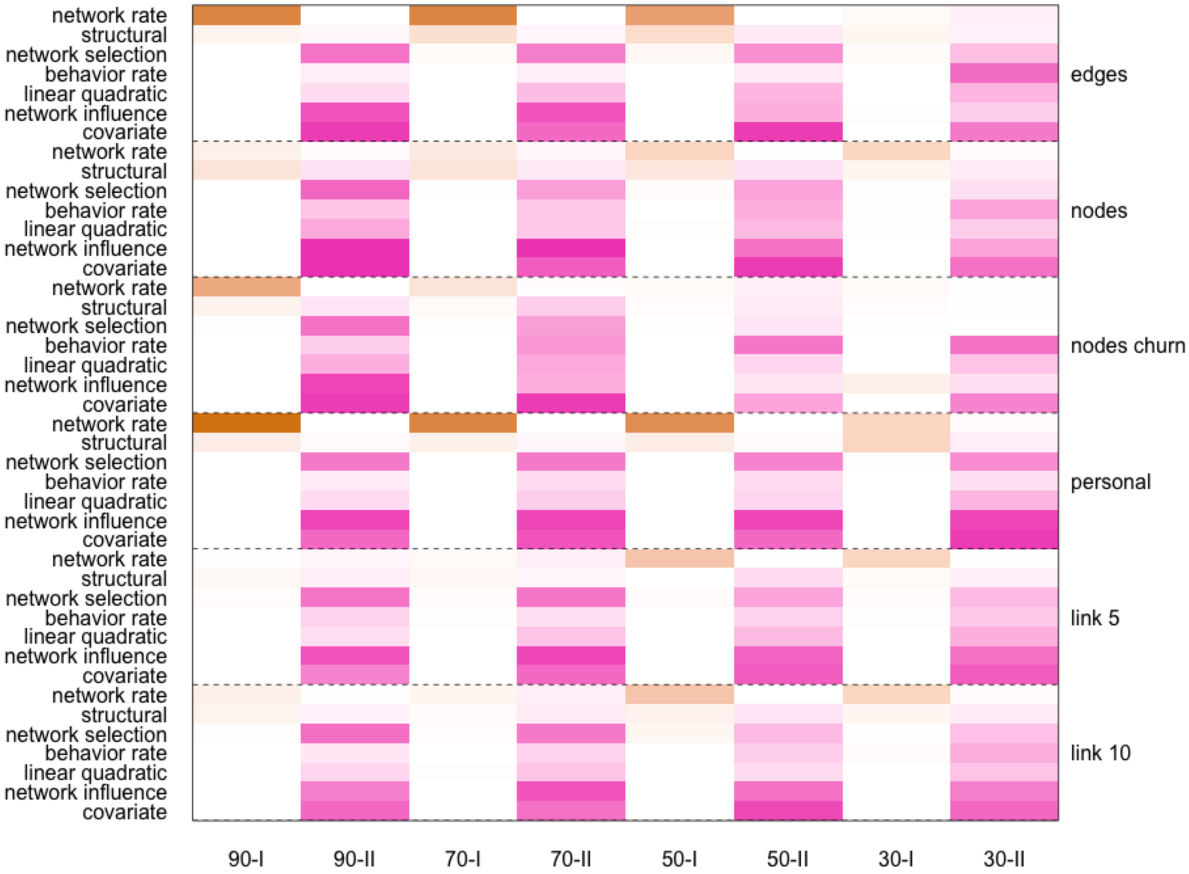


Figure 28: Summary across sampling approach of the proportion of errors of each type for each SIENA model parameter class, across level of missingness; errors grouped by type.

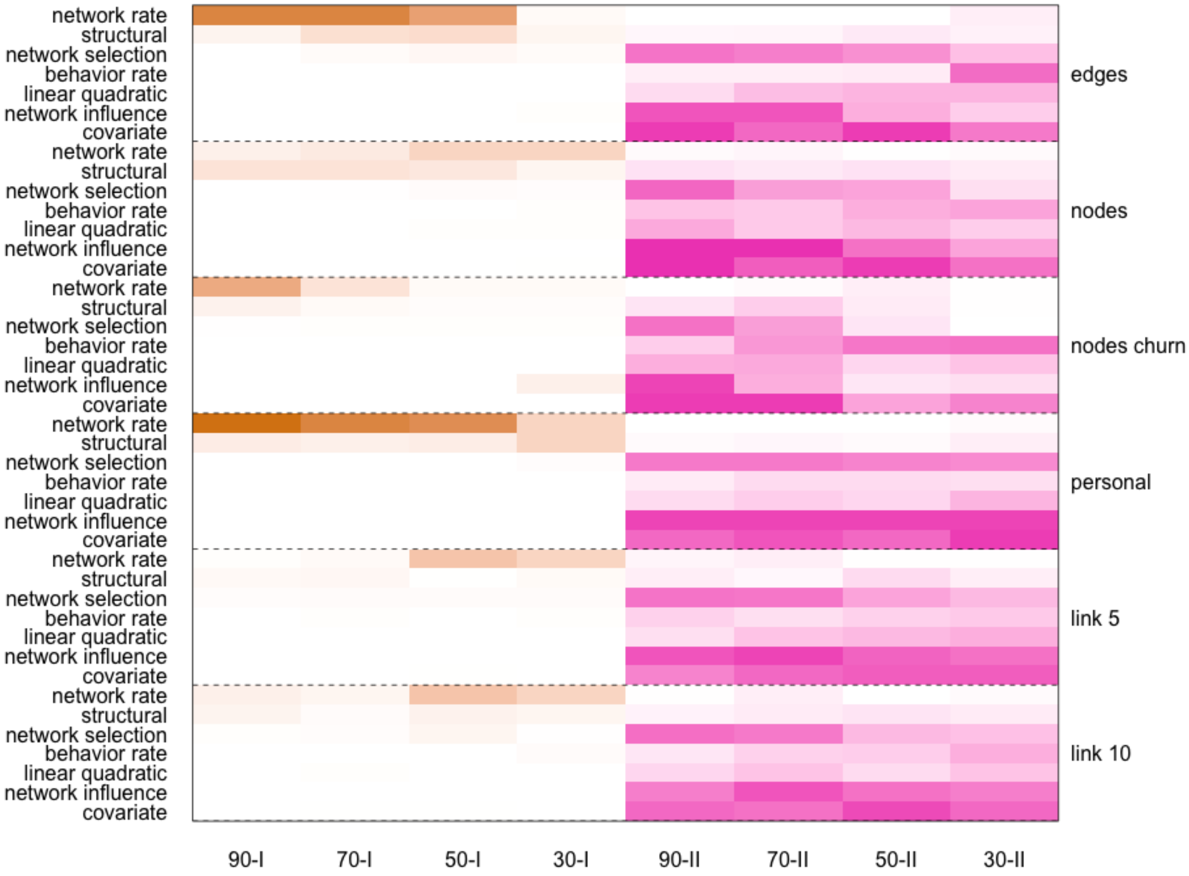


Figure 29: Deviation and significance level of parameters based on ERGM/p* models with 10% missing edges.

Figure 30: Deviation and significance level of parameters based on ERGM/p* models with 30% missing edges.

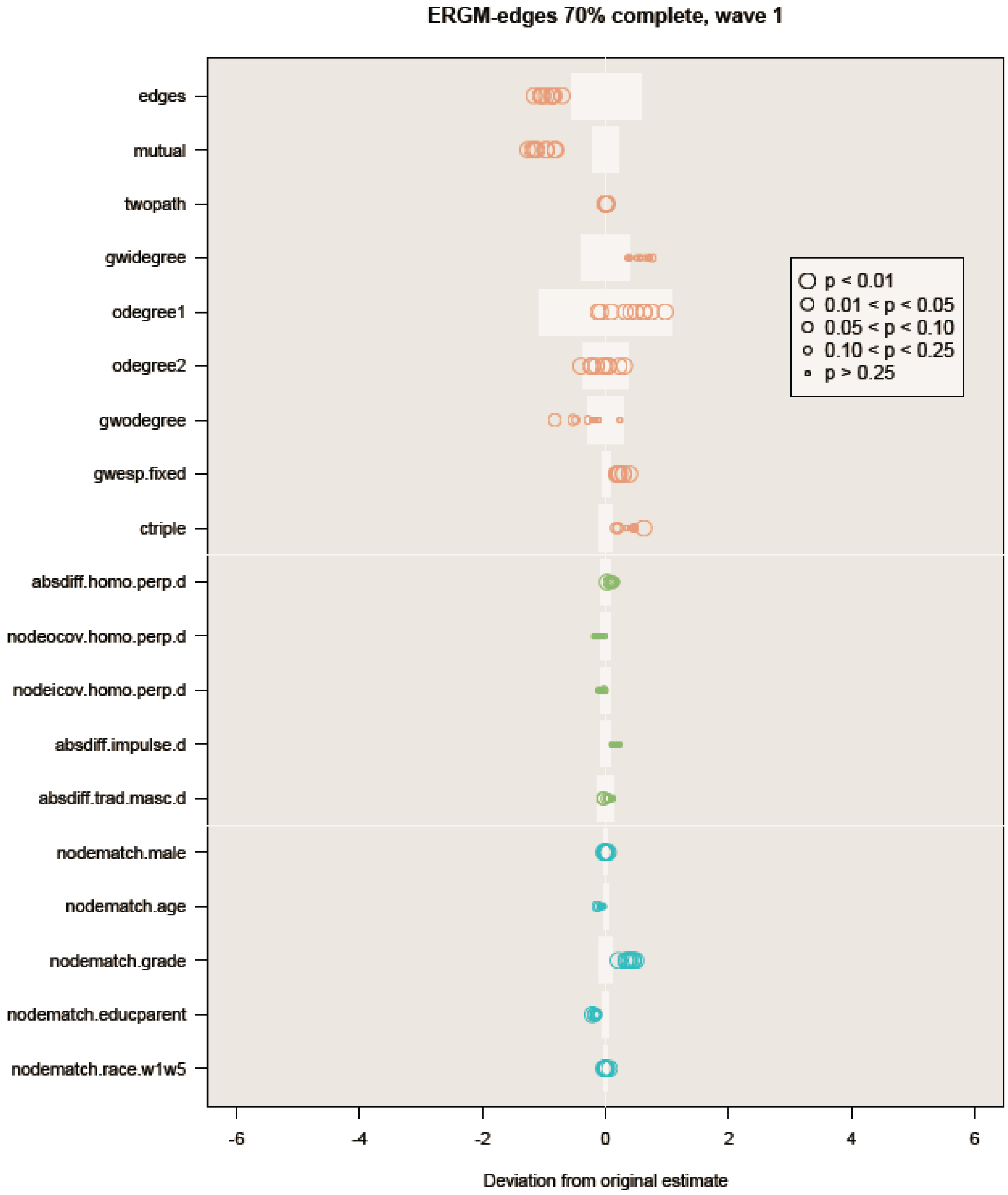


Figure 31: Deviation and significance level of parameters based on ERGM/p* models with 50% missing edges.

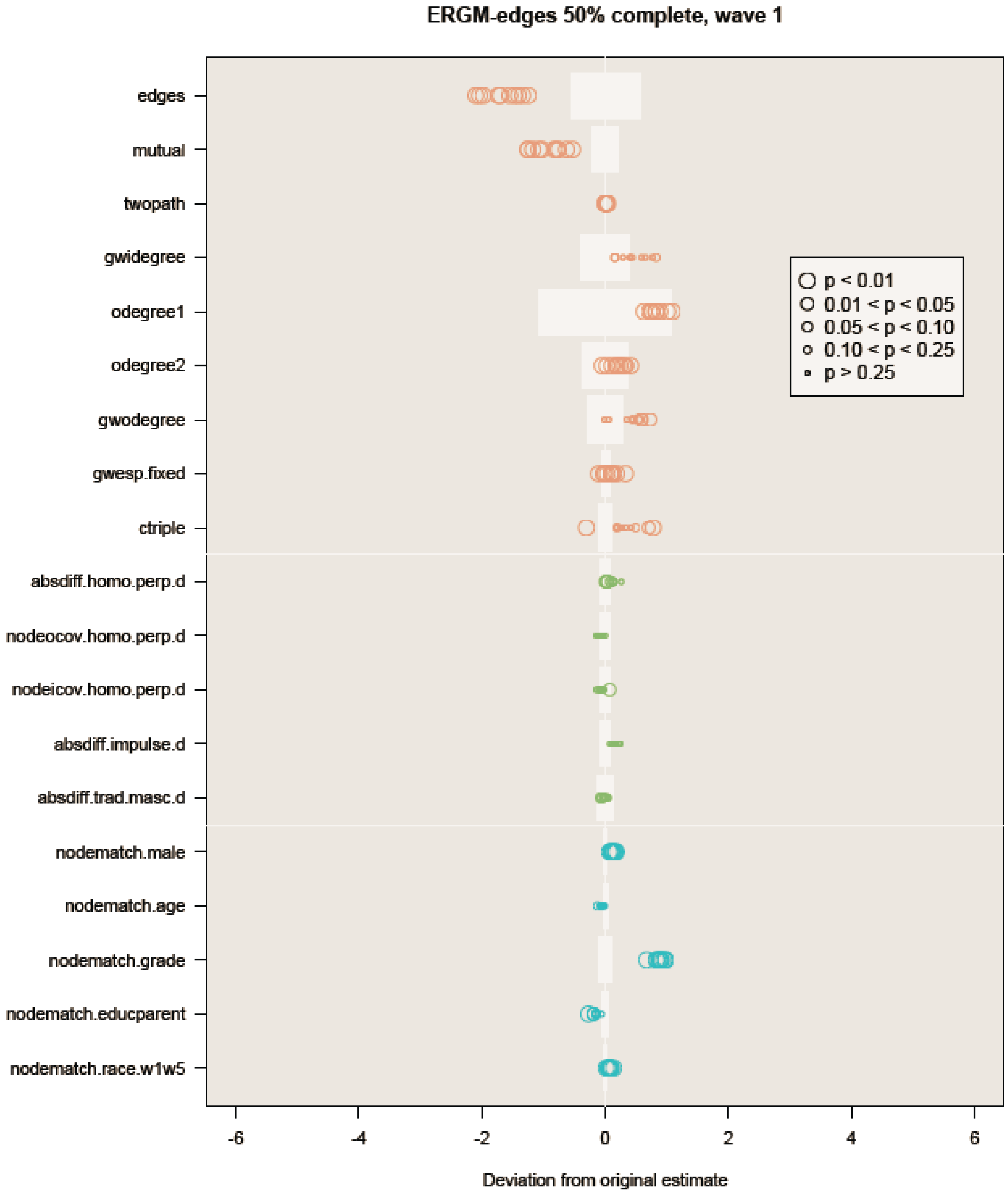


Figure 32: Deviation and significance level of parameters based on ERGM/P* models with 70% missing edges.

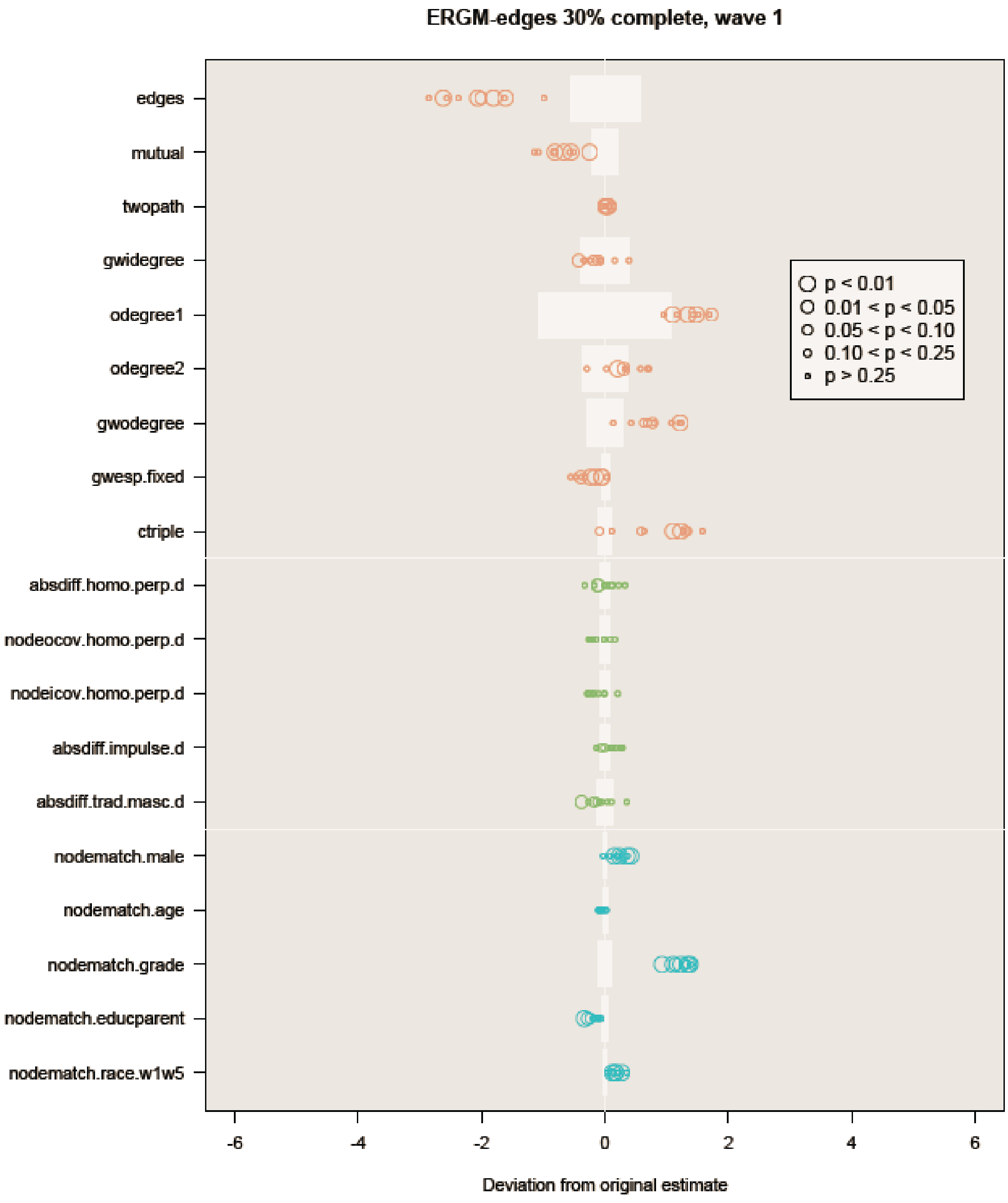


Figure 33: Deviation and significance level of parameters based on ERGM/P* models with 10% missing nodes.

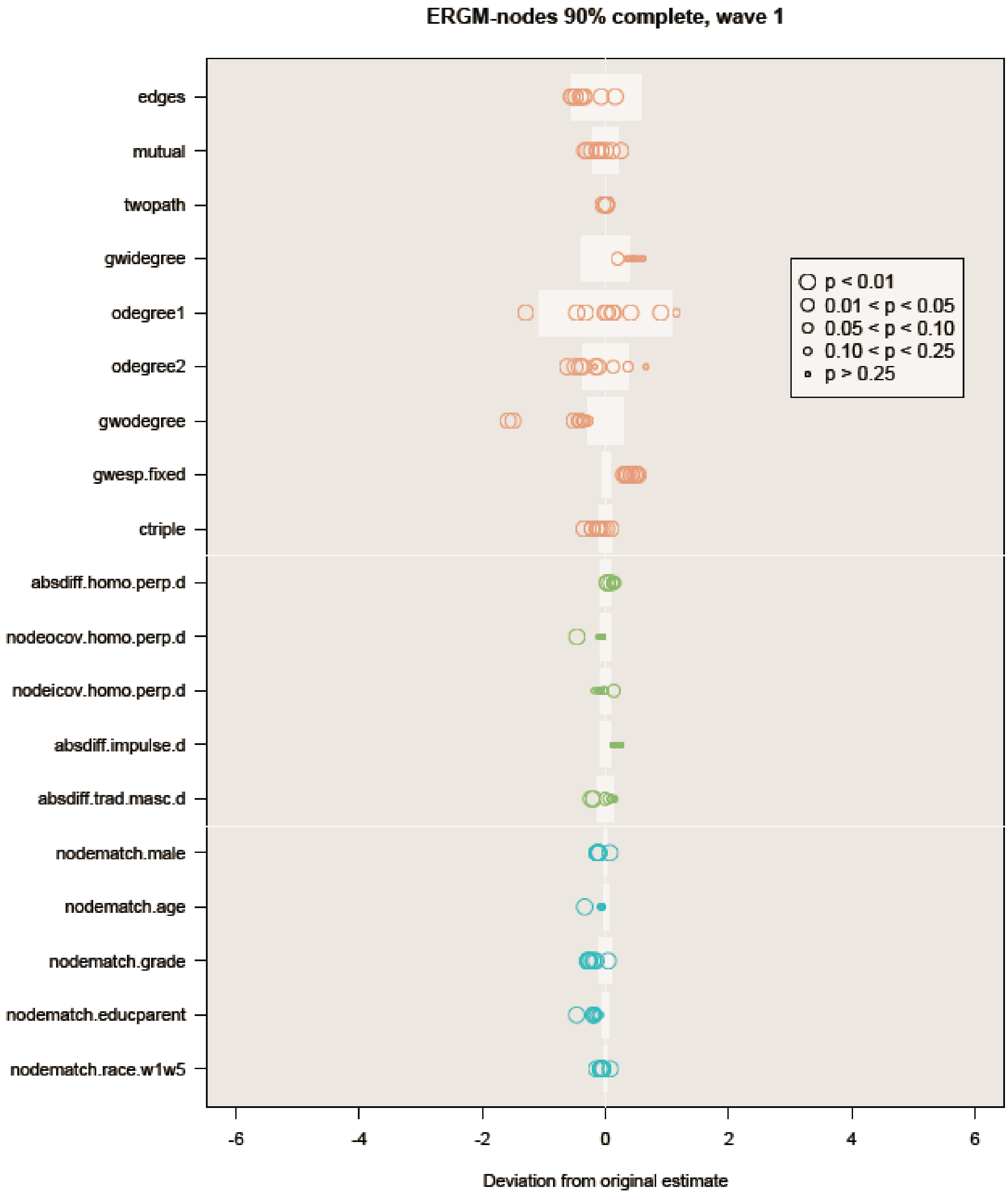


Figure 34: Deviation and significance level of parameters based on ERGM/P* models with 30% missing nodes.

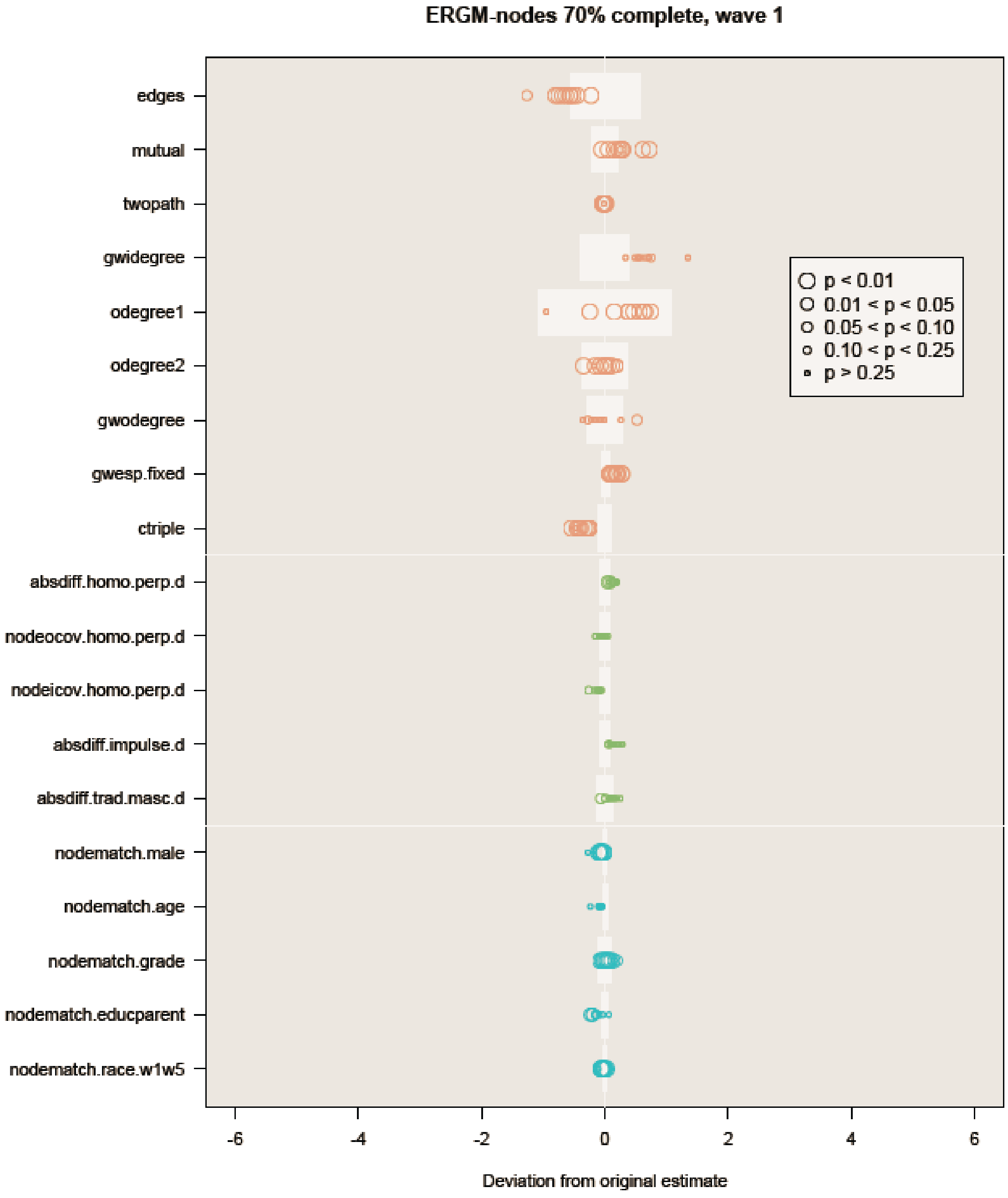


Figure 35: Deviation and significance level of parameters based on ERGM/P* models with 50% missing nodes.

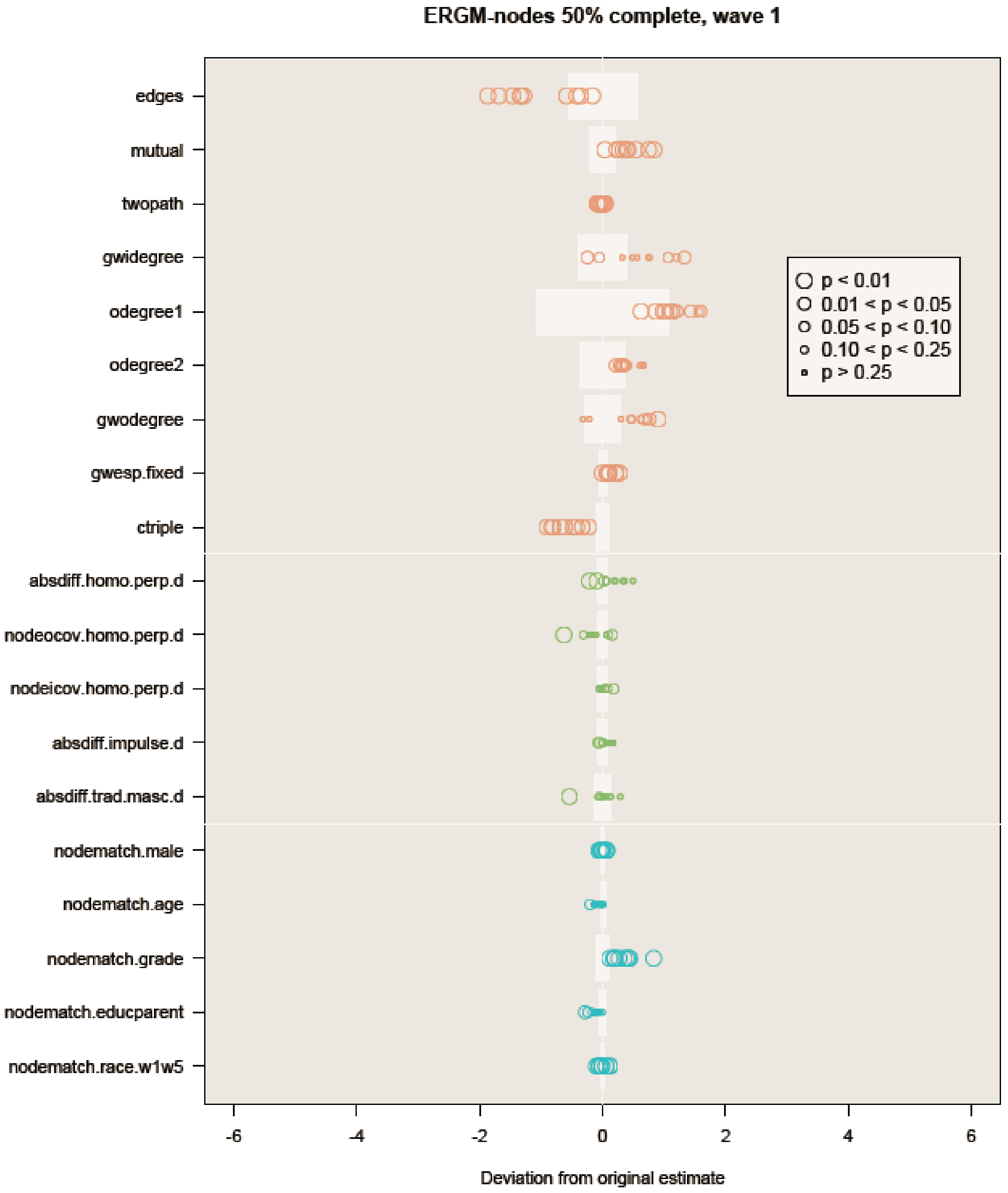


Figure 36: Deviation and significance level of parameters based on ERGM/P* models with 70% missing nodes.

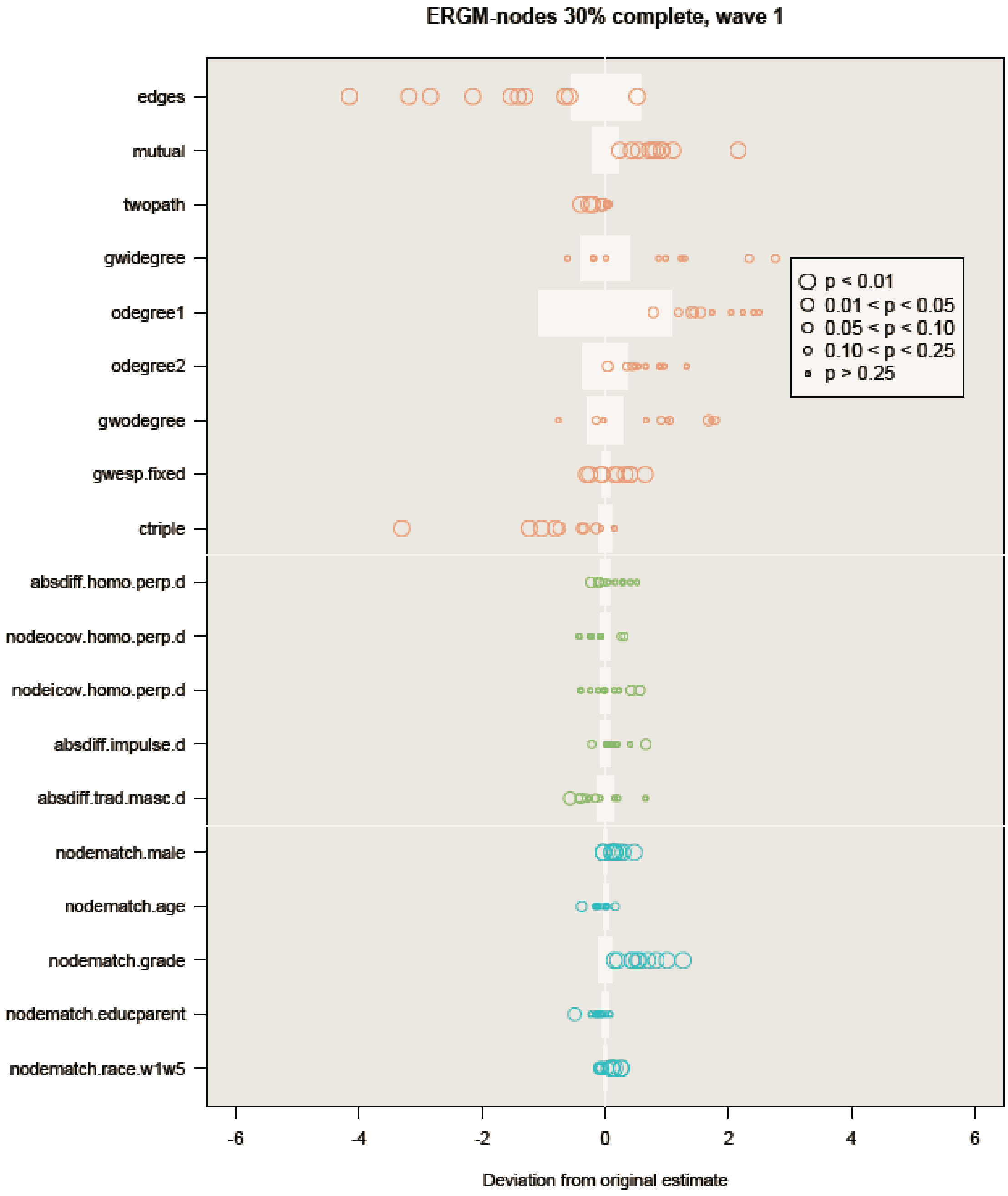


Figure 37: Deviation and significance level of parameters based on ERGM/P* models with 10% missing nodes, built from snowball sampling from randomly chosen seeds that account for 30% of the population.

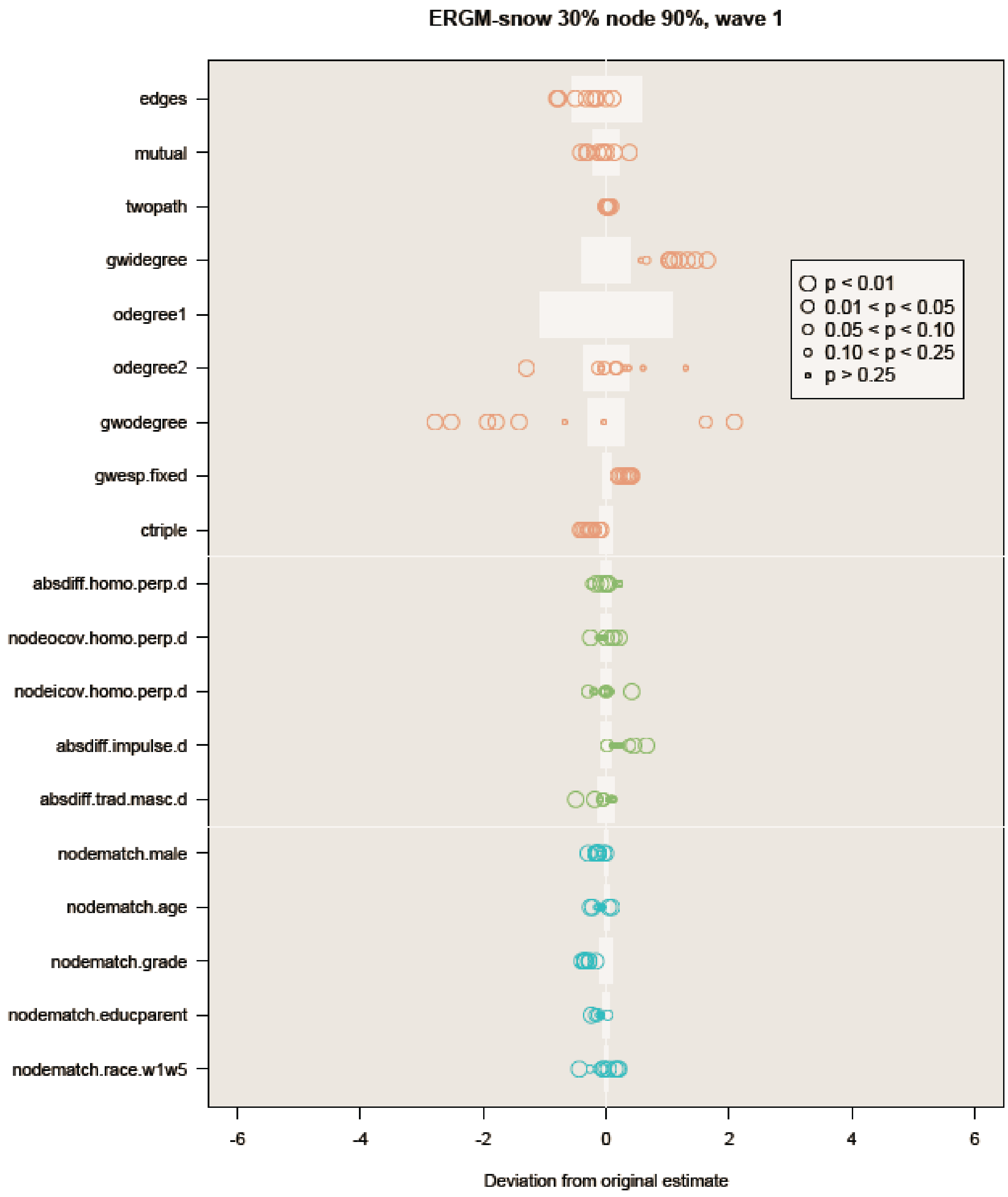


Figure 38: Deviation and significance level of parameters based on ERGM/P* models with 30% missing nodes, built from snowball sampling from randomly chosen seeds that account for 30% of the population.

Figure 39: Deviation and significance level of parameters based on ERGM/P* models with 10% missing edges, built from snowball sampling from randomly chosen seeds that account for 30% of the population.

Figure 40: Deviation and significance level of parameters based on ERGM/P* models with 30% missing edges, built from snowball sampling from randomly chosen seeds that account for 30% of the population.

Figure 41: Deviation and significance level of parameters based on ERGM/P* models with 10% missing nodes, built from snowball sampling from randomly chosen seeds that account for 10% of the population.

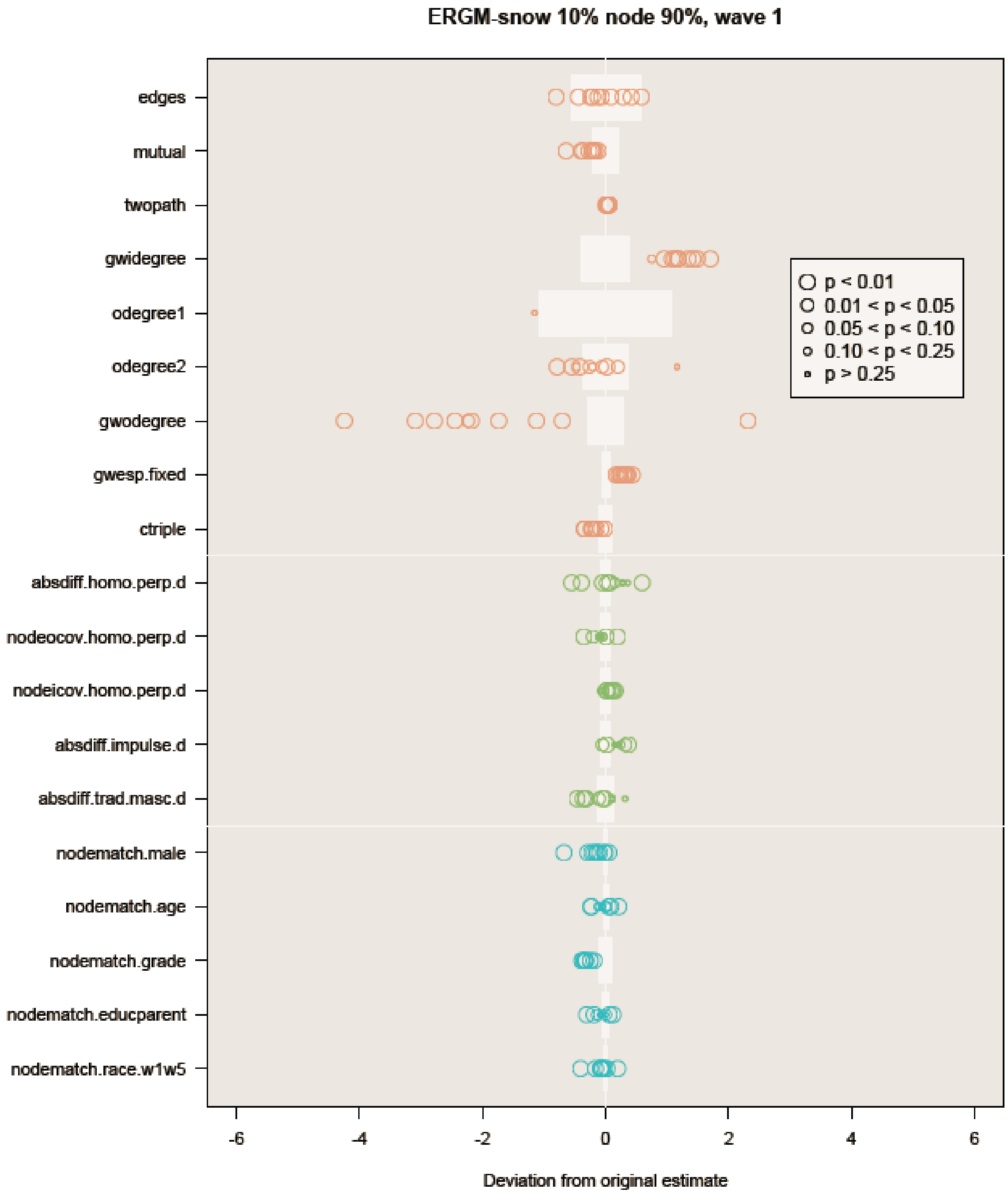


Figure 42: Deviation and significance level of parameters based on ERGM/P* models with 30% missing nodes, built from snowball sampling from randomly chosen seeds that account for 10% of the population.

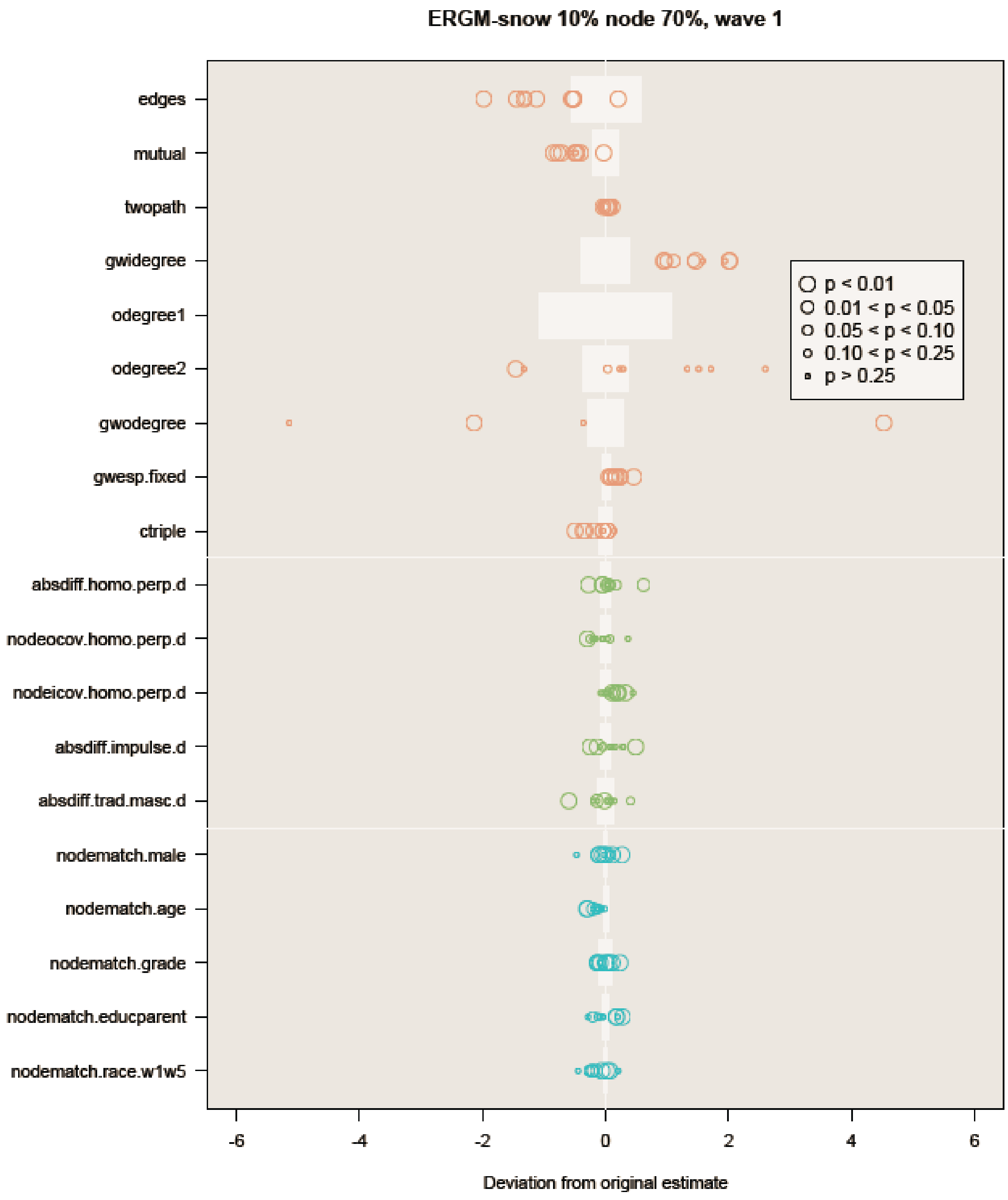


Figure 43: Deviation and significance level of parameters based on ERGM/P* models with 10% missing edges, built from snowball sampling from randomly chosen seeds that account for 10% of the population.

Figure 44: Deviation and significance level of parameters based on ERGM/P* models with 30% missing edges, built from snowball sampling from randomly chosen seeds that account for 10% of the population.

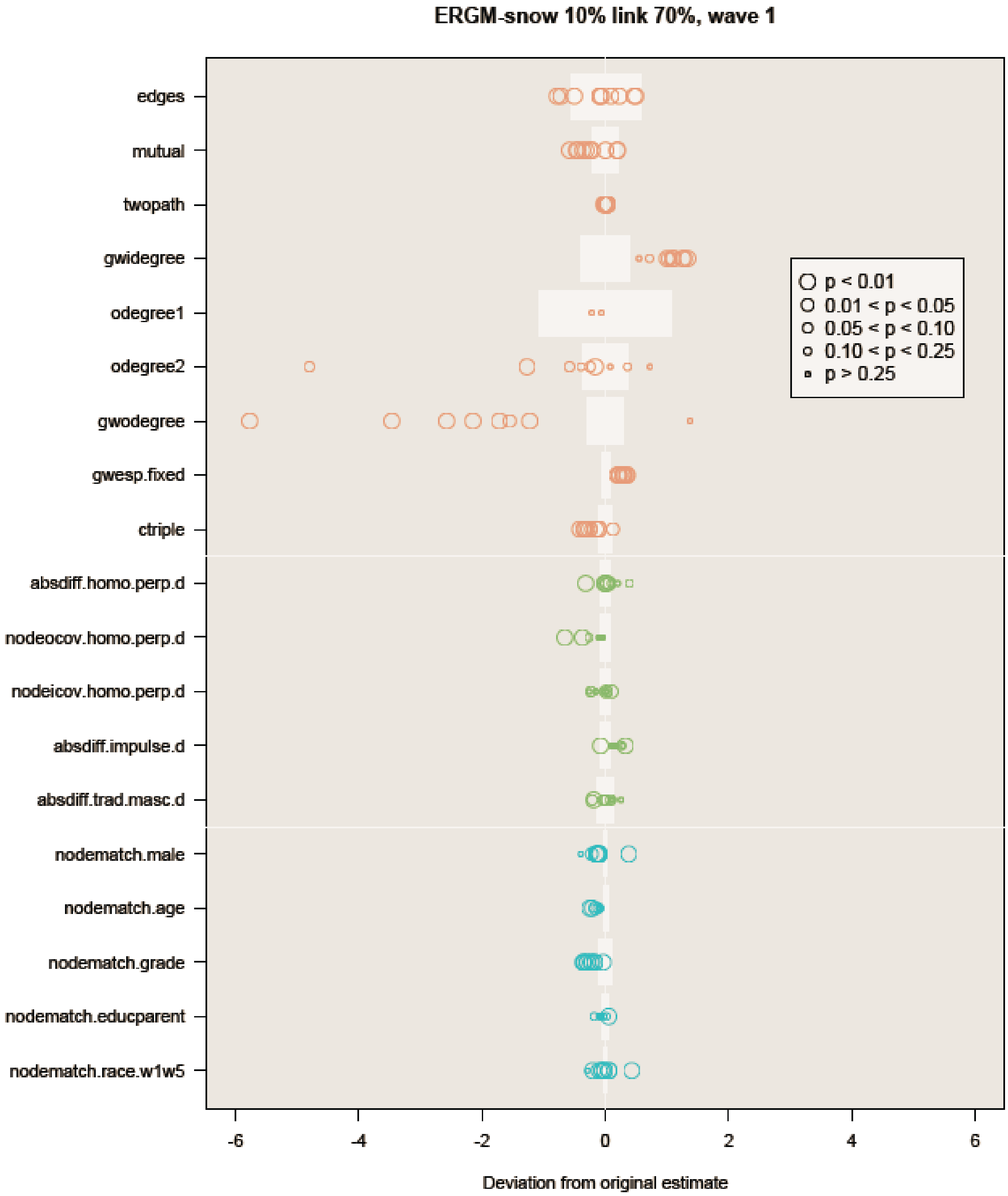


Figure 45: Deviation and significance level of parameters based on ERGM/P* models with 10% missing nodes, built from personal network sampling of the population.

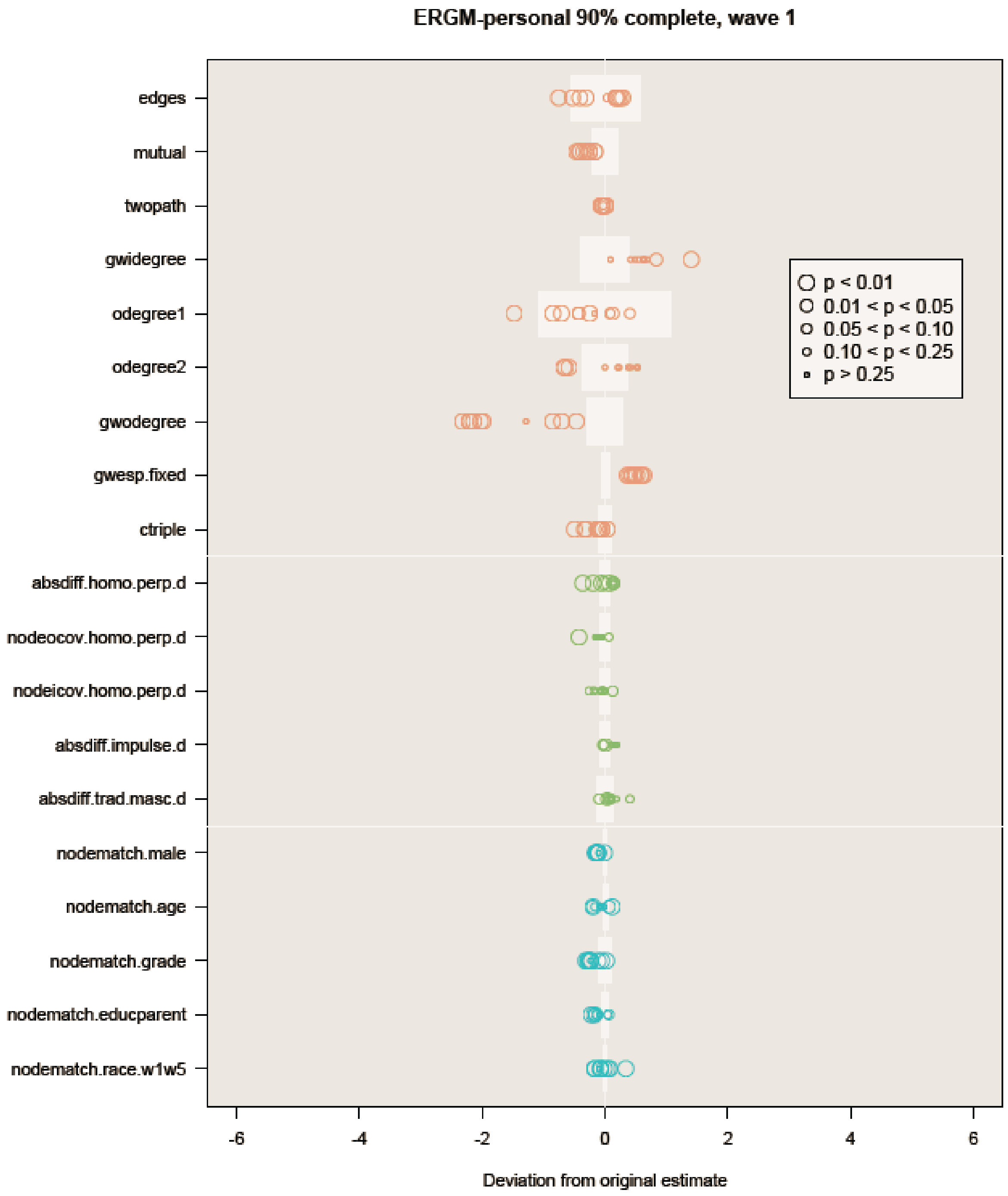


Figure 46: Deviation and significance level of parameters based on ERGM/P* models with 30% missing nodes, built from personal network sampling of the population.

Figure 47: Deviation and significance level of parameters based on ERGM/P* models with 50% missing nodes, built from personal network sampling of the population.

Figure 48: Deviation and significance level of parameters based on ERGM/P* models with 70% missing nodes, built from personal network sampling of the population.

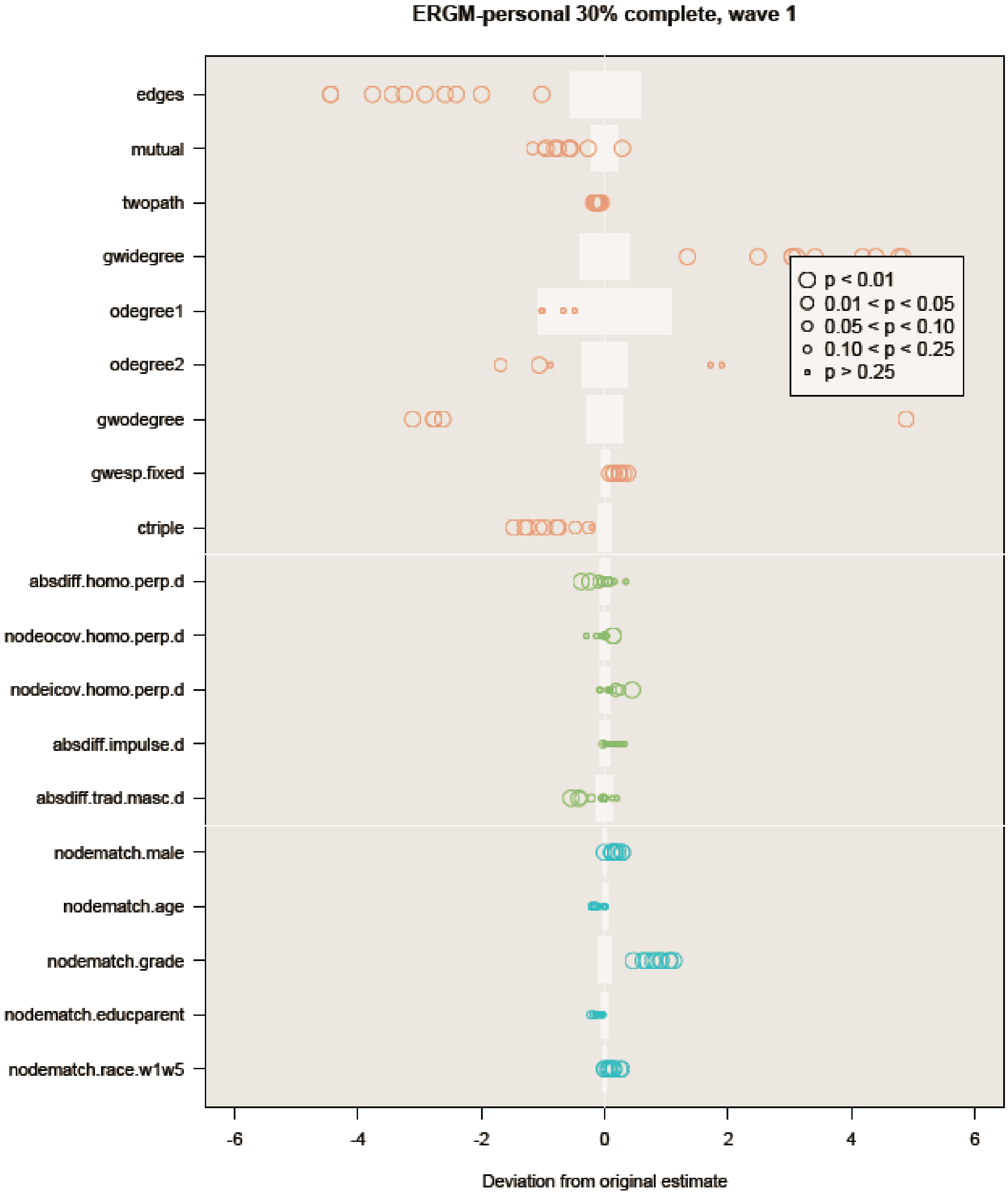


Figure 49: Summary across ERGM/P* model parameter classes of the proportion of errors of each type for each sampling approach, across level of missingness; errors grouped by level.

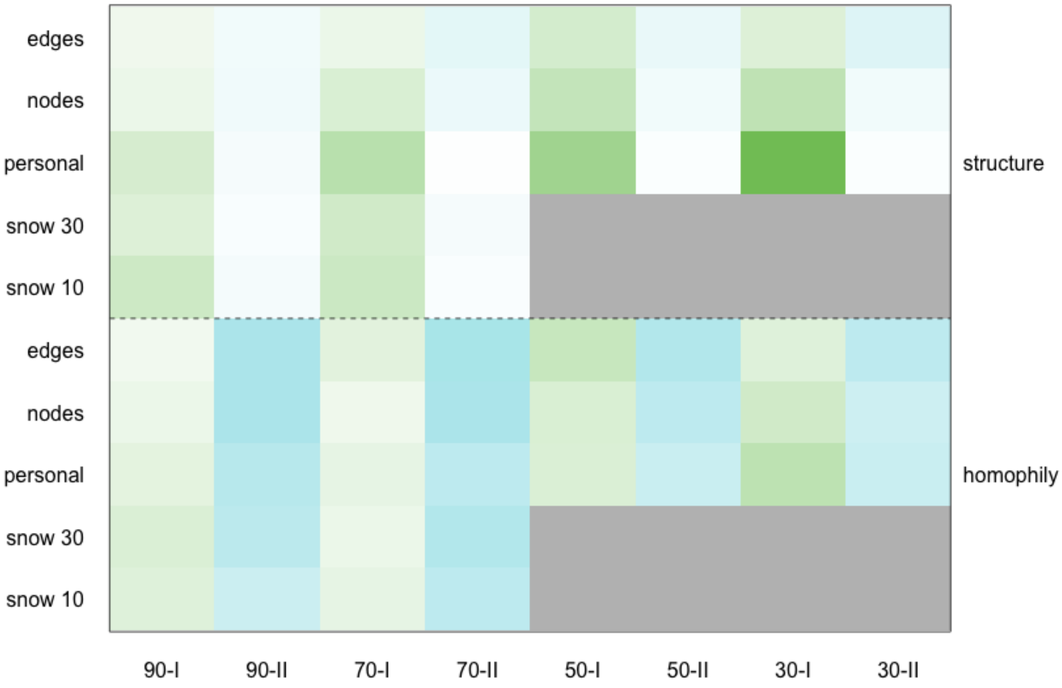


Figure 50: Summary across ERGM/P* model parameter classes of the proportion of errors of each type for each sampling approach, across level of missingness; errors grouped by type.

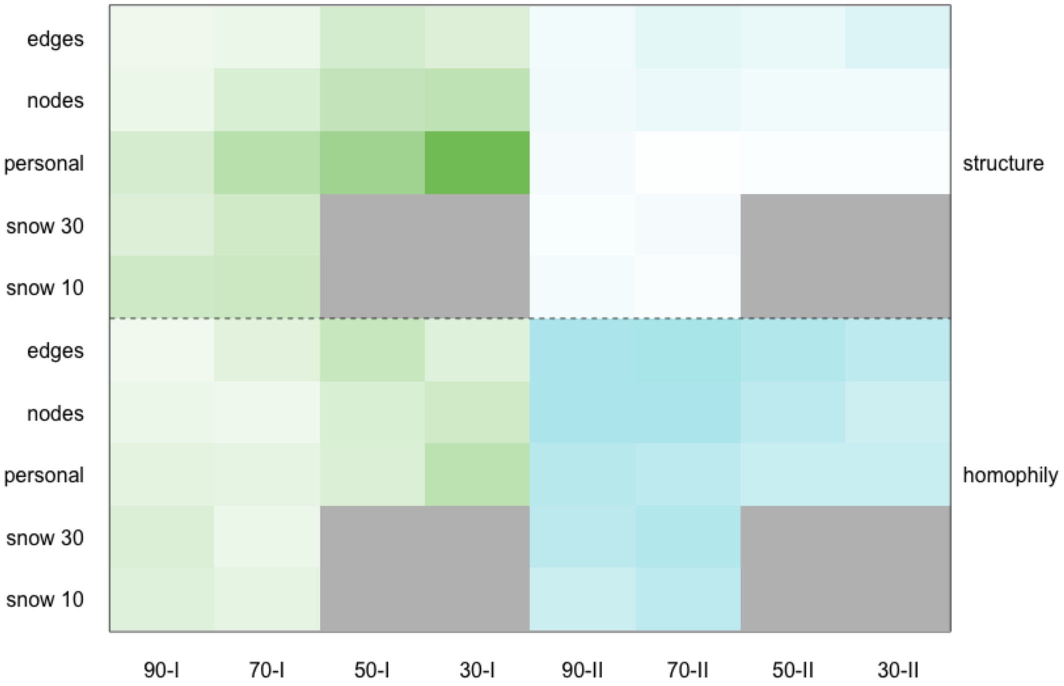


Figure 51: Summary across sampling approach of the proportion of errors of each type for each ERGM/P* model parameter class, across level of missingness; errors grouped by level.

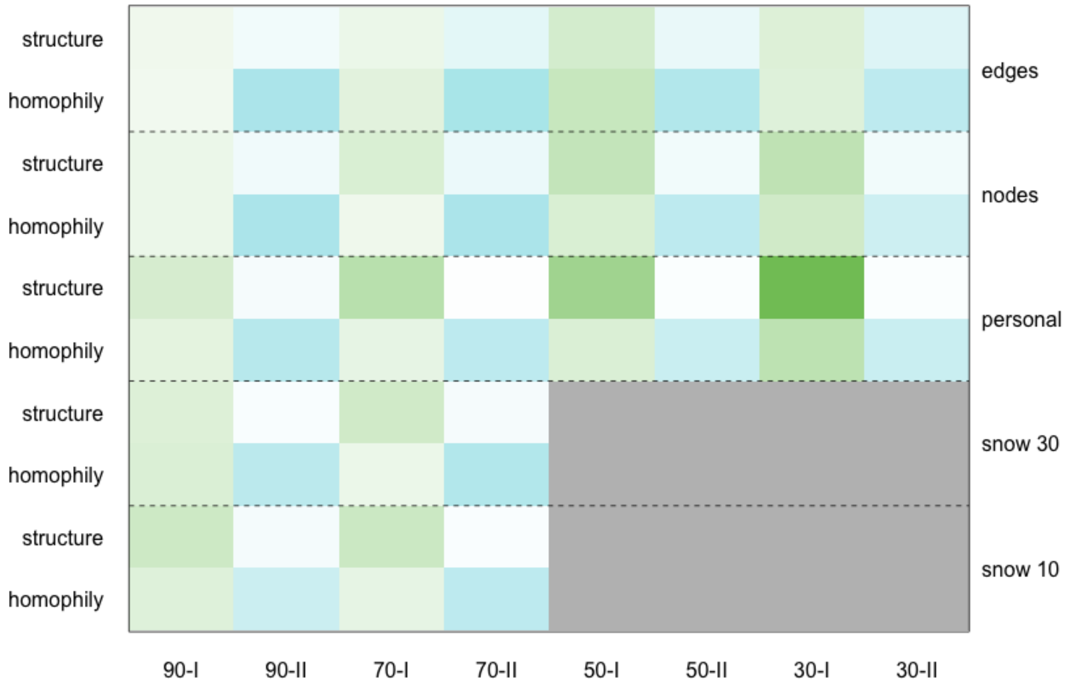
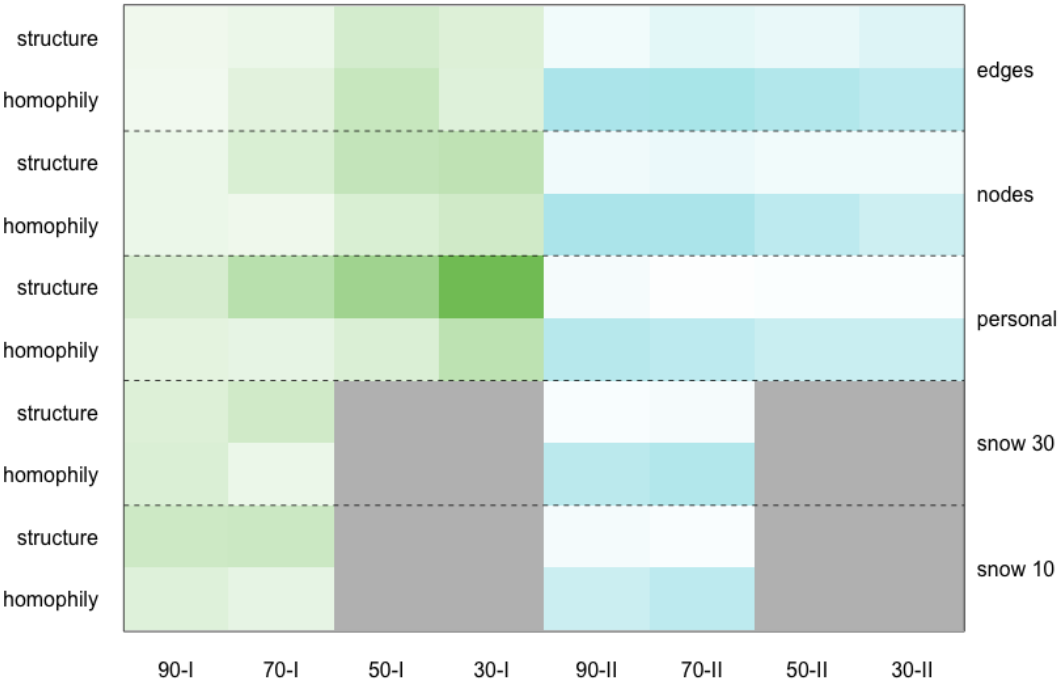
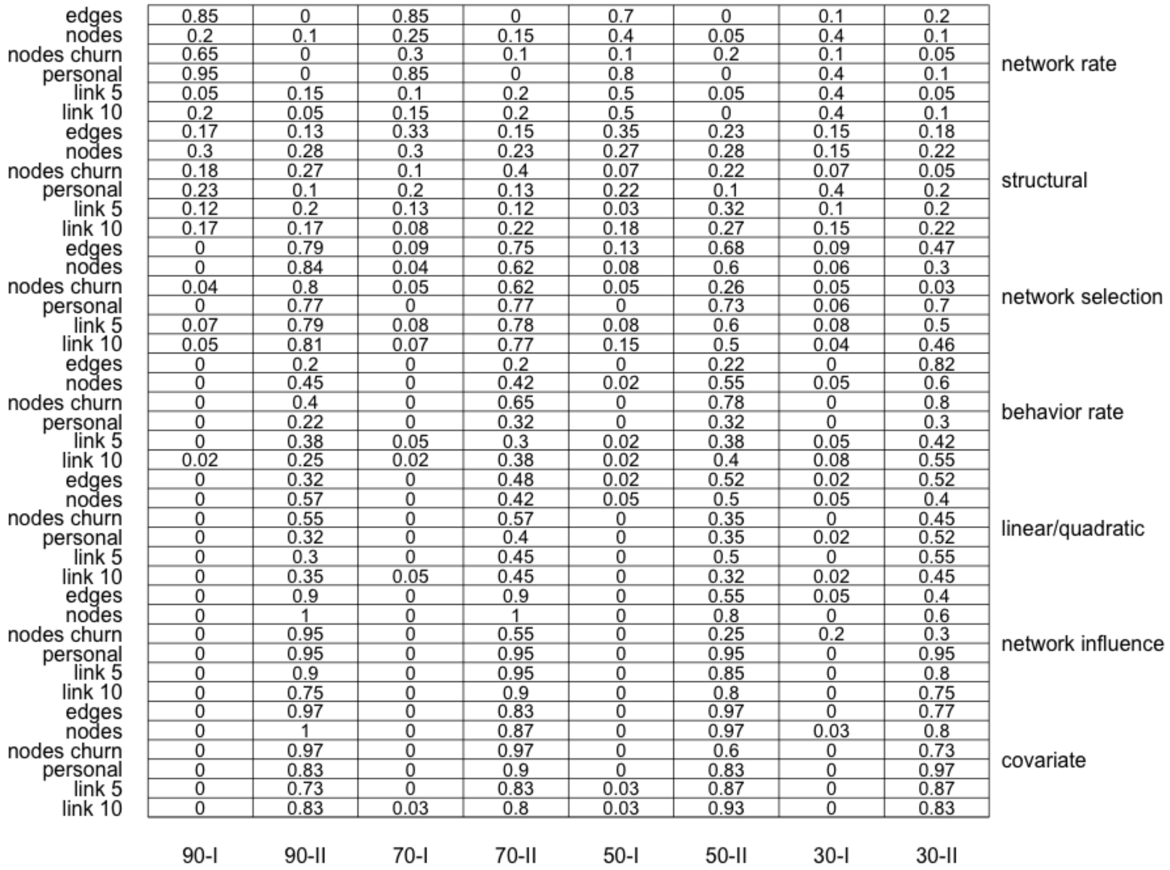


Figure 52: Summary across sampling approach of the proportion of errors of each type for each ERGM/P* model parameter class, across level of missingness; errors grouped by type.



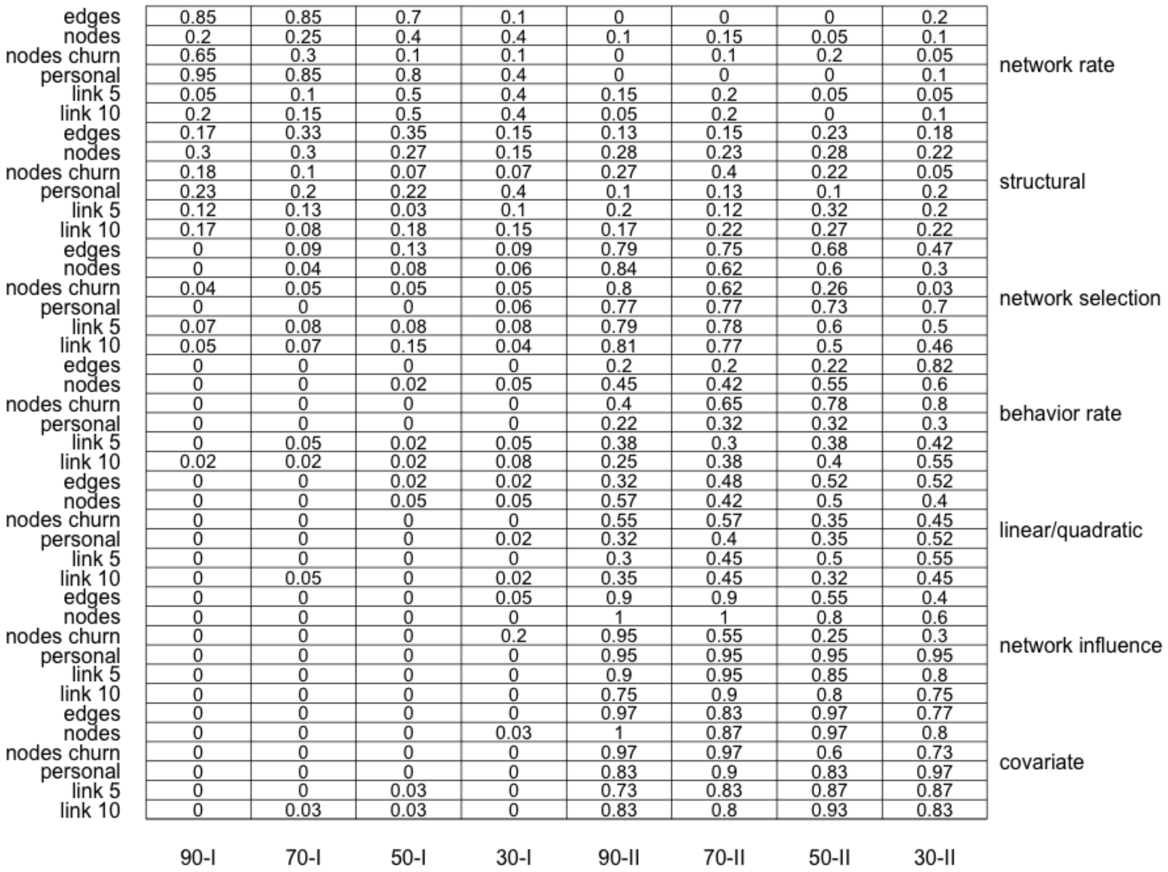
TABLES

Table 1: Summary across SIENA model parameter classes of the proportion of errors of each type for each sampling approach, across level of missingness; errors grouped by level.



	90-I	90-II	70-I	70-II	50-I	50-II	30-I	30-II	
edges	0.85	0	0.85	0	0.7	0	0.1	0.2	network rate
nodes	0.2	0.1	0.25	0.15	0.4	0.05	0.4	0.1	
nodes churn	0.65	0	0.3	0.1	0.1	0.2	0.1	0.05	
personal	0.95	0	0.85	0	0.8	0	0.4	0.1	
link 5	0.05	0.15	0.1	0.2	0.5	0.05	0.4	0.05	
link 10	0.2	0.05	0.15	0.2	0.5	0	0.4	0.1	
edges	0.17	0.13	0.33	0.15	0.35	0.23	0.15	0.18	structural
nodes	0.3	0.28	0.3	0.23	0.27	0.28	0.15	0.22	
nodes churn	0.18	0.27	0.1	0.4	0.07	0.22	0.07	0.05	
personal	0.23	0.1	0.2	0.13	0.22	0.1	0.4	0.2	
link 5	0.12	0.2	0.13	0.12	0.03	0.32	0.1	0.2	
link 10	0.17	0.17	0.08	0.22	0.18	0.27	0.15	0.22	
edges	0	0.79	0.09	0.75	0.13	0.68	0.09	0.47	network selection
nodes	0	0.84	0.04	0.62	0.08	0.6	0.06	0.3	
nodes churn	0.04	0.8	0.05	0.62	0.05	0.26	0.05	0.03	
personal	0	0.77	0	0.77	0	0.73	0.06	0.7	
link 5	0.07	0.79	0.08	0.78	0.08	0.6	0.08	0.5	
link 10	0.05	0.81	0.07	0.77	0.15	0.5	0.04	0.46	
edges	0	0.2	0	0.2	0	0.22	0	0.82	behavior rate
nodes	0	0.45	0	0.42	0.02	0.55	0.05	0.6	
nodes churn	0	0.4	0	0.65	0	0.78	0	0.8	
personal	0	0.22	0	0.32	0	0.32	0	0.3	
link 5	0	0.38	0.05	0.3	0.02	0.38	0.05	0.42	
link 10	0.02	0.25	0.02	0.38	0.02	0.4	0.08	0.55	
edges	0	0.32	0	0.48	0.02	0.52	0.02	0.52	linear/quadratic
nodes	0	0.57	0	0.42	0.05	0.5	0.05	0.4	
nodes churn	0	0.55	0	0.57	0	0.35	0	0.45	
personal	0	0.32	0	0.4	0	0.35	0.02	0.52	
link 5	0	0.3	0	0.45	0	0.5	0	0.55	
link 10	0	0.35	0.05	0.45	0	0.32	0.02	0.45	
edges	0	0.9	0	0.9	0	0.55	0.05	0.4	network influence
nodes	0	1	0	1	0	0.8	0	0.6	
nodes churn	0	0.95	0	0.55	0	0.25	0.2	0.3	
personal	0	0.95	0	0.95	0	0.95	0	0.95	
link 5	0	0.9	0	0.95	0	0.85	0	0.8	
link 10	0	0.75	0	0.9	0	0.8	0	0.75	
edges	0	0.97	0	0.83	0	0.97	0	0.77	covariate
nodes	0	1	0	0.87	0	0.97	0.03	0.8	
nodes churn	0	0.97	0	0.97	0	0.6	0	0.73	
personal	0	0.83	0	0.9	0	0.83	0	0.97	
link 5	0	0.73	0	0.83	0.03	0.87	0	0.87	
link 10	0	0.83	0.03	0.8	0.03	0.93	0	0.83	

Table 2: Summary across SIENA model parameter classes of the proportion of errors of each type for each sampling approach, across level of missingness; errors grouped by type.



	90-I	70-I	50-I	30-I	90-II	70-II	50-II	30-II	
edges	0.85	0.85	0.7	0.1	0	0	0	0.2	network rate
nodes	0.2	0.25	0.4	0.4	0.1	0.15	0.05	0.1	
nodes churn	0.65	0.3	0.1	0.1	0	0.1	0.2	0.05	
personal	0.95	0.85	0.8	0.4	0	0	0	0.1	
link 5	0.05	0.1	0.5	0.4	0.15	0.2	0.05	0.05	
link 10	0.2	0.15	0.5	0.4	0.05	0.2	0	0.1	
edges	0.17	0.33	0.35	0.15	0.13	0.15	0.23	0.18	structural
nodes	0.3	0.3	0.27	0.15	0.28	0.23	0.28	0.22	
nodes churn	0.18	0.1	0.07	0.07	0.27	0.4	0.22	0.05	
personal	0.23	0.2	0.22	0.4	0.1	0.13	0.1	0.2	
link 5	0.12	0.13	0.03	0.1	0.2	0.12	0.32	0.2	
link 10	0.17	0.08	0.18	0.15	0.17	0.22	0.27	0.22	
edges	0	0.09	0.13	0.09	0.79	0.75	0.68	0.47	network selection
nodes	0	0.04	0.08	0.06	0.84	0.62	0.6	0.3	
nodes churn	0.04	0.05	0.05	0.05	0.8	0.62	0.26	0.03	
personal	0	0	0	0.06	0.77	0.77	0.73	0.7	
link 5	0.07	0.08	0.08	0.08	0.79	0.78	0.6	0.5	
link 10	0.05	0.07	0.15	0.04	0.81	0.77	0.5	0.46	
edges	0	0	0	0	0.2	0.2	0.22	0.82	behavior rate
nodes	0	0	0.02	0.05	0.45	0.42	0.55	0.6	
nodes churn	0	0	0	0	0.4	0.65	0.78	0.8	
personal	0	0	0	0	0.22	0.32	0.32	0.3	
link 5	0	0.05	0.02	0.05	0.38	0.3	0.38	0.42	
link 10	0.02	0.02	0.02	0.08	0.25	0.38	0.4	0.55	
edges	0	0	0.02	0.02	0.32	0.48	0.52	0.52	linear/quadratic
nodes	0	0	0.05	0.05	0.57	0.42	0.5	0.4	
nodes churn	0	0	0	0	0.55	0.57	0.35	0.45	
personal	0	0	0	0.02	0.32	0.4	0.35	0.52	
link 5	0	0	0	0	0.3	0.45	0.5	0.55	
link 10	0	0.05	0	0.02	0.35	0.45	0.32	0.45	
edges	0	0	0	0.05	0.9	0.9	0.55	0.4	network influence
nodes	0	0	0	0	1	1	0.8	0.6	
nodes churn	0	0	0	0.2	0.95	0.55	0.25	0.3	
personal	0	0	0	0	0.95	0.95	0.95	0.95	
link 5	0	0	0	0	0.9	0.95	0.85	0.8	
link 10	0	0	0	0	0.75	0.9	0.8	0.75	
edges	0	0	0	0	0.97	0.83	0.97	0.77	covariate
nodes	0	0	0	0.03	1	0.87	0.97	0.8	
nodes churn	0	0	0	0	0.97	0.97	0.6	0.73	
personal	0	0	0	0	0.83	0.9	0.83	0.97	
link 5	0	0	0.03	0	0.73	0.83	0.87	0.87	
link 10	0	0.03	0.03	0	0.83	0.8	0.93	0.83	

Table 3: Summary across sampling approach of the proportion of errors of each type for each SIENA model parameter class, across level of missingness; errors grouped by level.

	90-I	90-II	70-I	70-II	50-I	50-II	30-I	30-II	
network rate	0.85	0	0.85	0	0.7	0	0.1	0.2	edges
structural	0.17	0.13	0.33	0.15	0.35	0.23	0.15	0.18	
network selection	0	0.79	0.09	0.75	0.13	0.68	0.09	0.47	
behavior rate	0	0.2	0	0.2	0	0.22	0	0.82	
linear quadratic	0	0.32	0	0.48	0.02	0.52	0.02	0.52	
network influence	0	0.9	0	0.9	0	0.55	0.05	0.4	
covariate	0	0.97	0	0.83	0	0.97	0	0.77	
network rate	0.2	0.1	0.25	0.15	0.4	0.05	0.4	0.1	nodes
structural	0.3	0.28	0.3	0.23	0.27	0.28	0.15	0.22	
network selection	0	0.84	0.04	0.62	0.08	0.6	0.06	0.3	
behavior rate	0	0.45	0	0.42	0.02	0.55	0.05	0.6	
linear quadratic	0	0.57	0	0.42	0.05	0.5	0.05	0.4	
network influence	0	1	0	1	0	0.8	0	0.6	
covariate	0	1	0	0.87	0	0.97	0.03	0.8	
network rate	0.65	0	0.3	0.1	0.1	0.2	0.1	0.05	nodes churn
structural	0.18	0.27	0.1	0.4	0.07	0.22	0.07	0.05	
network selection	0.04	0.8	0.05	0.62	0.05	0.26	0.05	0.03	
behavior rate	0	0.4	0	0.65	0	0.78	0	0.8	
linear quadratic	0	0.55	0	0.57	0	0.35	0	0.45	
network influence	0	0.95	0	0.55	0	0.25	0.2	0.3	
covariate	0	0.97	0	0.97	0	0.6	0	0.73	
network rate	0.95	0	0.85	0	0.8	0	0.4	0.1	personal
structural	0.23	0.1	0.2	0.13	0.22	0.1	0.4	0.2	
network selection	0	0.77	0	0.77	0	0.73	0.06	0.7	
behavior rate	0	0.22	0	0.32	0	0.32	0	0.3	
linear quadratic	0	0.32	0	0.4	0	0.35	0.02	0.52	
network influence	0	0.95	0	0.95	0	0.95	0	0.95	
covariate	0	0.83	0	0.9	0	0.83	0	0.97	
network rate	0.05	0.15	0.1	0.2	0.5	0.05	0.4	0.05	link 5
structural	0.12	0.2	0.13	0.12	0.03	0.32	0.1	0.2	
network selection	0.07	0.79	0.08	0.78	0.08	0.6	0.08	0.5	
behavior rate	0	0.38	0.05	0.3	0.02	0.38	0.05	0.42	
linear quadratic	0	0.3	0	0.45	0	0.5	0	0.55	
network influence	0	0.9	0	0.95	0	0.85	0	0.8	
covariate	0	0.73	0	0.83	0.03	0.87	0	0.87	
network rate	0.2	0.05	0.15	0.2	0.5	0	0.4	0.1	link 10
structural	0.17	0.17	0.08	0.22	0.18	0.27	0.15	0.22	
network selection	0.05	0.81	0.07	0.77	0.15	0.5	0.04	0.46	
behavior rate	0.02	0.25	0.02	0.38	0.02	0.4	0.08	0.55	
linear quadratic	0	0.35	0.05	0.45	0	0.32	0.02	0.45	
network influence	0	0.75	0	0.9	0	0.8	0	0.75	
covariate	0	0.83	0.03	0.8	0.03	0.93	0	0.83	

Table 4: Summary across sampling approach of the proportion of errors of each type for each SIENA model parameter class, across level of missingness; errors grouped by type.

	90-I	70-I	50-I	30-I	90-II	70-II	50-II	30-II	
network rate	0.85	0.85	0.7	0.1	0	0	0	0.2	edges
structural	0.17	0.33	0.35	0.15	0.13	0.15	0.23	0.18	
network selection	0	0.09	0.13	0.09	0.79	0.75	0.68	0.47	
behavior rate	0	0	0	0	0.2	0.2	0.22	0.82	
linear quadratic	0	0	0.02	0.02	0.32	0.48	0.52	0.52	
network influence	0	0	0	0.05	0.9	0.9	0.55	0.4	
covariate	0	0	0	0	0.97	0.83	0.97	0.77	
network rate	0.2	0.25	0.4	0.4	0.1	0.15	0.05	0.1	nodes
structural	0.3	0.3	0.27	0.15	0.28	0.23	0.28	0.22	
network selection	0	0.04	0.08	0.06	0.84	0.62	0.6	0.3	
behavior rate	0	0	0.02	0.05	0.45	0.42	0.55	0.6	
linear quadratic	0	0	0.05	0.05	0.57	0.42	0.5	0.4	
network influence	0	0	0	0	1	1	0.8	0.6	
covariate	0	0	0	0.03	1	0.87	0.97	0.8	
network rate	0.65	0.3	0.1	0.1	0	0.1	0.2	0.05	nodes churn
structural	0.18	0.1	0.07	0.07	0.27	0.4	0.22	0.05	
network selection	0.04	0.05	0.05	0.05	0.8	0.62	0.26	0.03	
behavior rate	0	0	0	0	0.4	0.65	0.78	0.8	
linear quadratic	0	0	0	0	0.55	0.57	0.35	0.45	
network influence	0	0	0	0.2	0.95	0.55	0.25	0.3	
covariate	0	0	0	0	0.97	0.97	0.6	0.73	
network rate	0.95	0.85	0.8	0.4	0	0	0	0.1	personal
structural	0.23	0.2	0.22	0.4	0.1	0.13	0.1	0.2	
network selection	0	0	0	0.06	0.77	0.77	0.73	0.7	
behavior rate	0	0	0	0	0.22	0.32	0.32	0.3	
linear quadratic	0	0	0	0.02	0.32	0.4	0.35	0.52	
network influence	0	0	0	0	0.95	0.95	0.95	0.95	
covariate	0	0	0	0	0.83	0.9	0.83	0.97	
network rate	0.05	0.1	0.5	0.4	0.15	0.2	0.05	0.05	link 5
structural	0.12	0.13	0.03	0.1	0.2	0.12	0.32	0.2	
network selection	0.07	0.08	0.08	0.08	0.79	0.78	0.6	0.5	
behavior rate	0	0.05	0.02	0.05	0.38	0.3	0.38	0.42	
linear quadratic	0	0	0	0	0.3	0.45	0.5	0.55	
network influence	0	0	0	0	0.9	0.95	0.85	0.8	
covariate	0	0	0.03	0	0.73	0.83	0.87	0.87	
network rate	0.2	0.15	0.5	0.4	0.05	0.2	0	0.1	link 10
structural	0.17	0.08	0.18	0.15	0.17	0.22	0.27	0.22	
network selection	0.05	0.07	0.15	0.04	0.81	0.77	0.5	0.46	
behavior rate	0.02	0.02	0.02	0.08	0.25	0.38	0.4	0.55	
linear quadratic	0	0.05	0	0.02	0.35	0.45	0.32	0.45	
network influence	0	0	0	0	0.75	0.9	0.8	0.75	
covariate	0	0.03	0.03	0	0.83	0.8	0.93	0.83	

Table 5: Summary across ERGM/p* model parameter classes of the proportion of errors of each type for each sampling approach, across level of missingness; errors grouped by level.

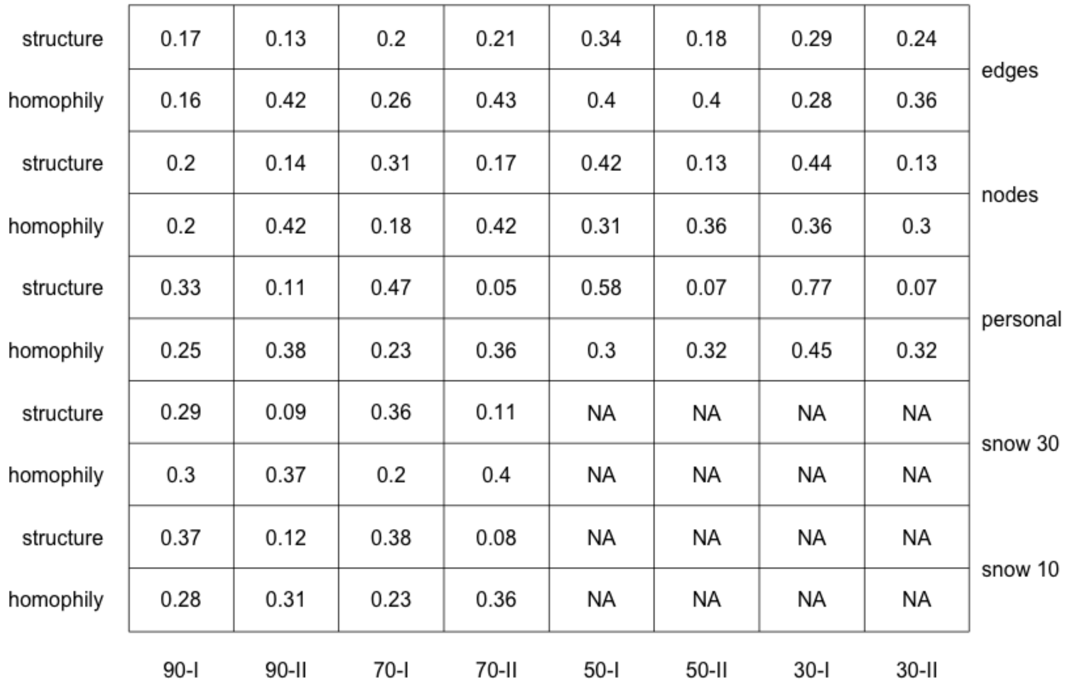
	90-I	90-II	70-I	70-II	50-I	50-II	30-I	30-II	
edges	0.17	0.13	0.2	0.21	0.34	0.18	0.29	0.24	
nodes	0.2	0.14	0.31	0.17	0.42	0.13	0.44	0.13	
personal	0.33	0.11	0.47	0.05	0.58	0.07	0.77	0.07	structure
snow 30	0.29	0.09	0.36	0.11	NA	NA	NA	NA	
snow 10	0.37	0.12	0.38	0.08	NA	NA	NA	NA	
edges	0.16	0.42	0.26	0.43	0.4	0.4	0.28	0.36	
nodes	0.2	0.42	0.18	0.42	0.31	0.36	0.36	0.3	
personal	0.25	0.38	0.23	0.36	0.3	0.32	0.45	0.32	homophily
snow 30	0.3	0.37	0.2	0.4	NA	NA	NA	NA	
snow 10	0.28	0.31	0.23	0.36	NA	NA	NA	NA	

.

Table 6: Summary across ERGM/p* model parameter classes of the proportion of errors of each type for each sampling approach, across level of missingness; errors grouped by type.

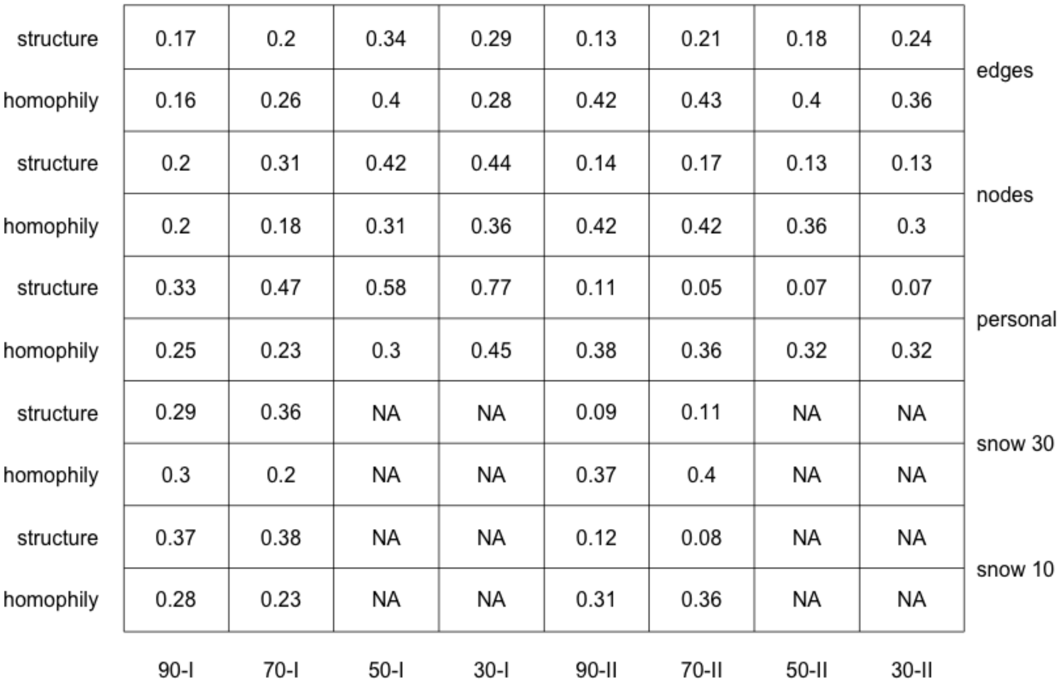
	90-I	70-I	50-I	30-I	90-II	70-II	50-II	30-II	
edges	0.17	0.2	0.34	0.29	0.13	0.21	0.18	0.24	structure
nodes	0.2	0.31	0.42	0.44	0.14	0.17	0.13	0.13	
personal	0.33	0.47	0.58	0.77	0.11	0.05	0.07	0.07	
snow 30	0.29	0.36	NA	NA	0.09	0.11	NA	NA	
snow 10	0.37	0.38	NA	NA	0.12	0.08	NA	NA	
edges	0.16	0.26	0.4	0.28	0.42	0.43	0.4	0.36	homophily
nodes	0.2	0.18	0.31	0.36	0.42	0.42	0.36	0.3	
personal	0.25	0.23	0.3	0.45	0.38	0.36	0.32	0.32	
snow 30	0.3	0.2	NA	NA	0.37	0.4	NA	NA	
snow 10	0.28	0.23	NA	NA	0.31	0.36	NA	NA	

Table 7: Summary across sampling approach of the proportion of errors of each type for each ERGM/p* model parameter class, across level of missingness; errors grouped by level.



		90-I	90-II	70-I	70-II	50-I	50-II	30-I	30-II
edges	structure	0.17	0.13	0.2	0.21	0.34	0.18	0.29	0.24
	homophily	0.16	0.42	0.26	0.43	0.4	0.4	0.28	0.36
nodes	structure	0.2	0.14	0.31	0.17	0.42	0.13	0.44	0.13
	homophily	0.2	0.42	0.18	0.42	0.31	0.36	0.36	0.3
personal	structure	0.33	0.11	0.47	0.05	0.58	0.07	0.77	0.07
	homophily	0.25	0.38	0.23	0.36	0.3	0.32	0.45	0.32
snow 30	structure	0.29	0.09	0.36	0.11	NA	NA	NA	NA
	homophily	0.3	0.37	0.2	0.4	NA	NA	NA	NA
snow 10	structure	0.37	0.12	0.38	0.08	NA	NA	NA	NA
	homophily	0.28	0.31	0.23	0.36	NA	NA	NA	NA

Table 8: Summary across sampling approach of the proportion of errors of each type for each ERGM/p* model parameter class, across level of missingness; errors grouped by type.



	90-I	70-I	50-I	30-I	90-II	70-II	50-II	30-II	
structure	0.17	0.2	0.34	0.29	0.13	0.21	0.18	0.24	edges
homophily	0.16	0.26	0.4	0.28	0.42	0.43	0.4	0.36	
structure	0.2	0.31	0.42	0.44	0.14	0.17	0.13	0.13	nodes
homophily	0.2	0.18	0.31	0.36	0.42	0.42	0.36	0.3	
structure	0.33	0.47	0.58	0.77	0.11	0.05	0.07	0.07	personal
homophily	0.25	0.23	0.3	0.45	0.38	0.36	0.32	0.32	
structure	0.29	0.36	NA	NA	0.09	0.11	NA	NA	snow 30
homophily	0.3	0.2	NA	NA	0.37	0.4	NA	NA	
structure	0.37	0.38	NA	NA	0.12	0.08	NA	NA	snow 10
homophily	0.28	0.23	NA	NA	0.31	0.36	NA	NA	

Table 9: parameter estimates for the baseline SIENA model used as a testbed for this simulation study.

parameter	estimate	sterr	pvalue
constant friend rate (period 1)	16.12	1.51	0
constant friend rate (period 2)	16.09	1.78	0
outdegree (density)	-2.36	0.18	0
reciprocity	2.24	0.17	0
transitive triplets	0.5	0.05	0
3-cycles	-0.35	0.06	0
indegree - activity (sqrt)	-0.53	0.17	0.002
outdegree - activity (sqrt)	0.23	0.06	0
male alter	0.04	0.09	0.622
male ego	-0.16	0.07	0.032
same male	0.4	0.06	0
same race	0.22	0.07	0.001
bully.beh alter	0.13	0.15	0.371
bully.beh squared alter	-0.11	0.09	0.25
bully.beh ego	-0.06	0.05	0.221
bully.beh similarity	-0.25	0.53	0.637
homo.perp alter	0.51	0.21	0.016
homo.perp squared alter	-0.32	0.14	0.019
homo.perp ego	0.08	0.06	0.159
homo.perp similarity	1.36	0.6	0.023
same grade	0.5	0.08	0
rate bully.beh (period 1)	3.43	0.72	0
rate bully.beh (period 2)	4.53	1.17	0
behavior bully.beh linear shape	-0.29	0.06	0
behavior bully.beh quadratic shape	0.09	0.07	0.22
behavior bully.beh average similarity	1.75	1.84	0.341
behavior bully.beh: effect from impulse	0.07	0.04	0.067
rate homo.perp (period 1)	10.62	4.29	0.013
rate homo.perp (period 2)	14.57	4.59	0.001
behavior homo.perp linear shape	-0.28	0.05	0
behavior homo.perp quadratic shape	0.13	0.03	0
behavior homo.perp average similarity	2.18	0.99	0.027
behavior homo.perp: effect from educparent	-0.06	0.04	0.071
behavior homo.perp: effect from impulse	0.04	0.02	0.032

Table 10: parameter estimates and standard errors for the ERGM/p* models uses as a testbed for this simulation study.

parameter	est.se	pvalue
edges	-4.927 (0.355)	0
mutual	1.856 (0.168)	0
twopath	-0.151 (0.015)	0
gwidegree	-0.139 (0.239)	0.561
odegree1	-2.224 (0.761)	0.003
odegree2	-1.375 (0.475)	0.004
gwodegree	-0.318 (0.246)	0.196
gwesp	1.723 (0.072)	0
ctriple	-0.206 (0.07)	0.003
absdiff.homo.perp.d	-0.251 (0.076)	0.001
nodeocov.homo.perp.d	0.034 (0.07)	0.625
nodeicov.homo.perp.d	0.09 (0.068)	0.188
absdiff.impulse.d	-0.091 (0.074)	0.22
absdiff.trad.masc.d	-0.23 (0.109)	0.036
nodematch.male	0.322 (0.028)	0
nodematch.age	0.028 (0.041)	0.498
nodematch.grade	1.025 (0.078)	0
nodematch.educparent	0.021 (0.047)	0.659
nodematch.race.w1w5	0.274 (0.03)	0

Note: Individual tables for each of the 44 comparisons run will be available as an electronic appendix upon request.

References

1. Pescosolido BA. Of pride and prejudice: The role of sociology and social networks in integrating the health sciences. *Journal of Health and Social Behavior.* Sep 2006;47(3):189-208.
2. Valente TW, Gallaher P, Mouttapa M. Using social networks to understand and prevent substance use: A transdisciplinary perspective. *Substance Use & Misuse.* 2004;39(10-12):1685-1712.
3. Barabasi AL. *Linked: How Everything is Connected to Everything Else and What It Means.* New York, NY: Plume; 2003.
4. Christakis NA, Fowler JH. *Connected: The Surprising Power of our Social Networks and How They Shape our Lives.* New York, NY: Little Brown; 2009.
5. Wasserman S, Faust K. *Social Network Analysis: Methods and Applications.* Cambridge, UK: Cambridge University Press; 1994.
6. Carrington PJ, Scott J, Wasserman S. *Models and Methods in Social Network Analysis.* New York, NY: Cambridge University Press; 2005.
7. Doreian P, Batagelj V, Ferligoj A. Positional Analyses of Sociometric Data. In: Carrington PJ, Scott J, Wasserman S, eds. *Models and Methods in Social Network Analysis*. New York, NY: Cambridge University Press; 2005:77-97.
8. Marsden PV. Recent Developments in Network Measurement. In: Carrington P, Scott J, Wasserman S, eds. *Models and Methods in Social Network Analysis*. New York: Cambridge University Press; 2005:8-30.
9. Snijders TAB, Steglich CEG, Schweinberger M. Modeling the Co-evolution of Networks and Behavior. In: van Montfort K, Oud J, Satorra A, eds. *Longitudinal Models in the Behavioral and Related Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates; 2007:41-71.
10. Thompson SK. *Sampling, Second Edition.* New York: Wiley; 2002.

11. Frank O. Network Sampling and Model Fitting. In: Carrington P, Scott J, Wasserman S, eds. *Models and Methods in Social Network Analysis*. New York: Cambridge University Press; 2005:31-56.

12. Goodman LA. Snowball sampling. *The Annals of Mathematical Statistics.* 1961;32(1):148-170.

13. Johnson JC, Boster JS, Holbert D. Estimating relational attributes from snowball samples through simulation. *Social Networks.* Jun 1989;11(2):135-158.

14. Gile KJ, Handcock MS. Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology.* 2010;40(1):285-327.

15. McCarty C. Structure in personal networks. *Journal of Social Structure.* 2002;3(1).

16. McCarty C, Killworth PD, Rennell J. Impact of methods for reducing respondent burden on personal network structural measures. *Social Networks.* May 2007;29(2):300-315.

17. Golinelli D, Ryan G, Green HD, Kennedy DP, Tucker JS, Wenzel SL. Sampling to reduce respondent burden in personal network studies and its effect on estimates of structural measures. *Field Methods.* Aug 2010;22(3):217-230.

18. Johnson JC, Orbach MK. Perceiving the political landscape: Ego biases in cognitive political networks. *Social Networks.* Jul 2002;24(3):291-310.

19. Holland PW, Leinhardt S. The structural implications of measurement error in sociometry†. *Journal of Mathematical Sociology.* 1973;3(1):85-111.

20. Robins G, Pattison P, Woolcock J. Missing data in networks: Exponential random graph (p*) models for networks with non-respondents. *Social Networks.* Jul 2004;26(3):257-283.

21. Handcock MS, Gile KJ. Modeling social networks from sampled data. *Annals of Applied Statistics.* Mar 2010;4(1):5-25.

22. Huisman M. Imputation of missing network data: Some simple procedures. *Journal of Social Structure.* 2009;10(1):1-29.

23. Koskinen JH, Robins GL, Pattison PE. *Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. Technical report.* Melbourne, Australia: Department of Psychology, School of Behavioural Science, University of Melbourne;2008.

24. Eudey L, Johnson JC, Schade E. Rankings versus ratings in social networks: Theory and praxis. *Journal of Quantitative Anthropology.* 1994;4(4):297-312.

25. Dawber TR, Kannel WB, Lyell LP. An approach to longitudinal studies in a community: The Framingham Study. *Annals of the New York Academy of Sciences.* May 1963;107:539-556.

26. Steglich C, Snijders TAB, Pearson M. Dynamic networks and behavior: Separating Selection from influence. *Sociological Methodology.* 2010;40:329-393.

27. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine.* Jul 2007;357(4):370-379.

28. Espelage DL (PI), Stein N (Co-PI). Middle School Bullying and Sexual Violence: Etiological Models and Moderators. *National Center for Injury Prevention and Control, Centers for Disease Control and Prevention*. Grant period: 9/1/2007 – 8/31/2010.

29. Green HD (PI), Tucker J(Co-PI). R01DA033280-01 - Social Network Effects in the Context of Adolescent Risk Behaviors. *National Institute on Drug Abuse*. Period of Grant: 2012-2016.

30. de la Haye K, Espelage D, Hoover M, Wasserman S, Green H. Differences in friendship network dynamics for bullying perpetration and homophobic bullying perpetration. *Society for Research on Adolescence Biennial Meeting, Austin, TX, USA.* March, 2014.