# Reporting Standards for Psychological Network Analyses in Cross-sectional Data

Julian Burger[1,2,3] *, Adela-Maria Isvoranu[1] *, Gabriela Lunansky[1], Jonas M.B. Haslbeck[1], Sacha Epskamp[1,2], Ria H.A. Hoekstra[1], Eiko I. Fried[4], Denny Borsboom[1], Tessa F. Blanken[1,5]

* The authors contributed equally to this manuscript.

[1]University of Amsterdam, Department of Psychology, Amsterdam, the Netherlands

[2]University of Amsterdam, Amsterdam Centre for Urban Mental Health, Amsterdam, the Netherlands

[3]University of Groningen, University Medical Center Groningen, University Center Psychiatry (UCP) Interdisciplinary Center Psychopathology and Emotion Regulation (ICPE)

[4]Leiden University, Department of Clinical Psychology, Leiden, the Netherlands

[5]Netherlands Institute for Neuroscience, Department of Sleep and Cognition, Amsterdam, the Netherlands

**Abstract**

Statistical network models describing multivariate dependency structures in psychological data have gained increasing popularity. Such comparably novel statistical techniques require specific guidelines to make them accessible to the research community. So far, researchers have provided tutorials guiding the *estimation* of networks and their accuracy. However, there is currently little guidance in determining what parts of the analyses and results should be *documented* in a scientific report. A lack of such reporting standards may foster researcher degrees of freedom and could provide fertile ground for questionable reporting practices. Here, we introduce reporting standards for network analyses in cross-sectional data, along with a tutorial and two examples. The presented guidelines are aimed at researchers as well as the broader scientific community, such as reviewers and journal editors evaluating scientific work. We conclude by discussing how the network literature specifically can benefit from such guidelines for reporting and transparency.

## 1. Introduction

Over the past decade, there has been a rapid increase in empirical contributions applying network analytic methods across many psychological disciplines. The increasing interest in networks (Barabási, 2012; Watts & Strogatz, 1998) led to empirical applications in various fields of psychology (Robinaugh, Hoekstra, Toner, & Borsboom, 2019) and resulted in a large number of special issues in journals such as *Psychometrika*, *The European Journal of Personality, The European Journal of Psychological Assessment*, *BMC Medicine,* and *The Journal of Traumatic Stress.* However, there is a lack of clear guidelines on how to report psychological network analyses. The present paper introduces such guidelines, aiming to enable researchers to identify all elements of their analyses that should be included in a scientific report. We argue that reporting guidelines can facilitate the evaluation of network contributions by the broader scientific community, including reviewers, editors, journalists, and science writers.

### 1.1 Questionable reporting practices and the benefit of reporting standards

While there are several tutorials on *estimating* networks from psychological data (Costantini et al., 2015; Epskamp, Borsboom, & Fried, 2018; Epskamp & Fried, 2018; Haslbeck, Bringmann, & Waldorp, 2020; Jones, Mair, & McNally, 2018; Williams & Mulder, 2020), as of yet, there is no guidance for how researchers should *report* the results of network analyses in a scientific paper. There are general reporting standards for statistical analyses, such as the *Journal Article Reporting Standards for Quantitative Research in Psychology* published by the APA Publications and Communications Board Task Force (Appelbaum et al., 2018). However, specific types of multivariate analyses contain explicit elements that go beyond the scope of generic reporting standards (Hoyle & Isherwood, 2013). For this reason, more tailored reporting standards do exist for other types of multivariate analyses, such as

structural equation modeling (Schreiber, Stage, King, Nora, & Barlow, 2006). At present, however, there are no explicated standards on how to report the results of network analyses.

A lack of clear reporting standards, in turn, may hinder rigorous scientific communication: Wigboldus and Dotsch (2016) highlight that a large part of the degrees of freedom in empirical research resulting in questionable research practices are in fact gray areas that pertain to questionable *reporting* practices. To this end, objective reporting standards for network analysis are an important contribution towards making empirical network studies more rigorous. Since such norms are not yet established in the network literature, the goal of the present paper is to explicate what we refer to as "minimal shared norms" in reporting psychological network analyses. By making these shared norms explicit, they can be extended and debated, and they will increase the replicability and reproducibility of network analysis, both of which will move the field of network psychometrics forward.

### 1.2 A brief introduction to psychological network analysis

While a detailed introduction to psychological network analysis is beyond the scope of this paper, in this section we briefly introduce this methodology as to keep the paper self-contained. A more extensive primer on network analyses in psychological science has recently been published (Borsboom et al., 2021), and a textbook dedicated to the emerging field of network psychometrics is currently in press (Isvoranu, Epskamp, Waldorp, & Borsboom, in press). In addition, we include a glossary that provides an overview over the most important network-specific concepts discussed in this paper.

A *network* is any system which can be represented with *nodes* (circles), which are connected by *edges* (lines) denoting a strength of connection between the nodes. In psychological networks, nodes represent observed variables, and edges are used to represent the strength of associations between two variables, typically after controlling for all other

variables in the dataset. This type of model is termed a *Markov Random Field*, which includes commonly used network models depending on the data used: Gaussian graphical models (GGM)—also termed partial correlation networks—for continuous data (Epskamp, Waldorp, Mõttus, & Borsboom, 2018; Lauritzen, 1996), Ising models for binary data (Epskamp, Maris, Waldorp, & Borsboom, 2018; Ising, 1925; Marsman et al., 2018; van Borkulo et al., 2014), and mixed graphical models (MGM) for mixed data (Haslbeck & Waldorp, 2020). Psychological networks can be estimated with (penalized) maximum likelihood estimation (Epskamp & Fried, 2018), Bayesian estimation (Williams & Mulder, 2020), or pseudo-likelihood estimation (i.e., nodewise regression) where each variable is regressed on all other variables, after which results are combined to form a network (Epskamp, Maris, et al., 2018; Haslbeck & Waldorp, 2020; Van Borkulo et al., 2014).

As is the case for statistical models in general, a crucial aspect of psychological network analysis is that estimated models are subject to sampling variation. As a result, edges may falsely be included while not being present in the true model, and differences in edge weights may be strong merely due to chance. To address such chance fluctuations, psychological network analyses should always include both model selection methods and checks for stability and accuracy. *Model selection* algorithms are diverse but generally fall under one of three categories (Blanken, Isvoranu, & Epskamp, in press): a) *pruning/thresholding* methods, which merely remove or hide edges that do not meet some criterion as defined by a classical statistical significance level or a lower Bayes factor; b) *model search* strategies, which use extensive model search methods to iteratively arrive at an optimal network structure, typically informed by an information criterion; and c) *regularization* methods, which use penalized maximum likelihood estimation to shrink parameters to zero, potentially removing them from the network. Each of these strategies has its pros and cons (Isvoranu & Epskamp, 2021). For example, regularization techniques (e.g.,

Meinshausen & Bühlmann, 2006; Ravikumar, Wainwright, & Lafferty, 2010; Tibshirani, 1996) may work well in retrieving an interpretable structure at low sample sizes, but may also feature a lower specificity rate than desired (Williams, Rhemtulla, Wysocki, & Rast, 2019). Furthermore, in such circumstances one must be careful to interpret the sparsity of the network, as this is, at least in part, a consequence of the estimation method used (Epskamp, Kruis, & Marsman, 2017). Checks for stability and accuracy usually involve the use of data-driven resampling methods such as bootstrapping (Epskamp et al., 2018) or Bayesian sampling methods (Williams & Mulder, 2020) to assess and visualize uncertainty around parameter estimates.

### 1.3 Scope of models and software

In this paper, whenever we refer to "network models," we intend to designate statistical models that are designed to capture pairwise statistical interactions between variables and that are estimated on cross-sectional data. Our focus lies on cross-sectional networks, because network analyses for this type of data account for the largest part of empirical network contributions over the past ten years (83% of the identified empirical papers between 2008 and 2018 as reported by Robinaugh, Hoekstra, Toner, & Borsboom, 2020). Of note, there are many other types of psychological network analyses than the ones we discuss here, including models estimated in panel data and time series data (Epskamp, 2020a; Gates & Molenaar, 2012; Haslbeck et al., 2020) or moderated network models (Haslbeck, Borsboom, & Waldorp, 2019). These are beyond the scope of the present paper as they require different reporting standards due to differences in data structure, estimation methods, and model assumptions.

Within the domain of cross-sectional network analysis, there is a wealth of software options. Depending on the choice of software, different reporting elements, such as specific

test statistics, might be required to ensure interpretability of the results. Here, we focus on software implemented in the open-source environment R (R Core Team, 2015), specifically on packages that have been most frequently used in the past decade in empirical, psychological network contributions (Robinaugh et al., 2020). An overview of the software packages that we cover in this paper can be found in Table 1. While we focus on a specific set of R-packages, most of the discussed reporting standards represent core elements of cross-sectional network analysis in psychological data. We therefore expect that the introduced reporting standards will also be applicable to other software, albeit not in regard to the specific test statistics included in this paper. For instance, reporting *parameter uncertainty* is not a unique standard of the packages discussed in this paper but should be included for any contribution that estimates partial correlation networks. Consequently, the listed packages should be seen as examples of how the core reporting standards introduced here can be applied to software that is frequently used in the literature, rather than restricting the domain of reporting standards to this type of software alone.

Lastly, the presented guidelines may also be applicable to some aspects of reporting simulation studies on network analyses. For example, simulation studies should include information on how networks were derived from the simulated data. However, simulation studies may require specific additional reporting elements, such as information on data-generating mechanisms and performance measures (e.g., bias or mean squared error). We therefore recommend considering additional guidelines for simulation studies, such as the guidelines provided by Morris, White, & Crowther (2019).

### 1.4 Organization of the proposed reporting standards

This paper adopts the typical structure of a psychological report according to APA standards (American Psychological Association, 2020) and can therefore be used as a reference for authors who prepare their work for submission to an APA journal. Of note, some of the recommendations discussed below, such as *reporting on the variable selection procedure*, are not unique reporting elements for network analyses. We included those elements for two reasons: First, to ensure that these guidelines are standalone readable, and second, because some more general elements deserve specific attention when using network analyses (e.g., variable selection is related to the problem of *topological overlap*, see box A).

We provide a reporting routine for both the *Methods* and the *Results* sections of an empirical APA report (sections 2 and 3, respectively), using the following structure:

I. ***General analysis routine -*** These sections contain reporting standards that are applicable to all analyses as defined above, independent of specific research questions. These routines include the reporting of general features of the data, the statistical approach, details about the sample and variables, as well as accuracy and stability checks. We recommend to always report these elements.

II. ***Analysis-specific routine -*** These sections contain reporting standards that apply only to specific research questions and analyses within the network analytic framework, such as reporting on group comparisons, centrality analyses, edge differences and visualization. Not all of these will be of interest for every empirical network contribution and are therefore only applicable if they align with their specific research question.

III. ***What to watch out for -*** The main focus of this paper lies on providing reporting standards and not interpretation guidelines. However, some reporting standards are closely related to interpretation. Therefore, in the *What to watch out for* boxes, we

discuss some of considerations that are important when applying network analyses to psychological data.

To illustrate these norms and reporting standards, we include two examples of network analyses on openly available data with two distinct research goals. Further, we include an overview of most network estimation packages and functions referred to in this paper, along with information on important arguments, current estimation defaults, applicable input data, and parameter interpretation (*Table 1*).

## 2. Reporting standards for the 'Methods' section

### General analysis routine

**2.1 Sample collection.** We recommend to specifically consider and report how and from which population the participants were recruited and whether a sub-population was included in the analyses (e.g., depressed patients; see Box A, *subsample selection*). Subsample selection can occur because of recruitment strategies (e.g., collecting data in clinical practice) or by selection after data collection (e.g., only include participants that scored higher than a certain cut-off). Make sure to report on subsamples in either case. Report the number of participants for whom data was collected and the number of participants that were included in the network analyses.

**2.2 Variable selection procedure**. As with any other study, it is important to precisely report what instruments were used to collect the data, as well as the versions of these instruments, if applicable (Flake & Fried, 2019). We recommend specifically considering the instrument, as some questionnaires might include multiple items that have the same relations to other nodes (i.e., topological overlap), which can lead to problematic inferences in networks (see Box A,

*instrument design*). With regard to network analyses, we recommend to additionally report on the number of variables on which data were collected. When the data are preprocessed before being included in the analyses (e.g., variable selection or transformation), report on these processing steps and indicate the number of variables included in the network analysis. Preprocessing choices concern, but are not limited to, collapsing variables (e.g., aggregating variables such as *loss* or *increase of appetite*), collapsing categories (e.g., binarization of Likert-scale data), data transformations (e.g., in case of violating assumptions; see Box A, *variable distribution*), and imputation or removal of missing data (e.g., listwise deletion of cases). An exhaustive list of choices that warrant justification is listed elsewhere (Flake & Fried, 2019). For the variables that are included in the network, we recommend comparing the distribution of the variables with the assumptions of the estimation method and checking any violations (e.g., skewness of the data; see Box A, *variable distribution*). If variables are removed/included following network stability analyses (see *2.5 Accuracy and stability of edge-estimates*), this should be reported as well.

**2.3 Deterministic relations between variables and skip-structures**. The manuscript should specifically report if the scale used to construct the network contains a so-called *skip-structure*, i.e., some questions in the questionnaire are skipped based on responses to previous questions. This can occur when participants are instructed to only answer one question or the other (e.g., report either on weight loss or weight gain) or when certain follow-up items are only administered to a subset of participants (e.g., only assessing nuanced depressive symptomatology if one of the core depression symptoms is present). This creates a missingness problem for the data that should be addressed, and the report should indicate precisely how this problem has been handled. This is important because some methods, such as imputing zeroes for skipped items, will induce dependency relationships in the data that

bias the network structure and can lead to faulty inference (see Borsboom et al., 2017). The latter problem will hold for any deterministic relationship included in the network (e.g., including a sum-score variable together with the components that make up the sum-score) and should be avoided. To our knowledge, no validated methods for handling such structures exist to date and therefore it is recommended not to analyze skip-structure questionnaires using network analysis. In the case of large diagnostic questionnaires (e.g., SCID, CIDI), one alternative could be to focus on the diagnostic category questions that all subjects have answered rather than on follow-up skip items.

**2.4 Estimation method**. We recommend to specifically mention in the manuscript how the data was modeled (i.e., continuous, ordinal, binary, etc.). The measurement level is linked to the estimation method used when performing a network analysis, which should always be reported as well (e.g., EBICglasso, IsingFit, MGM, etc.; see *Table 1* for a description of commonly used estimation techniques). In addition to the estimation method, mention any additional specifications. For example, when the networks are thresholded, report the chosen thresholds; when regularization is used, report the parameter specifying the search for appropriate regularization. Of note, even if researchers stick to default arguments (i.e., the standard settings that are used in the estimation procedure), we recommend reporting them, since defaults in software packages can change which in turn would make reproducing analyses difficult.[1] Finally, we advise considering the assumptions of each estimation method (see Box A, *variable distribution*), as well as how each estimation method handles missing data (see Box B, *missing data*).

[1] Within the *R* statistical software (R Core Team, 2015), the defaults of each package can be checked using the "?" + name of the function within a statistical package (e.g., ?*estimateNetwork*).

**2.5 Accuracy and stability of edge-estimates**. As with any procedure that involves parameter estimation, it is important to assess how accurate our estimates are (Fried, Epskamp, Veenman, & van Borkulo, in press). In the context of the currently most common estimation techniques in network analysis, accuracy can be assessed via a bootstrap procedure implemented in the R-package *bootnet* (Epskamp, Borsboom, et al., 2018; using the function *bootnet* and specifying the argument *type* as "nonparametric"). In this procedure, the model is estimated repeatedly under resampled or simulated data and statistics of interest (e.g., edge weights) are computed (Efron, 1979). As such, bootstrapping allows to approximate the sampling distribution of the parameters in the population. The sampling distribution can then be inspected visually (for details see e.g., Epskamp, et al., 2018). Specifically, in the methodology section of the manuscript, we advise reporting the number of bootstrap samples, as well as the type of bootstrap method employed (in the above case "nonparametric"). For methods that make use of Bayesian inference, such as *BGGM* (Williams & Mulder, 2020), there are equivalent measures to assess accuracy and stability, such as credibility intervals for estimates and convergence diagnostics.

**2.6 Statistical packages**. Finally, we recommend reporting the statistical software and packages that are used, including their versions. Full reproducibility is guaranteed only if this information is shared along with code and data, because statistical packages can change estimation defaults when they are updated (Epskamp, 2019). With this information, the reader can mimic the analyses under identical estimation settings and reproduce all results, for example using the *checkpoint* package in *R* (Ooi, de Vries, & Microsoft, 2020). We further recommend including any seed-settings in the code that have been used in conducting analyses (e.g., if estimation techniques based on cross-validation or the *Network Comparison Test* were used; Haslbeck & Waldorp, 2020; van Borkulo et al., 2017). Note, however, that

setting a seed does not fix results if parallel computing is used, as is often the case when drawing many bootstrap samples.

**Analysis-specific routine**

**2.7 Group comparisons**. If groups are compared, we recommend reporting which methods have been employed to compare groups (usually correlating weighted adjacency matrices; comparing networks using the *Network Comparison Test*, van Borkulo et al., 2017; comparisons based on the posterior predictive distribution or model selection in Bayesian GGMs; Williams, Rast, Pericchi, & Mulder, 2020; estimating moderated network models in *mgm*; Haslbeck, 2020; Haslbeck & Waldorp, 2020; or through using multi-group network modeling, Epskamp, Isvoranu, & Cheung, 2021). If groups are compared using multiple methods, we recommend reporting all comparisons that were made and in addition reflect on the consistency of the results. Of note, these methods are dependent on the sample size and identifying no differences may sometimes reflect power issues.

**2.8 Centrality indices**. One particular application of network analysis is to identify nodes that could be particularly influential, for example because they are well connected to other nodes. In graph theory and network analysis, the quantification of this relative influence based on the network flow is referred to as *centrality analysis*. Centrality metrics can be computed that quantify the role of each node in a network (Costantini et al., 2015; Jones, Ma, & McNally, 2019; Opsahl, Agneessens, & Skvoretz, 2010), for example via the *qgraph* package in R (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012; using the functions *centrality*, *centralityPlot*, or *centralityTable*), or via the *networktools* package in R (Jones, 2017; using the function *bridge*). If such inferences are of interest, we recommend carefully selecting centrality metrics that relate to the specific research question. For

example, if the research question involves identifying the most strongly connected nodes (as is the case in for example Elliott, Jones, & Schmidt, 2020), "strength centrality" may be most suited, whereas if the research question involves identifying nodes that bridge different clusters (as is the case in for example Levinson et al., 2018) "bridge centrality" measures may be most informative. There may also be research scenarios in which a combination of these metrics is of interest (as is the case in for example Isvoranu et al., 2021). We recommend reporting all centrality metrics that were computed, alongside the accuracy of their estimates (e.g., case-drop bootstrap in the *bootnet* package, using the function *bootnet* and argument *type* set to "case", for more information see Epskamp et al., 2018; see also Box A, *centrality*). Suppose the differences between node centralities are not robust. In that case, it cannot reliably be determined which node is "most central" (note that this does not imply the network was estimated with low accuracy; it is also possible that there simply are no differences in centrality between nodes; see Box B). In this case, we recommend only reporting that the centrality metric was computed, but that the centrality differences between nodes will not be further interpreted because these differences are not stable.

**2.9 Differences between edges within one network**. If edges within a network are compared with one another, we recommend reporting the method of comparison (e.g., the bootstrapped difference-test in the *R* package *bootnet*, using the *differenceTest* function; Epskamp et al., 2018). Further, if hypotheses are tested in a Bayesian context (Williams & Mulder, 2020), these should be stated explicitly (e.g., A – B > C – D).

**2.10 Clustering**. Clustering refers to the tendency of a network to exhibit groups of nodes that arise from their specific interconnections. If clustering of nodes is of interest, we recommend reporting which clustering method was employed when running the analyses (e.g., *Exploratory Graph Analysis*; Golino & Epskamp, 2017), why this particular method has been chosen (Hennig, 2015), as well as if and how the stability of the identified clusters was checked.

**What to watch out for, Box A**:

| | |
|---|---|
| Dataset | **Instrument design**. It is important to consider how the instrument used to gather the data was constructed. For instance, variables included in a network may come from a single questionnaire that was constructed to measure a latent variable, and is therefore intended to measure a single underlying construct. If a set of items does in fact depend on the same latent variable, but the items are interpreted as measuring distinct factors, possible distortions in e.g., centrality estimates should be taken into account (Fried, & Cramer, 2017). |
| Variable distribution | **Assumptions of estimation methods.** For each estimation method, model assumptions should be considered and violations of these assumptions should be addressed. Main assumptions include 1) independent cases; 2) the presence of (log) linear relationships and pairwise interactions only; 3) missing data are Missing (Completely) at Random (Rubin, 1976); 4) relevant distributional assumptions of the variables included in the network. |

| | |
|---|---|
| | **Variance.** Certain restrictions to variance, such as floor/ceiling effects or restrictions in range, can affect statistical relationships. This should be considered when interpreting edges and the importance of variables (e.g., suicidal ideation is typically restricted in variance but clinically relevant; see also 'centrality' below and Fried et al., 2018). Note that these artifacts not only pertain to networks estimated from continuous data but also to those estimated from binary data; for example, if symptoms are coded as present versus absent and most participants in the sample are healthy individuals without symptoms, floor effects may occur. |
| Subsample selection | **Biases due to subsample selection (e.g., Berkson's bias).** Sample selection is important because it can lead to unexpected patterns in the data. For example, if a sub-population (e.g., depressed patients) is recruited based on a cut-off on the total score of symptoms included in the network structure, one may find that, in that sub-population, many edges between symptoms are negative. The reason for this result is that the total score is composed of the individual item scores. As a simple example that illustrates the effect, suppose one throws coins A and B repeatedly and only selects cases in which only one of them falls heads (i.e., total score = 1). Within this set of throws (i.e., conditioning on the total score), the correlation between the outcomes of the tosses for the two coins will be negative because if coin A falls heads then, given a total score of 1, coin B must have fallen tails. This effect has been referred to as Berkson's bias (de Ron, Fried, & Epskamp, 2020). |

| | |
|---|---|
| | However, it has also been noted that Berkson's bias is but one of various effects of conditioning, and that these need not constitute bias in the statistical sense (Haslbeck, Ryan, & Dablander, 2020). Nevertheless, it is important for researchers to realize that creating subsamples based on functions of the variables in the network will often have strong effects on the network structures found in these subsamples. |
| Variable inclusion | **Variable selection**. The structure of network estimation results depends on which variables were included in the analysis. This is due to the fact that conditional dependencies are used in network estimation: conditioning on different sets of variables can therefore lead to different network structures. This implies that the network structure may change if variables are included in or excluded from the model.<br><br>**Item-scores versus sum-scores.** Depending on the research question, item-scores may sometimes be preferred, whereas sum-scores may be the best option at other times. For example, the general comorbidity of different psychopathologies can be shown at the sum-score level, but the specific symptoms that connect these clusters can only be identified at the more detailed item level. This is illustrated in the paper by Deserno et al. (2017), where the authors show how the relation between autism and well-being yields different information at different levels (item scores, subscale scores, sum-scores) and can be used to answer different research questions. Another option is to use latent network |

| | |
|---|---|
| | modeling, in which the indicators are modeled through the use of a latent node and independent measurement error (Epskamp, Rhemtulla, & Borsboom, 2017). Ultimately, what level to include in the network depends on the research question. The guiding principle should be to match the level of the included variables with the resolution at which inferences are ought to be made. |
| Centrality | **Local network properties**. Centrality is *not* a characteristic of a variable, but it is determined within the estimated network (see also 'variable distribution' and 'variable inclusion'; Bringmann et al., 2019; Fried et al., 2018). Thus, a variable that is peripheral in one network may be central in another. For instance, the symptom of insomnia may be on the periphery of a depression network and of a generalized anxiety network. At the same time, it may connect the depression network to the generalized anxiety network and thus may be highly central in the combined network. |

## 3. Reporting standards for the "Results" section

### General analysis routine

**3.1 Final sample size**. As with general statistical guidelines (e.g., Appelbaum et al., 2018), all information regarding sample size should be reported. This includes all operations that are relevant to the sample size, such as removal of outliers and missing data, data imputation, data transformations, split-half approaches, etc. For further details please refer to *Table 1* and Box B.

**3.2 Results of the accuracy and stability checks**. Results on how accurate parameters are estimated (e.g., Epskamp et al., 2018) should be reported. Usually, reports include plots giving information on bootstrapped confidence intervals (CIs), inclusion probabilities, or case-drop bootstraps, but which specific method to use is based on the choice of software. It is important to note that the bootstrapped confidence intervals discussed here cannot always be interpreted in the same manner as traditional confidence intervals (for detailed information, see Box B as well as Epskamp et al., 2018, and Fried et al., in press). Of note, which stability analysis to use is conditional on the research questions to be addressed (e.g., if centrality is not analyzed, reporting stability results for centrality may not be relevant). For most existing analyses and research questions, stability analyses are available.

**Analysis-specific routine**

**3.3 Network visualization**. When a network plot is included in the manuscript, we recommend using a colorblind-friendly theme, as well as reporting:

- What the edges represent (for example, partial correlations in the GGM or averaged logistic regression coefficients in the Ising model. In networks estimated using *mgm,* Haslbeck and Waldorp, 2020, edges between Gaussian variables can be interpreted as partial correlations, whereas relations that involve categorical variables can be interpreted in terms of (averaged) regression coefficients; for details on which type of coefficient is relevant, see *Table 1*);
- Information about the plot, such as the size of the smallest and largest edges in the network and whether any specific visualization tools were used (e.g., in *qgraph*; Epskamp et al., 2012; whether a *minimum*, *maximum* or *cut* value were used when plotting the network);
- How the layout of the network was set (e.g., manually or using a pre-defined algorithm).

**3.4 Network density and average absolute edge weights.** The network density refers to the number of estimated edges relative to the total number of possible edges and is used to give an indication of the sparsity of the network. If the overall network structure is of interest, we recommend reporting the network density and average absolute edge weights. When visualized with *qgraph* (Epskamp et al., 2012), parameters adjust the color saturation and width of an edge to the absolute weight and scale relative to the strongest weight of the graph. One cannot get a clear notion of the average edge weight from visualization alone (Epskamp et al., 2012), and thus reporting this is essential.

**3.5 Centrality indices.** If centrality is of interest (Costantini et al., 2015; Jones et al., 2019; Opsahl et al., 2010), we recommend including a supplementary table or appendix reporting the raw centrality scores in addition to visualizing raw centrality scores in the centrality plot itself,[2] as exact parameter values can often not be inferred from centrality plots with high precision. To assess the degree to which centrality estimates are subject to sampling error, we recommend reporting results of *centrality stability* (i.e., a case-drop bootstrap plot for the reported centrality indices), as well as the *correlation stability coefficient* (CS coefficient; Epskamp et al., 2018). In addition, the bootstrapped difference test allows to test for differences in centrality between two nodes, which should be reported in case a centrality comparison between two particular nodes is of interest. The bootstrapped difference tests can also be used to compare specific edge pairs in a network, see *3.7 Specific nodes and edges*.

**3.6 Predictability.** The predictability of a node quantifies how well that particular node can be predicted by all remaining nodes (Haslbeck & Fried, 2017; Haslbeck & Waldorp, 2018, 2020).

[2] The default behavior in qgraph up to version 1.6.9 provides z-scores instead of raw-scores. This, however, may inflate dissimilarity between centrality indices, and we therefore recommend to use raw scores instead.

If predictability of nodes is of interest, we recommend specifying which predictability measure was chosen for which type of variable (e.g., $R^2$), and including the predictability measures in the network plot. In addition, we recommend including a supplementary table or appendix reporting the raw predictability scores, as exact predictability values typically cannot be inferred from the visualization.

**3.7 Specific nodes and edges.** If more specific features of the network are of interest, such as a particular edge A – B, we recommend reporting the stability of that particular edge. Likewise, if specific nodes are of interest, say node A, it is important to report the stability of the edges between node A and its connecting nodes, as well as the stability of the centrality for that particular node (see also *3.5 Centrality indices*). When comparing the strength of two edges, we recommend reporting the results of the bootstrapped difference test. These may also be informative in other settings, e.g., if one is interested in the overall stability of the network structure. Finally, if clustering of nodes is of interest, we recommend reporting the number of resulting clusters, as well as the stability of the clusters.

**3.8 Group comparisons**. When interested in comparing the network structure between different groups, we recommend reporting:

- the sample size per group after data preprocessing choices (e.g., removal of outliers, removal of missing data, data imputation, data transformations);
- whether a particular statistical test was used to compare the groups: the resulting *p*-values or Bayes Factors, and whether these were adjusted for multiple testing;
- whether the chosen comparison method allows, the stability of each network structure should be reported alongside the network comparisons.

When comparing networks visually, arguments used for visualization become crucial (e.g., *minimum*, *maximum*, and *cut* values; whether the same layout was used, etc.), as well as the correlation between the weighted adjacency matrices of the two (or more) network structures. We thus recommend:

- using the same layout when comparing network structures. Note that merely comparing networks visually may be misleading and is not recommended in isolation (e.g., without also carrying out a statistical test), even if the layout is fixed across networks (e.g., equal layouts might suggest that network structures are more similar than they actually are).
- setting the same value as the strongest edge in both networks (e.g., in *qgraph* by setting the same *maximum* value) in both network structures.

**What to watch out for, Box B**:

| | |
|---|---|
| Features of the network structure | **Sparsity**. A central assumption of most of the models highlighted in the current manuscript is the *assumption of sparsity*, i.e., the true network structure can be expressed as a simplified, "sparse" network. If this assumption is violated, the performance of regularized estimation algorithms may be suboptimal (Epskamp, Kruis, et al., 2017), because many edges that are small but nonzero will be incorrectly set to 0. In this case, a nonregularized method (without model selection) can be used as an alternative (Williams et al., 2019), or the low-rank estimation approach proposed by Marsman and colleagues (2015).<br>**Collider structures**. Collider structures occur when a variable is a common effect of two or more variables. If a true causal collider structure (A -> B <- C) underlies the data and the variables A and C |

are marginally uncorrelated or weakly positively correlated, then the undirected network could feature an edge between the causes (A – C), which is negative if both causal effects are positive. As such, collider structures can produce strong and unexpected negative edges in the network structure, which may hamper the interpretation of results. While there is no principled way to detect collider structures, one way to detect at least potential collider structures is by comparing the partial correlations to marginal correlations. If a partial correlation is of a different sign (e.g., negative) than a marginal correlation (e.g., positive), then this can signal conditioning on a collider (in this case, also check whether the two variables are both strongly connected to a third, which may be a common effect).

**Network architecture.** When interpreting a network structure, it is important to keep an eye open for global features of the network. For instance, are there hubs in the network? How do these hubs influence the network structure? Is the network structure dense? Are there subnetworks? Global network aspects can inform and drive the interpretation of the network. Network architecture refers to the structure of the network as a whole; for instance, well-known architectures include small world, scale-free, and random graphs (Newman, 2018). Network architecture has been suggested to influence the recovery of the network structure (van Borkulo et al., 2014). For example, if a network features locally dense structures in the form of strong hubs (as in a scale free network), regularized estimation may have trouble recovering this (as it promotes sparsity).

| | |
|---|---|
| | In contrast, in a ring graph (as e.g. used by Epskamp & Fried, 2018) each node has only two neighbors, which a regularized estimation technique can easily recover. |
| Network visualization | **Plotting algorithms.** Network plots are always dependent on the chosen plotting settings, i.e., settings that determine the spatial position of nodes in the network. Some plotting algorithms, such as the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991), can be sensitive to small changes (e.g., small differences in edge weights). Although network plots are informative visual representations, the exact placement of nodes should not be interpreted as standing in a one-to-one relation with features of the data. In order to arrive at representations that optimally represent patterns in the data, one may utilize MDS-based algorithms (Jones et al., 2018). |
| Unstable network structures | **Accuracy and stability.** Network stability is typically assessed by investigating whether the same ordering of edge strengths or centrality estimates arises across random subsamples of the data. Importantly, an unstable network structure does not necessarily imply that the analysis failed and the network should be discarded. This is because there are two reasons why orderings of edges may be unstable under bootstrapping: a) there are estimation problems (e.g. *N* is too small), and b) all edges are equally strong so that there is no ordering in the first place (e.g., the network is a Curie-Weiss model; Marsman et al., 2018). However, unstable network structures do limit the interpretation of the network (e.g., if the centrality ordering is unstable |

for whatever reason, centrality differences should not be interpreted). In general, instability should be acknowledged, and findings from unstable network models should be presented with caution.

**Using bootstrapped confidence intervals.** Unless saturated (no model selection or regularization) maximum likelihood estimation is used, we argue against checking if bootstrapped CIs do (not) include 0, because the model selection methods themselves are already designed to put edges to zero. Therefore, doing additional checks on the CIs may lead to double thresholding. To this end, bootstrapped CIs of, for example, regularized network edges should never be used to assess for 'significance' of edges (Fried et al., in press), and seeing bootstrapped CIs that include zero is in no way evidence for instability or inaccuracy of parameter estimates. Rather, the width of CIs reflects the accuracy of parameter estimates, irrespective of whether they include 0 or not (Epskamp et al., 2018). Wide confidence intervals imply caution in interpretation, especially when interpreting the strength of edges, or the presence of weaker edges. While a clear definition of what *wide* represents is not established, this resolution can be driven by the specificity of a research question. For example, if the research question focuses on a specific edge (as for example done in Blanken, Borsboom, Penninx, & Van Someren, 2020), then it is particularly important to investigate the stability and accuracy of that edge: the wider the bootstrapped CI is for that edge, the less confidence we can attach to the estimate, and the more careful our inferences should be.

| | |
|---|---|
| | **Case-drop bootstrapping.** To assess the stability of centrality indices, an alternative method must be used, the *case-dropping bootstrap.* This is because centrality indices rely on absolute edge weights, and consequently, an edge weight of 0 is at the boundary of the parameter space. Bootstrapping parameters near the boundary of the parameter space is highly problematic and leads to false inferences. Since edge weights of 0 are to be expected in PMRFs, Epskamp et al. (2018) propose an alternative method to circumvent this problem by correlating the centrality indices from the whole sample with centrality indices obtained through estimating networks on subsets of the sample (i.e., the case-dropping bootstrap). Epskamp et al. (2018) term this *stability* (of the centrality rank order), as such correlations cannot say how *accurate* centrality estimates are. For example, suppose that all nodes in a network feature the exact same centrality. Then, any differences in centrality are due to chance, and we should expect these correlations then to be low even if the centrality measures are closely estimated to their true values (Borsboom et al., 2017). |
| Missing data | **Missing values**. It should be noted that not all estimators can handle missing data (see *Table 1*). Besides the use of (multiple) imputation strategies, which have not yet been studied in detail for network models, there are currently two ways for handling missing data when estimating GGMs. First, some estimators, such as *EBICglasso* and *ggmModSelect*, only require a correlation matrix as input, which can be estimated using pairwise observations. The *bootnet* package |

| | |
|---|---|
| | (Epskamp et al., 2018) does this by default for these estimators and will use the average of pairwise sample sizes as a proxy for the sample size (e.g., for BIC computation; Epskamp, 2020b). Specifically, the sample sizes used when estimating each pairwise correlation separately are computed, and the average of these is taken as the final sample size in the analyses. Second, the *psychonetrics* package includes full information maximum likelihood estimation (Epskamp, Isvoranu, & Cheung, 2020), which will only use observed data to estimate the network structure.<br>We recommend to include the portion of missing data, as well as to consider and report any potential source of systematic missingness. If such systematic influences are present, using any statistical strategy can lead to problematic inferences because accurate inferences will depend on strong assumptions regarding the missingness mechanism (e.g., that data are missing at random or missing completely at random Rubin, 1976). An example of such a systematic influence would be that missingness primarily occurs in participants with specific clinical features, such as high symptom levels. |
| Error rate | **Error rate**. The error rate, as well as the circumstances under which the error rate changes, should be considered. It is thus essential for researchers to consider whether they are favoring the sensitivity (true positive rate) or the specificity (true negative rate) of a model. Some estimation techniques (e.g., the *EBICglasso* algorithm; Epskamp & Fried, 2018) have high sensitivity but lower specificity. This means that weaker edges in the estimated network may be more prone to be |

false positives (i.e., Type I errors). Other estimation routines may be more conservative, retaining high specificity but featuring lower sensitivity (i.e., some edges may be missing from the network). As is typically the case in diagnostic situations, researchers face a trade-off between sensitivity and specificity: if one is more lenient to include edges in the estimated network, sensitivity will increase at the cost of specificity. Researchers can choose to err on the side of *discovery* (favor sensitivity over specificity) or to err on the side of *caution* (favor specificity over sensitivity). This choice is also driven by the research question. For example, in the study by Isvoranu and colleagues (2020), the aim was to identify edges between a polygenetic risk score and symptoms, which are generally weaker than edges between symptoms themselves. While good sensitivity is required to identify such small edges (and this was achieved in the paper as a result of a large sample size), high specificity is essential to justify interpreting the smaller edges in substantive terms. The authors therefore chose *ggmModSelect* as an estimator, which has been shown to have good specificity in large sample sizes (Isvoranu & Epskamp, 2021).

## 4. Illustrative examples

To illustrate the highlighted norms and reporting standards, we provide two examples of network analyses on openly available data, with two distinct research goals. Both examples contain the elements described under the general analysis routine, as well as analysis specific elements matched with the indicated research goal. For an overview of elements covered in both examples, see Table 2. This table may also be used as a summary checklist of the paper. First, using data from Burger and colleagues (2020), we aim to highlight the analysis specific routine on group comparisons, network visualization, and global network properties. Second, using open data (https://openpsychometrics.org/tests/TMAS/) collected on the *Taylor Manifest Anxiety Scale* (Taylor, 1953), we aim to highlight the analysis specific routine elements on centrality, differences between edges, network visualization, and local network properties.

### 4.1 Example 1: Relationships in later life

Data for the first example stem from the Swiss longitudinal study "Relationships in later life", which followed widowed and separated individuals after their loss experience and collected information on their psychosocial functioning, including depressive symptomatology. The data and project description can be found online (https://www.kpp.psy.unibe.ch/forschung/projekte/nccrlives/index_ger.html), and the results have been discussed in a previous paper (Burger, Stroebe, et al., 2020). The main research interest here lies in comparing depressive symptom networks between the widowed and separated individuals, specifically comparing how strongly they are connected and the overall structure of the two networks. Next to the *general analysis routine*, we therefore focus on *group comparison* (methods and results), *network visualization* (results), and *network density* (results).

## Methods

### General Analysis Routine

***Sample collection (2.1), Variable selection procedure (2.2), Deterministic relations between variables and skip-structures (2.3, not applicable here)***

For this analysis, we included data collected on the German version of the Center for Epidemiologic Studies Depression scale (CES-D; Radloff, 1977; German: Allgemeine Depressions-Skala, ADS-K; Hautzinger & Geue, 2016).

The dataset consists of 1276 married, 566 widowed, and 971 separated individuals. Participants were contacted via post mail and filled in a pen-and-paper questionnaire. To circumvent the issue that participants might be at different stages of adaptation to the adverse life event, we only included participants with a maximum distance of two years to the event (widowhood/separation). This resulted in 145 widowed and 217 separated individuals. To be able to include widowhood/separation as a node in the network, we added 145 married controls to the widowed sample, and 217 married controls to the separated sample[3]. This way, widowhood/separation is included as a binary node, indicating the presence versus absence of the respective life event.

In order to investigate conceptual overlap between variables, we examined bivariate correlations between all variables, and combined items if their content suggested strong conceptual similarity, and their bivariate correlation was $r \geq .50$. Accordingly, we combined the original items *mood*, *upset*, and *depressed* (new item "mood"), as well as the items *happy* and *enjoy* (new item "happy"). This resulted in 12 variables, each rated on an ordinal scale with four answer categories [1 = "rarely or none of the time (less than 1 day)", 2 = "some or a little of the time (1–2 days)", 3 = "occasionally or a moderate amount of time (3–4 days)", 4 = "most or all of the time (5–7 days)"].

[3]Note that adding control participants and including group membership in the network is only but one way to approach group comparisons. Many other techniques have been discussed recently, such as moderated network analysis (Haslbeck et al., 2019), or Bayesian approaches (Williams et al., 2020). For more detailed information on the approach used here, we advise to consider the original publication.

***Estimation method (2.4)***

We estimated partial correlation networks for both, the widowed and separated sample, using the *glasso* regularization and a tuning-parameter gamma set to 0.5 (Foygel & Drton, 2010). Due to the ordinal, non-normal nature of the data, we used Spearman's rank-correlation and pairwise complete observations to handle missing data. In total, of all variables included in the network analysis, 6.6% of the ratings were missing in the widowed/married sample and 5.1% in the separated/married sample. Here, we assume that these ratings are missing at random (Rubin, 1976).

***Accuracy and stability of edge-estimates (2.5)***

To assess accuracy of the edge estimates, we conducted the routine implemented in the *bootnet* package (Epskamp, Borsboom, et al., 2018; version 1.4.3), using nonparametric bootstrapping with 1,000 bootstrap samples.

***Statistical packages (2.6)***

The analyses have been conducted using *R*-version 3.5.2 on October 8th, 2020. For network estimation, we used the *estimateNetwork* function in the *bootnet* package (Epskamp, Borsboom, et al., 2018; version 1.4.3). Networks have been visualized using the *qgraph* package (Epskamp et al., 2012; version 1.6.5).

**Analysis-specific Routine**

***Group comparisons (2.7)***

Groups were compared by obtaining the difference in global strength within the *Network Comparison Test* (NCT; van Borkulo et al., 2017; version 2.2.1), using 2,000 iterations, and with seed set to '123'. This test assesses if the two networks differ in their overall level of connectivity. Since we are primarily interested in global differences in network connectivity, other tests available within the *NCT* were disregarded in the present analyses. Additionally, we correlated the weighted adjacency matrices of the two networks as an additional measure of similarity between the networks.

## Results

### General Analysis Routine

***Final sample size (3.1)***

The widowed network included 290 individuals (145 widowed and 145 married controls), and the separated network included 434 individuals (217 separated and 217 married controls).

***Results of accuracy and stability checks (3.2)***

Results of the nonparametric bootstrap analysis can be found in the supplementary materials (*Supplementary Figure 1*). In general, the confidence intervals were rather broad and overlapping. The order of edge estimates should therefore be interpreted with caution.

### Analysis-specific Routine

***Network visualization (3.3)***

The networks of widowed and separated individuals are visualized in *Figure 1*. Here, edges represent regularized partial correlations between symptoms. Edge weights in the widowed network ranged from 0.002 (*sad – getgo*) to 0.300 (*lonely – widowed*). Edge weights in the separated network ranged from 0.001 (*mood – unfriendly*) to 0.320 (*lonely – separation*). To facilitate interpretability, we used the colorblind-theme in *qgraph* (Epskamp et al., 2012), fixed the average layout between the two network plots using the *averageLayout* function, curved edges that would otherwise cross nodes, and made negative edges dashed (*Note: This is useful if printed without colors*). No specific minimum/maximum/cut values have been used for network visualization.

*Note: Any exploratory reporting of findings, such as relevant edges, will be specific to the given research context. The figures presented below are based on an adapted version of the publicly available code from the original article* (Burger, Stroebe, et al., 2020).

| | |
|---|---|
| ***Network density and average absolute edge weights (3.4)*** | Since we are interested in comparing the two networks with regard to their connectivity, we computed the density of the two networks by determining the ratio of detected edges to the total number of edges in a fully connected network. The network of widowed/married individuals had a density of .615 (48/78 edges), with a mean weight of 0.044, and the separated network had a density of .744 (58/78 edges), with a mean weight of 0.053. |
| ***Group comparisons (3.8)*** | While the global invariance test within the *Network Comparison Test* procedure indicated that there were some differences in the overall level of connectivity between the widowed and separated network ($p$ = .003), the weighted adjacency matrices showed a rather large correlation ($r$ = .750), indicating that the overall structure between the networks was similar. This shows that the networks differed in how strongly connected they are (sum of absolute edge weights, connectivity), while edges that were detected showed a similar pattern across the two networks (correlation of edges), i.e., edges that were large (small) in the separated network were generally also large (small) in the widowed network. |

### 4.2 Example 2: Taylor Manifest Anxiety Scale

Data for the second example data stem from the *openpsychometrics.org* project, using the *Taylor Manifest Anxiety Scale* (Taylor, 1953). The data and project description can be found online (https://openpsychometrics.org/tests/TMAS). Let us assume the main research interests here lie in the general network structure of anxiety, edge differences in the network structure, as well as in which items play a more central role in the network. Next to the *general analysis routine*, we therefore focus on *centrality results* (methods and results), *edge differences* (methods and results), *network visualization* (results), and *local network properties* (results).

## Methods

### General Analysis Routine

***Sample collection (2.1), Variable selection procedure (2.2), Deterministic relations between variables and skip-structures (2.3, not applicable here)***

For this analysis, we included data collected on the *Taylor Manifest Anxiety Scale* (Taylor, 1953). This data was collected online; at the end of the test users were asked if their answers were accurate and could be used for research. 76% said yes and data have been published on the *openpsychometrics.org* project. The dataset consisted of 5410 individuals.

The network model included all questions from the *Taylor Manifest Anxiety Scale* (Taylor, 1953), thus resulting in 50 nodes. Each item was rated on a binary scale with two answer categories [0 = FALSE, 1 = TRUE]. In addition, missing data was encoded as NA and we used listwise deletion for missing data points, as the chosen estimation algorithm does not allow for missing data. Data were assumed to be missing completely at random.

***Estimation method (2.4)***

We estimated the network structure using an Ising model (van Borkulo et al., 2014). An Ising model represents associations between dichotomous variables using pairwise log linear relationships, similar to partial correlation coefficients in a Gaussian Graphical Model (GGM; Epskamp, Waldorp, Mõttus, & Borsboom, 2018). To control for potential spurious associations, the estimation procedure here uses a penalized nodewise regression approach, specifically the eLasso penalty based on the Extended Bayesian Information Criterion (Ravikumar et al., 2010). Default values as set in the package were used, with the EBIC hypertuning parameter set to 0.25.

***Accuracy and stability of edge-estimates (2.5)***

To assess the accuracy of the edge weight estimates, we conducted the routine implemented in the *bootnet* package (Epskamp, Borsboom, et al., 2018; version 1.4.3), using nonparametric bootstrapping based on 1,000 bootstrap samples.

***Statistical packages (2.6)***

The analyses have been conducted using *R*-version 3.5.2 on October 12th, 2020. For network estimation, we used the *estimateNetwork* function in the *bootnet* package (Epskamp,

Borsboom, et al., 2018; version 1.4.3), using the *IsingFit* package (van Borkulo, Epskamp, & Robitzsch, 2014; version 0.3.1). The accuracy of estimates has been assessed using the *bootnet* function. Networks have been visualized using the *qgraph* package (Epskamp et al., 2012; 1.6.5).

**Analysis-specific Routine**

***Centrality Indices (2.8)*** To further quantify how well a node is directly connected to other nodes in the network structure, we investigated *strength* as a centrality measure (Costantini et al., 2015; Opsahl et al., 2010). To assess accuracy of the *strength* centrality estimates, we conducted the routine implemented in the *bootnet* package (Epskamp et al., 2018), using case-drop bootstrapping based on 1,000 bootstrap samples. Further, to ensure interpretable differences in centrality, we used the bootstrapped difference-test in the *bootnet package*.

***Differences between edges within one network (2.9)*** Finally, as we were interested in an exploratory fashion whether certain edges were stronger and stood out in the network structure, we carried out a bootstrapped difference-test using the *R* package *bootnet* (Epskamp et al., 2018).

## Results

**General Analysis Routine**

***Final sample size (3.1)*** Following removal of missing data, 4474 subjects were included in the current analyses.

***Results of accuracy and stability checks (3.2)*** In general, the confidence intervals were very narrow, indicating stable results. In addition, *strength* centrality estimates were stable, with a centrality stability coefficient of 0.75, indicating that 75% of the data could be dropped to retain with 95% certainty a correlation of 0.7 with the original dataset. Of note, while the most central items were more central than most other items in the network, they were not more central than each other

(see *Supplementary Figure 5*).

**Analysis-specific Routine**

| | |
|---|---|
| ***Network visualization (3.3)*** | The network visualization is presented in *Figure 2*. To facilitate interpretability, here we used the colorblind-theme in *qgraph* (Epskamp et al., 2012), included a legend with the description of each item, and used a *cut* value of 0. Edge weights ranged from –1.82 (Q47–Q50) to 2.28 (Q6–Q41). The layout used was the automatically generated layout based on the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991). Any exploratory reporting of findings, such as relevant edges, will be specific to the given research context. |
| ***Centrality indices (3.5)*** | *Supplementary Figure 2* presents the results of the centrality analyses. In addition, *Supplementary Table 1* presents the standardized and raw centrality indices. The three most central items were: Q27, Q31, and Q48. Of note, while these were more central than many other items in the network, differences between the items themselves were not robust (see *Supplementary Figure 5*). |
| ***Specific nodes and edges (3.7)*** | *Supplementary Figure 6* presents the results of the edge difference test. The labels are omitted for clarity. In general, the bootstrapped difference test identified several edges as significantly different from most other edges in the network. Of note, the two strongest edges in the current network structure were significantly different from each other and all other edges in the network. These are the edge between *Q6* and *Q41* and between *Q40* and *Q46*. |

## 5. Concluding remarks

As clear norms have not yet been established in the network literature, the current paper explicates minimal shared norms in reporting psychological network analyses. While network psychometrics is a relatively young field of research, we recognize that many norms discussed here have important implications for commonly used inferences. We therefore included two "what to watch out for" boxes, where we discussed important considerations for network analysis, as well as potential sources of misinterpretation of network structures.

It should be noted, however, that our description of validity threats is not exhaustive and subject to ongoing research. For example, although robustness analyses allow one to assess the uncertainty of claims based on the model (relative to sampling error), methods for assessing the goodness-of-fit of the model as a whole remain underinvestigated (although model fit assessment techniques are available for confirmatory network analyses; Epskamp, 2020a). Currently, operational network analysis techniques are better viewed as exploratory analysis and visualization tools in the tradition of Tukey (1977), or as phenomena-detection tools that can generate a starting point for theory formation (Borsboom, van der Maas, Dalege, Kievit, & Haig, 2021; Haig, 2005, 2014), than as confirmatory theory-testing approaches in the tradition of SEM (Hoyle, 2012). Hence, we currently advise against strong inferences based on network analyses alone, while noting that considerable methodological research opportunities are open to extending network analysis in this direction (Epskamp, 2020a).

Clear reporting standards for network psychometrics improve transparency, which is necessary for reproducibility. Only if the scientific community can follow exactly what analyses were conducted can we vet inferences drawn by respective authors. This is especially relevant in a field that is still fairly novel such as network psychometrics, where we encounter new challenges regularly. Overall, we trust the highlighted directions to aid researchers in

identifying elements of their analyses that are important to include in a scientific report, as well as to make empirical network studies more rigorous.

## Glossary

| | |
|---|---|
| **(weighted) Adjacency matrix** | A square matrix that encodes connections (or their weights) in a network. Each row and column represent a node in the network, and each cell represents the strength of the connection between the respective nodes. In this paper, we focus on undirected networks, which consist of symmetric adjacency matrices, i.e., the upper and lower triangle of the matrix are identical, and for which the diagonal elements are 0. |
| **Berkson's bias** | Unexpected connections that can arise when estimating models from a sub-sample of a population, where the sample has been selected as a function of the variables included in the model (de Ron et al., 2020). For example, Berkson's bias can arise when estimating a network of depressive symptoms from patients who score high on the sum-score of this depression scale. |
| **Bootstrapped difference test** | Significance test for investigating if the weight of two edges or the centrality of two nodes within the same network differs from one another (Epskamp, Borsboom, et al., 2018). The test is based on calculating the difference between the two bootstrap values (i.e., for the two edge weights or the two centrality indices), and subsequently testing if the bootstrapped confidence interval around this difference estimate includes 0. |
| **Bootstrap: Case-drop** | Resampling of different subsets of the data. In the context of network analysis, this allows to investigate the *stability of the centrality indices* for retaining different proportions of the original sample, e.g., 90%, 80%, etc. (Epskamp, Borsboom, et al., 2018). The stability of centrality can be calculated as the correlation between centrality indices established from the original sample with the ones established from the subsets. |
| **Bootstrap: Nonparametric** | Resampling data from the original sample (with replacement). In the context of network analysis, this allows to investigate the *accuracy of edge weight estimates*, based on the width of the bootstrapped confidence intervals around the estimate (Epskamp, Borsboom, et al., 2018). |
| **Centrality analysis** | Quantifies the projected influence of a node in terms of its direct and/or indirect connections to other nodes in the network. Different centrality metrics exist that differ in their approach to quantify this influence (e.g., strength/degree, betweenness, closeness, bridge centrality, etc.). |

| | |
|---|---|
| **Cluster analysis** | Quantifies the tendency of a network to exhibit groups of nodes ("clusters") that arise from their specific interconnections. |
| **Collider structure** | Refers to the causal structure of three variables, where one is the common effect of the other two (e.g., A → B ← C). If the *bivariate-correlation* between the two cause-variables (A, C) is zero or weakly positive, this structure may induce unexpected strong and negative relationships between them in a *partial-correlation* network (see also Berkson's bias). |
| **Density/sparsity (of a network)** | Quantifies how well connected a network is. In the context of statistical network models, density refers to the number of estimated edges relative to the number of edges if the network were fully connected. A dense network refers to a network structure with many connections, whereas a sparse network refers to a weakly connected structure. Estimation methods such as regularization and pruning assume a sparse true network structure. |
| **Edge** | Connection between two nodes. Edges can be weighted or unweighted, directed or undirected. The types of networks we discuss in this paper conceptualize edges as the strength of association between two nodes. This association is estimated from data, for example using (partial-) correlations. |
| **Model search/selection** | Algorithms that evaluate and select the best fitting model for the data, according to a criterium (e.g., penalized maximum likelihood estimation). |
| **Network comparison** | (Formal) comparison of two (or more) network structures consisting of the same set of nodes. It is possible to compare *global network properties* (e.g., the correlation between the adjacency matrices, structural invariance testing, etc.), as well as *local network properties* (e.g., comparing a specific edge between two nodes across the networks). |
| **Node** | Vertices ("circles") amongst which we aim to establish connections. In the types of networks discussed in this paper, nodes are observed variables in a dataset. |
| **(Pairwise) Markov Random Field** | Type of network that establishes *undirected* connections between variables. In this paper, we refer to different ways of estimating such undirected networks from cross-sectional data: *Gaussian Graphical Models* (or partial-correlation networks) for continuous data, *Ising models* for |

| | |
|---|---|
| | binary data, and *mixed graphical models* for data that consists of variables with mixed distributions. In contrast, so-called Bayesian networks, or Directed Acyclic Graphs (DAGs) establish *directed* connections between variables. |
| **Penalization/Regularization** | Method to prevent overfitting in highly parametrized models. Penalization/regularization approaches shrink model parameters towards zero, which includes removing some edges from the estimated network structure. |
| **Predictability** | Quantification of how well a node can be predicted by all remaining nodes, for example, by calculating the explained variance ($R^2$). |
| **Pruning/Thresholding** | Estimation method that removes or hides edges from the network according to some threshold (e.g., statistical significance of Bayes factor criterium) in order to achieve a sparse network structure. |
| **Seed settings** | Settings that determine the starting number ("seed") for routines that involve random sampling (e.g., as done in permutation tests such as the *Network Comparison Test*). Setting these transparently is essential for reproducibility because different seed settings will lead to slightly different results. |
| **Skip-structures** | Refers to instruments that skip certain items based on the response to previous questions, e.g., only asking for more nuanced depression symptoms if the central depression symptoms are present. |
| **Structure (of a network)** | The structure of a network is characterized by the presence/absence of its edges. Two networks with the same subset of nodes are equal in structure if they contain the exact same set of edges between the nodes. |
| **Topological overlap** | Two nodes exhibit topological overlap if they share the same relations to the other nodes in the network. |

## References


American Psychological Association. (2020). *Publication Manual of the American Psychological Association* (7th ed.).

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and Communications Board task force report. *American Psychologist*. https://doi.org/10.1037/amp0000191

Barabási, A. L. (2012). The network takeover. *Nature Physics*, *8*(1), 14–16. https://doi.org/doi.org/10.1038/nphys2188

Blanken, T. F., Borsboom, D., Penninx, B. W. J. H., & Van Someren, E. J. W. (2020). Network outcome analysis identifies difficulty initiating sleep as a primary target for prevention of depression: a 6-year prospective study. *Sleep*, *43*(5), zsz288.

Blanken, T. F., Isvoranu, A.-M., & Epskamp, S. (in press). Estimating Network Structures using Model Selection. In A.-M. Isvoranu, S. Epskamp, L. J. Waldorp, & D. Borsboom (Eds.), *Network Psychometrics with R: A Guide for Behavioral and Social Scientists.* Routledge, Taylor & Francis Group.

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., … Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*(1), 1–18. https://doi.org/https://doi.org/10.1038/s43586-021-00055-w

Borsboom, D., Fried, E. I., Epskamp, S., Waldorp, L. J., van Borkulo, C. D., van der Maas, H. L. J., & Cramer, A. O. J. (2017). False alarm? A comprehensive reanalysis of "Evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger. *Journal of Abnormal Psychology*, *26*(7), 989-999. https://doi.org/10.1037/abn0000306

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, 1–11. https://doi.org/10.1177/1745691620969647

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, W., Schoch, D., Wichers, M., & Snippe, E. (2019). What do centrality measures measure in psychological networks. *Journal of Abnormal Psychology*, *128*(8), 892–90. https://doi.org/10.1037/abn0000446

Burger, J., Isvoranu, A.-M., Lunansky, G., Haslbeck, J. M. B., Epskamp, S., Hoekstra, R. H. A., … Blanken, T. F. (2020). Reporting standards for psychological network analyses in cross-sectional data. *PsyArXiv (Preprint)*. https://doi.org/10.31234/osf.io/4y9nz

Burger, J., Stroebe, M. S., Perrig-Chiello, P., Schut, H. A., Spahni, S., Eisma, M. C., & Fried, E. I. (2020). Bereavement or breakup: Differences in networks of depression. *Journal of Affective Disorders*, *267*, 1–8. https://doi.org/10.1016/j.jad.2020.01.157

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13–29. https://doi.org/10.1016/j.jrp.2014.07.003

de Ron, J., Fried, E. I., & Epskamp, S. (2020). Psychological Networks in Clinical Populations: A tutorial on the consequences of Berkson's Bias. *Psychological Medicine*, *51*(1), 168–176. https://doi.org/10.1017/S0033291719003209

Deserno, M. K., Borsboom, D., Begeer, S., & Geurts, H. M. (2017). Multicausal systems ask for multicausal approaches: A network perspective on subjective well-being in individuals with autism spectrum disorder. *Autism*, *21*(8), 960–971.

Efron, B. (1979). Computers and the Theory of Statistics: Thinking the Unthinkable. *SIAM Review*, *21*(4), 460–480. https://doi.org/10.1137/1021092

Elliott, H., Jones, P. J., & Schmidt, U. (2020). Central symptoms predict posttreatment outcomes and clinical impairment in anorexia nervosa: A network analysis. *Clinical Psychological Science*, *8*(1), 139–154.

Epskamp, S. (2019). Reproducibility and replicability in a fast-paced methodological world. *Advances in Methods and Practices in Psychological Science*, *2*(2), 145–155. https://doi.org/https://doi.org/10.1177/2515245919847421

Epskamp, S. (2020a). Psychometric network models from time-series and panel data. *Psychometrika*, *85*, 206–231. https://doi.org/10.1007/s11336-020-09697-3

Epskamp, S. (2020b). Software updates: Bootnet 1.3 and psychonetrics 0.5, retrieved from http://psychonetrics.org/2020/01/26/software-updates-bootnet-1-3-and-psychonetrics-0-5/#New_default_behavior.

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*, 195–212. https://doi.org/10.3758/s13428-017-0862-1

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, *48*(4). https://doi.org/10.18637/jss.v048.i04

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617–634. https://doi.org/10.1037/met0000167

Epskamp, S., Isvoranu, A.-M., & Cheung, M. W.-L. (2021). Meta-analytic Gaussian Network Aggregation. *Psychometrika*. https://doi.org/https://doi.org/10.1007/s11336-021-09764-3

Epskamp, S., Isvoranu, A. M., & Cheung, M. W.-L. (2020). Meta-analytic Gaussian Network Aggregation. *PsyArXiv (Preprint)*. https://doi.org/10.31234/osf.io/236w8

Epskamp, S., Kruis, J., & Marsman, M. (2017). Estimating psychopathological networks: Be

careful what you wish for. *PLoS ONE*, *12*(6). https://doi.org/10.1371/journal.pone.0179891

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018). Network Psychometrics. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development* (pp. 953–986).

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized Network Psychometrics: Combining Network and Latent Variable Models. *Psychometrika*, *82*(4), 904–927. https://doi.org/https://doi.org/10.1007/s11336-017-9557-x

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian Graphical Model in Cross-Sectional and Time-Series Data. *Multivariate Behavioral Research*, *53*(4), 453–480. https://doi.org/10.1080/00273171.2018.1454823

Flake, J. K., & Fried, E. I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *PsyArXiv (Preprint)*. https://doi.org/10.31234/osf.io/hs7wm

Foygel, R., & Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*.

Fried, E. I., & Cramer, A. O. J. (2017). Moving forward: Challenges and directino for psychopathological network theory and methodology. *Perspectives on Psychological Science*, *12*(6), 999–1020. https://doi.org/10.17605/OSF.IO/BNEK

Fried, E. I., Eidhof, M. B., Palic, S., Costantini, G., Dijk, H. M. H., Bockting, C. L. H., … Karstoft, K. (2018). Replicability and Generalizability of Posttraumatic Stress Disorder (PTSD) Networks: A Cross-Cultural Multisite Study of PTSD Symptoms in Four Trauma Patient Samples. *Clinical Psychological Science*, *6*(3), 335–351.

https://doi.org/10.1177/2167702617745092

Fried, E. I., Epskamp, S., Veenman, M., & van Borkulo, C. D. (in press). Network Stability, Comparison, and Replicability. In A.-M. Isvoranu, S. Epskamp, L. J. Waldorp, & D. Borsboom (Eds.), *Network Psychometrics with R: A Guide for Behavioral and Social Scientists.*

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*. https://doi.org/10.1002/spe.4380211102

Gates, K. M., & Molenaar, P. C. M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, *63*(1), 310–319. https://doi.org/10.1016/j.neuroimage.2012.06.026

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLoS ONE*, *12*(6), e0174035. https://doi.org/10.1371/journal.pone.0174035

Haig, B. D. (2005). An Abductive Theory of Scientific Method. *Psychological Methods*, *10*(4), 371–388. https://doi.org/10.1037/1082-989X.10.4.371

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. MIT press.

Haslbeck, J. M. B. (2020). Estimating Group Differences in Network Models using Moderation Analysis. *PsyArXiv (Preprint)*. https://doi.org/10.31234/osf.io/926pv

Haslbeck, J. M. B., Borsboom, D., & Waldorp, L. J. (2019). Moderated Network Models. *Multivariate Behavioral Research*, 1–32. https://doi.org/10.1080/00273171.2019.1677207

Haslbeck, J. M. B., Bringmann, L. F., & Waldorp, L. J. (2020). A Tutorial on Estimating Time-Varying Vector Autoregressive Models. *Multivariate Behavioral Research*, *56*(1), 120–149. https://doi.org/10.1080/00273171.2020.1743630

Haslbeck, J. M. B., & Fried, E. I. (2017). How predictable are symptoms in psychopathological networks? A reanalysis of 18 published datasets. *Psychological Medicine*, *47*(16), 2767–2776. https://doi.org/10.1017/S0033291717001258

Haslbeck, J. M. B., Ryan, O., & Dablander, F. (2020). The Sum of All Fears: Comparing Networks Based on Symptom Sum-Scores. *Psychological Methods*, *accepted*.

Haslbeck, J. M. B., & Waldorp, L. J. (2018). How well do network models predict observations? On the importance of predictability in network models. *Behavior Research Methods*, *50*, 853–861. https://doi.org/10.3758/s13428-017-0910-x

Haslbeck, J. M. B., & Waldorp, L. J. (2020a). MGM: Estimating time-varying mixed graphical models in high-dimensional data. *Journal of Statistical Software*, *93*(8). https://doi.org/10.18637/jss.v093.i08

Haslbeck, J. M. B., & Waldorp, L. J. (2020b). MGM: Estimating time-varying mixed graphical models in high-dimensional data. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v093.i08

Hautzinger, M., & Geue, K. (2016). Allgemeine Depressionsskala. In K. Geue, B. Strauß, & E. Brähler (Eds.), *Diagnostische Verfahren in der Psychotherapie* (pp. 33–35). https://doi.org/10.1026/02700-00

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, *64*(15), 53–62. https://doi.org/10.1016/j.patrec.2015.04.009

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. Guilford press.

Hoyle, R. H., & Isherwood, J. C. (2013). Reporting results from structural equation modeling analyses in Archives of Scientific Psychology. *Archives of Scientific Psychology*, *1*(1), 14–22. https://doi.org/10.1037/arc0000004

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift Für Physik*, *31*, 253–258. https://doi.org/10.1007/BF02980577

Isvoranu, A.-M., & Epskamp, S. (2021). Which Estimation Method to Choose in Network Psychometrics? Deriving Guidelines for Applied Researchers. *Psychological Methods*.

Isvoranu, A.-M., Epskamp, S., Waldorp, L. J., & Borsboom, D. (Eds.). (in press). *Network Psychometrics with R: A Guide for Behavioral and Social Scientists*. Routledge, Taylor & Francis Group.

Isvoranu, A.-M., Guloksuz, S., Epskamp, S., van Os, J., Borsboom, D., Investigators, G., & others. (2020). Toward incorporating genetic risk scores into symptom networks of psychosis. *Psychological Medicine*, *50*(4), 636–643.

Isvoranu, A.-M., Ziermans, T., Schirmbeck, F., Borsboom, D., Geurts, H. M., de Haan, L., … others. (2021). Autistic Symptoms and Social Functioning in Psychosis: A Network Approach. *Schizophrenia Bulletin*.

Jones, P. J. (2017). *networktools: Assorted Tools for Identifying Important Nodes in Networks. R package version 1.0.0. https://CRAN.R-project.org/package=networktools*.

Jones, P. J., Ma, R., & McNally, R. J. (2019). Bridge centrality: A network approach to understanding comorbidity. *Multivariate Behavioral Research*, 1–15. https://doi.org/10.1080/00273171.2019.1614898

Jones, P. J., Mair, P., & McNally, R. J. (2018). Visualizing psychological networks: A tutorial in R. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01742

Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon Press.

Levinson, C. A., Brosof, L. C., Vanzhula, I., Christian, C., Jones, P., Rodebaugh, T. L., … others. (2018). Social anxiety and eating disorder comorbidity and underlying vulnerabilities: Using network analysis to conceptualize comorbidity. *International Journal of Eating Disorders*, *51*(7), 693–709.

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., … Maris, G. (2018). An Introduction to Network Psychometrics: Relating Ising Network Models

to Item Response Theory Models. *Multivariate Behavioral Research*, *53*(1), 15–35. https://doi.org/10.1080/00273171.2017.1379379

Marsman, Maarten, Maris, G., Bechger, T., & Glas, C. (2015). Bayesian inference for low-rank Ising networks. *Scientific Reports*, *5*(1), 1–7.

Meinshausen, N., Bühlmann, P., & others. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, *34*(3), 1436–1462. https://doi.org/10.1214/009053606000000281

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Newman, M. (2018). *Networks*. Oxford university press.

Ooi, H., de Vries, A., & Microsoft. (2020). *R-package checkpoint*. Retrieved from https://cran.r-project.org/web/packages/checkpoint/checkpoint.pdf

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, *32*(3), 245–251. https://doi.org/10.1016/j.socnet.2010.03.006

R Core Team. (2015). R Development Core Team. *R: A Language and Environment for Statistical Computing*, *55*, 275–286.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401. https://doi.org/10.1177/014662167700100306

Ravikumar, P., Wainwright, M. J., & Lafferty, J. D. (2010). High-dimensional ising model selection using ℓ1-regularized logistic regression. *Annals of Statistics*, *38*(3), 1287–1319. https://doi.org/10.1214/09-AOS691

Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R., & Borsboom, D. (2019). The network

approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine*. https://doi.org/10.1080/01559982.2019.1584953

Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R., & Borsboom, D. (2020). The network approach to psychopathology: A review of the literature 2008-2018 and an agenda for future research. *Psychological Medicine*, *50*(3), 353–366. https://doi.org/10.1017/S0033291719003404

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Schreiber, J. B., Stage, F. K., King, J., Nora, A., & Barlow, E. A. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *Journal of Educational Research*, *99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Taylor, J. A. (1953). A personality scale of manifest anxiety. *Journal of Abnormal and Social Psychology*, *48*(2), 285–290. https://doi.org/10.1037/h0056264

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. Retrieved from http://www.jstor.org/stable/2346178

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, Mass.

Van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific Reports*, *4*. https://doi.org/10.1038/srep05918

van Borkulo, C. D., Boschloo, L., Kossakowski, J. J., Tio, P., Schoevers, R. A., Borsboom, D., & Waldorp, L. J. (2017). Comparing network structures on three aspects: A permutation test. *Manuscript Submitted*. https://doi.org/10.13140/RG.2.2.29455.38569

van Borkulo, C. D., Epskamp, S., & Robitzsch, A. (2014). IsingFit: Fitting Ising models

using the eLasso method. *R Package Version 2.0*.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature*, *393*(6684), 440–442.

Wigboldus, D. H. J., & Dotsch, R. (2016). Encourage Playing with Data and Discourage Questionable Reporting Practices. *Psychometrika*, *81*, 27–32. https://doi.org/10.1007/s11336-015-9445-1

Williams, D. R., & Mulder, J. (2020). BGGM: Bayesian Gaussian Graphical Models in R. *Journal of Open Source Software*, *5*(51), 2111. https://doi.org/10.21105/joss.02111

Williams, D. R., Rast, P., Pericchi, L. R., & Mulder, J. (2020). Comparing Gaussian Graphical Models With the Posterior Predictive Distribution and Bayesian Model Selection. *Psychological Methods*, *25*(5), 653–672. https://doi.org/10.1037/met0000254

Williams, D. R., Rhemtulla, M., Wysocki, A. C., & Rast, P. (2019). On Nonregularized Estimation of Psychological Networks. *Multivariate Behavioral Research*, 719–750. https://doi.org/10.1080/00273171.2019.1575716

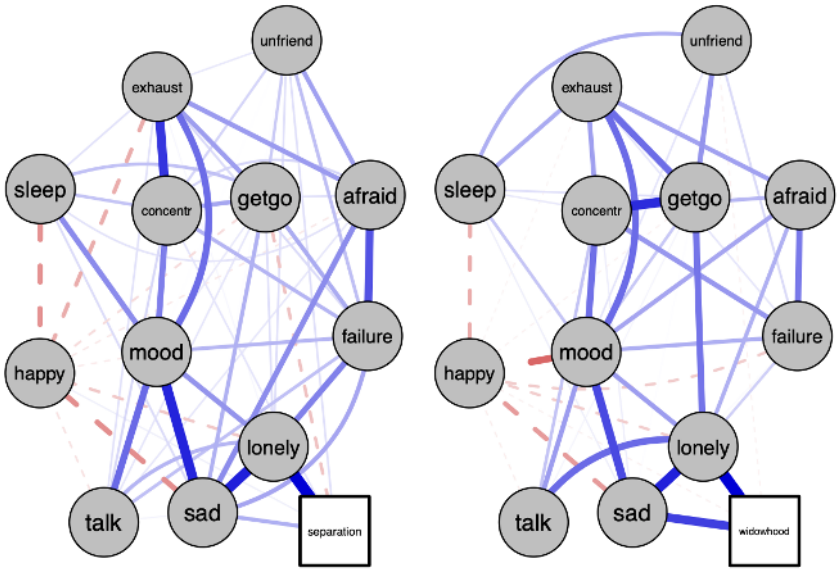


**Figure 1.** Example 1: Regularized partial-correlation networks (tuning-parameter gamma = 0.5) for the separated (left) and the widowed sample (right). Solid-blue edges represent positive, regularized partial-correlations, dashed-red edges represent negative, regularized partial-correlations. No specific minimum/maximum/cut values have been used. Edge weights in the separated network ranged from 0.001 (*mood – unfriendly*) to 0.320 (*lonely – separation*). Edge weights in the widowed network ranged from 0.002 (*sad – getgo*) to 0.300 (*lonely – widowed*).

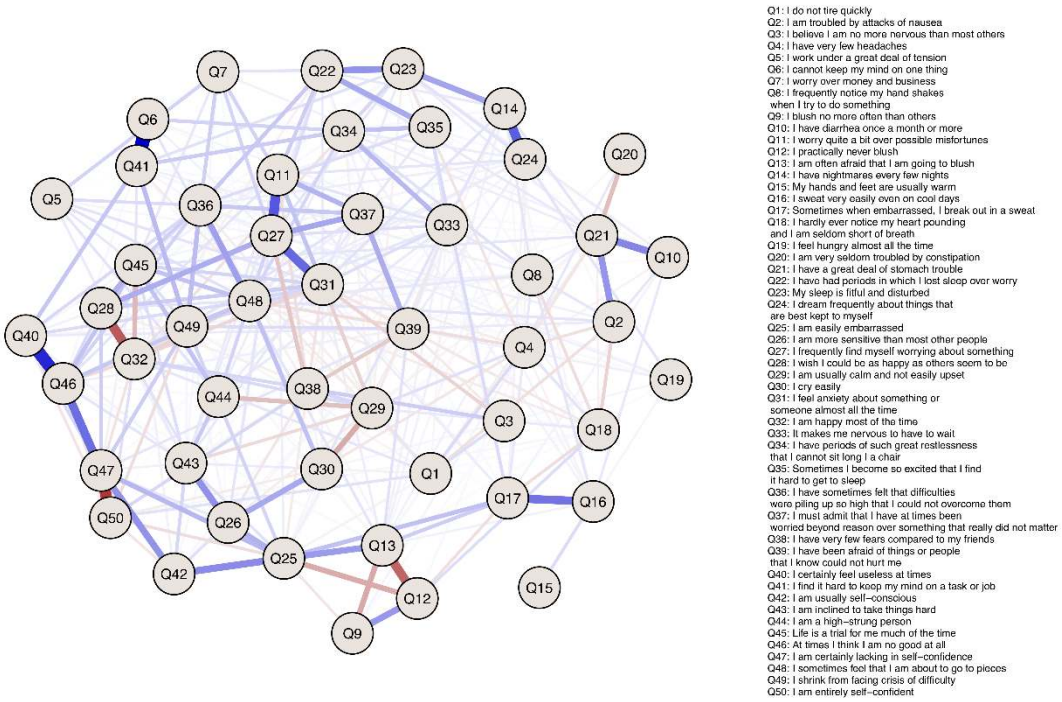


**Figure 2.** Example 2: Regularized log-linear relations. Blue edges represent positive relations, red edges represent negative relations. The cut argument has been set to 0. Edge weights ranged from –1.82 (Q47–Q50) to 2.28 (Q6–Q41).

Table 1. Overview and detailed information of commonly applied estimation routines in R.

| Model (data) | Parameter interpretation | <package name>::<main function> | Description | Input type* | Main defaults | Bootnet default set | Bootnet default differences | Missing data handling | Notes |
|---|---|---|---|---|---|---|---|---|---|
| Ising Model (binary) | Logistic regression coefficients / loglinear interactions | `IsingFit::IsingFit` | Regularized estimation nodewise logistic regressions and EBIC model selection | Raw data (0/1 encoded) | Gamma hyperparameter is set to 0.25 and an AND-rule is used (both regression estimates required to be nonzero). | `"IsingFit"` | Automatic missing data removal (listwise) and automatic median split if input data is not binary | None (rows with missing data need be removed before analysis) | Regularization via *glmnet* package. |
| | | `IsingSampler::EstimateIsing` | Unregularized estimation using psuedolikelihood, loglinear modeling or univariate logistic regressions | Raw data (any binary encoding) | Pseudolikelihood estimation (`method = "pl"`). | `"IsingSampler"` | Loglinear model used for up to 20 nodes and univariate logistic regressions for >20 nodes | None (rows with missing data need be removed before analysis) | Saturated or pre-defined model only (bootstrap threshold with `bootnet::bootThreshold`) |
| | | `psychonetrics::Ising` | Maximum likelihood estimation | Raw data (any binary encoding) or summary statistics (means + covariance matrix) | psychonetrics model** | N/A | N/A | Listwise deletion (pairwise covariance matrix can potentially be used as input) | Not possible with many (over 20) nodes. |
| Gaussian Graphical Model (normal or ordinal data) | The parameters (i.e., edges) represent the unique association among two variables, after conditioning on all other variables in the network**** | `qgraph::EBICglasso` | Regularized estimation using glasso and EBIC Model selection. | Variance-covariance / correlation matrix | Gamma hyperparameter is set to 0.5 | `"EBICglasso"` | Bootnet correlation defaults used when raw data is used as input*** | Pairwise deletion (sample size can be set to average of sample sizes for each pair of variables) | Poor performance in large sample sizes with dense network structures |
| | | `qgraph::qgraph(..., graph = "pcor")` | Unregularized network (saturated). | Variance-covariance / correlation matrix | Saturated model (all edges included). | `"pcor"` | Bootnet correlation defaults used when raw data is used as input*** | Pairwise deletion (sample size can be set to average of sample sizes for each pair of variables) | Edges that are not significant (based on *p*-values or bootstraps) can be hidden, but no model selection is performed. |
| | | `qgraph::ggmModSelect` | Unregularized estimation using extensive model search | Variance-covariance / correlation matrix | Gamma hyperparameter is set to 0 (BIC) | `"ggmModSelect"` | Bootnet correlation defaults used when raw data is used as input*** | Pairwise deletion (sample size can be set to average of sample sizes for each pair of variables) | Slow with many nodes (>30) unless `stepwise = FALSE` is used. Note that this setting, however, also has its drawbacks (Isvoranu & Epskamp, 2021), and the decision should not only be based on the complexity of the network but also on the implications for the algorithm. |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | `psychonetrics::ggm` | (Full information) maximum likelihood estimation | Raw data or summary statistics (means + covariance matrix) | psychonetrics model** | N/A | N/A | Missing data handling through full information maximum likelihood is supported with estimator = "FIML" | Ordinal data supported with `ordered = TRUE` (uses weighted least squares estimation). |
| Mixed Graphical Model (normal / categorical / count) | (logistic / linear / multinomial) regression weights based on standardized data | `mgm::mgm(..., lambdaSel = "EBIC")` | Regularized nodewise regressions with EBIC model selection, potentially with interaction effects (3-way, 4-way, etcetera) | Raw data | Gamma hyperparameter is set to 0.25 | `"mgm" (criterion = "EBIC")` | Gamma hyperparameter is set to 0.5, type and level arguments are automatically set, edges are automatically signed if possible, listwise deletion automatically applied to data. | None (rows with missing data need be removed before analysis) | Default when using *bootnet* but not when using *mgm*. Reduces to Ising model with only binary variables. Edge weights between continuous variables are **not** partial correlation coefficients (but have the same interpretation) |
| | | `mgm::mgm(..., lambdaSel = "CV")` | Regularized nodewise regressions with *k*-fold cross-validation model selection, potentially with interaction effects (3-way, 4-way, etcetera) | Raw data | Number of folds is set to 10 | `"mgm" (criterion = "CV")` | type and level arguments are automatically set, edges are automatically signed if possible, listwise deletion automatically applied to data. | None (rows with missing data need be removed before analysis) | Default when using mgm but not when using *bootnet*. Reduces to Ising model with only binary variables. Edge weights between continuous variables are **not** partial correlation coefficients (but have the same interpretation) |
| Correlation Network (any) | Bivariate marginal correlations | `qgraph::qgraph(..., graph = "cor")` | Bivariate estimation | Variance-covariance / correlation matrix | Saturated model (all edges included). | `"cor"` | Bootnet correlation defaults used when raw data is used as input*** | Pairwise deletion (sample size can be set to average of sample sizes for each pair of variables) | Edges that are not significant (based on *p*-values or bootstraps) can be hidden, but no model selection is performed. |
| | | `psychonetrics::corr` | Maximum likelihood estimation | Raw data or summary statistics (means + covariance matrix) | psychonetrics model** | N/A | N/A | Missing data handling through full information maximum likelihood is supported with estimator = "FIML" | Ordinal data supported with `ordered = TRUE` (uses weighted least squares estimation). |
| Relative Importance Network (continuous) | Normalized *lmg* metric relative importance masures | `relaimpo::calc.relimp` | Relative importance | Raw data | Saturated model (all edges included). | `"relimp"` | No automated function outside of bootnet wrapper | None (rows with missing data need be removed before analysis) | Returns directed (not causal) network |

* Input for bootnet estimateNetwork function must always be raw data
** Models in psychonetrics are saturated by default. Functions such as prune, stepup and modelsearch can be used for exploratory model search.
*** Bootnet will automatically correlate the data if raw data is used (Pearson correlations from Version 1.4 onwards). For ordinal data, polychoric correlations or Spearman correlations can be used as input.
**** In cases in which a transformation is applied (e.g., huge; not standardized), the parameters should be interpreted as precision matrix elements.

Table 2. Overview of routines covered in the two examples.

| | **Example 1** | **Example 2** |
|---|---|---|
| ***Data description*** | **Relationships in later life** (data and results from *Bereavement or breakup: Differences in networks of depression*; Burger et al., 2020). | **Taylor Manifest Anxiety Scale** (data and results from an online offering of the *Taylor Manifest Anxiety Scale*; Taylor, 1953). |
| | **Methods: General analysis routine** | |
| ***2.1 Sample collection*** | ✓ | ✓ |
| ***2.2 Variable selection procedure*** | ✓ | ✓ |
| ***2.3 Deterministic relations between variables and skip-structures*** | ✓ (not applicable) | ✓ (not applicable) |
| ***2.4 Estimation method*** | ✓ | ✓ |
| ***2.5 Accuracy and stability of edge-estimates*** | ✓ | ✓ |
| ***2.6 Statistical packages*** | ✓ | ✓ |
| | **Methods: Analysis-specific routine** | |
| ***2.7 Group comparison*** | ✓ | |
| ***2.8 Centrality indices*** | | ✓ |
| ***2.9 Differences between edges*** | | ✓ |
| ***2.10 Clustering*** | | |

Table 2. (*cont*)

| | | |
|---|---|---|
| **Results: General analysis routine** | | |
| ***3.1 Final sample size*** | ✓ | ✓ |
| ***3.2 Results of the accuracy and stability checks*** | ✓ | ✓ |
| **Results: Analysis-specific routine** | | |
| ***3.3 Network visualization*** | ✓ | ✓ |
| ***3.4 Network density and average absolute edge weights*** | ✓ | |
| ***3.5 Centrality indices*** | | ✓ |
| ***3.6 Predictability*** | | |
| ***3.7 Specific nodes and edges*** | | ✓ |
| ***3.8 Group comparisons*** | ✓ | |