False alarm? A comprehensive reanalysis of "evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger (2017)

Borsboom, D.; Fried, E.I.; Epskamp, S.; Waldorp, L.J.; van Borkulo, C.D.; van der Maas, H.L.J.; Cramer, A.O.J.

DOI
10.1037/abn0000306

Publication date
2017

Document Version
Submitted manuscript

Published in
Journal of Abnormal Psychology

False alarm? A comprehensive reanalysis of "Evidence that psychopathology symptom networks have limited re....

Article *in* Journal of Abnormal Psychology · October 2017

DOI: 10.1037/abn0000306

CITATIONS
8

READS
361

7 authors, including:

Denny Borsboom
University of Amsterdam
195 PUBLICATIONS **8,138** CITATIONS
SEE PROFILE

Eiko Fried
University of Amsterdam
70 PUBLICATIONS **804** CITATIONS
SEE PROFILE

Sacha Epskamp
University of Amsterdam
49 PUBLICATIONS **2,183** CITATIONS
SEE PROFILE

Lourens Waldorp
University of Amsterdam
97 PUBLICATIONS **2,169** CITATIONS
SEE PROFILE

False alarm?

A comprehensive reanalysis of "Evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger.

Denny Borsboom

Eiko I. Fried

Sacha Epskamp

Lourens J. Waldorp

Claudia D. van Borkulo

Han L. J. Van der Maas

University of Amsterdam

Angélique O. J. Cramer

Tilburg University

Word count:
Main text: 7061
Abstract: 223

Author note

We would like to thank Richard McNally, Jeroen Vermunt, Claudi Bockting, and Helma van den Berg for their comments on an earlier draft of this paper, and Ria Hoekstra for her help in gathering and processing data used in this manuscript. Denny Borsboom, Eiko Fried, Claudia van Borkulo, and Lourens Waldorp are supported by European Research Council Consolidator Grant no. 647209. Angélique Cramer is supported by Veni grant no. 451-14-002 awarded by the Netherlands Organisation for Scientific Research (NWO).

Correspondence concerning this article should be addressed to Denny Borsboom, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, The Netherlands, email: dennyborsboom@gmail.com.

Abstract

Forbes, Wright, Markon, and Krueger (2017) state that "psychopathology networks have limited replicability" and that "popular network analysis methods produce unreliable results". These conclusions are based on an assessment of the replicability of four different network models for symptoms of major depression and generalized anxiety across two samples; in addition, Forbes et al. (2017) analyze the stability of the network models within the samples using split-halves. Our re-analysis of the same data with the same methods led to results directly opposed to those of Forbes et al. (2017): All network models replicate very well across the two datasets and across the split-halves. We trace the differences between Forbes et al.'s (2017) results and our own to the fact that they did not appear to accurately implement all network models, and used debatable metrics to assess replicability. In particular, Forbes et al. (2017) deviate from existing estimation routines for relative importance networks, do not acknowledge the fact that the skip-structure used in the interviews strongly distorted correlations between symptoms, and incorrectly assume that network structures and metrics should not only be expected to be the same across the different samples, but also across the different network models used. In addition to a comprehensive re-analysis of the data, we end with a discussion of best practices concerning future research into the replicability of psychometric networks.

General scientific summary

This commentary presents a reanalysis of the data presented in the target paper by Forbes, Wright, Markon, and Krueger (2017), which shows that, contrary to their conclusions, network models replicate well.

Introduction

Network modeling is quickly gaining ground as a promising way of understanding psychopathological phenomena. As both the theoretical framework and the statistical modeling routines have seen rapid development over the past few years, recent papers have begun to take stock of what has been achieved and to evaluate which new directions psychopathological network research should take (Fried & Cramer, *in press*; Fried, van Borkulo, Cramer, Boschloo, Schoevers, & Borsboom, 2017). The reproducibility of network research ranks firmly among the top priorities: as Epskamp, Borsboom, and Fried (2017) state, "[t]he current replication crisis in psychology stresses the crucial importance of obtaining robust results, and we want the emerging field of psychopathological networks to start off on the right foot". Similarly, replicability was recently highlighted as one of the five core challenges that the psychopathological network discipline is facing (Fried & Cramer, *in press*).

Thus, the importance of assessing stability and replicability of network structures stands beyond doubt. Upon reading FWMK's conclusions, therefore, our immediate reaction was one of concern about some of the network analysis methodologies currently in use; a response we expect many readers to share, especially because FWMK do not thread lightly in their assessment of psychopathology networks. Even though their analysis is limited to just two datasets, they do not hesitate to draw general conclusions and state that "popular network analysis methods produce unreliable results" (General Scientific Summary, p. 1), have "poor replicability" (p. 18) and "limited utility" (p. 18), so that "novel results originating from

psychopathology networks should be held to higher standards of evidence before they are ready for dissemination or implementation in the field" (p. 18).

However, after we had acquired access to the datasets FWMK analyze and used the appropriate network analyses, we found that many of the numerical results from our statistical analyses turned out vastly different from those of FWMK, and support the exact opposite of FWMK's conclusion: psychopathology networks replicate very well. We were able to trace the diverging results to a number of inaccuracies in FWMK's analyses. First, contrary to their claims, FWMK do not accurately implement state-of-the-art network analyses, as we will show below. Second, FWMK's methodology for assessing replication uses debatable measures of replicability. Third, the correlation matrices used by FWMK are distorted due to the presence of a skip structure in the interview.

In the present commentary, we will illustrate how these issues have led FWMK to underestimate the quality of network methodology. In addition, we discuss best practices to most effectively conduct research into the reproducibility of psychopathology networks.

Evidence that psychopathology networks replicate well

When we set out to reproduce FWMK's results using the same analyses on the same NCS-R and NSMHWB data and split-halves[1], we found that networks replicated well. Table 1 shows a summary of these results for Ising models, relative importance networks, and Directed Acyclic Graphs (DAGs). We do not report results for association networks, first

[1] We would like to thank FWMK for providing us with the exact splits of the data used in the split-half analyses.

because FWMK do not challenge the replicability of association networks, and second because we encountered major issues with the correlation matrices that we discuss in the next section. In addition to the replicability metrics used by FWMK, we report additional metrics to facilitate assessment of the degree to which networks replicate[2]. The most intuitive and important of these metrics, in our view, is the correlation between the network connections in the NCS-R and NSMHWB datasets. This correlation measures the correspondence between the strength of network connections found in both datasets. If the correlation equals one, network connections in the networks are perfectly linearly related across samples, meaning that the networks have essentially the same structure; if it equals zero, the networks have no detectable linear correspondence; if it equals minus one, the networks are exact opposites.

Table 1 shows that the correlations between network connection strengths are all well above .9, indicating that the networks found in the datasets under consideration are highly similar. Figure 1 shows this high correspondence between the network structures by representing them using the same layout; this is advisable because even when plotting two exactly identical networks with different layouts, it is impossible to tell visually how similar networks are. Our split-half analyses, using the same splits as used by FWMK, show comparable results: all parametric network models show correlations between network connection parameters of well over 0.9[3]. We shortly discuss these results, after which we will turn to the question why FWMK reach conclusions opposite from ours.

---

[2] All analyses we report are performed using R version 3.3.1 and the relevant packages on platform x86_64-w64-mingw32. All code is available at https://osf.io/akywf, with the exception of the NSMHWB dataset which is not publicly accessible; an instructive summary of our analyses with a subset of sample code can be consulted in Appendix A.

[3] Results of the split-half analyses are included in Appendix B.

Insert Table 1 about here

Table and appendices are supplied below the manuscript

The Ising model. The Ising model (Van Borkulo et al., 2014) is arguably the most important of the models fitted by FWMK, as it represents state-of-the-art regularized network model estimation for Pairwise Markov Random Fields (PMRFs; Epskamp, 2017) in dichotomous data. Tallying all networks that are reported in the literature at the moment of writing this comment, 62% used a variant of the PMRF, and this percentage is growing quickly because the PMRF has become the default network modeling technique. It is complemented by robustness analyses in *bootnet* (Epskamp, Borsboom, & Fried, 2017) as well as statistical tests for network invariance (Van Borkulo et al., 2016), which are powerful tools in assessing network estimation quality and testing the equivalence of network models in different populations, as we will illustrate in this comment.

As FWMK note themselves, and as Figure 1 (left panels) shows, estimated Ising networks are nearly identical: node threshold parameters correlate .93 across the datasets, while network connection parameters (edge weights) show a correlation of .95 (Spearman correlations equal .85 and .88, respectively). Even though the absolute position of nodes in centrality orders is not invariant, as also reported by FWMK, their relative positions are strongly aligned: the centrality metrics of strength, betweenness and closeness correlate 0.94, 0.94, and 0.76, respectively, across the two datasets. The only sign of non-replication

concerns the presence of three weak negative edges in the NCS-data that were absent in the NSMHWB data; however, this difference across samples was not statistically significant (see below). Split-half analyses, as reported in Appendix B, show similar results and indicate high stability of the Ising model.

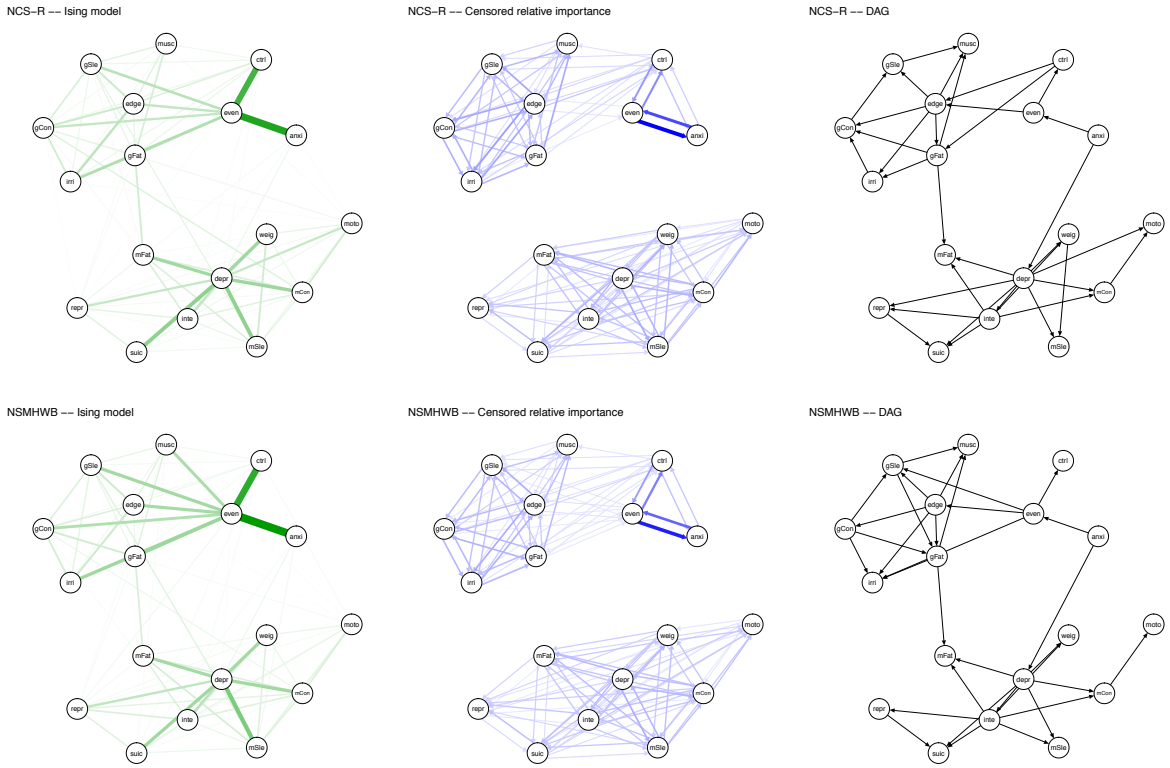

Figure 1. Network structures estimated with the Ising model (left panels), censored relative importance networks (middle panels), and Directed Acyclic Graphs (DAGs; right panels) for the NCS-R (top panels) and NSMHWB (bottom panels) data.

Moving beyond descriptive measures, and in contrast to FWMK, we used the Network Comparison Test (NCT) to statistically evaluate the similarity of the Ising models estimated on the NCS-R and NSMHWB data using permutation testing (Van Borkulo et al., 2016). The NCT results also indicate that the network structures of NCS-R and NSMHW replicate very well. First, a test for invariance of network structures, which tests the null hypothesis that all edges are precisely identical across the samples, was not significant (M=2.66, p=0.121). Second, testing for the invariance of individual edges revealed that none of the edges differed significantly across the two datasets. Thus, despite the high power to detect differences given the two large samples (N~9000 per sample), we could not reject the null hypothesis that the NCS-R and NSMHWB networks are precisely identical at the level of the populations from which these samples were drawn.

Relative importance networks. As shown in Figure 1 (middle panels), relative importance networks, which were estimated by exactly following the original methodology described in Robinaugh et al. (2014), replicated even better than the Ising models. Uncensored relative importance networks featured a correlation of 0.99 between the estimated edge weights in the two datasets, as well as between split-halves of the same datasets (see Appendix B). These findings deviate significantly from those of FWMK; we will explain this divergence in the next section.

DAG analysis. Replication results for DAGs were good, although not as excellent as the results for the Ising models and relative importance networks. This is not surprising because

DAGs require stronger assumptions[4], which are less likely to be met in these data. As Table 1 and Figure 1 show, 27 out of 34 DAG edges replicated from the NCS-R to the NSMHWB dataset (79.4%), which indicates that the results do converge. In addition, in- and outdegree of nodes featured correlations of .62 and .87 respectively. Visual inspection of Figure 1 (right panels) shows that the same bridge symptoms, which connect MDE to GAD, are identified in the two datasets. Of note, two edges (gFat - gCon and gCon - irri) switch direction between the datasets.

Cross-method replicability. FWMK count how often edges show up in different network estimation routines. It is clear from the way they interpret the resulting findings that they assume that one should expect these different networks to converge to 100%. This, however, is not true. For instance, suppose the data arose from the DAG A → B ← C → D. Then one would *not* expect to find the Ising model to return the network A – B – C – D, because B is a common effect of A and C and therefore A and C must be conditionally dependent given B[5] (Pearl, 2011). Instead, one expects the network to also include a direct relation between A and C. In addition, given this network structure, one would *never* expect *any* correlations to be nonzero in the association network: because all variables are connected, one instead expects a fully connected association network. Thus, counting how often individual edges

[4] For example, DAG analysis assumes that the causal graph contains no cycles and that there are no independence relations in the data that are not a function of the causal relations coded in the DAG (faithfulness); see e.g. Pearl (2009) for an extensive treatment.
[5] This is because, if A and C are independent causes of B, then knowing that B is present means that, if A is not present and thus did not produce B, then C must have been the cause of B.

replicate across these different network structures is of limited utility, because it is implausible to expect them to be the same.

In addition, network estimation techniques differ in sensitivity and specificity (Van Borkulo et al., 2014), meaning that some techniques more often err on the side of caution, and as such identify fewer edges, which should be accommodated in assessing replicability. For instance, in relative importance networks all connections are estimated, while Ising models only estimate connections that improve the fit of the model (van Borkulo et al., 2014). Similarly, given the stronger causal interpretation of edges in a DAG opposed to Ising models, it is sensible that DAG estimation methods should be more conservative than Ising model estimation methods, leading DAGs to be sparser. Thus, in addition to principled differences between the edges the methods should detect in the first place, there are also differences in sensitivity and specificity that should be accounted for.

Therefore, rather than counting how many edges are present in different networks, one should investigate a *nesting relationship*: a sparser network should not estimate edges that are absent from the denser network, and a denser network should not leave out edges that are present in the sparser network. When assessing this nesting relation, we found that 100% of the edges in the NCS-DAG (the sparser network) were present in the NCS-Ising model (the denser network). The same holds for the NSMHWB data. Strikingly, when we compare DAGs and Ising models *across* datasets, 97% of the NCS DAG-edges are included in the NSMHWB Ising model, and 100% of the NSMHWB DAG-edges are included in the NCS Ising model. In addition, we found that 100% of the edges that are missing in the Ising models are also missing in the DAGs. This is the case both in the split-half analysis and in the

replication analysis. Cross-method replication could hardly be better[6].

Why did FWMK underestimate the replicability of psychopathology networks?

It is remarkable that our results differ so much from those of FWMK, especially given the strong conclusions FWMK draw. After studying their methodology in detail, we argue that the different conclusions originate from two sources. First, FWMK's analyses contain several statistical inaccuracies. With "statistical inaccuracies", we mean to identify statistical computations that we expect FWMK to acknowledge, upon reflection, as yielding a suboptimal representation of the relations in the data[7]. Unfortunately, these inaccuracies have had strong impact on the results. Second, their results rest on debatable methodologies. With "debatable methodologies" we mean to identify issues that we see as problematic, but that can be legitimately disputed depending on one's point of view on what psychopathology networks should deliver or even on one's underlying philosophy of science. We discuss these issues in turn.

Statistical inaccuracies

When studying their methodology, we found that FWMK do not adopt the standard estimation procedure for relative importance networks introduced by Robinaugh et al. (2014),

[6] We have not investigated the cross-method replicability including relative importance networks, as these do not feature careful edge selection methods.

[7] One of these inaccuracies was already acknowledged: the reader may note that the DAGs in Forbes et al. (2017) are different from the widely circulated version of their paper in April 2017 that was accepted for publication in the *Journal of Abnormal Psychology*, and that we were asked to comment on. The difference is due to an error in the implementation of DAGs that FWMK caught in time to correct the paper between acceptance and publication.

nor any other published procedure[8]. It is unclear why they deviate from the standard procedure that is used in Robinaugh et al. (2014) and, to the best of our knowledge, in all other papers that have used relative importance networks (Bryant et al., 2017; Heeren & McNally, 2016; Hoorelbeke, Marchetti, De Schryver, & Koster, 2016; McNally, 2016; McNally et al., 2015).

First, FWMK use non-normalized instead of normalized estimates for the *lmg* metric to assess relative importance. While the optimal choice here is debatable, this poses a deviation from standing methodology that should have been acknowledged. Second, FWMK strictly threshold networks by permanently excluding edges under 0.05 from the network, while Robinaugh et al. (2014) removed these edges for visualization, but not in the computation of centrality measures. Third, and most importantly, FWMK deviate from existing work by introducing a thresholding procedure that has extreme consequences: whenever an edge between two nodes (e.g., A → B) does not have a weight of at least 0.005 points higher than the corresponding reciprocal edge for the same two nodes (i.e., A ← B), FWMK remove that edge from the network. This thresholding rule has not been used anywhere else in the literature, and for good reason. For suppose A explains 50% of the variance in B, and B explains 50% of the variance in A: even though these could be the strongest edges in the network, *they would be both removed* because neither of these edges is >0.005 points higher than the other.

The consequences of FWMK's procedure are illustrated in Figure 2. When the

[8] If we deviate in the same way from the literature, we can reproduce their reported results and hence we are certain this deviation is the source of the differences; see Figure 2 for details and https://osf.io/2t7qp/ for code replicating both our and FWMK's analyses.

relative importance network is computed on the NCS-R data as described by Robinaugh et al. (2014), the resulting network retains 118 edges (left panel). Using non-normalized *lmg* with the same threshold results in a network that retains 99 edges (middle panel). Finally, applying the deviant thresholding procedure used by FWMK duplicates their analysis, leaving only 31 out of the original 118 edges (right panel; the red edges in the middle panel network indicate those removed by FWMK's thresholding rule). Occasionally this procedure indeed deletes both edges between two nodes; e.g., both edges between *even* and *ctrl* are deleted, as one can see by comparing the correctly computed network (Figure 2, left panel) to the network reported by FWMK (Figure 2, right panel).

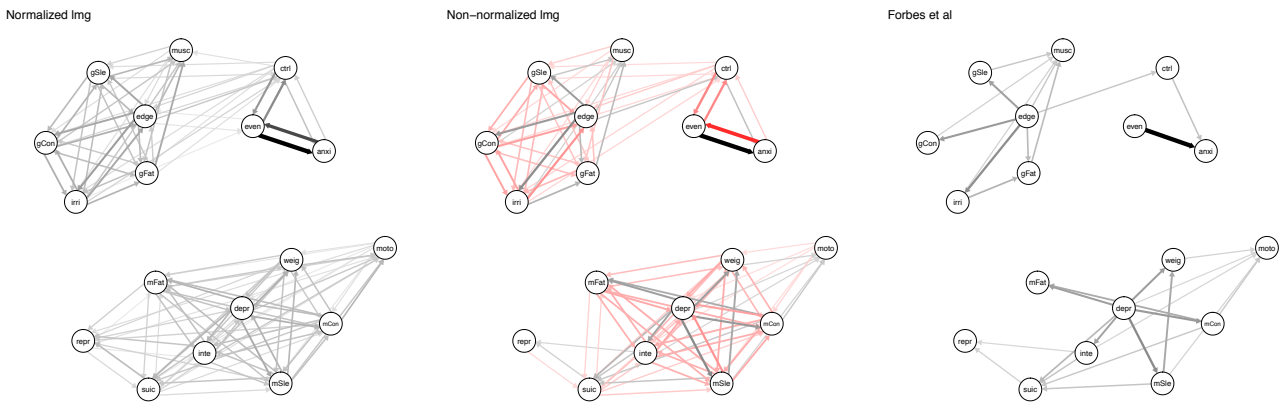

Figure 2. Relative importance networks (estimated on the NCS-R data) using normalized *lmg* (left, as used by Robinaugh et al. 2014; 118 edges) and non-normalized *lmg* (middle and right). Red edges in the middle panel (99 edges) indicate edges that are removed by the thresholding rule used by Forbes et al. (2017), and the right panel shows the network they reported (31 edges).

Thus, these analyses do not replicate the standard procedure introduced by Robinaugh et al. (2014) or any other procedure currently in the literature, and introduce a thresholding rule that causes many edges, including some of the strongest, to be deleted. As a result, we suggest that the conclusions presented by FWMK that pertain to relative importance networks are not trustworthy, and that our results, as presented in Table 1, should be consulted instead. It should be noted that these results should still be interpreted with care, as it is unclear whether relative importance networks, as used by Robinaugh et al. (2014) on continuous data, generalize well to the binary data analyzed here in the first place; relative importance networks are computed using linear regressions, which introduces an inappropriate distributional assumption. However, in contrast to the above inaccuracies, we were not able to resolve this in the current work; hence the reader should keep in mind that both FWMK's paper and our reanalysis are based on an incorrect distributional assumption insofar as relative importance networks are concerned.

A second issue that we consider to qualify as a statistical inaccuracy concerns FWMK's use of a distorted tetrachoric correlation matrix, which underlies both their factor analyses and their association networks. To see why this correlation matrix is distorted, first note that the Composite International Diagnostic Interview (CIDI), which yielded the symptom data, involves a skip-structure. This means that the full symptomatology of MDE is only interrogated if at least one of the core symptoms of *depressed mood* and *loss of interest* was present; the full symptomatology of GAD is only interrogated if the interviewee reported the presence of *anxiety*, *anxiety about multiple events* and *loss of control about the worry*. As such, both the NCS-R and the NSMHWB data contain a high percentage of missing values.

In both datasets, FWMK impute zeros for these missing values. This practice assumes Guttman scale properties for the skipped symptoms, i.e., if one does not have the symptom of *feeling sad* over a period of two weeks (a symptom that acts as a gateway in the skip structure), one cannot have the symptom of *insomnia* (a non-gateway item). This practice is acceptable in many contexts, and although the procedure can strongly affect all network models, it does not necessarily invalidate their results. For instance, as can be seen in Figure 1, the GAD skip structure translates to the sequence *anx* → *eve* → *ctrl* in the DAG, with *eve* being the most important gateway item connecting to the other symptoms, while the MDE skip structure translates to the sequence *depr* → *inte* in MDE. These sequences accurately reflect the actual order of the symptoms in the interview, and thus the DAGs correctly pick up the skip structure, which we know is a true causal structure in the data (see also Borsboom & Cramer, 2013, Figure 7).

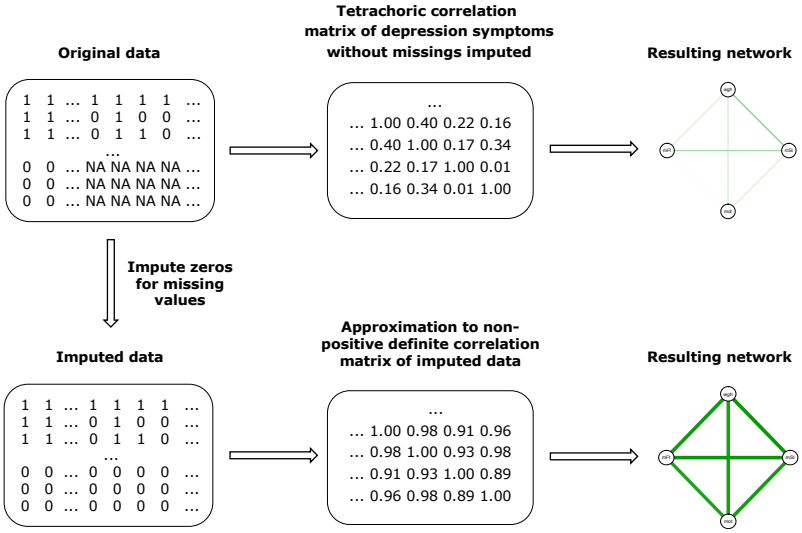

Figure 3. The effect of the imputation strategy used by FWMK on the network structure of the NCS-R data. The figure shows how imputation alters the tetrachoric correlation between depression symptoms, and the resulting networks. The correlations shown represent actual NCS-R correlations between four non-skip items (weight problems, sleep problems, psychomotor problems, and fatigue) before and after imputation.

Unfortunately, however, imputing zeros for missing values is not advisable when the goal is to analyze or represent a correlation matrix. This is because it alters the correlations in the data enormously, as is graphically represented in Figure 3. To give an indication of how serious these distortions are, we note that the *average* correlation between depression symptoms in the correlation matrix as used by FWMK equals .94 for the NCS-R data and .96 for the NSMHWB data. This is unrealistically high and nowhere near the average tetrachoric correlation of .33 that characterizes the data if missing values are handled with, for example, pairwise deletion. Also, these values do not resemble correlations typically found for these kinds of symptoms (e.g. see Beard et al., 2016).

In addition, the imputation process introduces deterministic dependencies in the data, which in this case leads the correlation matrices for both the NCS-R and the NSMHWB data to become non-positive definite (this means that the matrices do not have the characteristics every proper correlation matrix should have and, therefore, should not be used in standard statistical analyses). As a result, these correlation matrices are untrustworthy, and unrepresentative of the associations present in the data. Because of this, the results of both association networks and factor analyses reported by FWMK are unreliable. Note that the effects of the imputation strategy are visible in all analyses that FWMK report, and that they affect our analyses in the same way. At present we are unaware of an analytic strategy that could address this issue satisfactorily.

It is important to recognize that, because of the problems outlined above, all statistics reported by FWMK that pertain to association networks and relative importance networks are either inaccurate or corrupted to an unknown extent by FWMK's imputation strategy. This

has direct consequences for FWMK's findings with respect to cross-method replicability. For example, FWMK's abstract presents, as a main result, that "only 13-21% of the edges were consistently estimated across these networks". These percentages are uninformative, not only because one does not in fact expect different networks to converge upon the same structure, as explained in the previous section, but also because the underlying computations are compromised by statistical inaccuracies, as identified in this section. In fact, the only interpretable results on cross-method replicability that FWMK *could* have obtained pertain to the comparison between Ising models and DAGs, because these are the only models that they estimated without problems[9]. With respect to this comparison, however, FWMK claim that 41 edges of the NCS-R DAG were also present in their NCS-R Ising model (see their Footnote 7). Unfortunately, their NCS-R DAG only contained 34 edges, which means it is impossible that 41 edges would replicate. We therefore have no other option than to conclude that none of the statistics on cross-method replicability reported by FWMK are accurate.

Debatable methodology for assessing replicability

After pointing out the statistical inaccuracies in FWMK's analyses, this section covers the methodology used to evaluate the replicability of network models. In contrast to the issues mentioned in the previous section, one can have legitimately different points of view on the appropriateness of the measures in question and the importance of the problems they encounter. In our view, the main problem with FWMK's assessment of replicability is that

[9] The reader should take care to interpret this statement as applying to FWMK's published paper and not to their widely circulated preprint, which did not implement DAGs correctly.

they do not use any measures that would seem of immediate relevance to any such analysis (e.g., correlations between the edge weights across samples, as reported in Table 1, or statistical tests such as NCT), and instead rely on several replicability and stability measures that have not been validated, and that are problematic for reasons explained below.

First, FWMK compute the *percentage* of change of the value of a parameter from one dataset to the next, and then average this percentage over all parameters. This percentage is relative to the original size of the edge. This means that small changes in parameters very close to zero can result in huge differences: for instance, when the same parameter is 0.00001 in one dataset and 0.00003 in the second, the computations of FWMK convert this into a 300% change, which may be entirely inconsequential for the interpretation of the network structure. Figure 4 (left panel) illustrates, for the Ising model, how it is possible for parameter values to feature an average 30% change across datasets, even though the network parameters are in fact nearly identical. The reason is indeed that large percentage changes are much more likely to occur in small edge weights: strong edge weights hardly change at all. As a result, the correlation between edge weights remains extremely high (Figure 4, right panel).

To show that this problem arises in latent variable models as well as networks, we also computed FWMK's measure for the parameters of a two-dimensional IRT model fitted on the NCS data; when replicating this model on the NMSHWB data, the percentage parameter change equals 44%, while the correlation between the discrimination parameters in the two samples equals 0.96. Moreover, a small simulation in which we simulated data from a two-factor model and applied FWMK's measure resulted in an average parameter change no less than 483%, even though the parameters of the model correlate 0.99 across samples.

Thus, factor models show roughly the same behavior as network models with this measure.

We conclude that it is inadvisable to attach normative evaluations to the absolute estimates of this metric, as FWMK do when they interpret the percentage differences in parameter estimates ("these are all substantial changes in the context of a model that is promoted for its specificity", p. 14). The average parameter change metric may be productively used in various methodological investigations (e.g., to compare different models or estimation routines in simulation studies), but it is unfit to serve as an arbiter of replicability.

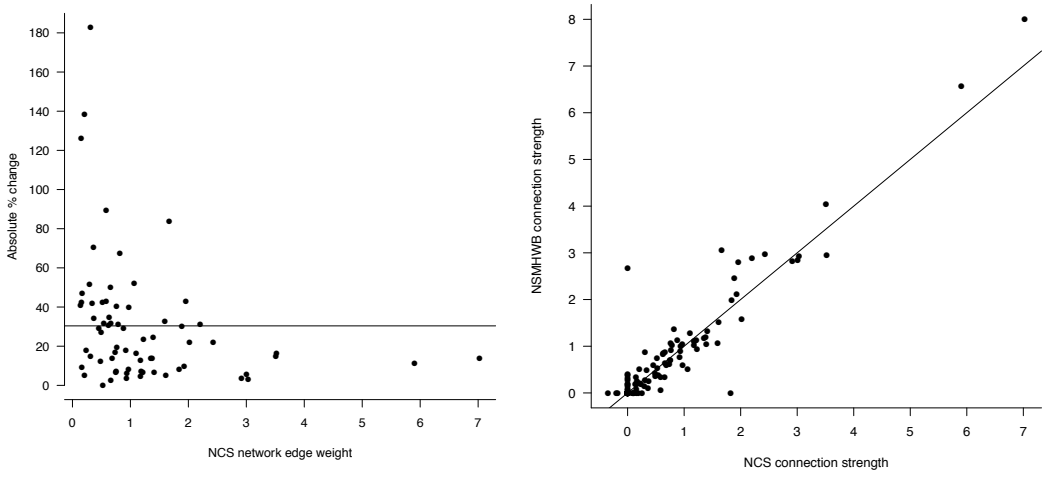

Figure 4. The absolute percentage change in edge weights across datasets relative to the size of the edge weights (left panel). This panel shows that smaller edge weights show larger changes expressed as a percentage of the original weight. The right panel shows that these changes are mostly irrelevant: the strong linear relation (r=.95) between edge weights in the NCS-R and NSMHWB data (right panel) is unaffected by the parameter changes.

Second, FWMK consider how well the *absolute* position of nodes in the centrality ordering replicates, i.e. the question whether a node that ranks 6th in one dataset also ranks 6th in the other. Since edge weights and centrality measures are, as all other statistics, affected by sampling error, nodes can shift positions in the rank ordering due to sampling fluctuations. How strongly sampling fluctuations affects these statistics depends on (a) the sample size, and (b) the differences between nodes in terms of centrality at the population level (i.e., the network structure). Epskamp, Borsboom and Fried (2017) give the extreme example where, at a population level, there are no differences in centrality at all (i.e., all nodes are equally central). In this case, one should not expect that order to replicate at all, because any absolute ordering differences in a given sample must be due to sampling error.

Therefore, instead of expecting the orderings to replicate by default, one should inspect both the network structure and the *sampling variability* of centrality measures, which shows how reliably they are estimated and whether differences between them are statistically significant. Fortunately, the R-package *bootnet* (Epskamp, Borsboom & Fried, 2017) can be used for this purpose. Running *bootnet* on the Ising model results obtained by FWMK shows that most of the edge weights, which are the basis of centrality calculations, are estimated reliably (see Appendix 1); however, the edges related to the gateway items used in the skip structure (especially item 11, which is the symptom of *being anxious about multiple events*) are much less reliable, which is likely due to structural zeros in the contingency tables for these items, as induced in FWMK's treatment of missing values. Inspecting the robustness of the centrality ordering itself reveals that while strength centrality is estimated stably, closeness and betweenness were much less stable. Appendix C explains that this is due to a

particularity in the data that likely results from the skip structure; hence, one should hesitate to generalize this result to other datasets or modeling contexts. We advise that, in future research, investigators do well to interpret centrality results in the context of a robustness analysis using *bootnet*.

In addition, correspondence of the absolute positions of nodes in the centrality ordering across samples, to which FWMK attach primary significance, is extremely strict as a primary measure of replicability. To see this, suppose we have twenty-six nodes, corresponding to the letters in the alphabet, for which centrality measures induce the ordering A,B,…, Z in dataset 1. Now one executes the same analysis in another dataset, which yields the ordering Z, A, B, C, … Y. Because none of the variables occupy the exact same place in the ordering, FWMK would interpret this as evidence that psychopathology networks do not replicate (in fact, there would be no correspondence at all in this case). However, only Z changed position from least to most central, and although no node occupies the exact same absolute position, one should at the same time conclude that the centrality order does replicate to a large degree, since the *relative* positioning is nearly entirely preserved. This does not invalidate FWMK's measure of correspondence in absolute position, which can still be useful, but it does mean that this metric should be viewed with caution and, importantly, should always be assessed a) in the light of stability of the relative positioning of nodes as assessed by the correlation between centrality scores of nodes across samples (e.g., 0.94 for strength and betweenness, and 0.76 for closeness in the Ising model) and b) in the light of sampling variability.

Third, FWMK's express concern over the fact that different centrality measures

identify different nodes as central. However, just as the various network estimation methods get at different aspects of the data and should not be expected to yield the same network solution, centrality measures such as strength, betweenness and closeness are not interchangeable measures that will converge on "the most influential node", as FWMK suggest (p. 5). Instead, they are indices that assess different *kinds* of centrality. Thus, if strength centrality is highest for *depressed mood* but *fatigue* shows the highest score on closeness, or when *anxiety about multiple events* has the highest strength in the Ising model but *depressed mood* has the highest strength in the DAG, that signals neither a problem nor a cause for concern. Instead, these results, if robust across samples and assessment methods, should be viewed as potentially important clues about the structure of a psycho(patho)logical construct under consideration.

So what about measurement error?

Since the various network models replicate very well across datasets, the reader may wonder how this fits in with FWMK's explanation of the supposed poor replication results in terms of measurement error. That is, FWMK hypothesize that, because edges between two nodes are controlled for other nodes in the network, networks primarily work on residual variances that are largely composed of measurement errors. The results of our reanalysis provide a direct refutation of this theory: if FWMK's explanation were correct, one should expect bad replicability, but our analyses in fact show replicability to be good. Also, if FWMK's explanation were correct, one would expect simulation studies and robustness analyses to show that network models produce unreliable results, which is not the case (van Borkulo et

al., 2014; Epskamp et al., 2017).

Indeed, despite the suggestive Venn-diagrams used in their paper, the psychometric intuitions that underlie FWMK's argumentation are inaccurate. The following thought experiment may help elucidate why this is the case. Suppose one encountered a situation in which all systematic relations between depression symptoms were due to a latent variable, and everything else was pure random measurement error. If FWMK were correct, this would imply that a network model should be expected to return a spurious network without any robust connections: after all, because in their view partial correlations are largely correlations between measurement errors, and measurement errors are not structurally related, there is nothing real for the network to go on. However, this is not what one would find: if a latent variable model gave rise to all correlations between variables, then we would not find an empty network but a fully connected one (Epskamp et al., *in press*; Ellis & Junker, 1997). Thus, a latent variable model corresponds to a dense network of systematic relations (Marsman, Maris, Bechger, & Glas, 2015), and not to an empty or spurious network, as FWMK's theory would suggest.

More generally, one can prove that every latent variable structure implies a specific network structure, as Molenaar (2003) already suspected and as Maris and his co-workers have been recently able to formally prove (Epskamp et al., *in press*; Marsman et al., 2015; Kruis & Maris, 2016). Thus, even though network models do not explicitly represent shared variance in a separate node that renders the other nodes conditionally independent (i.e., a latent variable), they do imply the presence of shared variance in sets of connected nodes. In fact, given that the known mathematical equivalence relations between the models implies

that they produce the same joint probability distribution for the items, the models should not be expected to differ in this respect. This has the somewhat ironic consequence that, if network structures replicated badly across two datasets, then this would imply that factor structures (i.e., the configuration of loadings in exploratory factor models) would replicate badly as well. Measurement error has little to do with this, because both latent variable models and network models operate on the same systematic relations in the data.

Despite this, however, we do note that additional methodological research is necessary to systematically study the replication properties of different models under various conditions, as these would likely be influenced by various factors such as the overall fit of the model, the number of parameters (and an important caveat of network models is that they typically do require many parameters to be estimated), and the strength of the associations in the data. Psychometric intuition, however, is an unreliable guide in this respect. Thus, mathematical analyses and simulation studies are required to study these issues, especially when making critical generalized claims about an entire psychometric field based on the analysis of two datasets.

Best practices for future research

Despite the inadequacy of the data and analyses used by FWMK, we stress again that we consider both stability and replicability of networks to be extremely important topics. Therefore, we commend FWMK for taking up these issues. Regarding stability, we agree with FWMK that model stability should be tested in all statistical models, including both network and factor models. Thus, we hope that FWMK's paper—together with the *bootnet*

R-package and the accompanying tutorial paper (Epskamp, Borsboom, and Fried, 2017)—will shift the attention of both applied and technical researchers to this topic. Regarding replicability, we offer a roadmap for network replication studies in this section that may aid future researchers in obtaining more objective and trustworthy results.

The method: Replication as a non-empirical question. First, we address a central issue in the design of FWMK: they confound evidence for replication problems that concern a *particular estimated model* with evidence for problems *of the model in general*. This is a non-sequitur. For suppose that one fitted a specific regression model to two different samples, and the regression coefficients were different from each other. Nobody would conclude from such a result that "regression analysis has limited replicability". The problem with equating "not the same result in two data sets" to "method does not work" is that we do not know whether the 'true' relationship between variables is the same across samples. In the absence of this knowledge, we cannot know for sure if differences in results are due to differences in sample characteristics, or to a flawed method.

One may think that this problem is circumvented in FWMK's evaluation of split-half results, which are based on the correspondence of networks within the same sample. However, this only partly addresses the problem. First, because one does not know whether split-half performance with this particular kind of data (here: MDE-GAD symptom data obtained with interviews containing skips) generalizes to other kinds of data, as is necessary for blanket statements like "popular network analysis methods produce unreliable results", as touted in FWMK's General Scientific Summary. Second, because even in a given sample one

does not know whether any given network model is true, let alone which one, and in the absence of this knowledge it is impossible to assess which part of model instability arises from defects in the methodology, and which part arises from model misfit, population heterogeneity, violations of distributional assumptions, etc.

Thus, if the primary aim of research is to assess the general methodological adequacy of a method, the evaluation of two specific empirical datasets is of limited use. Putting a network *method* to the test requires that one knows the 'true' network structure and this can only be done by (a) establishing mathematical proof that the method converges on the true structure in the long run (as, e.g., Meinshausen and Bühlmann, 2006, have done for the Gaussian graphical model and Ravikumar et al., 2010 for the Ising model) or (b) simulating such 'true' network structures and, subsequently, assess the capability of a method, in a variety of settings, to accurately estimate that 'true' network structure (as executed by Van Borkulo et al., 2014 for the Ising model). This motivates the rule that *methodological adequacy should be established on methodological grounds.*

The network structure of a psychological construct: Replication as an empirical question. Once a particular method is proven to accurately retrieve a 'true' network structure using methodological studies, there is another question of replicability that is empirical in nature; namely, what is the particular network structure of a psychological construct such as major depression, or generalized anxiety disorder? Answering this question *does* entail the comparison of network structures across many data sets and many participants. As we have shown above, the design used by FWMK is suboptimal in this respect, and this raises the

question what kind of methodological design would be needed to properly assess replicability in network analysis. Although the below list is not meant to be exhaustive (see Anderson & Maxwell, 2016, for additional issues in replication research), we suggest the following best practices:

1. *No skip structure.* If one desires a replicability assessment that is not confounded by methodological design, one needs data that do not contain a skip structure. We realize this may be a challenge given that many data sets, such as NCS-R and NSMHBW, do contain a skip structure. We also realize that we are guilty as charged in this respect since we, too, used NCS-R data, albeit it for illustration or hypothesis-generating purposes (Borsboom & Cramer, 2013; Cramer et al., 2010). Also, in certain cases there is no other option than to use a design with skip structures (e.g., one cannot ask a person who does not drink whether they got into legal problems because of drinking; Rhemtulla et al., 2016). Future studies in datasets without skip structure will enable us to gauge the replicability of psychopathology networks, and we are glad to see that such studies are already on the way (e.g., network replicability across four large clinical PTSD datasets: https://osf.io/2t7qp/).
2. *Open access data and code*. Reproducibility studies should themselves be reproducible. The NSMHWB data used in this research, however, are not publicly accessible, which means that third parties cannot replicate either our results or those of FWMK without engaging in a lengthy, cumbersome, and costly procedure to gain access to the data (we were charged 947 USD just to be able to

check the veracity of FWMK's analyses). This is highly undesirable. Replication studies are different from other studies in that their consequences may be more far-reaching, because they can discredit or invalidate whole research programs. Therefore, we need to be sure that the analyses and reported results are sound. The only way interested third parties can verify this is through free access to the data used. We acknowledge that freely available data sets containing clinical patient data may be challenging, for example due to issues concerning extending informed consent of patients to third parties. However, we feel encouraged by a recent paper about replicability in clinical science, which contains a multitude of valuable recommendations, that important progress is forthcoming (Tackett et al., *in press*). Analysis code should naturally always be available, as it is needed to replicate and verify reported analyses – the current report illustrates how important this is – and we commend FWMK for sharing their code.

3. *Preregistration of analyses.* Replication research differs from the exploratory designs in which network analyses are most often used, because researchers have a clear idea of the hypothesis to be tested: replication across samples. In addition, especially in replication research, the selection of measures used to gauge replicability is important: after the data are in, it is always possible to come up with a particular selection of measures that emphasizes evidence for or against replicability. To minimize the influence of subjective choices made after the data are in, we encourage any replicability effort to be preregistered, for example at the Open Science Framework (OSF: https://osf.io). Preregistration has an additional

advantage, because interested researchers are able to check 1) a-priori hypotheses and 2) the analysis plan. The Open Science Framework also allows for uploading the code that was used for the analyses, so other researchers can check the veracity of the reported results *before* the paper is even submitted for review. This reduces the probability of submitting or even publishing papers that later turn out to be ill-founded. In the current study, such a procedure would have safeguarded against the statistical inaccuracies manifest in Forbes et al. (2017).

Conclusion

We think that practically all researchers are united by a common goal: the pursuit of scientific knowledge. As such, we stress the importance of expanding our knowledge about psychopathological networks and acknowledge the challenges ahead (Fried & Cramer, *in press*). If one day we were to find out that networks are either not replicable, or that they cannot be suitable candidate models for explaining psychopathology, then we would consider this a victory for clinical science — despite our investment in these models. Falsification is an essential component of the scientific enterprise and the burden of doing so should befall on all of us and on all our theories and hypotheses.

In our comprehensive re-analysis, however, we have shown that FWMK's devastating conclusions are not licensed by their analysis. We conclude that the main conclusion of FWMK that "popular network analysis methods produce unreliable results", is a strongly overstated generalization that is not warranted on the basis of their research design and statistical analyses. The replicability issue, however, is not settled with the publication of

either FWMK's paper nor our commentary. It is for this reason that we have formulated best practices to investigate the important replication issue properly, by using adequate data and optimal analysis designs. Our hope is that future work will lead towards the robust and replicable scientific knowledge that we should all be looking for.

References

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: beyond statistical significance. *Psychological Methods*, *21*, 1-12.

Beard, C., Millner, A. J., Forgeard, M. J. C., Fried, E.I., Hsu, K. J., Treadway, M., Leonard, C. V., Kertz, S., & Björgvinsson, T. (2016) Network analysis of depression and anxiety symptom relations in a psychiatric sample. *Psychological Medicine*, *46*, 3359–3369.

Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, *62*, 495-523.

Epskamp, S. (2017). *Network Psychometrics*. Doctoral dissertation. Retrieved from http://sachaepskamp.com/Dissertation, June 14th 2017.

Epskamp, S., Borsboom, D. & Fried, E.I. (2017). Estimating psychological networks and their accuracy: a tutorial paper. *Behavior Research Methods*. doi: 10.3758/s13428-017-0862-1

Fried, E. I. & Cramer, A. O. J. (*in press*). Moving forward: challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science*. Retrieved from https://osf.io/bnekp, June 11th 2017.

Fried, E. I., van Borkulo, C. D., Cramer, A. O. J., Lynn, B., Schoevers, R. A., Borsboom, D. (2016). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*. doi: 10.1007/s00127-016-1319-z

Forbes, M., Wright, A., Markon, K., & Krueger, R. (2017). Evidence that psychopathology symptom networks have limited replicability. *Journal of Abnormal Psychology*.

Kruis, J., & Maris, G. (2016). Three representations of the Ising model. *Scientific Reports*, *6*,

34175.

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2015). Bayesian inference for low-rank Ising networks. *Scientific Reports*, *5*, 9050.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*, 1436–1462.

Molenaar, P. C. M. (2003). State space techniques in structural equation modeling. Retrieved from http://bit.ly/2ssau1K, June 14th 2017.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, UK: Cambridge University Press.

Ravikumar, P., Wainwright, M. J., & Lafferty, J. D. (2010). High-dimensional Ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, *38*, 1287–1319.

Santos, H. J., Fried, E. I., Asafu-Adjei, J., & Ruiz, J. (2017). Network of perinatal depressive symptoms in Latinas: Relationship to stress-related and reproductive biomarkers. *Research in Nursing and Health*, 1–11. http://doi.org/10.1002/nur.21784

Rhemtulla, M., Fried, E. I., Aggen, S. H., Tuerlinckx, F., Kendler, K. S., & Borsboom, D. (2016). Network analysis of substance abuse and dependence symptoms. *Drug and Alcohol Dependence*, *161*, 230–237. http://doi.org/10.1016/j.drugalcdep.2016.02.005

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*. doi: 10.1038/srep05918

Van Borkulo, C. D., Boschloo, L., Borsboom, D., Penninx, B. W. J. H., Waldorp, L. J., &

Schoevers, R. A. (2015). Association of symptom network structure with the course of depression. *JAMA Psychiatry*, *72*, 1219.

Van Borkulo, C. D., Boschloo, L., Kossakowski, J., Tio, P., Schoevers, R. A., Borsboom, D., & Waldorp, L. J. (2016). Comparing network structures on three aspects: A permutation test. (Submitted for publication). doi: 10.13140/RG.2.2.29455.38569.

Table 1. Replication results of comparing the networks for the NCS-R and NSMHWB data. In addition to the metrics discussed by Forbes et al. (see their Table 2 for detailed explanations), the table reports Pearson correlations between network parameters in the two samples (all > .9), replication statistics for censored and uncensored relative importance networks as implemented in accordance with Robinaugh et al. (2014), and most central nodes for different centrality measures. See Table 1 of Forbes et al. (2017) for node abbreviations.

	Ising models		*Relative importance networks (censored)*		*Relative importance networks (uncensored)*		DAGs	
	NCS-R	NSMHWB	NCS-R	NSMHWB	NCS-R	NSMHWB	NCS-R	NSMHWB
Network characteristics[a]								
Nr. of edges (% of possible)	80 (52.3%)	79 (51.6%)	118 (38.6%)	124 (40.5%)	306 (100%)	306 (100%)	34 (22.2%)	33 21.6%)
Density (as in Forbes et al.)	1.08	1.17	0.13	0.12	0.06	0.06	N/A	N/A
Quality of replication								
Correlation between all edges	0.95		0.98		0.99		N/A	
Correlation for non-zero edges	0.97		0.98		0.99		N/A	
Jaccard index[b]	0.77		0.92		1.00		0.68	
% change in edge weights[a]	30.4%		8.3%		22.2%		N/A	
Replicated edges[a]	69 (86.3%)		116 (98.3%)		306 (100%)		27 (79.4%)	
Non-replicated edges[a]	11 (13.8%)		2 (1.7%)		0 (0%)		7 (20.6%)	
Edges unique to replication set[a]	10 (12.7%)		8 (6.5%)		0 (0%)		6 (18.2%)	
Node centrality correlations								
Strength/Outstrength/Outdegree	0.94		0.94		0.98		0.87	
Instrength/Indegree	N/A		0.76		N/A		0.62	
Closeness	0.76		N/A		0.98		1.00	
Betweenness	0.94		0.84		0.92		0.79	
Most central nodes[c]								
Strength/Outstrength/Outdegree	even	even	depr	depr	depr	depr	depr	depr;inte
Instrength/Indegree	N/A	N/A	inte	weig	tie (15 nodes)	mFat	tie (4 nodes)	irri
Closeness	depr	mFat	N/A	N/A	mFat	mSle	anxi	anxi
Betweenness	depr	even	ctrl	even	gFat	gFat	edge	depr
Rank-order correspondence[a]	Correlation (Kendall tau-b)	Matches in Rank-Order	Correlation (Kendall tau-b)	Matches in Rank-Order	Correlation (Kendall tau-b)	Matches in Rank-Order	Correlation (Kendall tau-b)	Matches in Rank-Order
Strength/Outstrength/Outdegree	0.69	3 (16.7%)	0.82	9 (50%)	0.8	4 (22.2%)	0.75	14 (77.8%)
Instrength/Indegree	N/A	N/A	0.39	2 (11.1%)	N/A	N/A	0.57	16 (88.9%)
Closeness	0.71	3 (16.7%)	N/A	N/A	0.87	6 (33.3%)	1.00	18 (100%)
Betweenness	0.77	11 (61.1%)	0.84	14 (77.8%)	0.57	9 (50%)	0.66	10 (55.6%)

[a] Computed following the methodology of Forbes et al. (2017).
[b] The Jaccard index is the proportion of shared edges relative to the total number of edges in both networks (shared and non-shared).
[c] Computed following the methodology of Forbes et al., but for single centrality measures.

Appendix A: Example codes

R-code to execute the NCS-R network analyses

This appendix describes the R-code used to run the analyses that give rise to the network models reported, using the NCS-R dataset. Codes to perform all analyses are present at https://osf.io/akywf/?view_only=a0634a2b063c4538abca59e5f18c2baf, with the exception of the NMSHWB data, which are unfortunately not public. To replicate the full analyses, one needs to purchase access to the data from the the Australian Bureau of Statistics.

Estimating network structures

Step 1: Load packages and data.

```
# Packages to load:
library("bootnet")
library("bnlearn")
library("qgraph")
library("relaimpo")
library("IsingFit")

# Set the random seed:
set.seed(123)

# Load NCS dataset:
DataNCS <- read.csv("NCS.csv", header = TRUE)

# Node labels:
Labels <- c("depr", "inte", "weig", "mSle", "moto", "mFat", "repr", "mCon",
            "suic", "anxi", "even", "ctrl", "edge", "gFat", "irri", "gCon",
            "musc", "gSle")

# Variables:
Vars <- names(DataNCS)
```

Step 2: Replicate Forbes et al.'s Ising model analyses. These analyses replicate Forbes et al. (2017) exactly.

```
res_Ising_NCS <- IsingFit(DataNCS)
Ising_NCS <- res_Ising_NCS$weiadj
```

Step 3: Replicate Forbes et al.'s DAG analyses. These analyses can replicate Forbes et al. (2017) exactly, although using certain R versions or operating systems might lead to slightly different results due to randomness in the bootstrap (even though set.seed is used).

E.g., we obtained 35 edges in the NCS DAG network often. For the commentary, we used a laptop that replicated the 34 edge network reported by Forbes et al using the codes below.

```
# Make data categorical:
DataNCScat <- DataNCS
for (i in 1:ncol(DataNCScat)){
  DataNCScat[,i] <- as.factor(DataNCScat[,i])
}

# Using codes from McNally, R. J., Mair, P., Mugno, B. L., & Riemann, B. C. (
2017). Co-morbid obsessive-compulsive disorder and depression: a Bayesian net
work approach. Psychological Medicine, 1-11.
set.seed(123)
bnlearnRes_NCS <- boot.strength(DataNCScat, R = 1000, algorithm = "hc", algor
ithm.args = list(restart = 5, perturb = 10), debug = TRUE)

# Edges with strength > 0.85:
DAG_NCS <- amat(averaged.network(bnlearnRes_NCS, threshold = 0.85))
```

Step 4: Run relative importance networks using Robinaugh et al.'s (2014) procedure, using normalized lmg. Two networks are generated, one in which no edges are censored (as analyzed by Robinaugh et al) and one in which edges below 0.05 are removed (as shown by Robinaugh et al).

```
# Empty matrix:
relimp_NCS_uncensored <- matrix(0, 18, 18)

# For every node, compute incomming relative importance (normalized lmg, as u
sed by Robinaugh et al (2014)):
# Robinaugh, D. J., LeBlanc, N. J., Vuletich, H. A., & McNally, R. J. (2014).
Network analysis of persistent complex bereavement disorder in conjugally ber
eaved adults. Journal of abnormal psychology, 123(3), 510-522.
for (i in 1:18){
  formula <- as.formula(paste0(Vars[i]," ~ ",paste0(Vars[-i],collapse=" + "))
)
  res <- calc.relimp(formula, DataNCS, rela = TRUE)
  relimp_NCS_uncensored[-i,i] <- res@lmg
}

# Censor (note, Robinaugh et al (2014) only hide edges under 0.05, they do no
t censor):
relimp_NCS_censored <- ifelse(relimp_NCS_uncensored < 0.05 ,0, relimp_NCS_unc
ensored)
```

Replicating errors in Forbes

Error 1: Relative Importance Networks These codes replicating the relative importance network error by Forbes et al and generate Figure 2

```
### Replicating Forbes et al' error:
# Compute relative importance networks using non-normalized lmg:
relimp_NCS_nonnormalized <- matrix(0,18,18)

for (i in 1:18){
  formula <- as.formula(paste0(Vars[i]," ~ ",paste0(Vars[-i],collapse=" + "))
)
  resNCS <- calc.relimp(formula, DataNCS, rela = FALSE)
  relimp_NCS_nonnormalized[-i,i] <- resNCS@lmg
}

# Threshold according to Forbes et al:
relimp_NCS_nonnormalized_censored <- ifelse(
  relimp_NCS_nonnormalized > 0.05 & # Retain edges above 0.05
    (relimp_NCS_nonnormalized > (t(relimp_NCS_nonnormalized) + 0.005)), # Ret
ain *only* edges that are 0.005 stronger than edge in transpose
  relimp_NCS_nonnormalized,0)

# Number of edges:
sum(relimp_NCS_nonnormalized_censored!=0)
# 31: same as reported by Forbes et al

# Figure 2:
g_robinaugh <- relimp_NCS_censored
g_nonNorm <- ifelse(relimp_NCS_nonnormalized > 0.05, relimp_NCS_nonnormalized
, 0)
g_Forbes <- relimp_NCS_nonnormalized_censored

# Layout:
L <- averageLayout(g_robinaugh, g_nonNorm, g_Forbes)

# Plot figure:
layout(t(1:3))
qgraph(g_robinaugh, layout = L, title = "Normalized lmg", labels = Labels,
       asize = 4, edge.color = "black", parallelEdge = TRUE)
box("figure")
qgraph(g_nonNorm, layout = L, title = "Non-normalized lmg", labels = Labels,
       asize = 4, edge.color = ifelse(g_nonNorm != g_Forbes, "red","black"),
       parallelEdge = TRUE)
box("figure")
qgraph(g_Forbes, layout = L, title = "Forbes et al", labels = Labels,
       asize = 4, edge.color = "black", parallelEdge = TRUE)
box("figure")
# Shape of curve and placement of nodes might differ using different qgraph v
ersions
```

Error 2: Implausible correlation matrix due to imputation method.

Establish correlation matrices using different methods for handling missing data. R gives a warning saying that the correlation matrix is not positive definite. We examine (a) the non-

positive definite tetrachoric correlations, and (b) the nearest positive definite matrix. This shows that the main problem lies in the imputation of zeroes and not in the fact that the nearest positive definite matrix is used.

```
# Association network as comuted by Forbes et al:
tetrachorNearPD_NCS <- cor_auto(DataNCS)

## Variables detected as ordinal: V1; V2; V3; V4; V5; V6; V7; V8; V9; V10; V1
1; V12; V13; V14; V15; V16; V17; V18

## Warning in cor_auto(DataNCS): Correlation matrix is not positive definite.
## Finding nearest positive definite matrix

# Gives a warning searching for nearest positive definite matrix. This can be
disabled:
tetrachor_NCS <- cor_auto(DataNCS,forcePD = FALSE)

## Variables detected as ordinal: V1; V2; V3; V4; V5; V6; V7; V8; V9; V10; V1
1; V12; V13; V14; V15; V16; V17; V18

# Both produce implausibly high correlations (these are all depression sympto
ms):
round(tetrachorNearPD_NCS[1:9,1:9],2)

##      V1   V2   V3   V4   V5   V6   V7   V8   V9
## V1 1.00 0.99 0.98 0.99 0.93 0.98 0.93 0.99 0.95
## V2 0.99 1.00 0.97 0.98 0.91 0.97 0.92 0.98 0.94
## V3 0.98 0.97 1.00 0.98 0.91 0.96 0.88 0.97 0.93
## V4 0.99 0.98 0.98 1.00 0.93 0.98 0.91 0.99 0.95
## V5 0.93 0.91 0.91 0.93 1.00 0.89 0.79 0.94 0.82
## V6 0.98 0.97 0.96 0.98 0.89 1.00 0.90 0.98 0.92
## V7 0.93 0.92 0.88 0.91 0.79 0.90 1.00 0.91 0.88
## V8 0.99 0.98 0.97 0.99 0.94 0.98 0.91 1.00 0.94
## V9 0.95 0.94 0.93 0.95 0.82 0.92 0.88 0.94 1.00

round(tetrachor_NCS[1:9,1:9],2)

##      V1   V2   V3   V4   V5   V6   V7   V8   V9
## V1 1.00 0.99 0.99 1.00 0.96 0.99 0.96 0.99 0.98
## V2 0.99 1.00 0.97 0.98 0.91 0.97 0.92 0.98 0.94
## V3 0.99 0.97 1.00 0.98 0.91 0.96 0.87 0.97 0.93
## V4 1.00 0.98 0.98 1.00 0.93 0.98 0.91 0.98 0.95
## V5 0.96 0.91 0.91 0.93 1.00 0.89 0.78 0.94 0.81
## V6 0.99 0.97 0.96 0.98 0.89 1.00 0.89 0.98 0.92
## V7 0.96 0.92 0.87 0.91 0.78 0.89 1.00 0.91 0.87
## V8 0.99 0.98 0.97 0.98 0.94 0.98 0.91 1.00 0.94
## V9 0.98 0.94 0.93 0.95 0.81 0.92 0.87 0.94 1.00
```

Stability Analyses

These codes use bootnet to establish stability assessments of the Ising model, such as reported in Appendix A. Shown here are results based on 100 bootstrap samples

```
network <- estimateNetwork(DataNCS, default="IsingFit")

# Bootstraps ran on 24-core supercomputer. Reduce nCores to present number of
cores
boota <- bootnet(network, nBoots = 5000, nCores = 20)
bootb <- bootnet(network, nBoots = 5000, type = "case",  nCores = 20, caseN =
25)

# Plot edge weight CI
plot(boota, labels = FALSE, order = "sample")

# Centrality stability
plot(bootb)

# Edge weights diff test
plot(boota, "edge", plot = "difference", onlyNonZero = TRUE, order = "sample"
)

# Centrality diff test
plot(boota, "strength", order="sample")

# Centrality stability coefficient
corStability(bootb)
```

Appendix B: Split-half comparisons

This appendix contains two tables that present summary results for the split-half comparisons for NCS-R and NSMHWB samples, after running the analyses with accurately implemented relative importance networks (see main text for details) on the same splits used by FWMK. In addition, these tables give general information about the relation between the models arrived at in both splits in addition to the results presented in Forbes et al. (2017): the correlation between edge weights, the correlation between non-zero edge weights, Jaccard index, and correlations across split halves for centrality measures. Matches in rank order across split-halves appear to have been assessed by hand in Forbes et al. (2017); we wrote an automated R-script to assess these across split-halves, which gives slightly different results.

Table B1. Summary of split-half comparisons for the NCS-R data. This table matches the analysis reported in Table 3 of Forbes et al. (2017). In addition to the metrics discussed by FMWK (see their Table 2 for detailed explanations), the table reports Pearson correlations between network parameters in the two samples (all > .9), and replication statistics for censored and uncensored relative importance networks as implemented in accordance with Robinaugh et al. (2014).

	Ising models		*Relative importance networks (censored)*		*Relative importance networks (uncensored)*		DAGs	
	First half	Second half	First half	Second half	First half	Second half	First half	Second half
Network characteristics								
Connectivity (% of possible)	46.7% (45.1-48.4)	47.1% (44.4-49.7)	38.6% (37.9-39.2)	38.6% (37.9-38.9)	100% (100-100)	100% (100-100)	17% (16.3-19)	17.3% (15.7-18.3)
Density (as in Forbes et al.)	1.14 (1.11-1.17)	1.12 (1.08-1.19)	0.13 (0.13-0.13)	0.13 (0.13-0.13)	0.06 (0.06-0.06)	0.06 (0.06-0.06)	N/A	N/A
Quality of replication								
Correlation between all edges	0.95 (0.93-0.97)		0.99 (0.99-0.99)		0.99 (0.99-1)		N/A	
Correlation for non-zero edges	0.96 (0.95-0.97)		0.99 (0.98-0.99)		0.99 (0.99-1)		N/A	
Jaccard index	0.76 (0.69-0.82)		0.98 (0.96-0.98)		1 (1-1)		0.61 (0.49-0.66)	
% change in edge weights	35.6% (27.7-41.8)		6.8% (5.5-8.5)		10.4% (7.9-15.4)		N/A	
% replicated edges	86% (81.9-91.4)		98.3% (96.7-100)		100% (100-100)		74% (64.3-80)	
% non-replicated edges	14% (8.6-18.1)		1.7% (0-3.3)		0 % (0-0)		26% (20-35.7)	
Edges unique to replication set	15% (8.8-18.1)		1.3% (0-2.5)		0 % (0-0)		27.7% (16-33.3)	
Node centrality correlations								
strength/outstrength/outdegree	0.97 (0.93-0.99)		0.99 (0.98-0.99)		0.99 (0.97-0.99)		0.89 (0.82-0.94)	
Instrength/Indegree	N/A		0.93 (0.89-0.95)		N/A		0.51 (0.29-0.66)	
Closeness	0.76 (0.43-0.89)		N/A		0.97 (0.85-0.98)		N/A	
Betweenness	0.82 (0.54-0.96)		0.99 (0.98-1)		0.73 (0.39-0.82)		0.87 (0.44-0.96)	
Rank-order correspondence	Correlation (Kendall tau-b)	Matches in Rank-Order	Correlation (Kendall tau-b)	Matches in Rank-Order	Correlation (Kendall tau-b)	Matches in Rank-Order	Correlation (Kendall tau-b)	Matches in Rank-Order
Strength/outstrength/outdegree	0.8 (0.63-0.87)	27.8% (5.6-50)	0.88 (0.87-0.93)	44.4% (33.3-55.6)	0.91 (0.87-0.96)	58.3% (33.3-72.2)	0.68 (0.53-0.76)	55.6% (44.4-72.2)
Instrength/Indegree	N/A	N/A	0.71 (0.54-0.79)	16.7% (5.6-38.9)	N/A	100% (All 1)	0.42 (0.18-0.61)	55.6% (11.1-72.2)
Closeness	0.58 (0.41-0.71)	19.4% (11.1-27.8)	N/A	100% (All 0)	0.84 (0.79-0.95)	47.2% (11.1-66.7)	N/A	100% (88.9-100)
Betweenness	0.61 (0.21-0.77)	50% (38.9-66.7)	0.87 (0.62-1)	100% (77.8-100)	0.63 (0.36-0.75)	44.4% (33.3-61.1)	0.46 (0.27-0.66)	61.1% (27.8-88.9)

Table B2. Summary of split-half comparisons for the NSMHWB data. This table matches the analysis reported in Table 4 of Forbes et al. (2017). In addition to the metrics discussed by FMWK (see their Table 2 for detailed explanations), the table reports Pearson correlations between network parameters in the two samples (all > .9), and replication statistics for censored and uncensored relative importance networks as implemented in accordance with Robinaugh et al. (2014).

	Ising models		*Relative importance networks (censored)*		*Relative importance networks (uncensored)*		DAGs	
	First half	Second half	First half	Second half	First half	Second half	First half	Second half
Network characteristics								
Connectivity (% of possible)	47.7% (43.1-48.4)	45.8% (43.1-48.4)	40.2% (39.5-41.5)	40.5% (39.2-41.8)	100% (100-100)	100% (100-100)	14.7% (12.4-15)	14.7% (13.7-18.3)
Density (as in Forbes et al.)	1.17 (1.14-1.25)	1.22 (1.12-1.33)	0.12 (0.12-0.12)	0.12 (0.12-0.12)	0.06 (0.06-0.06)	0.06 (0.06-0.06)	N/A)	N/A
Quality of replication								
Correlation between all edges	0.93 (0.92-0.96)		0.99 (0.99-0.99)		0.99 (0.99-0.99)		N/A	
Correlation for non-zero edges	0.95 (0.93-0.97)		0.98 (0.97-0.99)		0.99 (0.99-0.99)		N/A	
Jaccard index	0.74 (0.68-0.77)		0.96 (0.94-0.98)		1 (1-1)		0.47 (0.41-0.55)	
% change in edge weights	48.4% (36.8-68.7)		7.1% (5.9-8.3)		9.6% (8.3-12.3)		N/A	
% replicated edges	83.4% (78.1-89.4)		98.4% (94.5-100)		100% (100-100)		68.2% (56.5-73.7)	
% non-replicated edges	16.6% (10.6-21.9)		1.6% (0-5.5)		0% (0-0)		31.8% (26.3-43.5)	
Edges unique to replication set	13% (11.9-16.9)		2.4% (0-5.5)		0% (0-0)		37.8% (27.3-48.1)	
Node centrality correlations								
strength/outstrength/outdegree	0.98 (0.96-0.99)		0.99 (0.98-0.99)		0.99 (0.99-0.99)		0.82 (0.62-0.94)	
Instrength/Indegree	N/A		0.87 (0.79-0.95)		N/A		0.39 (0.01-0.78)	
Closeness	0.74 (0.59-0.91)		N/A		0.97 (0.92-0.98)		N/A	
Betweenness	0.73 (0.47-0.89)		0.95 (0.82-0.99)		0.44 (0.06-0.79)		0.82 (0.03-0.93)	
Rank-order correspondence	Correlation (Kendall tau-b)	Matches in Rank-Order	Correlation (Kendall tau-b)	Matches in Rank-Order	Correlation (Kendall tau-b)	Matches in Rank-Order	Correlation (Kendall tau-b)	Matches in Rank-Order
Strength/outstrength/outdegree	0.78 (0.61-0.84)	33.3% (16.7-44.4)	0.9 (0.84-0.95)	38.9% (27.8-55.6)	0.9 (0.86-0.93)	50% (33.3-66.7)	0.61 (0.36-0.79)	66.7% (50-77.8)
Instrength/Indegree	0.58 (0.39-0.8)	16.7% (0-27.8)	0.59 (0.4-0.76)	27.8% (11.1-44.4)	N/A	100% (All 1)	0.38 (0.12-0.76)	44.4% (11.1-72.2)
Closeness	N/A	100% (All 0)	0.78 (0.7-0.87)	86.1% (66.7-	0.83 (0.69-0.87)	27.8% (22.2-50)	N/A	100% (100-100)
Betweenness	0.57 (0.44-	55.6% (38.9-			0.42 (0.26-	22.2% (11.1-	0.51 (0.16-	66.7% (44.4-

	0.81)	72.2)	100)	0.76)	44.4)	0.68)	83.3)

Appendix C: Stability & Accuracy Analyses

This document contains the results of the bootstrapping pipeline using the R-package *bootnet* explained in detail in:

Epskamp, S., Borsboom, D., & Fried, E. I. (2017). Estimating Psychological Networks and their Accuracy: A Tutorial Paper. *Behavior Research Methods*, 1–34. DOI 10.3758/s13428-017-0862-1.

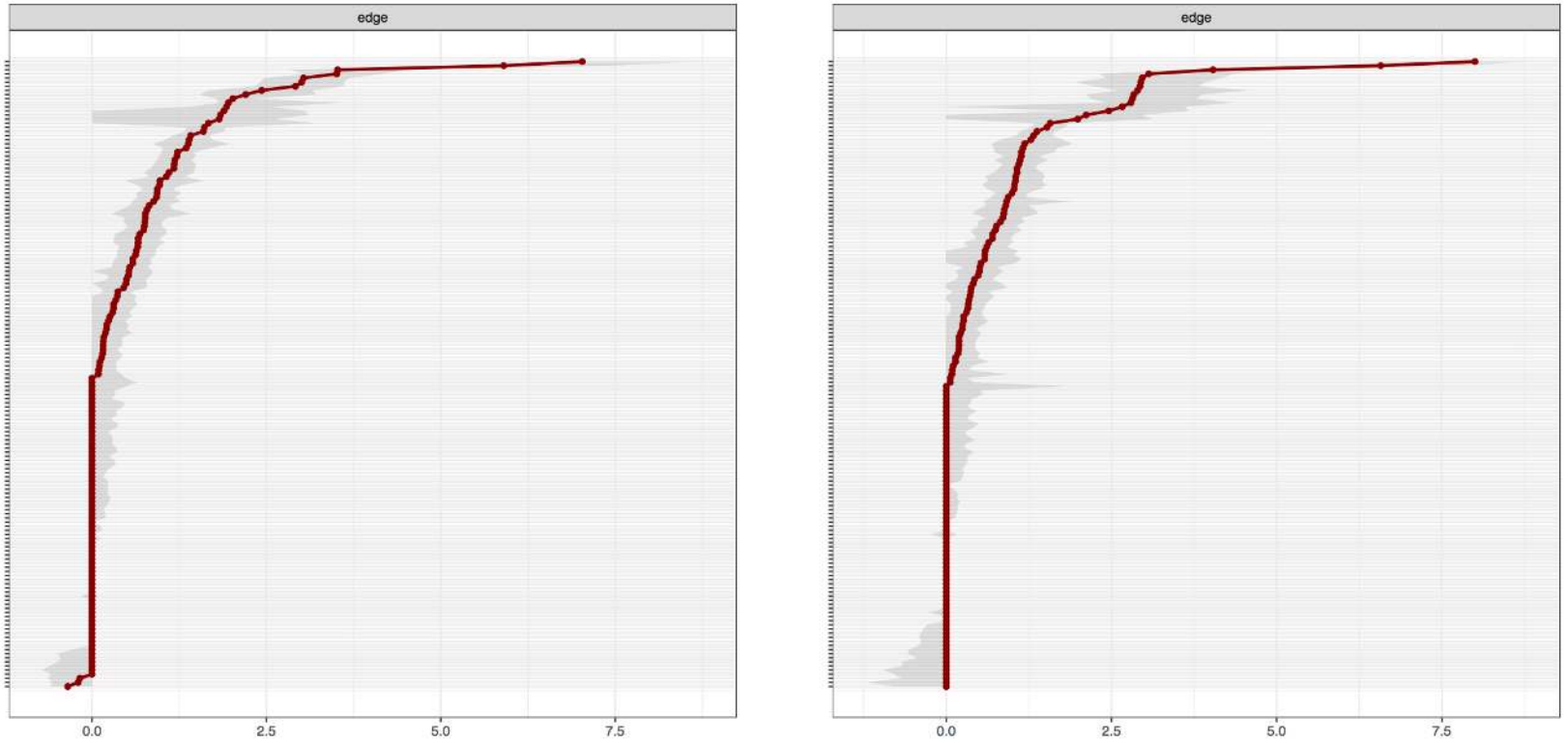

Figure S1. Bootstrapped edge-weights. The red line depicts point estimates of the edge weights, the grey bar 95% confidence intervals. Left: Ising Model estimated on NCS data. Right: Ising Model estimated on NSMHWB data.

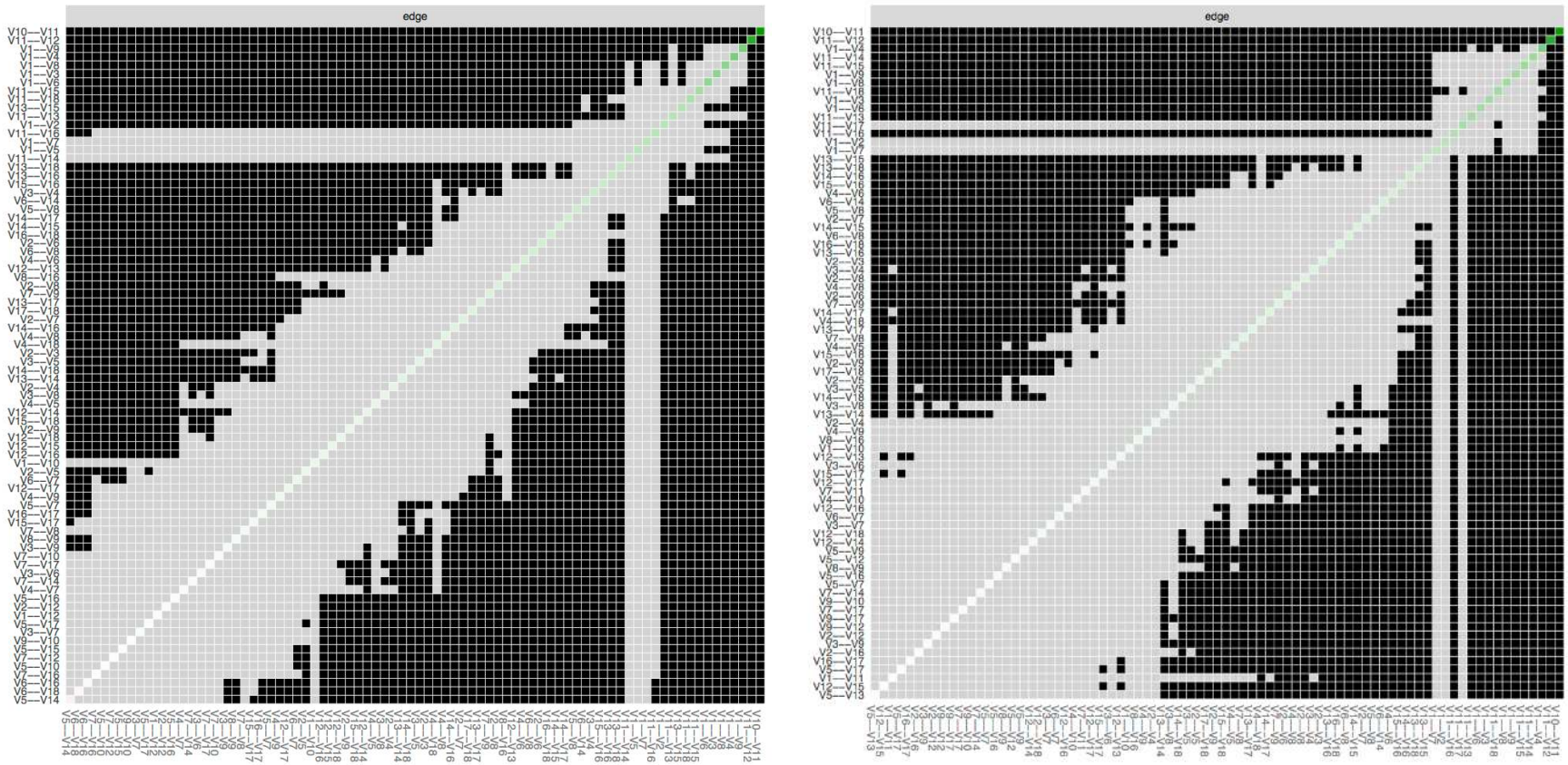

Figure S2. Bootstrapped significance (α = 0.05) between edges. Each row and column indicates an edge. Black boxes represent significant differences and gray boxes represent non-significant differences. The color in the diagonal corresponds with the edge colors in the original network figures. Left: Ising Model estimated on NCS data. Right: Ising Model estimated on NSMHWB data.

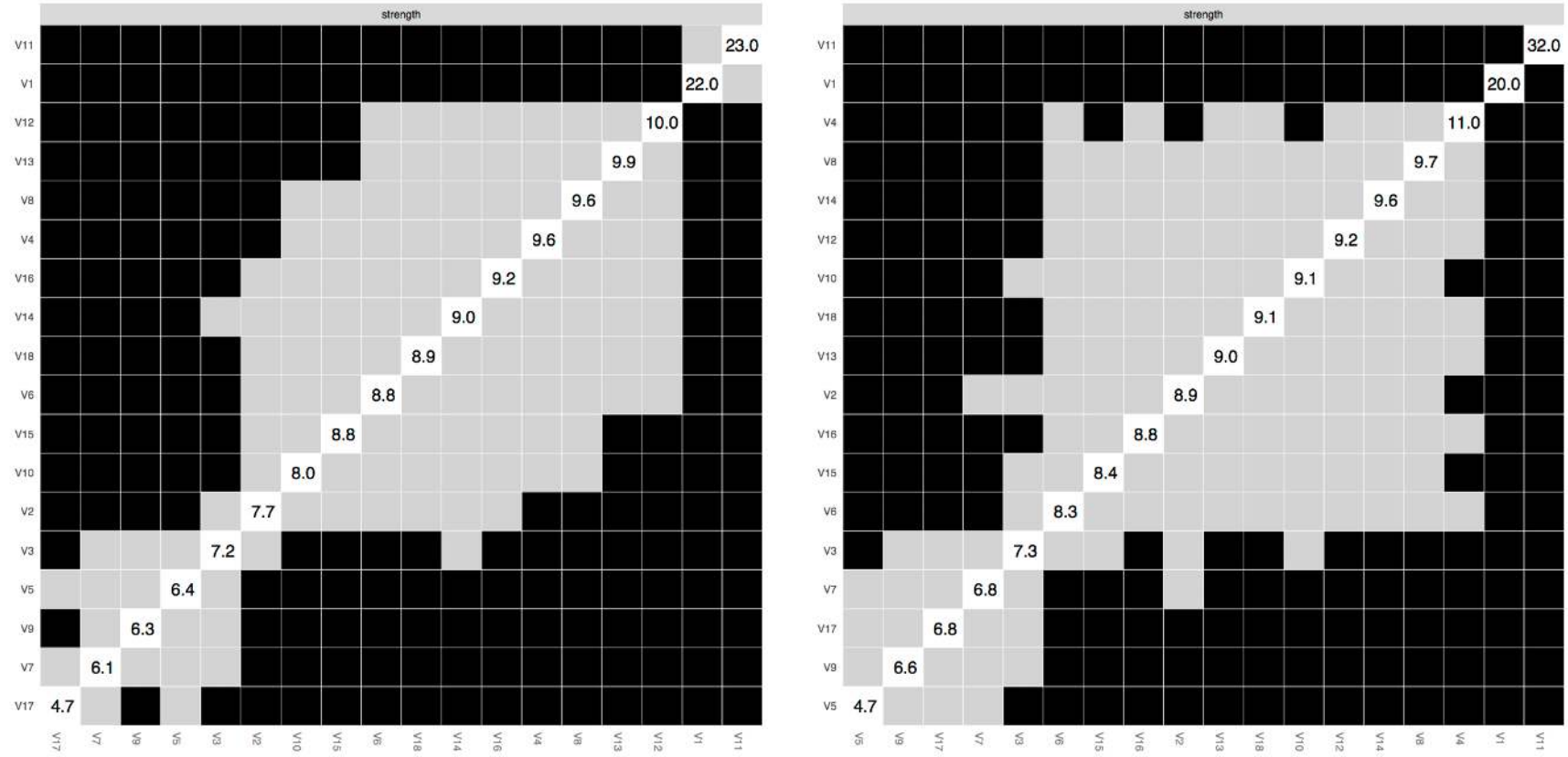

Figure S3. Bootstrapped significance (α = 0.05) between strength centrality metric of the networks. Each row and column indicates a node. Black boxes represent significant differences and gray boxes represent non-significant differences. The value in the diagonal corresponds with the strength of a node. Left: Ising Model estimated on NCS data. Right: Ising Model estimated on NSMHWB data.

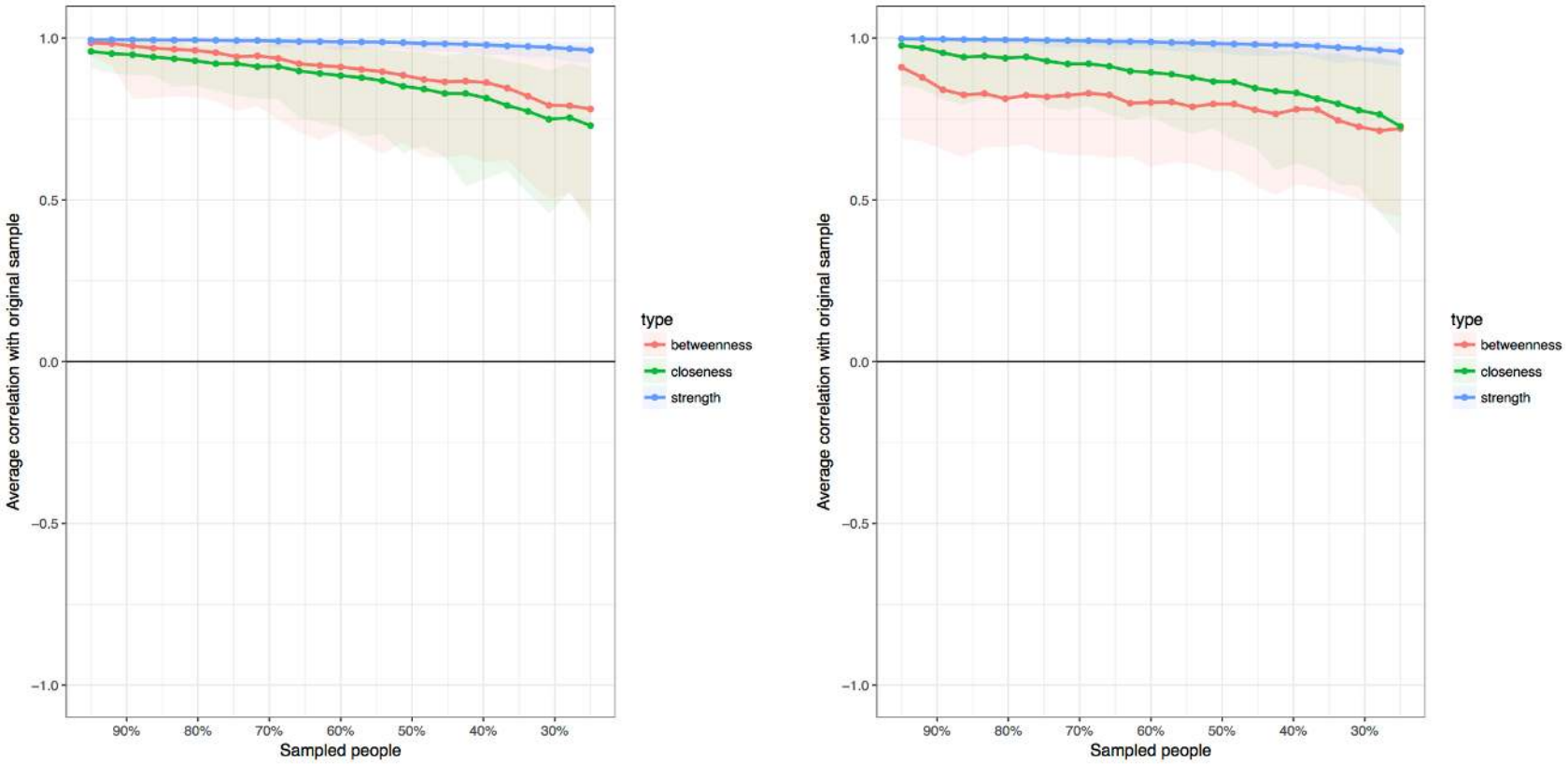

Figure S4. The correlation between the original centrality index and the centrality index after dropping a percentage of subjects at random from the data. Left: Ising Model estimated on NCS data. Right: Ising Model estimated on NSMHWB data. Stability centrality coefficients (i.e. % of cases that can be dropped to retain with 95% certainty a correlation of 0.7 of centrality between network estimated on original data and network estimated on subsampled data): Betweenness NCS=0.49, NSMHWB=0; Closeness NCS=0.46, NSMHWB=0.55; Strength NCS=0.75, NSMHWB=0.75.

Explanation of diverging betweenness centrality results across the two datasets

There was a strong difference between the CS coefficients for betweenness (cf. Figure S4): 0.49 for the Ising Model estimated in the NCS data, but 0 for the Ising Model estimated in the NSMHWB data. Given the similarity of the datasets and networks, this is surprising. We further investigated this difference. Below is a figure showing betweenness of all sampled datasets with only 5% of the cases dropped:

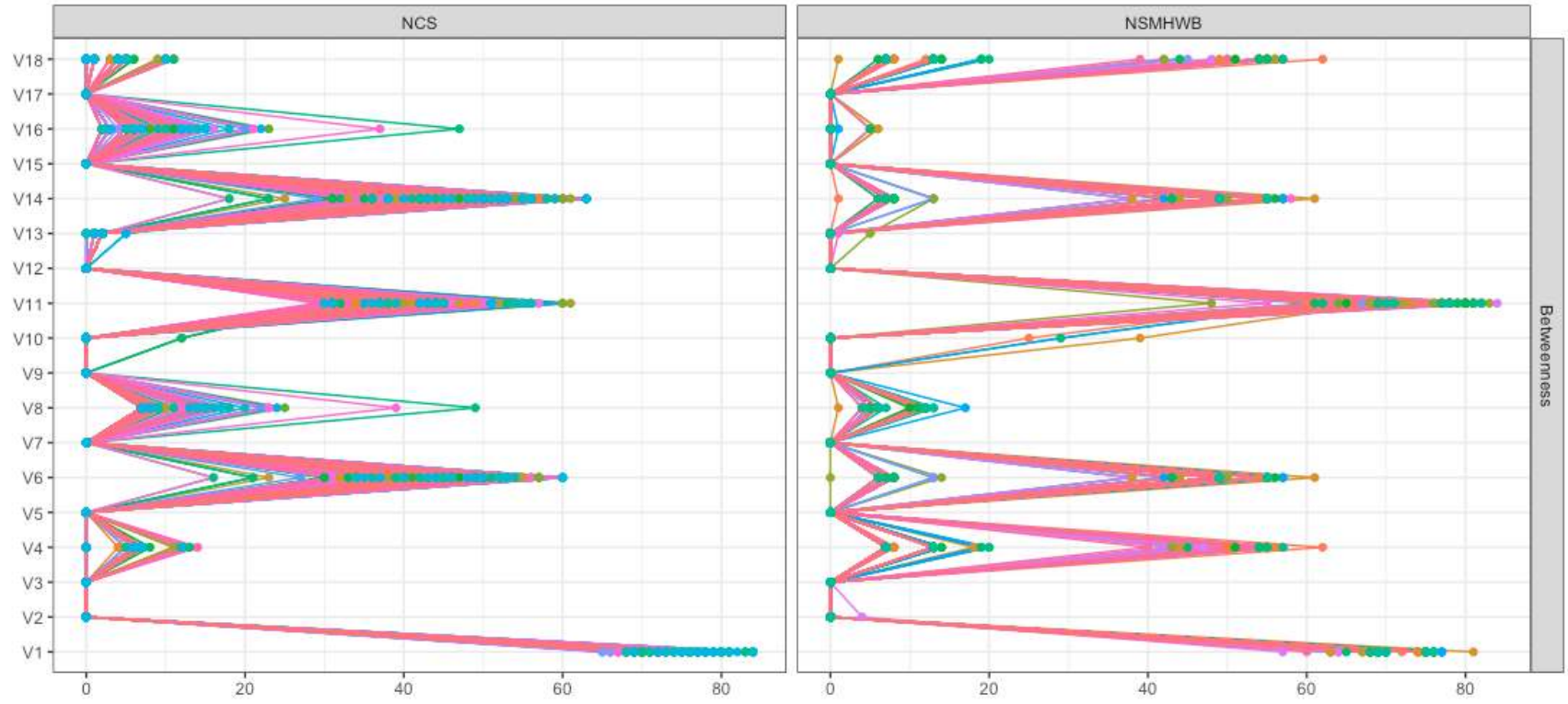

Figure S5. Investigation of differences in betweenness CS coefficients across the Ising Models in NCS and NSMHWB data.

We can see that the NSMHWB dataset showed some bifurcations in betweenness centrality estimates, whereas the NCS does not. This is especially pronounced in nodes 4, 6, 14, and 8, which are the bridge symptoms that connect anxiety and depression symptom clusters. In NCS, the edge 6—14 is slightly stronger than 4—18, leading shortest paths between the two clusters (on which betweenness centrality is estimated) to more consistently (irrespective of the particular participants included in the sample) go through nodes 6 and 14. In the NSMHWB dataset, however, 6—14 and 4—8 are nearly identical, and slight variations due to sampling (i.e. bootstrapping) lead to the shortest paths between clusters go through one of the two edges, resulting in high betweenness for the pair of nodes through which all shortest paths go—e.g. 6 and 14—and low betweenness for the other two nodes, e.g. 4 and 18. This leads to a betweenness CS coefficient of 0 because the shortest path differs strongly with very small fluctuations of participants in the data.