American Psychological Association

Journal of Abnormal Psychology

© 2019 American Psychological Association
ISSN: 0021-843X

http://dx.doi.org/10.1037/abn0000496

The Replicability and Generalizability of Internalizing Symptom Networks Across Five Samples

Carter J. Funkhouser and Kelly A. Correa
University of Illinois at Chicago and Northwestern University

Stephanie M. Gorka
University of Illinois at Chicago

Brady D. Nelson
Stony Brook University

K. Luan Phan
The Ohio State University

Stewart A. Shankman
University of Illinois at Chicago and Northwestern University

The popularity of network analysis in psychopathology research has increased exponentially in recent years. Yet, little research has examined the replicability of cross-sectional psychopathology network models, and those that have used single items for symptoms rather than multiitem scales. The present study therefore examined the replicability and generalizability of regularized partial correlation networks of internalizing symptoms within and across 5 samples (total N = 2,573) using the Inventory for Depression and Anxiety Symptoms, a factor analytically derived measure of individual internalizing symptoms. As different metrics may yield different conclusions about the replicability of network parameters, we examined both global and specific metrics of similarity between networks. Correlations within and between nonclinical samples suggested considerable global similarities in network structure (r_ss = .53–.87) and centrality strength (r_ss = .37–.86), but weaker similarities in network structure (r_ss = .36–.66) and centrality (r_ss = .04–.54) between clinical and nonclinical samples. Global strength (i.e., connectivity) did not significantly differ across all 5 networks and few edges (0–5.5%) significantly differed between networks. Specific metrics of similarity indicated that, on average, approximately 80% of edges were consistently estimated within and between all 5 samples. The most central symptom (i.e., dysphoria) was consistent within and across samples, but there were few other matches in centrality rank-order. In sum, there were considerable similarities in network structure, the presence and sign of individual edges, and the most central symptom within and across internalizing symptom networks estimated from nonclinical samples, but global metrics suggested network structure and symptom centrality had weak to moderate generalizability from nonclinical to clinical samples.

General Scientific Summary
There were considerable similarities in network structure, the presence and sign of individual edges, and the most central symptom within and across internalizing symptom networks estimated from undergraduate or community samples. Although the presence and sign of individual edges and the most central symptom were similarly consistent between nonclinical samples and a clinical sample, the generalizability of global network characteristics was generally weak to moderate.

Keywords: network, replicability, internalizing, depression, anxiety

Supplemental materials: http://dx.doi.org/10.1037/abn0000496.supp

Carter J. Funkhouser and Kelly A. Correa, Department of Psychology, University of Illinois at Chicago, and Department of Psychiatry and Behavioral Sciences, Northwestern University; Stephanie M. Gorka, Department of Psychiatry, University of Illinois at Chicago; Brady D. Nelson, Department of Psychology, Stony Brook University; K. Luan Phan, Department of Psychiatry and Behavioral Health, The Ohio State University; Stewart A. Shankman, Department of Psychology and Department of Psychiatry, University of Illinois at Chicago, and Department of Psychiatry and Behavioral Sciences, Northwestern University.

Ethical approval for this study was obtained from the UIC Office for the Protection of Research Subjects (Protocols 2012-0646 and 2005-0626). This work was supported by National Institute of Mental Health Grants R01 MH098093 (PI: Stewart A. Shankman) and R01 MH101497 (PI: K. Luan Phan).

Correspondence concerning this article should be addressed to Stewart A. Shankman, Department of Psychiatry and Behavioral Sciences, Northwestern University, North Lake Shore Drive, Chicago, IL 60611. E-mail: stew.shankman@northwestern.edu

A psychiatric disorder has traditionally been conceptualized as a set of co-occurring symptoms that are the result of a "common cause" or underlying latent variable. The network theory of psychopathology conceptually contrasts with this perspective, as it proposes that symptoms *causally interact* in dynamic networks (Borsboom, 2017). For example, in contrast to the common cause model's assertion that insomnia, fatigue, and concentration difficulties share a common underlying mechanism, the network approach to psychopathology suggests that insomnia may cause fatigue, which causes concentration difficulties, and so forth.

In psychopathology research, the network approach has often been examined using pairwise Markov random field (PMRF) models that estimate pairwise relationships between symptoms after statistically controlling for all other symptoms in the network (e.g., partial correlation coefficients). PMRF models of cross-sectional data may highlight direct and potentially causal bivariate relationships among symptoms (Beard et al., 2016; Epskamp & Fried, 2018) as well as elucidate mechanisms of comorbidity (Cramer, Waldorp, van der Maas, & Borsboom, 2010). The resulting model estimates can be visualized as a network of nodes (e.g., symptoms) that are connected by edges (i.e., pairwise relationships between nodes). PMRF networks can also reveal which symptoms are most *central* (i.e., most strongly connected to other symptoms). Centrality measures in undirected networks do not provide information about the direction of relationships. It is possible that a central symptom leads to other symptoms, in which case the central symptom may be involved in the onset and/or maintenance of other symptoms and represent a viable target for therapeutic intervention (e.g., Beard et al., 2016; Fried et al., 2017).[1] However, it is also possible that other symptoms lead to the central symptom, in which case intervening on the central symptom would have no effect on other symptoms.

Replicability and Generalizability of Cross-Sectional Psychopathology Networks

Simulation studies have examined the methodological validity of network analytic methods by investigating the range of conditions in which the models adequately converge on the "true" network structure (e.g., van Borkulo et al., 2014). However, determining the extent to which PMRF models of psychopathology replicate and generalize requires the comparison of network structures across samples (Borsboom, Robinaugh, Rhemtulla, & Cramer, 2018; Forbes, Wright, Markon, & Krueger, 2017a; Fried & Cramer, 2017). The few studies that have examined the replicability and generalizability of psychopathology networks have sparked debate as to which metrics are most appropriate for assessing similarities between samples. Some have argued that network interpretations and clinical implications focus on the presence, sign, and strength of specific edges and the rank-order of node centrality (i.e., which symptom is most central, second most central, third most central, etc.). An examination of these characteristics within and across two samples of major depressive disorder (MDD) and generalized anxiety disorder (GAD) symptoms revealed that 83.4–86.6% of edges replicated within and between networks but only 16.7–55.6% of individual nodes' centrality rank-order matched within and between samples (Forbes et al., 2017a), leading to the conclusion that "popular network analysis methods produce unreliable results" (Forbes et al., 2017a, p. 969; see also Forbes, Wright, Markon, & Krueger, 2019).[2]

Others have critiqued these specific metrics and argued that network replicability should be assessed using more global metrics such as the "coefficient of similarity" (i.e., spearman correlation between edge lists; Borsboom et al., 2017; Fried et al., 2018), the Network Comparison Test (NCT; van Borkulo, Epskamp, & Millner, 2016), and Spearman correlations of centrality indices between networks (Borsboom et al., 2017). Reanalyzing Forbes, Wright, Markon, and Krueger's (2017a) data using these metrics, Borsboom et al. (2017) came to the opposite conclusion as Forbes et al. (2017a)—specifically, that the networks under consideration were "highly similar" (Borsboom et al., 2017, p. 990). Three recent studies examining the replicability and generalizability of posttraumatic stress disorder (PTSD) symptom networks similarly found that coefficients of similarity and correlations of centrality strength (i.e., the sum of all absolute edge weights connected to a node) replicated strongly. However, NCTs identified some differences in structure and global strength between networks (Benfer et al., 2018; Fried et al., 2018; Knefel et al., 2019). Additionally, two studies examining global network strength as a potential prognostic indicator of depression treatment response came to different conclusions (although the effects were of comparable size and in the same direction; Schweren, van Borkulo, Fried, & Goodyer, 2018; van Borkulo et al., 2015). In sum, the extent to which psychopathology networks replicate and generalize across samples remains unclear.

Limitations of Prior Psychopathology Network Replicability and Generalizability Studies

The existing research on the replicability of psychopathology networks (and the psychopathology network literature more broadly) has several noteworthy limitations. First, many network analyses (e.g., Forbes et al., 2017a) used zero-imputation to account for a skip structure in their data (e.g., if a participant did not endorse the cardinal symptom(s) of a disorder [e.g., depressed mood or loss of interest for MDD], the remaining symptoms of that disorder were not assessed). Zero-imputation is problematic because it can substantially alter the correlation matrices upon which networks are based (Borsboom et al., 2017). This alteration of correlation matrices is likely to increase the observed replicability because it alters the correlation matrices the same way in both samples, and likely contributed to the unexpectedly high replicability found in the previously mentioned study of MDD and GAD symptom networks (Borsboom et al., 2017; Forbes et al., 2017a). Second, many studies have assessed symptoms (i.e., nodes) using

[1] This centrality hypothesis is not an implication of network theory and, although it is popular in the psychopathology network literature, the utility of centrality measures for identifying intervention targets in psychopathology networks has been questioned (Bringmann et al., 2019; Dablander & Hinne, 2019).

[2] Steinley, Hoffman, Brusco, and Sher (2017) provided a commentary on Forbes et al.'s (2017a) article in which they introduced a new method termed fixed-margin sampling for examining whether network model results differ from what would be expected by chance. However, because the appropriateness of this method as a measure of network replicability has come into question (Epskamp, Fried, et al., 2018), we do not discuss this technique in further detail.

single items from a questionnaire or interview, which tend to be less reliable than aggregations of multiple items (Cicchetti & Prusoff, 1983). Third, the few studies that have examined psychopathology network replicability examined replicability within one (e.g., PTSD) or two *DSM* disorders.

The Present Study

To address the limitations of prior psychopathology network replicability studies, the present study aimed to examine the replicability and generalizability of regularized partial correlation networks of internalizing symptoms across five samples (two undergraduate samples, two community samples, and a clinical sample). The Inventory of Depression and Anxiety Symptoms (IDAS; Watson et al., 2007) was utilized to assess symptoms because it (a) does not contain skip-outs, (b) assesses each symptom using multiple items, and (c) assesses multiple nonoverlapping domains of internalizing psychopathology.

Due to the debate regarding which metrics are most suitable for assessing replicability (Borsboom et al., 2017; Forbes et al., 2017a; Forbes, Wright, Markon, & Krueger, 2017b), our analyses were agnostic and examined (a) the more global indices (e.g., the coefficient of similarity, the three permutation tests implemented by the NCT, Spearman correlations of centrality indices) advocated for by some (Borsboom et al., 2017; van Borkulo et al., 2016) as well as (b) the more specific metrics (e.g., the presence/absence and sign of individual edges and matches in centrality rank-order) advocated for by others (Forbes et al., 2017a, 2017b). This study was primarily exploratory in nature, but our hypotheses were twofold. First, we hypothesized that more global metrics would be more replicable and generalizable than more detailed metrics, as found by Forbes et al. (2017a, 2017b) and Borsboom et al. (2017). Second, we expected metrics to be more similar between networks estimated from random split-halves of the same sample or from two samples from the same population (e.g., between the two undergraduate networks and between the two community networks), and less similar between samples from different populations (e.g., between an undergraduate network and a clinical network).

Method

Participants

Participants (total N = 2,573)[3] were taken from five samples recruited in the United States. Demographic characteristics of each sample are provided in Table S1 in the online supplementary material. The first sample (henceforth called “Undergraduate Sample 1”) consisted of 1,176 unselected undergraduates attending a large, suburban university in the northeast (Distefano et al., 2018; Nelson & Hajcak, 2017). The second sample (“Undergraduate Sample 2”) included 578 unselected[4] undergraduates attending a large, urban university in the midwest (Altman, Campbell, Nelson, Faust, & Shankman, 2013; Gorka et al., 2013). Participants in the two undergraduate samples were recruited through the psychology department subject pools and received course credit for their participation.

The third (“Community Sample 1”) and fourth (“Community Sample 2”) samples consisted of participants recruited from the community to participate in a family study of neurophysiological vulnerability factors for internalizing psychopathology (Correa, Liu, & Shankman, 2019; Funkhouser, Correa, Carrillo, Klemballa, & Shankman, 2019; Shankman et al., 2018). There were no diagnosis-based inclusion criteria; however, participants were required to be between the ages of 18 and 30 and have at least one full biological sibling who was eligible to participate. More detailed information regarding the study’s inclusion and exclusion criteria and procedures is available elsewhere (Shankman et al., 2018). To avoid nonindependence of observations due to the presence of sibling pairs, we randomly assigned one sibling from each sibling pair to Community Sample 1 and assigned the other sibling to Community Sample 2. Participants who did not participate with a sibling (n = 97) were randomly allocated to one of the two community samples, resulting in two samples each consisting of biologically unrelated individuals (n = 277 and 276, respectively). As these two community samples were taken from a single study, they should be considered split-halves of one sample rather than two independent samples. However, we refer to them as separate samples to ease interpretation.

The fifth sample (“Clinical Sample”) consisted of 266 adults aged 18–65 who were either seeking treatment for a range of internalizing problems (n = 133; Lieberman, Gorka, Funkhouser, Shankman, & Phan, 2017) or recruited for a study examining psychophysiological processes in individuals who met *DSM–IV* diagnostic criteria for MDD and/or panic disorder (PD; n = 133; Shankman et al., 2013).[5] Data from treatment-seekers used in the present study were collected prior to treatment. Further details regarding inclusion and exclusion criteria and study procedures are described elsewhere (Lieberman, Gorka, DiGangi, Frederick, & Phan, 2017; Shankman et al., 2013). Although treatment-seekers differed from those recruited based on MDD and/or PD diagnosis in their severity of several IDAS scales, we combined them to maximize power for network estimation in light of recommendations that sample sizes be at least three times the number of estimated parameters (e.g., Fried & Cramer, 2017).

Assessment of Internalizing Symptoms

The IDAS was used to assess internalizing symptoms in all five samples. The IDAS includes 11 factor-analytically derived subscales[6] representing empirically distinct symptoms of internalizing psychopathology over the past 2 weeks. The well-being subscale was reverse coded and renamed “low well-being” so that higher scores indicate greater symptom severity for all subscales. Internal

[3] Four participants (0.2% of the total sample) had incomplete data and were excluded from all analyses to simplify the reporting of results. The total N of 2,573 represents the number of participants who had complete data and were included in the network analyses.

[4] The vast majority of participants in Undergraduate Sample 2 were unselected, but 109 (18.8%) participants were selected to be female with either high or low OCD or bulimia tendencies (Altman et al., 2013).

[5] Both studies also recruited a group of healthy controls, but these individuals were excluded so that the fifth sample could be an exclusively clinical sample.

[6] The IDAS also includes a 12th subscale entitled general depression, but this scale was not included in the present study because it contains items from several other subscales and thus is an aggregation of multiple symptoms. Including this subscale would also distort the covariance matrices due to conditioning on a collider (i.e., Berkson’s Bias; de Ron et al., 2019).

consistencies of the 11 subscales were adequate across all five samples and are presented in Table S2.

Data Analyses

Network estimation. A nonparanormal transformation was applied to all data sets prior to network estimation to reduce the assumption of normality (Liu, Lafferty, Wasserman, & Wainwright, 2009), in line with recommendations for estimating networks on continuous, non-normal data (Epskamp, Borsboom, & Fried, 2018; Epskamp & Fried, 2018).[7] We then individually[8] estimated partial correlation networks based on Pearson correlations for each sample. Networks were regularized using the "graphical least absolute shrinkage and selection operator" (GLASSO) with the extended Bayes information criterion (EBIC; $\gamma = 0.5$; Foygel & Drton, 2010) to identify edges that are likely to be spurious (i.e., close to zero) and shrink their edge weights to exactly zero. GLASSO network estimation and visualization were performed using the R package *qgraph* (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012). GLASSO is widely used in psychopathology network analysis of continuous data and has been thought to result in high specificity, meaning that it should not estimate edges that are not in the "true" network (Epskamp & Fried, 2018). However, recent work suggests GLASSO may have poorer specificity when the network is dense and many edges are near zero (Williams & Rast, 2019), which may, in turn, affect replicability. As suggested by a reviewer, we also used the *mgm* package to estimate networks using nodewise regression, which may have higher specificity than GLASSO (Williams & Rast, 2019). Identical layouts (i.e., the average of the Fruchterman-Reingold algorithm layouts for each GLASSO network) and the maximum value of the edge weights were imposed on the network plots to facilitate visual comparisons across networks and estimation methods.

Although previous studies have often examined symptom centrality using *strength* (i.e., the sum of all absolute edge weights connected to a node), *closeness* (i.e., the inverse of the summed length of all shortest edges between a node and all other nodes), and *betweenness* (i.e., the number of times a node lies on the shortest connecting edge between two other nodes), the latter two indicators have relatively weaker conceptual interpretability in the context of symptom networks (Bringmann et al., 2019; Forbes et al., 2017a) and have exhibited poor reliability (Beard et al., 2016; Epskamp, Borsboom et al., 2018). We therefore focus on centrality strength and report closeness and betweenness in the online supplementary materials. Lastly, we used the R package *mgm* to calculate symptom *predictability* (i.e., the amount of variance in a node explained by neighboring nodes, or R^2; Haslbeck & Fried, 2017).

Accuracy and stability of network parameters. We investigated the accuracy and stability of network parameters using two bootstrapping approaches as implemented by the R package *bootnet* (Epskamp, Borsboom, et al., 2018). First, we computed 95% confidence intervals (CIs) around edge weights via nonparametric bootstrapping using 1,000 iterations. These CIs provide a measure of edge weight accuracy. Second, we used case-drop bootstrapping to estimate correlation stability (conditional stimulus [CS]) coefficients to determine the stability of the rank-order of centrality indices. Simulation studies indicate that CS-coefficients above .25 imply moderate stability and CS-coefficients above .50 reflect strong stability (Epskamp, Borsboom, et al., 2018). Edge weights difference tests and centrality difference tests were calculated to test whether edges or node centralities significantly differ (see Epskamp, Borsboom, et al., 2018 for further description of these methods), and results are reported in Figures S1 and S2 in the online supplementary materials.

Within-sample comparisons. We examined the within-sample consistency of the two undergraduate samples by randomly dividing them each into two equally sized split-halves, applying the nonparanormal transformation to both split-halves, estimating a network for each split-half, calculating global network characteristics for each network and similarities between the two networks, and repeating this process 10 times (Forbes et al., 2017a). We did not examine split-halves of the two community samples or the clinical sample, as the split-halves likely would not have had sufficient power for network estimation (Fried & Cramer, 2017), and Community Samples 1 and 2 were essentially split-half samples to begin with. Global network characteristics extracted from each split-half network included global strength, the number of nonzero, positive, and negative edges, mean node predictability, and the most central node.

Global metrics. Global metrics of similarity between split-halves included the coefficient of similarity, spearman correlations of centrality indices between split-halves, and three permutation tests implemented by the NCT using 5,000 iterations. First, the *omnibus test of network structure invariance* tested whether the overall structures (i.e., matrices of the edge weights) of the two networks being compared were identical. Second, the *global strength invariance test* examined whether global strength estimates (i.e., the sum of absolute edge weights) significantly differed between networks. Third, the *individual edge invariance test* quantified the number of edges that differed between networks by testing the null hypothesis that each edge was identical between networks (using the Holm-Bonferroni method to correct for multiple comparisons).

Specific metrics. Specific network characteristics examined included the percentage of edges that were consistently present and absent across split-halves and matches in rank-order centrality as calculated by Forbes et al. (2017a) and Borsboom et al. (2017) to allow for multiple simultaneous ranks in the case of ties. Some psychopathology network analyses have interpreted the strongest edges (e.g., Beard et al., 2016) and a recent study examined the replicability of the strongest and most stable edges, defined as those with bootstrapped 95% CIs that did not include zero (Forbes et al., 2019). However, bootstrapped CIs do not test whether edges significantly differ from zero (Epskamp, Borsboom, et al., 2018) and thus are not an optimal measure for identifying the strongest and most stable edges in a network. To facilitate cross-study comparison, we therefore included edges with CIs excluding zero in the tables summarizing the results, but do not otherwise discuss them.

[7] Other recommended approaches for handling non-normal data include estimating networks derived from polychoric or Spearman correlations (Epskamp & Fried, 2018). In the present study, networks derived from polychoric or Spearman correlations were highly similar to networks estimated using the nonparanormal transformation ($rs > .88$).

[8] Although a recently developed extension of the graphical lasso called the fused graphical lasso allows for the joint estimation of multiple partial correlation networks, we report results of individually estimated networks because individually and jointly estimated networks were nearly identical ($rs > .98$).

In the pairwise comparisons examining whether edges were consistently present or absent across split-halves, we focused on whether edges present in the *sparser* network (i.e., network with fewer edges) were also present in the *denser* network because a denser network should not omit edges estimated in a sparser network if networks are estimated with high specificity (Borsboom et al., 2017). Although the specificity of GLASSO has been questioned, nodewise regression has demonstrated high specificity (Williams & Rast, 2019). Similarly, we examined whether edges that were absent in the *denser* network were also absent in the *sparser* network because a sparser network should not estimate edges that are absent in a denser network. This nesting approach accounts for differences in sparsity across networks due to different sample sizes. Rather than reporting the results of each of the 10 iterations, we report the median and range of the global network characteristics and between-network comparisons across the 10 split-halves for each of the two undergraduate samples.

Cross-network comparisons. The same metrics that were used to examine consistency between random split-halves of the two undergraduate samples were used to make cross-network comparisons. For specific metrics, we first investigated metrics across all five networks (e.g., how many edges were consistently estimated across all five networks?) and then between each pair of networks (e.g., how many of the edges estimated in Network A were also estimated in Network B?). When comparing the presence or absence of individual edges between networks, we again focused on whether edges present in the sparser network were also present in the denser network and whether edges that were absent in the denser network were also absent in the sparser network.

In addition to evaluating replicability across networks, we used the replicationSimulator function in *bootnet* to estimate the *expected* replicability of each network's structure. This is important because the expected replicability rate of a characteristic would not be 100% even if the characteristic were present in the "true" network. In traditional significance tests, the expected replicability rate equals the statistical power of the test. The exact expected replicability rate is difficult to compute for complex multivariate models such as network models, and depends on several factors (e.g., sample size, network structure). The replicationSimulator function estimates expected replicability by simulating new data based on a network model and evaluating how well one should expect this network structure to replicate in other samples if the estimated network were true at the population level. Although this approach is not ideal because it ignores estimation error in the original network and compares edge lists and centrality using Pearson correlations rather than spearman correlations, it provides an estimate of the expected replicability rate for each network structure. Expected replicability results are presented in Figure S3 in the supplementary materials.

Availability of Data and Materials

This study was not formally preregistered. However, in line with recommendations to improve reproducibility and replicability in clinical psychological research (Tackett et al., 2017) and in psychopathology network analysis specifically (Borsboom et al., 2017; Guloksuz, Pries, & van Os, 2017), we have made all data and our analytic code available in the online supplementary materials.

Results

Symptom severities for each sample are reported in Table S3. The samples differed in symptom severity such that the clinical sample had higher severity than the other four samples for all symptoms except appetite gain. There were also some differences in symptom severities among the undergraduate and community samples, but these differences were smaller in magnitude and not consistent across symptoms. There were also sample differences in sex, $\chi^2(4) = 24.17$, $p < .001$, and age, $F(4, 2547) = 297.38$, $p < .001$. There were more females in Undergraduate Sample 2 than in Undergraduate Sample 1 ($p < .001$) and Community Sample 1 ($p < .001$), but no other significant differences in sex between samples. All five samples differed from each other on age (with one exception—the two community samples did not differ, $p = .971$). The clinical sample was the oldest, followed by the two community samples, Undergraduate Sample 2, and Undergraduate Sample 1.

Network Estimation

The five networks are visualized in Figure 1. Visual comparisons across networks indicated that many edges were consistent across all five networks, including relatively strong edges such as *dysphoria-lassitude* and *dysphoria-social anxiety*. There were also some edges that were inconsistent across the five networks, however. For example, the edge *appetite loss-lassitude* was only present in three of the five samples.

Accuracy and Stability of Network Parameters

Bootstrapped CIs around the edge weights were small to moderate (see Figure S4) and, as expected, were smaller for networks estimated from larger samples (e.g., Undergraduate Networks 1 and 2). The CS-coefficient for strength exceeded the recommended cutoff of .50 (Epskamp, Borsboom, et al., 2018) in in all five networks (see Table S4 and Figure S5).

Within-Sample Comparisons

Global metrics. The analyses of the within-sample consistency of the two undergraduate networks indicated that the median global characteristics of the networks estimated from the split-halves extracted from each of the two samples were generally similar within and across the two samples (see Table 1). Comparisons between the 10 pairs of random split-halves for each of the two undergraduate samples are summarized in Table 2. The omnibus test of network structure invariance and global strength invariance test were not significant in either sample (*p*s = .459 and .405) and the median number of significantly different edges across split-halves was zero for both samples. These tests failed to reject the hypotheses that there were differences in the overall structure, global strength, or individual edges of the networks for the split-halves. Edge weights (r_ss = .87 and .78), node predictabilities (r_ss = .93 and .94), and centrality strength (r_ss = .85 and .86) were consistently strongly correlated between split-halves.

Specific metrics. We examined whether estimated edges replicated across split-halves for each sample, and found that a median of 84.8–89.9% of nonzero edges in the sparser split-half network replicated (i.e., were nonzero and had the same sign) in

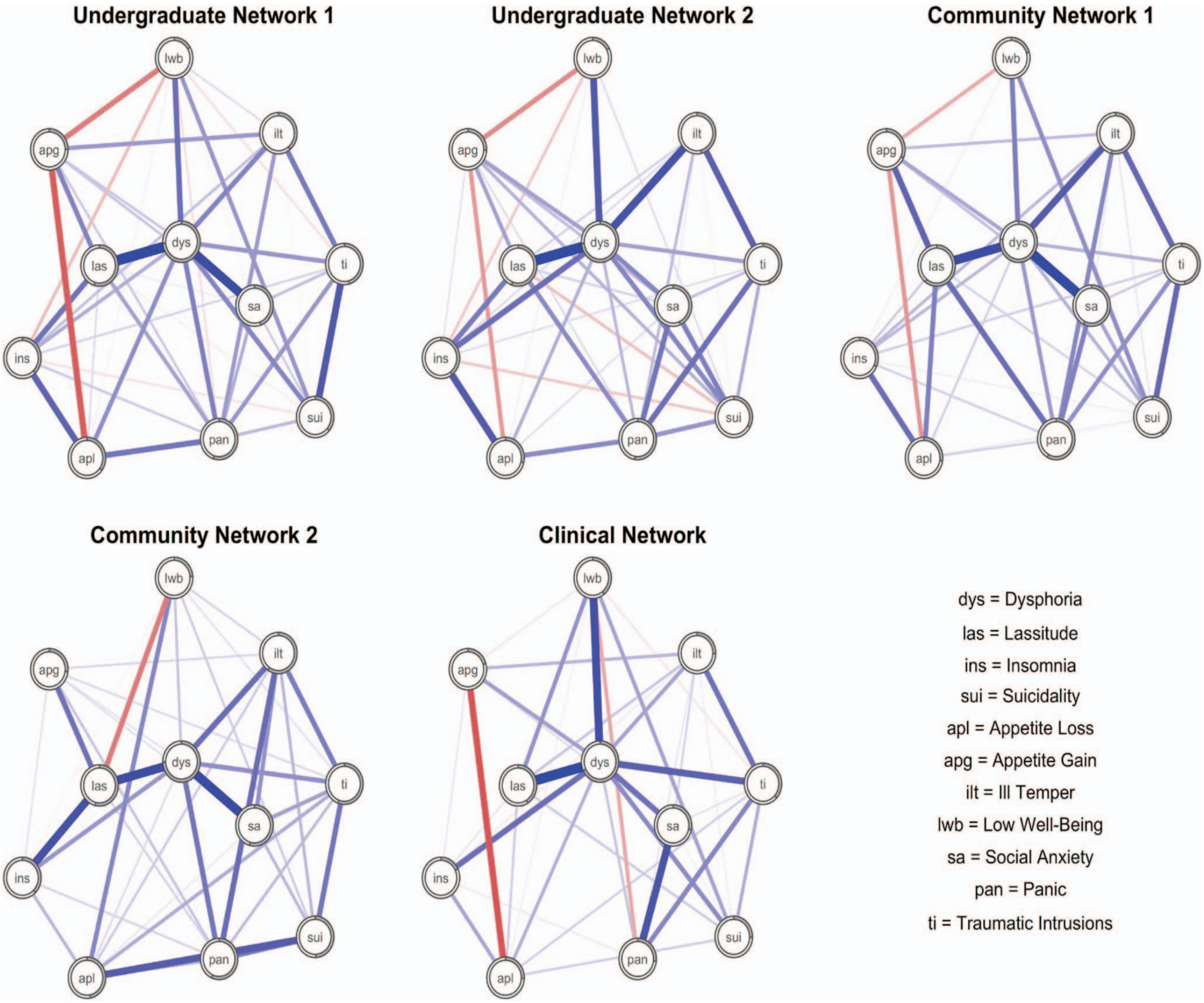

Figure 1. Pairwise Markov random field (PMRF) networks for each sample. Edge thickness represents the strength of the partial correlation. Blue edges indicate positive relationships, and red edges indicate negative relationships. The gray area in the rings around each node indicates predictability (i.e., the variance explained by neighboring nodes). See the online article for the color version of this figure.

the denser split-half network. Examination of consistencies in absent edges revealed that 58.0–69.2% of the absent edges in the denser split-half network replicated in the sparser split-half network. When examining matches in centrality rank-order, dysphoria was identified as the most central symptom in all split-halves. Across all 11 symptoms, however, exact matches in centrality strength rank-order between split-halves were much less frequent (36.4% and 45.5%).

Cross-Sample Comparisons

Global metrics. Global characteristics of the five networks are summarized in Table 3. Consistent with prior research showing that global network strength increases with symptom reduction[9] (Beard et al., 2016; Bos et al., 2018), the clinical network had the lowest global strength and was the sparsest of the five networks. Results of the pairwise network comparisons are shown in Table 4. Similarities between the two undergraduate samples were generally comparable in strength to similarities within the two undergraduate samples (i.e., between random split-halves). Omnibus tests of network structure invariance were statistically significant for three of the 10 pairwise comparisons and indicated that the overall network structure of Community Network 2 significantly differed from that of the two undergraduate networks and the

[9] Interestingly, this finding is inconsistent with the network theory, which suggests that networks should be more strongly connected in clinical samples (Borsboom, 2017). This phenomenon has been observed in multiple studies examining depression networks and is in need of explanation.

Table 1
Median (Range) of Global Network Characteristics of the Ten Pairs of Random Split-Halves From the Two Undergraduate Samples

	Undergraduate sample 1		Undergraduate sample 2	
Network characteristic	First half	Second half	First half	Second half
Global strength	4.75 (4.55–5.00)	4.86 (4.42–5.21)	4.52 (4.26–4.77)	4.59 (4.11–4.90)
% of nonzero edges	72.7% (67.3–76.4)	76.4% (65.5–81.8)	70.9% (60.0–80.0)	70.9% (67.3–76.4)
% of zero edges	27.3% (23.6–32.7)	23.6% (18.2–34.5)	29.1% (20.0–40.0)	29.1% (23.6–32.7)
% of positive edges	89.7% (87.5–94.6)	88.8% (86.0–94.9)	91.2% (89.5–97.4)	90.4% (87.2–97.4)
% of negative edges	10.3% (5.4–12.5)	11.3% (5.1–14.0)	8.9% (2.6–10.5)	9.7% (2.6–12.8)
% of edges with 95% CIs that excluded zero	67.6% (61.0–69.2)	62.8% (52.3–75.0)	48.7% (43.2–62.9)	41.3% (37.8–46.2)
Mean node predictability (R^2)	.43 (.42–.44)	.43 (.42–.44)	.44 (.41–.45)	.42 (.41–.44)
Most frequent central node (strength)	dysphoria	dysphoria	dysphoria	dysphoria

clinical network. None of the global strength invariance tests were significant, however, meaning that the null hypothesis that global strength was equal across networks was not rejected. Additionally, few edges were identified by the individual edge invariance test as being significantly different between networks (maximum number of significantly different edges between any pair of networks = 3 [5.5%]). Coefficients of similarity between nonclinical networks ranged from .53 (Community Network 2 vs. Undergraduate Network 2 and Community Network 1) to .84 (Undergraduate Network 1 vs. Undergraduate Network 2). Coefficients of similarity between the clinical network and nonclinical networks ranged from .36 (vs. Community Network 2) to .66 (vs. Community Network 1). Symptom predictability was strongly correlated among the undergraduate and community networks (r_ss $\geq .83$), but was less consistent between the clinical network and the other four networks with r_ss ranging from .39 to .52 (see Table S5). Centrality strength for each of the five networks is plotted in Figure 2. Spearman correlations of strength between nonclinical networks ranged from .37 (Undergraduate Network 1 vs. Community Network 2) to .73 (Undergraduate Network 1 vs. Undergraduate Network 2), and comparisons between the clinical network and nonclinical networks ranged from .04 (vs. Undergraduate Network 2) to .54 (vs. Community Sample 2).

Specific metrics. Examination of consistencies in specific network parameters revealed that 53 (96.4%) of the 55 total possible edges were estimated (i.e., nonzero) in at least one of the five networks and 21 (38.2%) were consistently estimated in all five networks.[10] All but one (95.3%) of the 21 edges estimated in all five networks were positive in all five networks. The remaining edge (*ill temper-suicide*) was negative in Undergraduate Sample 2 but positive in the other four networks. Comparing nonzero edges in the sparsest network (the clinical network) to those in the denser networks and absent edges in the densest network (Undergraduate Network 1) to those in the sparser networks, we found that 55.3% of the 38 estimated edges in the sparsest network were estimated in all four of the other networks, and two (16.7%) of the 12 absent edges in the densest network were absent in all four other networks. Across the pairwise comparisons between networks, 73.7–85.4% of the nonzero edges in the sparser network replicated (i.e., were nonzero and had the same sign) in the denser network and 33.3–75.0% of the absent edges in the denser network replicated (i.e., were also absent) in the sparser network.

Consistent with the results of the random split-halves of the two undergraduate networks, dysphoria was consistently identified as the most central node across all five networks. Although the most central node was consistent across networks, the number of matches in other rank-orders (e.g., the second, third, etc., most central node) was consistently poor and ranged from zero to two (18.2%). Centrality difference tests (see Figure S2) indicated that there were few *significant* differences in centrality strength between symptoms with the exception that dysphoria was significantly more central than all other symptoms in all five networks. In fact, excluding dysphoria, there was only one case across all five networks in which the centrality strength of two symptoms that were adjacent in rank-order (e.g., second and third most central) significantly differed.

Results of Models Estimated Using Nodewise Regression

The nodewise regression networks are plotted in Figure S8, and their accuracy, stability, characteristics, and replicability and generalizability are presented in detail in Figures S9–S10 and Tables S8–S10. Nodewise regression yielded networks that were highly correlated with (r_ss $> .73$), but sparser than, the GLASSO networks. Coefficients of similarity ranged from .28 to .82, and NCTs indicated four significant differences in global structure, two significant differences in global strength, and zero to two significant differences in individual edges between networks. Cross-network correlations of centrality strength ranged from .29 to .86. Specific metrics of replicability indicated that 50.0–96.0% of individual edges were consistent in each pair of networks and matches in centrality rank-order ranged from one (9.1%) to three (27.3%).

Discussion

This study aimed to conduct an agnostic examination of the replicability and generalizability of PMRF networks of internalizing symptoms across five samples. As there has been debate surrounding the replicability and generalizability of psychopathology network models (Borsboom et al., 2017; Forbes et al., 2017a, 2017b, 2019; Jones, Williams, & McNally, 2019) which partially revolves around disagreement regarding which measures are most appropriate for evaluating similarities across networks, we exam-

[10] The 21 edges estimated consistently in all five networks may represent the core of the generalizable network structure of internalizing symptoms, and are plotted in Figure S7.

Table 2
Median (Range) of Network Comparisons Between the Ten Pairs of Random Split-Halves From the Two Undergraduate Samples

Split-half network comparisons	Undergraduate sample 1		Undergraduate sample 2	
NCT results				
Omnibus test of network structure invariance *p*-value	.459 (.237–.968)		.405 (.017–.980)	
Global strength invariance test *p*-value	.585 (.113–.961)		.692 (.494–.987)	
% of significantly different edges in the individual edge invariance test	0% (0–0)		0% (0–1.8)	
Edges				
r_s between all edges	.87 (.83–.91)		.78 (.66–.85)	
r_s between nonzero edges	.78 (.65–.84)		.64 (.46–.85)	
r_s between node predictabilities	.94 (.85–.98)		.93 (.87–.97)	
Jaccard Index[a]	.76 (.69–.84)		.69 (.58–.74)	
% of nonzero edges in the sparser split-half that were also nonzero in the denser split-half	89.9% (82.5–94.9)		84.8% (78.8–89.5)	
% of nonzero edges in the sparser split-half that were nonzero *and* had the same sign in the denser split-half	81.6% (79.1–90.2)		78.5% (68.4–84.6)	
% of edges with 95% CIs that did not include zero in the sparser network that also had CIs that excluded zero in the denser network	86.1% (75.0–100)		72.8% (57.9–81.2)	
% of absent edges in the denser split-half that were also absent in the sparser split-half	69.2% (50.0–85.7)		58.0% (53.3–69.2)	
Node centrality	% matched	Most central node	% matched	Most central node
Same most central node in both split-halves (strength)	100%	dysphoria	100%	dysphoria
Rank-order correspondence	r_s	Matches (%)[b]	r_s	Matches (%)[b]
Strength	.85 (.72–.95)	36.4% (27.3–63.6)	.86 (.61–.94)	45.5% (18.2–63.6)

Note. NCT = Network Comparison Test. [a] The proportion of shared edges relative to the total number of edges in both networks. [b] As calculated by Forbes, Wright, Markon, and Krueger (2017a) and Borsboom et al. (2017) to allow for multiple simultaneous ranks.

ined the more global measures advocated for by some groups (e.g., coefficients of similarity, correlations of centrality, NCTs; Borsboom et al., 2017) as well as the more specific metrics advocated for by others (e.g., replicability of individual edges, matches in rank-order centrality; Forbes et al., 2017a, 2017b).

Global Metrics of Network Similarity

Broader metrics of similarity generally suggested moderate to strong similarities between random split-halves of the two undergraduate samples and across the four nonclinical networks, with weaker global similarities between the clinical and nonclinical networks. Among the four nonclinical networks, coefficients of similarity indicated moderate to strong similarities in global network structure and correlations of centrality strength were generally in the moderate range. However, edges and centrality strength tended to be more highly correlated within samples or between networks from the same population than between networks from different populations (e.g., undergraduate vs. clinical network). Taken together, these metrics suggest that internalizing symptom network characteristics (at least as measured by the IDAS) had stronger replicability than generalizability.

Specific Metrics of Network Similarity

Compared with broader metrics of consistency, the replicability and generalizability of specific metrics of network similarities was weaker both within and across samples. Within- and cross-sample comparisons of individual edges revealed that approximately four in five estimated edges replicated, and the percentage of individual edges that generalized from nonclinical networks to the clinical network (73.7–81.6%) was similar to the percentage of individual edges that replicated between nonclinical networks (75.0–85.4%). In other words, our results suggest that if internalizing symptom networks are estimated from two samples from any of the populations sampled in this study, a relationship between two symptoms that is present in one network (e.g., dysphoria-insomnia) has an approximately 80% chance of replicating in the other network.

Table 3
Global Characteristics of the Five Networks

Network characteristic	Undergraduate network 1	Undergraduate network 2	Community network 1	Community network 2	Clinical network
Global strength	5.24	5.10	4.71	4.78	4.19
Number of nonzero edges (% possible)	43 (78.2%)	41 (74.5%)	41 (74.5%)	40 (72.7%)	38 (69.1%)
Number of zero edges (% possible)	12 (21.8%)	14 (25.5%)	14 (25.5%)	15 (27.3%)	17 (30.9%)
Number of positive edges (% total)	35 (81.4%)	35 (85.4%)	39 (95.1%)	38 (95%)	34 (89.5%)
Number of negative edges (% total)	8 (18.6%)	6 (14.6%)	2 (4.9%)	2 (5.0%)	4 (10.5%)
Number of edges with 95% CIs that excluded zero (% total)	32 (74.4%)	27 (65.9%)	19 (46.3%)	17 (42.5%)	13 (34.2%)
Mean node predictability (R^2)	.42	.42	.46	.49	.32
Most central node (strength)	dysphoria	dysphoria	dysphoria	dysphoria	dysphoria

Table 4

Results of Pairwise Comparisons Between the Five Networks

Network characteristic	Network 1 vs. 2	Network 1 vs. 3	Network 1 vs. 4	Network 1 vs. 5	Network 2 vs. 3	Network 2 vs. 4	Network 2 vs. 5	Network 3 vs. 4	Network 3 vs. 5	Network 4 vs. 5
NCT results										
Omnibus test of network structure invariance *p*-value	.079	.770	.003	.194	.153	.014	.505	.171	.139	<.001
Global strength invariance test *p*-value	.733	.216	.353	.121	.253	.761	.134	.852	.334	.287
No. (% total) of significantly different edges in the individual edge invariance test	1 (1.8%)	0 (0%)	3 (5.5%)	0 (0%)	2 (3.6%)	2 (3.6%)	1 (1.8%)	1 (1.8%)	0 (0%)	2 (3.6%)
Edges										
r_s between all edges	.84	.83	.61	.60	.73	.53	.58	.53	.66	.36
r_s between nonzero edges	.77	.75	.63	.73	.72	.44	.73	.38	.66	.22
r_s between node predictabilities	.96	.92	.85	.52	.93	.87	.49	.83	.51	.39
Jaccard Index[a]	.75	.83	.63	.65	.74	.62	.61	.62	.65	.62
No. (%) of nonzero edges in the sparser network that were also nonzero in the denser network	36 (87.8%)	38 (92.7%)	32 (80.0%)	32 (84.2%)	35 (85.4%)[b]	31 (77.5%)	30 (78.9%)	31 (77.5%)	31 (81.6%)	30 (78.9%)
No. (%) of nonzero edges in the sparser network that were nonzero *and* had the same sign in the denser network	35 (85.4%)	34 (82.9%)	31 (77.5%)	31 (81.6%)	31 (75.6%)	30 (75.0%)	28 (73.7%)	30 (75.0%)	31 (81.6%)	28 (73.7%)
No. (%) of edges with 95% CIs that excluded zero in the sparser network that also had CIs that excluded zero in the denser network	26 (96.3%)	18 (94.7%)	12 (70.6%)	12 (92.3%)	17 (89.5%)	12 (70.6%)	11 (84.6%)	9 (52.9%)	10 (76.9%)	7 (53.8%)
No. (%) of absent (i.e., zero) edges in the denser network that were also absent in the sparser network	7 (58.3%)	9 (75.0%)	4 (33.3%)	6 (50.0%)	8 (57.1%)[b]	5 (35.7%)	6 (42.9%)	5 (35.7%)	7 (50.0%)	7 (46.7%)
Node centrality strength										
r_s	.73	.51	.37	.42	.65	.41	.04	.80	.44	.54
No. (%) of rank-order matches[c]	3 (27.3%)	2 (18.2%)	3 (27.3%)	1 (9.1%)	2 (18.2%)	2 (18.2%)	1 (9.1%)	1 (9.1%)	1 (9.1%)	3 (27.3%)

Note. NCT = Network Comparison Test; Network 1 = Undergraduate Network 1; Network 2 = Undergraduate Network 2; Network 3 = Community Network 1; Network 4 = Community Network 2; Network 5 = Clinical Network.

[a] The proportion of shared edges relative to the total number of edges in both networks. [b] Both networks had the same number of nonzero edges, so we randomly selected Community Network 1 to be the "sparser" network and Undergraduate Network 2 to be the "denser" network. [c] As calculated by Forbes et al. (2017a) and Borsboom et al. (2017) to allow for multiple simultaneous ranks.

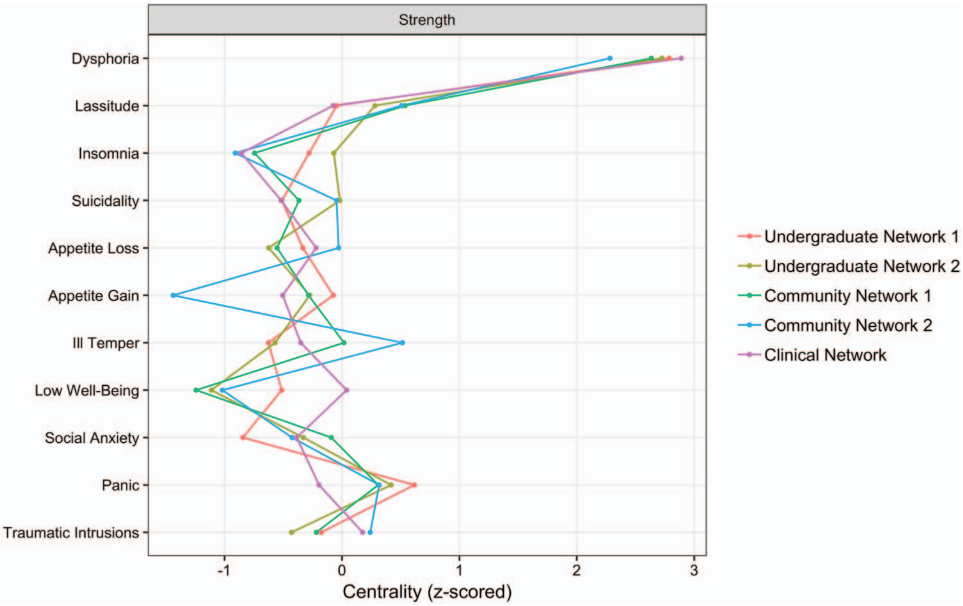

Figure 2. Standardized centrality strength for each network. See the online article for the color version of this figure.

When comparing rank-order centrality, dysphoria was consistently the most central symptom and shared a positive edge with all other symptoms in all five networks (with the exception of Community Network 2, in which all edges except for *dysphoria-suicidality* were present). This result is consistent with the tripartite model's assertion that negative affect is the common component across depression and anxiety disorders (Clark & Watson, 1991; Shankman & Klein, 2003). The consistency of exact centrality rank-order of the other symptoms was generally poor, but was generally poorer between the clinical network and a nonclinical network than between two nonclinical networks. However, there were few *significant* differences in symptom centrality (see Figure S2), suggesting that differences in centrality rank-order (with the exception of dysphoria being most central) were likely reflective of sampling variability rather than true variability (i.e., nonreplication; Borsboom et al., 2017; Jones et al., 2019).

Implications, Caveats, and Recommendations for Regularized Cross-Sectional Psychopathology Network Analysis

Psychopathology networks have had a notable impact on psychopathology research (Fried et al., 2017), but are not without their criticisms and controversies (Bringmann & Eronen, 2018; Forbes et al., 2017a, 2017b; Guloksuz et al., 2017). This controversy has extended to network replicability, and the question of which replicability metrics are most relevant or appropriate has been a particular source of disagreement. More global network characteristics such as overall network structure and global network strength have been the focus of several studies, including studies examining whether network characteristics predict treatment response (Schweren et al., 2018; van Borkulo et al., 2015), and these characteristics replicated well across networks in the present study. Specific characteristics such as individual edges have also been examined extensively in the psychopathology network literature (e.g., to understand the role of bridging edges in comorbidity; Cramer et al., 2010), and individual edges replicated and generalized approximately 80% of the time on average in the present study. More advanced methods for evaluating the replicability of individual edges such as Bayesian equivalence testing (Williams & Mulder, 2019) may be informative in future replicability work. Additionally, many studies have sought to identify the most central symptom in a network (e.g., Beard et al., 2016), and dysphoria was consistently the most central symptom within and across samples in the present study. However, researchers interested in whether a symptom is the second most central, third most central, and so forth should note the few significant within-network differences in centrality and the poor cross-network consistency of exact centrality rank-order.

As different network metrics answer different questions, we recommend that researchers examine multiple replicability metrics when appropriate and select metrics that align with their research question(s). For example, a global metric such as the coefficient of similarity may be useful for quantifying the degree of overall similarity between two networks, but can obfuscate meaningful differences in specific network characteristics (e.g., individual edges). It is also important to note that network structure can influence network models and their replicability in several ways. First, network structure can affect the appropriateness of replicability metrics. For example, one would only expect centrality rank-order to replicate if there were true differences in centrality. If all nodes were equally central (e.g., see the Curie-Weiss network; Kac, 1968), any observed rank-order would purely reflect sampling error. The dependence of replicability on details of the data-generating mechanism is not specific to network models, but rather applies to all multivariate techniques. Second, network structure can influence model estimation performance (Epskamp, Kruis, & Marsman, 2017; van Borkulo et al., 2014). This study primarily focused on GLASSO-regularized networks because this estimation method is widely used in psychopathology network analyses of continuous data. However, GLASSO may have poorer specificity than other methods when the network structure is dense (Williams & Rast, 2019; Williams, Rhemtulla, Wysocki, & Rast, 2019), which is often the case in psychopathology networks. Estimating networks using a nodewise estimation approach rather than GLASSO did not appreciably increase replicability in the present study, but may be helpful in other studies, especially those aiming to detect many small effects. In light of these considerations, future studies examining psychopathology networks and their replicability should evaluate replicability metrics on a case-by-case basis and take network structure and its impact on model estimation methods into consideration (e.g., by using bootstrapping and/or the replicationSimulator function).

There are also several unresolved issues pertaining to psychopathology network analysis. First, defining and quantifying network replicability remains an ongoing challenge (Borsboom et al., 2017; Forbes et al., 2017a, 2019). As different definitions can yield different conclusions regarding whether causally distinct networks are structurally equivalent (Schieber et al., 2017), further research is needed to examine the validity of network replicability metrics under various conditions. Second, the extent to which networks replicate across different measures of the same disorder(s) is unclear. Recent studies suggest that networks may be influenced by symptom severity thresholds (Hoffman, Steinley, Trull, & Sher, 2018) but relatively robust to different raters (Moshier et al., 2018). However, the effect of other differences between measures (e.g., lack of content overlap; Fried, 2017) on symptom networks and their replicability is unknown. Third, symptoms are typically conceptualized and modeled as manifest variables in symptom networks, but at least some symptoms may be indicators of latent constructs (McFarland & Malta, 2010). If this is the case, the extent to which conclusions drawn from network analyses of observed symptoms generalize to the level of latent constructs may be overestimated (Westfall & Yarkoni, 2016). Modeling symptoms in a network as latent variables offers one potential solution to this problem (Epskamp, Rhemtulla, & Borsboom, 2017).

Strengths and Limitations

This study had numerous strengths, including the inclusion of five samples and the examination of replicability within samples and across samples from the same population as well as generalizability to samples from different populations. Recent work suggests Berkson's Bias, which occurs when relationships in a subpopulation (e.g., a clinical sample for which a clinical severity cutoff was an inclusion criterion) differ from those in the general population, is a concern when interpreting networks estimated from clinical samples (de Ron, Fried, & Epskamp, 2019). The multisample approach and comparison of clinical and nonclinical samples allowed us to detect the extent to which Berkson's Bias occurred, and this may explain why coefficients of similarity and correlations were stronger between nonclinical networks than between the clinical network and nonclinical networks. The strong psychometric properties of the factor analytically derived IDAS was also a strength, as well as the fact that the IDAS assessed each symptom using multiple items using nonoverlapping subscales. This is an important feature because it is unclear how psychopathology network analyses should handle "topographically overlapping" nodes (e.g., "feeling blue" and "sad mood"; however, see the goldbricker function in the networktools R package [Jones, 2018] for a proposed tool to aid in identifying topologically overlapping nodes).

This study also had several limitations. First, our results were specific to internalizing symptoms and the reported replicability and generalizability results may not extend to other domains of psychopathology (e.g., externalizing symptoms). Second, the present study only examined centrality strength, closeness, and betweenness, and results should not be extended to other centrality measures. Third, this study focused on LASSO-regularized and nodewise regression networks and results should not be generalized to other types of networks or estimation methods. Therefore, researchers may benefit from consideration of other network estimation methods if the hypothesized "true" network is dense. Fourth, the sizes of the community samples and the clinical sample were somewhat smaller than those examined in other studies of network replicability and generalizability (Forbes et al., 2017a; Fried et al., 2018). Although these sample sizes are consistent with those commonly studied in the psychopathology network literature, they prevented us from examining consistencies between random split-halves of the two community samples and the clinical sample. Fifth, sample characteristics may have influenced the results. The clinical sample was comprised of treatment-seekers for internalizing problems and individuals with MDD and/or PD, and this heterogeneity may have reduced the degree to which the clinical network approximated the "true" population network (Fried & Cramer, 2017). Additionally, although the random assignment of siblings to either Community Samples 1 or 2 avoided nonindependence of observations, it is possible that this procedure artificially inflated the similarity of correlation matrices between the two community samples given the familial aggregation of psychopathology. This may have led to overestimated replicability between the two community networks, but, importantly, would not have affected any other pairwise network comparisons. Sixth, we were unable to examine the replicability of network characteristics between clinical samples because there was only one clinical sample.

Conclusion

Replicability and generalizability are crucial issues for the field of psychopathology network analysis. This study examined similarities within and across networks of internalizing symptom networks and found substantial global similarities across networks. Approximately 80% of individual edges replicated within and across samples and the most central symptom (i.e., dysphoria) was consistent within and across samples. In sum, both global and specific metrics generally indicated considerable replicability within and across undergraduate and community samples, but network structure and symptom centrality had weak to moderate generalizability to a clinical sample.

References

Altman, S. E., Campbell, M. L., Nelson, B. D., Faust, J. P., & Shankman, S. A. (2013). The relation between symptoms of bulimia nervosa and obsessive-compulsive disorder: A startle investigation. *Journal of Abnormal Psychology, 122,* 1132–1141. http://dx.doi.org/10.1037/a0034487

Beard, C., Millner, A. J., Forgeard, M. J. C., Fried, E. I., Hsu, K. J., Treadway, M. T., . . . Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological Medicine, 46,* 3359–3369. http://dx.doi.org/10.1017/S0033291716002300

Benfer, N., Bardeen, J. R., Cero, I., Kramer, L. B., Whiteman, S. E., Rogers, T. A., . . . Weathers, F. W. (2018). Network models of posttraumatic stress symptoms across trauma types. *Journal of Anxiety Disorders, 58,* 70–77. http://dx.doi.org/10.1016/j.janxdis.2018.07.004

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry, 16,* 5–13. http://dx.doi.org/10.1002/wps.20375

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 (2017). *Journal of Abnormal Psychology, 126,* 989–999. http://dx.doi.org/10.1037/abn0000306

Borsboom, D., Robinaugh, D. J., Rhemtulla, M., & Cramer, A. O. J. (2018). Robustness and replicability of psychopathology networks. *World Psychiatry, 17,* 143–144. http://dx.doi.org/10.1002/wps.20515

Bos, F. M., Fried, E. I., Hollon, S. D., Bringmann, L. F., Dimidjian, S., DeRubeis, R. J., & Bockting, C. L. H. (2018). Cross-sectional networks of depressive symptoms before and after antidepressant medication treatment. *Social Psychiatry and Psychiatric Epidemiology, 53,* 617–627. http://dx.doi.org/10.1007/s00127-018-1506-1

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., . . . Snippe, E. (2019). What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology*. Advance online publication.

Bringmann, L. F., & Eronen, M. I. (2018). Don't blame the model: Reconsidering the network approach to psychopathology. *Psychological Review, 125,* 606–615. http://dx.doi.org/10.1037/rev0000108

Cicchetti, D. V., & Prusoff, B. A. (1983). Reliability of depression and associated clinical symptoms. *Archives of General Psychiatry, 40,* 987–990. http://dx.doi.org/10.1001/archpsyc.1983.01790080069009

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 100,* 316–336. http://dx.doi.org/10.1037/0021-843X.100.3.316

Correa, K. A., Liu, H., & Shankman, S. A. (2019). The role of intolerance of uncertainty in current and remitted internalizing and externalizing psychopathology. *Journal of Anxiety Disorders, 62,* 68–76. http://dx.doi.org/10.1016/j.janxdis.2019.01.001

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences, 33,* 137–150. http://dx.doi.org/10.1017/S0140525X09991567

Dablander, F., & Hinne, M. (2019). Node centrality measures are a poor substitute for causal inference. *Scientific Reports, 9,* 6846. http://dx.doi.org/10.1038/s41598-019-43033-9

de Ron, J., Fried, E. I., & Epskamp, S. (2019). *Psychological networks in clinical populations: A tutorial on the consequences of Berkson's Bias*. http://dx.doi.org/10.31234/OSF.IO/5T8ZW

Distefano, A., Jackson, F., Levinson, A. R., Infantolino, Z. P., Jarcho, J. M., & Nelson, B. D. (2018). A comparison of the electrocortical response to monetary and social reward. *Social Cognitive and Affective Neuroscience, 13,* 247–255. http://dx.doi.org/10.1093/scan/nsy006

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods, 50,* 195–212. http://dx.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*. Advance online publication. http://dx.doi.org/10.18637/jss.v048.i04

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

Epskamp, S., Fried, E. I., van Borkulo, C. D., Robinaugh, D. J., Marsman, M., Dalege, J., . . . Cramer, A. O. J. (2018). Investigating the utility of fixed-margin sampling in network psychometrics. *Multivariate Behavioral Research, 12,* 1–15. http://dx.doi.org/10.1080/00273171.2018.1489771

Epskamp, S., Kruis, J., & Marsman, M. (2017). Estimating psychopathological networks: Be careful what you wish for. *PLoS ONE, 12,* e0179891. http://dx.doi.org/10.1371/journal.pone.0179891

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika, 82,* 904–927. http://dx.doi.org/10.1007/s11336-017-9557-x

Forbes, M. K., Wright, A. G. C., Markon, K. E., & Krueger, R. F. (2017a). Evidence that psychopathology symptom networks have limited replicability. *Journal of Abnormal Psychology, 126,* 969–988. http://dx.doi.org/10.1037/abn0000276

Forbes, M. K., Wright, A. G. C., Markon, K. E., & Krueger, R. F. (2017b). Further evidence that psychopathology networks have limited replicability and utility: Response to Borsboom et al. (2017) and Steinley et al. *Journal of Abnormal Psychology, 126,* 1011–1016. http://dx.doi.org/10.1037/abn0000313

Forbes, M. K., Wright, A. G. C., Markon, K. E., & Krueger, R. F. (2019). Quantifying the reliability and replicability of psychopathology network characteristics. *Multivariate Behavioral Research*. Advance online publication. http://dx.doi.org/10.1080/00273171.2019.1616526

Foygel, R., & Drton, M. (2010). Extended Bayesian information criteria for gaussian graphical models. *Advances in Neural Information Processing Systems, 1,* 604–612.

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders, 208,* 191–197. http://dx.doi.org/10.1016/j.jad.2016.10.019

Fried, E. I., & Cramer, A. O. J. (2017). Moving forward: Challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science, 12,* 999–1020. http://dx.doi.org/10.1177/1745691617705892

Fried, E. I., Eidhof, M. B., Palic, S., Costantini, G., Huisman-van Dijk, H. M., Bockting, C. L. H., . . . Karstoft, K.-I. (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,* 335–351. http://dx.doi.org/10.1177/2167702617745092

Fried, E. I., van Borkulo, C. D., Cramer, A. O. J., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: A review of recent insights. *Social Psychiatry and Psychiatric Epidemiology, 52,* 1–10. http://dx.doi.org/10.1007/s00127-016-1319-z

Funkhouser, C. J., Correa, K. A., Carrillo, V. L., Klemballa, D. M., & Shankman, S. A. (2019). The time course of responding to aversiveness in females with a history of non-suicidal self-injury. *International Journal of Psychophysiology, 141,* 1–8. http://dx.doi.org/10.1016/j.ijpsycho.2019.04.008

Gorka, S. M., Nelson, B. D., Sarapas, C., Campbell, M., Lewis, G. F., Bishop, J. R., . . . Shankman, S. A. (2013). Relation between respiratory sinus arrythymia and startle response during predictable and unpredictable Threat. *Journal of Psychophysiology, 27,* 95–104. http://dx.doi.org/10.1027/0269-8803/a000091

Guloksuz, S., Pries, L.-K., & van Os, J. (2017). Application of network methods for understanding mental disorders: Pitfalls and promise. *Psychological Medicine, 47,* 2743–2752. http://dx.doi.org/10.1017/S0033291717001350

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

Hoffman, M., Steinley, D., Trull, T. J., & Sher, K. J. (2018). Criteria definitions and network relations: The importance of criterion thresholds. *Clinical Psychological Science, 6,* 506–516. http://dx.doi.org/10.1177/2167702617747657

Jones, P. J. (2018). *networktools: Assorted tools for identifying important nodes in networks*. Retrieved from https://cran.r-project.org/package=networktools

Jones, P. J., Williams, D. R., & McNally, R. J. (2019). *Sampling variability is not nonreplication: A Bayesian reanalysis of Forbes, Wright, Markon, & Krueger*. http://dx.doi.org/10.31234/osf.io/egwfj

Kac, M. (1968). Mathematical mechanism of phase transition. In M. Chretien, E. P. Gross, & S. Deser (Eds.), *Statistical physics, phase transitions, and superfluidity* (pp. 241–305). New York, NY: Gordon & Breach.

Knefel, M., Lueger-Schuster, B., Bisson, J., Karatzias, T., Kazlauskas, E., & Roberts, N. P. (2019). A cross-cultural comparison of ICD-11 complex posttraumatic stress disorder symptom networks in Austria, the United Kingdom, and Lithuania. *Journal of Traumatic Stress*. Advance online publication. http://dx.doi.org/10.1002/jts.22361

Lieberman, L., Gorka, S. M., DiGangi, J. A., Frederick, A., & Phan, K. L. (2017). Impact of posttraumatic stress symptom dimensions on amygdala reactivity to emotional faces. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 79,* 401–407. http://dx.doi.org/10.1016/j.pnpbp.2017.07.021

Lieberman, L., Gorka, S. M., Funkhouser, C. J., Shankman, S. A., & Phan, K. L. (2017). Impact of posttraumatic stress symptom dimensions on psychophysiological reactivity to threat and reward. *Journal of Psychiatric Research, 92,* 55–63. http://dx.doi.org/10.1016/j.jpsychires.2017.04.002

Liu, H., Lafferty, J., Wasserman, L., & Wainwright, M. J. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research, 10,* 2295–2328.

McFarland, D. J., & Malta, L. S. (2010). Symptoms as latent variables. *Behavioral and Brain Sciences, 33,* 165–166. http://dx.doi.org/10.1017/S0140525X1000066X

Moshier, S. J., Bovin, M. J., Gay, N. G., Wisco, B. E., Mitchell, K. S., Lee, D. J., . . . Marx, B. P. (2018). Examination of posttraumatic stress disorder symptom networks using clinician-rated and patient-rated data. *Journal of Abnormal Psychology, 127,* 541–547. http://dx.doi.org/10.1037/abn0000368

Nelson, B. D., & Hajcak, G. (2017). Defensive motivation and attention in anticipation of different types of predictable and unpredictable threat: A startle and event-related potential investigation. *Psychophysiology, 54,* 1180–1194. http://dx.doi.org/10.1111/psyp.12869

Schieber, T. A., Carpi, L., Díaz-Guilera, A., Pardalos, P. M., Masoller, C., & Ravetti, M. G. (2017). Quantification of network structural dissimilarities. *Nature Communications, 8,* 13928. http://dx.doi.org/10.1038/ncomms13928

Schweren, L., van Borkulo, C. D., Fried, E., & Goodyer, I. M. (2018). Assessment of symptom network density as a prognostic marker of treatment response in adolescent depression. *Journal of the American Medical Association Psychiatry, 75,* 98–100. http://dx.doi.org/10.1001/jamapsychiatry.2017.3561

Shankman, S. A., Funkhouser, C. J., Klein, D. N., Davila, J., Lerner, D., & Hee, D. (2018). Reliability and validity of severity dimensions of psychopathology assessed using the Structured Clinical Interview for *DSM–5* (SCID). *International Journal of Methods in Psychiatric Research, 27,* e1590. http://dx.doi.org/10.1002/mpr.1590

Shankman, S. A., & Klein, D. N. (2003). The relation between depression and anxiety: An evaluation of the tripartite, approach-withdrawal and valence-arousal models. *Clinical Psychology Review, 23,* 605–637. http://dx.doi.org/10.1016/S0272-7358(03)00038-2

Shankman, S. A., Nelson, B. D., Sarapas, C., Robison-Andrew, E. J., Campbell, M. L., Altman, S. E., . . . Gorka, S. M. (2013). A psychophysiological investigation of threat and reward sensitivity in individuals with panic disorder and/or major depressive disorder. *Journal of Abnormal Psychology, 122,* 322–338. http://dx.doi.org/10.1037/a0030747

Steinley, D., Hoffman, M., Brusco, M. J., & Sher, K. J. (2017). A method for making inferences in network analysis: Comment on Forbes, Wright, Markon, and Krueger (2017). *Journal of Abnormal Psychology, 126,* 1000–1010. http://dx.doi.org/10.1037/abn0000308

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., . . . Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science, 12,* 742–756. http://dx.doi.org/10.1177/1745691617690042

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,* 5918. http://dx.doi.org/10.1038/srep05918

van Borkulo, C., 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 longitudinal depression. *Journal of the American Medical Association Psychiatry, 72,* 1219–1226. http://dx.doi.org/10.1001/jamapsychiatry.2015.2079

van Borkulo, C. D., Epskamp, S., & Millner, A. J. (2016). *Network comparison test: Statistical comparison of two networks based on three invariance measures*. Retrieved from https://cran.r-project.org/web/packages/NetworkComparisonTest/index.html

Watson, D., O'Hara, M. W., Simms, L. J., Kotov, R., Chmielewski, M., McDade-Montez, E. A., . . . Stuart, S. (2007). Development and validation of the Inventory of Depression and Anxiety Symptoms (IDAS). *Psychological Assessment, 19,* 253–268. http://dx.doi.org/10.1037/1040-3590.19.3.253

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE, 11,* e0152719. http://dx.doi.org/10.1371/journal.pone.0152719

Williams, D. R., & Mulder, J. (2019). *BGGM: A R package for Bayesian Gaussian graphical models*. Retrieved from http://dx.doi.org/10.31234/osf.io/3b5hf

Williams, D. R., & Rast, P. (2019). Back to the basics: Rethinking partial correlation network methodology. *British Journal of Mathematical & Statistical Psychology*. Advance online publication. http://dx.doi.org/10.1111/bmsp.12173

Williams, D. R., Rhemtulla, M., Wysocki, A. C., & Rast, P. (2019). On nonregularized estimation of psychological networks. *Multivariate Behavioral Research, 54,* 719–750. http://dx.doi.org/10.1080/00273171.2019.1575716

Received March 13, 2019
Revision received September 12, 2019
Accepted October 23, 2019 ■