Learning to Live with Sampling Variability: Expected Replicability in Partial Correlation Networks

Donald R. Williams

University of California, Davis

Corresponding author:

Donald R. Williams

Department of Psychology

University of California, Davis

drwwilliams@ucdavis.edu

Author Note

DRW was supported by a National Science Foundation Graduate Research Fellowship under Grant No. 1650042.

Abstract

The topic of replicability has recently captivated the emerging field of network psychometrics. Although methodological practice (e.g., *p*-hacking) has been identified as a root cause of unreliable research findings in psychological science, the statistical model itself has come under attack in the partial correlation network literature. In a motivating example, I first describe how sampling variability inherent to partial correlations can *merely* give the appearance of unreliability. For example, when going from zero-order to partial correlations there is necessarily more sampling variability that translates into reduced statistical power. I then introduce novel methodology for deriving expected network replicability (ENR), wherein replication is modeled with the Poisson-binomial distribution. This analytic solution can be used with the Pearson, Spearman, Kendall, and polychoric partial correlation coefficient. I first employed the method to estimate ENR for a variety of datasets from the network literature. Here it was determined that partial correlation networks do not have inherent limitations, given current estimates of replicability were consistent with ENR. I then highlighted sources that can reduce replicability, that is, when going from continuous to ordinal data with few categories and employing a multiple comparisons correction. To address these challenges, I described a strategy for using the proposed method to plan for network replication. I end with recommendations that include the importance of the network literature repositioning itself with gold-standard approaches for assessing replication, including explicit consideration of type I and type II error rates. The method for computing ENR is implemented in the R package **GGMnonreg**.

Keywords: replicability, partial correlation network, power, error rates, frequentist inference

Learning to Live with Sampling Variability: Expected Replicability in Partial Correlation Networks

In the social-behavioral sciences, network theory has emerged as an increasingly popular framework for understanding psychological constructs and mental disorders (Borsboom, 2017; Jones, Heeren, & McNally, 2017). The underlying rationale is that a group of observed variables, say, self-reported symptoms, are a dynamic system that mutually influence and interact with one another (Borsboom & Cramer, 2013). The observed variables are "nodes" and the featured connections between nodes are "edges." This work focuses on partial correlation networks, wherein the edges represent conditionally dependent nodes–pairwise relations that have controlled for the other nodes in the network. This powerful approach has resulted in an explosion of research; for example, network analysis has been used to shed new light upon a variety of disorders including depression (Mullarkey, Marchetti, & Beevers, 2018), post-traumatic stress (McNally et al., 2015), and substance abuse (Rhemtulla et al., 2016).

Recently, the topic of replication has surfaced in the network literature (Forbes, Wright, Markon, & Krueger, 2017, 2019; Fried et al., 2018; Jones, Williams, & McNally, 2019). This has aligned network analysis with emerging traditions in psychological science that emphasize replicability. However, there are important distinctions between, say, large scale replication efforts (e.g., the many labs projects; Klein et al., 2014), and those in the network literature. Perhaps the most striking contrast is where the focus has been placed, especially as it relates to the root cause of unreliable research findings. While the "credibility revolution" has brought much needed attention to improving methodological *practice* (Nelson, Simmons, & Simonsohn, 2018; Vazire, 2018), for example by highlighting the pitfalls associated with *p*-hacking or cultivating a "garden of forking paths" (Gelman & Loken, 2014), the statistical model itself has been targeted in the network literature. That is, limited replicability has been attributed to supposed deficiencies of conditional dependence models.

The most prominent examples that have argued networks lack replicability, largely due to the statistical model, can be found in Forbes et al. (2017) and Forbes et al. (2019). In particular, the latter proposed metrics to quantify replicability. These were termed "direct measures of consistency" and they were argued to provide advantages compared to customary approaches used in the network literature (e.g., the network comparison test; van Borkulo et al., 2016). I focus on their proposed method for detecting individual edges. The basic idea was to simply tally the presence and absence of each edge in estimated networks of the same construct. The results were then summarized by noting the number of edges that were detected in the respective networks:

> Looking at the four networks all together, a total of 114 edges were estimated (95% of the 120 possible edges) and only 39 (34.2%) were estimated consistently in all four networks. Further, five (4.4%) edges reversed in sign and 26 (22.8%) edges were estimated in only a single network (p. 11, Forbes et al., 2019).

This formed their impetus for arguing network models are unreliable. Although descriptives of this nature can provide some food for thought, their ultimate utility depends on the *expected* level of replicability: since they attempted to replicate hundreds of effects, consistently estimating 30% in four networks *could* be rather impressive!

It is commonplace in replication-oriented research to have a wide spectrum of perspectives–often arising from the same data. This also applies to the network literature. For example, using the same methodology and data as in Forbes et al. (2019), Fried et al. (2018) concluded that "Despite differences in culture, trauma type, and severity of the samples, considerable similarities emerged" (p. 335). Furthermore, a recent response to Forbes et al. (2019) highlighted the role of *naturally* occurring sampling variability. In other words, failing to detect an effect may not necessarily indicate a lack of replicability or inherent limitations of the model, but rather it could be *expected* (Jones et al., 2019). This provides the foundation from which this work is built: I provide a tool that allows for quantifying expected replicability for individual edges in psychological networks.

To answer the question of expected network replicability, an important distinction must be made between the *model* and the *method* used to estimate that model. I explicitly adopt a frequentist perspective and focus on long run error rates. This also forms the underlying logic of large scale replication efforts, where there is meticulous planning aimed towards controlling both type I and II errors (e.g., by following a pre-registration protocol; Lakens, 2019). Key to this endeavour is assuming that the employed *procedure* has a defined error rate. For example, what could be said about replicability if a larger sample and/or effect size resulted in an inflated false positive rate? The most popular approach for network estimation, which employs ℓ_1-regularization, has an ill-defined error rate that depends on just those factors (Figure 5 and 6 in Williams, Rhemtulla, Wysocki, & Rast, 2019). There is a related limitation that also contributes to ℓ_1-regularization being less than ideal for assessing replication: the sparsity and bias inducing properties result in a distorted sampling distribution. This presents challenges for computing valid confidence intervals and *p*-values (e.g., coverage probabilities equal to $1 - \alpha$ and a uniform distribution of *p*-values under the null hypothesis; Bühlmann, Kalisch, & Meier, 2014). The proposed methodology depends on these properties and thus a non-regularized *method* is employed to estimate the *model* (Williams & Rast, 2019; Williams et al., 2019).

This work is organized as follows. I first present a motivating example that highlights “issues” specific to assessing replicability in partial correlation networks. In the next section, I extend the motivating example and provide an analytic expression for computing expected network replicability. I then use the method to highlight additional sources that can reduce replicability, that is, when employing a multiple comparison correction and going from continuous to ordinal data with few categories. The next section discusses sample size planning in network replicability studies. I conclude with recommendations and limitations.

The Gaussian Graphical Model

For multivariate normal data, the Gaussian graphical model (GGM) captures conditional relationships that are typically visualized to infer the underlying dependence structure (i.e., the partial correlation "network"; Højsgaard, Edwards, & Lauritzen, 2012; Lauritzen, 1996). There is an undirected graph that is denoted $G = (V, E)$, which includes a vertex set $V = \{1, \ldots, p\}$ and an edge set $E \subset V \times V$. The former refers to "nodes" and the set represents, say, items in a questionnaire, whereas the latter contains the estimated network structure. Let $\mathbf{y} = (y_1, ..., y_p)^\top$ be a random vector indexed by the graphs vertices that is assumed to follow a multivariate normal distribution, $\mathbf{y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite covariance matrix. Further, without loss of information, the data are considered centered with mean vector 0. The undirected graph is obtained by determining which off-diagonal elements of the precision matrix, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, are non-zero. That is, (i, j) $\in E$ when node i and j are determined to be conditionally dependent and set to zero otherwise. Note that the edges (or "connections") in a GGM are non-zero partial correlations $\rho_{ij \cdot \boldsymbol{z}}$. These can be computed directly from $\boldsymbol{\Theta}$ as

$$\rho_{ij \cdot \boldsymbol{z}} = \frac{-\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}. \tag{1}$$

Hence, estimating partial correlation networks can be accomplished with classical hypothesis testing to determine which relations are "significantly" different from zero. This is described in Drton and Perlman (2004) and Williams and Rast (2019), both of which employed the Fisher-z transformation for the sample partial correlations. This transformation can be written as

$$\xi_{ij \cdot \boldsymbol{z}} = F(r_{ij \cdot \boldsymbol{z}}) = 0.5 \left[\log\left(\frac{1 + r_{ij \cdot \boldsymbol{z}}}{1 - r_{ij \cdot \boldsymbol{z}}} \right) \right], \tag{2}$$

where $\xi_{ij \cdot \boldsymbol{z}}$ denotes the Fisher transformed correlation. The resulting sampling distribution is (approximately) normal, which is one key advantage of employing (2). With $\xi_{ij \cdot \boldsymbol{z}}$ in hand, all that is needed to test

$$\begin{aligned} \mathcal{H}_0 &: \rho_{ij \cdot \boldsymbol{z}} = 0 \\ \mathcal{H}_1 &: \rho_{ij \cdot \boldsymbol{z}} \neq 0 \end{aligned} \tag{3}$$

is the sample size (n) and the number of nodes (p). This is a result of the standard error being defined as $\frac{1}{\sqrt{n-3-c}}$, where $c = (p-2)$ is the number of nodes conditioned on (i.e., those included in $\boldsymbol{z}$). This result is described in Fisher (1924). Wald confidence intervals can then be constructed to see whether zero is excluded (Williams & Rast, 2019). Together, determining E reduces to the case of null hypothesis significance testing with the customary z-statistic. This has far reaching implications for quantifying replication, in that key tenets (e.g., controlling type I and II errors) of replication efforts in psychological science can be applied to network analysis.

Partial versus Zero-Order Correlations

At this point, it is worth emphasizing an underlying perspective of this work: because partial correlation networks are relatively new to psychological science, contrasts to the more familiar case of the bivariate, or zero-order correlation, provides an informative reference point for thinking about replicability. Indeed, in my experience it is common to juxtapose association (bivariate) versus conditional dependence (partial) networks (see for example Forbes, Wright, Markon, & Krueger, 2020). The notion that bivariate and partial correlations are *not* "apples and oranges" is also reflected in a seminal work outlining the statistical foundations of multivariate analysis:

> ...all statistical inference procedures for the simple population

> correlation can be used for the partial correlation. The procedure for the partial correlation is exactly the same... (Anderson, 2003, pp. 143 - 144)

The one modification, as described in the text surrounding Equation (3), is that the degrees of freedom for a partial correlation reflects the number of variables in the network–otherwise the procedure is the same. There are, however, what are perhaps underappreciated distinctions that have a direct bearing upon evaluating network replicability. These are demonstrated below (Section Motivating Examples).

In Forbes et al. (2019), the GGM (and more generally "conditional independence networks") was suggested to have inherent limitations that contribute to unreliable findings compared to, say, zero-order correlations $\xi_{ij} = F(r_{ij})$ (a.k.a "association networks"). Indeed, it was specifically noted that "The edges in the association networks were by far the most replicable" (Forbes et al., 2017). The key difference is that zero-order correlations are not conditioned on any variables and partial correlations are conditioned on $p-2$ variables. Although not mentioned in Forbes et al. (2019), controlling for variables does present two hurdles for detecting edges, which is the key ingredient for evaluating replicability:

(1) For nodes i and j, the partial correlation $r_{ij\cdot\mathbf{z}}$ will necessarily have *more* sampling variability than the corresponding zero-order correlation r_{ij}. This a direct result of the standard error derivation, which, in the case of ξ_{ij}, is defined as $\frac{1}{\sqrt{n-3}}$. Because $c = (p-2)$ has been removed, the discrepancy between the two is directly related to the number of nodes in the network. That is, even when holding all else constant (e.g., the "population" value), the sampling distribution of a partial correlation will have more variance than a zero-order correlation.

(2) In psychological applications, it is reasonable to assume that $\rho_{ij\cdot\mathbf{z}}$ will often be *smaller* than ρ_{ij}. This is not guaranteed for the individual case, but especially

when considering all the relations in the network, the partial correlations will typically be (much) smaller. As an example, Figure 1 includes the distribution of sample zero-order and partial correlations computed from 16 post-traumatic stress symptoms (sample 1 from Fried et al., 2018).

Both factors necessarily result in a reduction in statistical power. I emphasize that this does not make GGMs inherently unreliable. Rather, if these two points are not considered, this can merely give the appearance of unreliability due to incurring many type II errors (i.e., in Equation 3 failing to reject $\mathcal{H}_0$ when $\mathcal{H}_1$ is true). But this could be (natural) sampling variability, which can be gleaned from adopting a frequentist perspective.

Motivating Examples

In most sections of this work, I assume that the data are continuous and normally distributed. Accordingly, I rely heavily upon the Pearson partial correlation which allows for estimating the graph and deriving expected network replicability with an analytic solution. Importantly, this does not limit the generality of this work, in that all of the ideas can seamlessly be applied to Spearman's rank (Kim, 2015), the so-called Gaussian rank estimator (i.e., based on Van Der Waerden scores, see references in Boudt, Cornelissen, Croux, & Boudt, 2012), and Kendall's tau based partial correlations (Johnson, 1979), each of which are commonly used in the Gaussian graphical modeling literature (Hoff, 2007; H. Liu, Han, Yuan, Lafferty, & Wasserman, 2012; Mohammadi & Wit, 2015a). This far reaching applicability is due to there being an analytic solution for each measure of (partial) association.

A Frequentist Perspective

Recall that the "direct measures" of Forbes et al. (2019) quantify network replicability by tallying success and failure for detecting the relations. This can be understood in reference to the frequentist testing strategy formalized by Neyman and Pearson (Neyman

& Pearson, 1933). In this framework, there is a binary loss function, that is, the so-called 0–1 loss, wherein a given test (e.g., Equation 3) results in either 1 (acceptance of $\mathcal{H}_0$) or 0 (rejection of $\mathcal{H}_0$). Note that the idea of a "loss function" has its roots in decision theory. Although a full discussion is beyond the scope of this work[1], it suffices to know that the associated risks in classical testing "are exactly the type I and type II errors underlying the Neyman-Pearson theory" (p. 81, Robert, 2007). These error rates are long-run probabilities. For example, when $\mathcal{H}_0$ is "true" the associated risk is α (the type I error rate) and when $\mathcal{H}_1$ is "true" the associated risk is the type II error rate, that is,

$$\beta = \Phi\Big(Z_{\alpha/2} - \frac{F(\rho_{ij\cdot z})}{1/\sqrt{(n-3-c)}}\Big), \tag{4}$$

where $Z_{\alpha/2}$ is the critical value for a two-sided significance test and Φ is the cumulative density of a standardized normal curve. Importantly, $1-\beta$ corresponds to statistical power. Furthermore, the only difference for zero-order correlations is c (i.e., $p-2$), such that, when holding all else constant in (4), power will *always* be lower when controlling for variables. Consequently, there will be more type II errors in partial correlation than association networks. Said another way, when deciding to classify (yes/no) whether an edge exists, there will be more "risk" associated with that decision. Again, this does not indicate that the GGM is unreliable, but, rather, there needs to be more thought about the role of sampling variability.

Example 1: Binomial Distribution

Natural sampling variability for detecting an edge (or more generally an "effect") can be modeled with the binomial distribution. Importantly, the binomial distribution has been used for purposes similar in spirit to investigating network replicability. For example,

[1] Interested readers are referred to Pratt (1977).

it provides the foundation for the "test of excessive significance" (Ioannidis & Trikalinos, 2007). The basic idea is to test for the presence of publication bias, as indicated by there being more successful trials (or "significant" results) than expected. On the other hand, Lakens and Etz (p. 875, 2017) used the binomial distribution to gain "A better understanding of the probability to observe mixed results...," which is the basic idea of this example.

Suppose that there are four network models under consideration. It is possible to compute the probability of observing a number of successes out of a given number of independent Bernoulli trials. The key is that each trial, or attempt to detect an edge, results in a success or failure which is essentially the "direct measures" introduced in Forbes et al. (2019). In this example, success is defined as a statistically "significant" result (assuming $\mathcal{H}_0$ is false). The desired probability can be computed as

$$Pr(s|\mathcal{A}) = \frac{\mathcal{A}!}{s!(\mathcal{A}-s)!}\,(1-\beta)^s(\beta)^{\mathcal{A}-s}, \tag{5}$$

where s is the number of successful replications in $\mathcal{A}$ attempts. Because $\mathcal{H}_1$ is assumed to be true, the probability of success is $1-\beta$ (Equation 4). Note that when $\mathcal{H}_0$ is "true," the probability of a successful trial that results in *incorrectly* rejecting $\mathcal{H}_0$ is the type I error rate (i.e., α). It is clear from Equations (4) and (5), that the probability of detecting a given edge in, say, three out of four networks (or all four) can be computed without simulation. This is advantageous for models, such as the GGM, that include many relations.

Procedure. To emphasize the first point (partial correlation have more sampling variance; section Partial versus Zero-Order Correlations), I investigated the influence of controlling for variables on replicability. This was computed analytically. The population value for both a zero-order ρ_{ij} and partial correlation $\rho_{ij\cdot\mathbf{z}}$ was set to 0.20. I then computed the probability of replicating that edge in each of two, four, and eight networks, assuming sample sizes ranging from 150 to 500 (increments of 1) and the number of control variables, c, ranging from 0 to 100 (increments of 20). Note that the type II error rate, β, in Equation (5) was computed by plugging the assumed values (e.g., c and n) into Equation (4).[2]

Results. Figure 1 (panel A) includes the results. Recall that the population value was held constant for both a partial and zero-order correlation. Hence, all that differed was the number of control variables, which, as discussed above, necessarily results in more sampling variability. This can be seen in each panel. The probability of replicating the edge diminished when the partial correlation network size increased. On the other hand, for a zero-order correlation, the number of nodes does not directly contribute to the sampling variance for computing the corresponding test-statistic.

These results also highlight the nature of attempting to replicate an effect several times. It is clear that, with each attempt, the probability of successful replication in all networks decreases with more attempts. For the zero-order correlation ($\rho = 0.20$), the probability of replicating the edge twice was nearly 0.50, whereas the probability of replicating the edge four times reduced to 0.20. Of course, replicability was even lower for the partial correlation. This discrepancy is attributed to natural sampling variability: replicability would be the same if the sample size for the partial correlation was merely increased by c.

Example 2: Simulation

In this example, the population value was not held constant. This elucidates point 2 (Section Partial versus Zero-Order Correlations), where I conjectured that a partial correlation will typically be smaller than the corresponding zero-order correlation. Furthermore, rather than focus on replicating one edge, simulation was used to investigate the case of replicating several edges.

[2] All computations in this work were done with the `R` programming language. Reproducible code can be found here.

Procedure. The true networks include four nodes in total and the sample size was set to 800. All six partial correlations were set to 0.10 (approximately 80 % power with $\alpha = 0.05$). This translates into a fully connected GGM with six edges. When solving for the zero-order correlations, this resulted in population values of 0.125 (≈ 0.94 power with $\alpha = 0.05$). I then tallied the number of replicated edges in two, three, and four networks. The probability that say, three edges, were replicated in two networks was computed as the proportion of the 5,000 simulation trials that this result occurred.

Results. Figure 1 (panel A) includes the results. Recall that the partial correlations were smaller than the zero-order correlations, in addition to the former having more sampling variability. This translated into striking differences between association and partial correlation networks. For example, even though there was 80 % power to detect each individual partial correlation, the probability of replicating more than three edges in four networks was around 0.25. On the other hand, the probability was around 0.80 for the association networks. It is clear that replicating an entire partial correlation network is improbable with sample sizes common to the network literature (see Table 1 in Wysocki & Rhemtulla, 2019).

Summary

These motivating examples were meant to provide intuition for thinking about partial network replicability. Natural, and expected sampling variability, can present challenges for the ambitious goal of replicating several edges. I emphasize that this does not suggest partial correlations lack replicability or that Gaussian graphical models are unreliable. Rather, we would *expect* to see mixed results that can ultimately be mitigated by properly planning for replicability.

Expected Network Replicability: Methodology

An important first step towards investigating network replicability is coming to terms with realistic expectations, given, say, the sample sizes or even the type of partial correlation. Accordingly, in this section, I extend the general idea of expected replicability to settings that closely mirror the partial correlation network literature. This requires overcoming two obstacles. First, the binomial distribution in Equation (5) assumes a constant success probability. This was accomplished by holding the population values constant in the motivating examples. In order to compute expected network replicability (ENR), however, this assumption must be relaxed to allow for varying edge sizes (or effect sizes). Second, simulation quickly becomes cumbersome when the networks are large and/or when ENR is desired for polychoric partial correlations (which would require computationally intensive bootstrapping). Hence, an analytic solution for computing ENR would be quite useful.

Poisson-Binomial Distribution

The binomial distribution provides a foundation from which to compute expected network replicability. This is because it is a special case of the Poisson-binomial distribution, wherein the success probabilities are assumed to be equal (Hong, 2013). This assumption can be relaxed, which affords varying edge weights and an analytic solution. With varying success probabilities, that corresponds to different effect sizes, a Poisson-binomial random variable, $\mathbf{P}_b$, is the sum of random indicators, that is,

$$\mathbf{P}_b = \sum_{e=1}^{n_e} I_e \tag{6}$$

$$\mathbf{I}_e \sim \text{Bernoulli}(1 - \beta_e),\ e = 1, .., n_e,$$

where n_e denotes the number of edges and $1 - \beta$ is the success probability, or statistical power, for the eth edge. Note that β is given in Equation (4). The key difference from Equation (5) is the subscript e, which effectively allows for varying success probabilities. In this context, this translates into varying power $(1 - \beta_e)$, that, in turn, permits varying edge weights in the *true* network. This leads to an analytic solution for computing ENR, in addition to the variance of $\mathbf{P}_b$, that is,

$$\begin{aligned}\mathrm{E}[\mathbf{P}_b] &= \sum_{e=1}^{n_e}(1-\beta_e) \\ \mathrm{Var}[\mathbf{P}_b] &= \sum_{e=1}^{n_e}\beta_e \cdot (1-\beta),\end{aligned} \tag{7}$$

where β_e is the type II error rate for the eth edge. In words, the expectation is simply the sum of statistical power for each edge in a given network. Note that this expectation and variance applies to the *number* and not the proportion of replicated edges. Furthermore, Equation 7 applies to one network and the extension to, say, three networks, entails using the probability of detecting a given edge in all networks under consideration.

It is also important to consider the cumulative density function (CDF) of the Poisson-binomial distribution. This allows for computing, say, the probability of replicating more than 80 % of the edges, as well as the probability of replicating between 80 and 100 % of the edges. However, directly evaluating the CDF is only feasible with few edges. A variety of methods have been developed to overcome this issue. In this work, I use the discrete Fourier transform (DFT) method that was described in Hong (2013) and implemented in the `R` package **poibin**.

A Note of Independent Trials. To this point, I have not discussed the notion of independent Bernoulli trials, which is assumed to be the case for both the binomial and the Poisson-binomial distributions. This was not an issue in the motivating examples, as the first focused on one edge in independent networks and the second used simulation. However, as suggested by a reviewer, edges within a network are *dependent*. Indeed, the issue of dependence can be thought of in at least three ways:

1. The most common form of dependence relates to sampling *without* replacement, whereas the binomial distribution assumes that sampling occurs *with* replacement. When this assumption is violated, the probability of success is not

the same for each trial which then requires the hyper-geometric distribution (Thomopoulos, 2017). This does not seem to apply to computing ENR.

2. An additional form of dependence relates to serial correlation between the trials. For example, in an experiment that has repeated trials, the outcome at trial t might be related to the outcome at $t - 1$. A variety of methods have been proposed to overcome this issue (see the references in Ladd, 1975). This also does not seem to apply to computing ENR.

3. The final form of dependence relates to possible covariance among the edges (and test statistics) themselves (Drton & Perlman, 2005), that is, the probability of success for a given edge might affect the probability of success of another edge. This does seem to present some potential challenges for deriving ENR analytically.

To my knowledge, there are no methods that address the issue of using dependent test-statistics with the Poisson-binomial distribution. However, it would be quite useful if this assumption has a negligible impact on computing the mean (and the distribution more generally). For example, it is possible to covert the expected number of "significant" edges into a proportion, that is, $n_e^{-1} \cdot \sum_{e=1}^{n_e}(1 - \beta_e), \ e = 1, \ldots, n_e$. In the case of one network, and assuming independence, this equivalent to *sensitivity* or the proportion of true edges that were detected. This insight would extend Williams and Rast (2019), where it was noted that $1 - \alpha$ (the false positive rate) is *specificity*. Together, by defining both α and β, this would lead to an analytic solution for both sensitivity and specificity, in addition to ENR.

Simulation. To determine the degree to which dependence matters, I compared the proposed derivation based on the Poisson-binomial distribution to simulation. To this end, I followed the approach in Hong (2013) and focused on the entire cumulative density function.

Procedure. Denote the CDF of $\mathbf{P}_b$ as $F_{\mathbf{P}_b}(k) = \Pr(\mathbf{P}_b \leq k), k = 0, 1, \ldots, n_e$, that gives the probability of having less than or equal to k successes out of the total number of edges. The mean absolute error (MAE) was computed as

$$\text{MAE} = n_e^{-1} \cdot \sum_{k=0}^{n_e} |F_{sim}(k) - F_{DFT}(k)| \tag{8}$$

where $F_{sim}(k)$ is the simulation based CDF and $F_{DFT}(k)$ is the discrete Fourier transformation based CDF. To gain insight into the worst-case scenario, the maximum difference between the CDFs was also computed. This simulation included one, two, and four networks, with three network sizes of 10, 20, and 30 nodes and sample sizes of 100, 250, 500, and 1,000. Additionally, three network structures were examined including random, cluster, and scale-free (see pg 17 in Mohammadi & Wit, 2015b). The probability of replicating a given edge was set to $(1 - \beta_e)^{n_n}$ (Equation 7, $\alpha = 0.05$), where n_n is the number of networks. The graphs were generated with the `R` package **BDgraph** using the default settings (Mohammadi & Wit, 2015b) and $F_{sim}(k)$ was obtained from 10,000 iterations.

Results. Figure 3 includes these results. The MAE (panel A) indicates that the analytical derivation is accurate compared to simulation. Recall that the former assumes the trials are independent (when they are not). The MAE was largest for the smallest sample size or when the p/n ratio was not sufficiently small. And even then, the error does not seem large. This quickly dissipated with largest sample size, where the MAE was below 0.0025 (or less than a probability of 1%) for all networks sizes and structures. The maximum difference also indicates that the proposed method is accurate. For example, with $n = 250$, the *maximum* difference between CDFs was around 0.04. This translates into suggesting the probability of replicating more than a specified number of edges is 0.70 when it is closer to 0.66. This again diminished with larger sample sizes to ultimately be negligible for all network structures. Although not included here, note that estimating the mean (i.e., the expected number of replicated edges) was also very accurate.

Summary

The proposed method becomes quite accurate with large sample sizes, although, in my opinion, the error also does not seem problematic with small sample sizes. Importantly, the question of network replicability inherently requires large sample sizes to answer. This can be inferred from Figure 2 (panel B): replicating only six edges in four networks was nearly impossible, even though there was 80% power to detect each individual edge. Accordingly, I argue that the Poisson-binomial distribution can be used to model network replicability. As an added bonus, it is possible to analytically derive sensitivity for the case of one network. These are major contributions to the network literature.

Expected Network Replicability: Application

In this section, ENR was computed for a variety of datasets from the network literature. When ENR is used to evaluate replicability (or lack thereof) in practical applications, a pressing question is whether to correct for multiple comparisons. This is because determining the edge set, E, requires potentially hundreds of tests. One approach to address is to employ a Bonferroni corrected alpha level (e.g., Drton & Perlman, 2004). However, the motivating example points to a tradeoff between reducing spurious associations and replicability. On the one hand, a stringent alpha level will necessarily reduce statistical power and thus replicability. On the other hand, when tallying replications, it becomes increasingly unlikely that a type I error will emerge in each replication attempt. To investigate this tradeoff, the following includes ENR for both an uncorrected alpha level and a Bonferroni correction for multiple comparisons.

Illustrative Data

Dataset 1. This dataset come from the `R` package **psych** (Revelle, 2019). There are 25 self-reported items ($p = 25$) that measured personality in 2236 subjects. The items (scored from one to six) were taken from the international personality item pool. The majority of subjects were between 20 and 60 years old (M = 29.5 years, SD = 10.6 years). 68% were females.

Dataset 2. This dataset comes from a network study using the Post-traumatic Stress Disorder Checklist (PCL; $p = 20$) for DSM–5 in 221 United States military veterans (Armour, Fried, Deserno, Tsai, & Pietrzak, 2017). The items are scored from zero to five. Further details about the PCL can be found in Davis (1983). The subjects were between 21 and 89 years old (M = 54.0 years, SD = 14.68 years). 86.7 % were male.

Dataset 3. The final dataset was used in Forbes et al. (2019) to argue that networks lack replicability. This dataset comes from a study using the Patient Health Questionnaire (PHH-Q) to measure depression with nine items and the Brief Measure for Assessing Generalized Anxiety Disorder (GAD-7) that includes seven items (a total of 16 nodes). The items were scored from 0 to 3. Note that there are two waves of data and I use the first wave. The subjects had an average age of 32 (SD = 12) and 78.2 % of the 403 participants were female.

Procedure

For each dataset, ENR was computed with the following steps:

1. Estimate the partial correlation matrix and then set absolute values less than 0.05 to zero (Epskamp, 2016; Williams et al., 2019). This served as the true network structure.

2. Compute statistical power, with sample sizes of 500, 1000, and 2,500, for each edge as $1 - \beta_e$ (Equation 4). Note that the number of control variables, c, was set to $p - 2$. The uncorrected and Bonferroni corrected alpha levels were set to 0.05 and $\alpha = 0.05/n_t$, respectively. n_t denotes the number of tests $\frac{1}{2}p(p-1)$

3. With statistical power in hand, compute the probability mass function of a Poisson-binomial distribution for $k = 0, 1, \ldots, n_e$, where k is the number of successful replications and n_e is the number of edges. Note that I made the

simplifying assumption of equal sample sizes in 2 and 4 networks. Hence the power for replicating a given edge was set to $(1 - \beta_e)^{n_n}$ (Equation 7), where n_n is the number of networks.

Results

Figure 4 includes the estimated ENR. Before describing the results, note that the densities can most easily be interpreted in reference to the x-axis. For example, if a density is completely below 50% of the edges, this indicates that there is essentially no chance of replicating more than half of the edges. Also, when the density is centered around, say, 75%, this indicates that there is a 50% chance of replicating more (less) than 75% of the edges.

These results highlight a central aspect of this paper: natural sampling variability present in partial correlations networks can give the appearance of unreliability. With $n = 500$ and four networks, for example, the probability of replicating more than 50% of the edges was essentially zero. In fact, the probability density was mostly contained within 25% and 50% which indicates that ENR is in that range. This generalized to each dataset. Hence, it seems reasonable to assume that, with $n = 500$ and four networks, ENR will be below 50%.

To replicate more than 75% of the edges, again in four networks, this seems to require sample sizes of at least 2,000. Importantly, this is much smaller than the sample sizes used to investigate replicability in Forbes et al. (2019), where the largest was 956. These estimates for ENR were again consistent across the datasets. This suggests that all previous investigations of network replicability have not been adequately powered.

Correcting for multiple comparisons had a striking effect on ENR. For example, without correction for multiple comparison ($n = 500$ and four networks), we would expect to replicate more than 25 % of the edges, whereas, with a Bonferroni correction, we expect to replicate less than 25% of the edges. Indeed, for dataset 1 the most likely value was to replicate 32% of the edges for the uncorrected and less than 10% for the corrected. In other words, by avoiding a multiple comparison correction, this resulted in a three-fold increase in ENR.

Finally, these results are also consistent with the motivating examples (i.e., Figure 1, panel B). Replicability necessarily decreases when there are more networks under consideration. For example, replicability was much higher for two networks than four networks. This is a direct result of the probability of something happening many times in a row, say, detecting a given edge, becoming less likely with each attempt. This of course was also influenced by the Bonferroni adjustment, such that ENR will decrease with more networks and substantially so when using a corrected alpha level.

Planning for Network Replication

As demonstrated throughout this work, network replicability requires careful consideration of type I and type II error rates. An important aspect of the proposed methodology is that it can be used to explicitly plan for replicating a network. At first blush, this may seem to be an overly ambitious goal, as it entails defining the magnitude of potentially hundreds of edges in the *true* network. To overcome this issue, an alternative strategy is to define the smallest effect (edge) size of interest (SESOI, Lakens, 2014). The basic idea is to consider the minimal edge size deemed important and thus worth replicating. And then ensuring that there is adequate power to detect the SESOI.

To utilize the SESOI within the ENR framework, it first needs to be reframed in terms of the Poisson-binomial distribution. Recall that the expected number of replicated edges is the sum of statistical power for those edges (Equation 7). When expressed as a proportion of the number of edges, as noted above (Section Expected Network Replicability: Methodology), this is approximately equal to the average power. Hence, it follows that statistical power for the SESOI provides a lower bound for replicability. This is because, after all, there will be more power for edges larger than the SESOI.[3] In turn, this translates into the proportion of replicated edges being larger than power for detecting the SESOI. Hence, by defining the SESOI, in addition to α and β, the required sample size for achieving a desired level of replicability can be computed analytically.

Sample Size Determination. The sample size needed to detect the SESOI in one network is given as

$$N = c + 3 + \left(\frac{Z_{\alpha/2} + Z_{\beta_{\text{SESOI}}}}{F(\rho_{\text{SESOI}})} \right)^2, \tag{9}$$

where $Z_{\alpha/2}$ is the quantile of a standard normal distribution that corresponds to the critical value, $Z_{\beta_{\text{SESOI}}}$ is the upper quantile of a standard normal distribution that corresponds to the desired type II error rate, c is the number of control variables ($p - 2$), and $F(\rho_{\text{SESOI}})$ is the SESOI. The extension beyond one network then requires computing β_{SESOI}, such that there is, say, a probability of 0.80 for detecting the SESOI in any number of networks. This can be computed as (1 - $\beta_{\text{SESOI}})^{1/n_n}$, where n_n is the number of networks.[4] Together, this provides the required sample size needed to obtain a defined lower bound for replicability.

Illustrative Example

Procedure. To demonstrate the utility of defining a SESOI in the context of replicability, I sampled 60 edges from a uniform distribution with a lower and upper bound of 0.10 and 0.15, respectively. This mimics setting $\rho_{\text{SESOI}} = 0.10$ in Equation 9. The power $(1 - \beta_{\text{SESOI}})$ for detecting the SESOI in one network was initially to 0.70, 0.80, 0.90. Next, to ensure that there was the same power for detecting the SESOI in two networks (assuming equal sample sizes), each value was then set to $(1 - \beta_{\text{SESOI}})^{1/2}$. Plugging these values into Equation 9 resulted in sample sizes of 880, 1045, and 1303 ($\alpha = 0.05$) that were

[3] This assumes that there are edges larger than the SESOI in the *true* network.

[4] To have a 80 % success probability of detecting ρ_{SESOI} in two networks, for example, statistical power is set to $0.80^{1/2} \approx 0.894$ (conversely $0.894^2 \approx 0.80$).

used to compute ENR, that is, the probability mass function for a Poisson-binomial distribution.

Results. Figure 5 includes these results, where it can be seen that the proposed approach provides a lower bound for replicability. This can be inferred by noting that the distributions are contained within the shaded region, which corresponds to replicating more edges than statistical power for ρ_{SESOI}. Hence, for assessing replicability in two networks, the sample size needs to be rather large for replicating most edges greater than, in this case, 0.10 (what is often consider a small effect size). The sample sizes would need to be larger (smaller) if the SESOI was smaller (larger). Furthermore, as shown in Figure 4, employing a multiple comparison correction would entail even large sample sizes. For example, if the goal was to replicate 70% of the edges with a Bonferroni correction, the sample size would need to be over 1,800 (as opposed to $n = 880$ for an uncorrected α level of 0.05).

Discussion

In this work, I tackled the important topic of replicability in partial correlation networks. I highlighted *expected* decreases in replicability, wherein power to detect (and thus replicate) the edges diminishes. For example, when moving from a zero-order to the corresponding partial correlation, the latter is typically smaller and there is an increase in sampling variability that is a function of the number of nodes in the network. Furthermore, I demonstrated that applying a multiple comparisons corrections can substantially impact replicability. Together, this points translate into reduced statistical power that can give the mere appearance of unreliability when they are not considered. This does not imply partial correlation networks are inherently unreliable, as these “issues” can be addressed in the design phase of future studies investigating network replicability.

Moreover, I introduced novel methodology that allows for quantifying expected network replicability (ENR). This can be used for two purposes. First, computing ENR can provide realistic expectations about the number of edges that can be replicated, given the number of nodes, sample size, edge weights, and type of data. This provides a meaningful reference point that, to date, is absent from the literature. Second, ENR can be used to carefully design studies with the goal of investigating network replicability. For this purpose, I showed that defining the smallest edge size of interest (SESOI) can be used to ensure an adequate number of edges are replicated. As a bonus, this methodology can be used to perform power analyses for one network (e.g., Figure 3).

An additional contribution of this work is that I merged the topic of network replicability with replication-oriented research in the social-behavioral sciences. As I argued, controlling type I errors and minimizing type II errors is of central importance for investigating replication. In other words, in my view, it is not possible to argue that networks lack replicability without considering statistical power. Yet, in the network literature, methodology (ℓ_1- regularization) is regularly used that does not have defined error rates. Accordingly, none of the calculations presented in this work are possible with regularized partial correlations. I demonstrated that utilizing statistical methods that have controlled error rates not only provides invaluable insights, but it also allows for weaving concepts, such as the SESOI, into the very fabric of network psychometrics.

Are Partial Correlations Unreliable?

Gaussian graphical models do not have inherent limitations that make them unreliable. Rather, as this work demonstrated, there is simply more sampling variability than zero-order correlations. This is baked right into the mathematical expression for the Fisher-z standard error. This is exacerbated by the fact that partial correlation networks include many small effects (Figure 1). Hence, as a result of sampling variability in combination with small effect sizes, very large sample sizes are needed to minimize the type II error rate. This does not make GGMs unreliable, as this is easily overcome by explicitly planning for replicability. To my knowledge, however, a concerted effort has never been made to plan for replicating a psychological network. The introduced methodology can be used to either explicitly plan for replicability or to provide a reference point for expected network replicability.

Furthermore, Forbes et al. (2020) recently noted that

> To see whether statistical control is increasing the reliability of our parameter estimates here, we can examine whether the confidence intervals of the parameter estimates are narrower or wider when comparing the unpartialled (zero-order) correlation for each symptom pair to the fully partialled correlation that underlies the edge weight...The confidence intervals for the partial correlations were 20-21% wider on average, and 89-92 % of the confidence intervals were wider for the partial correlation, compared to the corresponding zero-order correlation's at each wave.

This was taken to be evidence that partial correlations are unreliable, especially compared to zero-correlations. However, this is quite confusing because we already know that, by definition, partial correlations will have a wider sampling distribution (holding all else constant). This is a mathematical result that goes back nearly a century and it was highlighted in the motivating example. Rather than an inherent deficiency, this information can be harnessed to properly plan for replicability. In fact, rather than focus on type I and II error rates, an alternative approach would focus on the sample size required to achieve a desired confidence interval width (e.g., Rothman & Greenland, 2018). This is again easily computed by plugging numbers into the Fisher-z based standard error (i.e., $1\sqrt{n-3-c}$; $c = p - 2$).

Although I disagree with the evidence Forbes et al. (2019) used to argue networks are unreliable, an alternative interpretation of their results does provide valuable information. In my view, they unknowingly highlighted the reality of estimating models with many effects, which, in the case of partial correlation networks, are expected to be small. The importance of this cannot be understated, as network replicability does not entail detecting one small effect, but to detect many edges *simultaneously*. It could be argued that the inferences are therefore unreliable Forbes et al. (see for example 2020). However, meaningful substantive interpretation depends upon considering sample to sample variability. Rather than this being some deficiency of GGMs, it suggests that researchers should be careful not to think they have actually estimated *the network*, especially when the sample size is not large.

A Note on Interpretation

To ensure ENR is properly interpreted, it is important to keep in mind the following. The term "probability" was used often in this work. Although this usage is correct, it explicitly refers to a *long-run* probability and it does not apply to an individual case. Consider, for example, an estimate of ENR suggesting that there is a 0.20 probability of replicating more than half of the edges. This means that in, say, 100 replication attempts, more than half of the edges will be replicated roughly 20 times. In other word, frequentist probability refers to a proportion that arises with (hypothetical) repeated sampling.

Alternative Definitions of Replicability

This work focused exclusively on replicating individual edges or "edge recovery." However, it is important to note that there are alternative ways to define replicability. For example, replicating the rank ordering of edge weights or centrality indices may be of interest. Furthermore, in Borsboom (2017), correlations between edge sets were examined. Although extending the proposed method in those directions would be challenging, the more general idea of considering sampling variability can indirectly shed light upon replicability. Consider the centrality index strength, that is the *sum* of edge weights for each node. The very nature of summing partial correlations can substantially increase sampling variability. Hence, it could turn out that replicating, say, the rank order of strength, requires even larger sample sizes than detecting the individual edges.

What Other Factors May Affect Replicability?

It is important to note that this work focused exclusively on the role of sampling variability. The underlying assumption is that nothing else affects network replicability. In other words, *all* assumptions of the proposed model are satisfied. However, in reality, there is assuredly some degree of measurement error which is known to attenuate correlations. Furthermore, the samples could be comprised of various sub-populations, resulting in heterogeneity. In both cases, ENR will likely be overly optimistic but it still provides a useful reference point for thinking about replicability and sampling variability. These factors can be mitigated by making a concerted effort to satisfy the assumptions, including attention to scale reliability and clearly defining the population.

Extending the Methodology

The presented methodology can be extended in several ways. First, it is common to estimate polychoric partial correlations in the psychological literature. This is perhaps the most straightforward extension, as all that is needed to compute ENR is an estimate of the standard error. This can be obtained from bootstrapping. This extension is described in the Appendix. Second, it would be useful to compute ENR for the Ising model (Marsman et al., 2017). Because the Ising model is commonly estimated with logistic regression, power analyses for multiple logistic regression could potentially be used to compute ENR. The idea of defining the SESOI has also been used for the Ising model.

Limitations

There are several limitations. First, although the assumption of independent trials did not seem to matter all that much, a more rigorous proof was not provided. It would thus be premature to assume that the Poisson-binomial distribution can be used beyond partial correlation networks. Furthermore, what did seem to matter was the number of nodes in the network, especially when the sample size was small (i.e., the p/n ratio).

Caution is therefore warranted in these situations, although it is important to note they are not commonly encountered (see Table 1 in Wysocki & Rhemtulla, 2019). Second, I only considered ordinal data in which each category had an equal probability. The results can thus be considered a best case scenario and hence replicability could be much lower in heavily skewed data wherein one or two levels are inflated (a further loss of information). In the package **GGMnonreg**, it is possible to simulate the sampling distribution for skewed ordinal data. Third, although not the primary focus of this work, I noted the presented methodology can be used for rank partial correlations that are sometimes used in the network literature. However, the partial Spearman correlation has been critiqued as being "ad-hoc" and there is some indication that it does not always correspond to conditional dependence (Gripenberg, 1992; Q. Liu, Li, Wanga, & Shepherd, 2018).

Recommendations

I recommend that network researchers rely more heavily on frequentist reasoning and inference, especially in the context of replicability. As I demonstrated, there are straightforward analytic expressions for computing ENR that are a direct result of considering error rates. In network analysis, researchers should start defining the smallest edge size of interest (i.e., the lower bound of an "important" effect). This allows for determining the sample size required to achieve a desired level of replicability. And this also would greatly improve network analysis in general. I did not discuss type I errors much in this work. Based on the results in Figure 4, I encourage network researchers to avoid arbitrarily applying a harsh multiple comparisons correction. Note that making *at least* one false positive is essentially guaranteed with so many effects. I argue that this is not necessarily a problem, as we also know the expected number of false positives. With, say, 20 nodes, 50 % connectivity, and $\alpha = 0.05$, we would expect roughly five false positives (on average). And observing the same false positive in two networks would be rare (i.e., $0.05^2 = 0.0025$). The results also point towards employing psychometric scales specifically with partial correlation networks in mind. For example, if there are two potential scales power (and thus replicability) can be maximized by choosing the one with the most ordinal categories (Figure A1). Together, these recommendations share a common theme–there needs to be more planning in network analysis.

Conclusion

I introduced a novel method for quantifying network replicability. In several examples, it was used to highlight sources of underappreciated, natural sampling variability, that are inherent to partial correlation networks. The methodology and ideas discussed in this work can provide a foundation from which to build network replication projects. The method for computing expected network replicability implemented in the `R` package **GGMnonreg**.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis.* London: Wiley-Interscience.

Armour, C., Fried, E. I., Deserno, M. K., Tsai, J., & Pietrzak, R. H. (2017). A network analysis of DSM-5 posttraumatic stress disorder symptoms and correlates in U.S. military veterans. *Journal of Anxiety Disorders*, *45*(May 2013), 49–59. doi: 10.1016/j.janxdis.2016.11.008

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. doi: 10.1002/wps.20375

Borsboom, D., & Cramer, A. O. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. doi: 10.1146/annurev-clinpsy-050212-185608

Boudt, K., Cornelissen, J., Croux, C., & Boudt, K. (2012). The Gaussian rank correlation estimator: robustness properties. *Stat Comput*, *22*(2), 471–483. doi: 10.1007/s11222-011-9237-0

Bühlmann, P., Kalisch, M., & Meier, L. (2014). High-Dimensional Statistics with a View Toward Applications in Biology. *Annual Review of Statistics and Its Application*, *1*(1), 255–278. doi: 10.1146/annurev-statistics-022513-115545

Davis, M. H. (1983). A Multidimensional Approach to Individual Differences in Empathy. *JSAS Catalog of Selected Documents in Psychology*.

Drton, M., & Perlman, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika*, *91*(3), 591–602. doi: 10.1093/biomet/91.3.591

Drton, M., & Perlman, M. D. (2005). Multiple Testing and Error Control in Gaussian Graphical Model Selection. , *22*(3), 430–449. doi: 10.1214/088342307000000113

Efron, B., & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors , Confidence Intervals , and Other Measures of Statistical Accuracy. *Statistical Science*, *1*(1), 54–75.

Epskamp, S. (2016). Regularized Gaussian Psychological Networks: Brief Report on the Performance of Extended BIC Model Selection. *arXiv*, 1–6.

Fisher, R. (1924). *The distribution of the partial correlation coefficient.* (Vol. 3).

Forbes, M. K., Wright, A. G., Markon, K., & Krueger, R. (2020). On unreplicable inferences in psychopathology symptom networks and the importance of unreliable parameter estimates. *psyarxiv.* doi: 10.31234/OSF.IO/BVU84

Forbes, M. K., Wright, A. G., Markon, K. E., & Krueger, R. F. (2017). Evidence that psychopathology symptom networks have limited replicability. *Journal of Abnormal Psychology*, *126*(7), 969–988. doi: 10.1037/abn0000276

Forbes, M. K., Wright, A. G., Markon, K. E., & Krueger, R. F. (2019). Quantifying the Reliability and Replicability of Psychopathology Network Characteristics. *Multivariate Behavioral Research.* doi: 10.1080/00273171.2019.1616526

Fried, E. I., Eidhof, M. B., Palic, S., Costantini, G., Dijk, H. M. H.-v., 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*(3), 335–351. doi: 10.1177/2167702617745092

Gelman, A., & Loken, E. (2014). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Psychological bulletin*, *140*(5), 1272–1280. doi: dx.doi.org/10.1037/a0037714

Gripenberg, G. (1992, 6). Confidence Intervals for Partial Rank Correlations. *Journal of the American Statistical Association*, *87*(418), 546. doi: 10.2307/2290289

Hoff, P. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, *1*(1), 265–283. doi: 10.1214/07-aoas107

Højsgaard, S., Edwards, D., & Lauritzen, S. (2012). *Graphical Models with R.* doi: 10.1007/978-1-4614-2299-0

Hong, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics and Data Analysis*, *59*(1), 41–51. doi: 10.1016/j.csda.2012.10.006

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245–253. doi: 10.1177/1740774507079441

Johnson, N. S. (1979). Nonnull Properties of Kendall's Partial Rank Correlation Coefficient. *Biometrika*, *66*(2), 333. doi: 10.2307/2335667

Jones, P. J., Heeren, A., & McNally, R. J. (2017). Commentary: A network theory of mental disorders. *Frontiers in Psychology*, *8*, 1305. doi: 10.3389/fpsyg.2017.01305

Jones, P. J., Williams, D. R., & McNally, R. J. (2019). Sampling Variability is not Nonreplication: A Bayesian Reanalysis of Forbes, Wright, Markon, &amp; Krueger. *PsyArXiv*. doi: 10.31234/OSF.IO/EGWFJ

Kim, S. (2015). ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications*, *22*, 665–674. doi: 10.5351/CSAM.2015.22.6.665

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Stepan, B., Bernstein, M. J., … Nosek, B. A. (2014, 5). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. doi: 10.1027/1864-9335/a000178

Ladd, D. W. (1975). An algorithm for the binomial distribution with dependent trials. *Journal of the American Statistical Association*, *70*(350), 333–340. doi: 10.1080/01621459.1975.10479867

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. doi: 10.1002/ejsp.2023

Lakens, D. (2019). The Value of Preregistration for Psychological Science: A Conceptual Analysis. *PsyArXiv*, 1–14.

Lakens, D., & Etz, A. J. (2017, 11). Too True to be Bad. *Social Psychological and*

Personality Science, *8*(8), 875–881. doi: 10.1177/1948550617693058

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

Liu, H., Han, F., Yuan, M., Lafferty, J., & Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, *40*(4), 2293–2326. doi: 10.1214/12-aos1037

Liu, Q., Li, C., Wanga, V., & Shepherd, B. E. (2018). Covariate-adjusted Spearman's rank correlation with probability-scale residuals. *Biometrics*, *74*(2), 595–605. doi: 10.1111/biom.12812

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., … Maris, G. (2017). An Introduction to Network Psychometrics: Relating Ising Network Models to Item Response Theory Models. *Taylor & Francis*, *53*(1), 15–35. doi: 10.1080/00273171.2017.1379379

McNally, R. J., Robinaugh, D. J., Wu, G. W. Y., Wang, L., Deserno, M. K., & Borsboom, D. (2015). Mental Disorders as Causal Systems. *Clinical Psychological Science*, *3*(6), 836–849. doi: 10.1177/2167702614553230

Mohammadi, A., & Wit, E. C. (2015a). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, *10*(1), 109–138. doi: 10.1214/14-BA889

Mohammadi, A., & Wit, E. C. (2015b, 1). BDgraph: An R Package for Bayesian Structure Learning in Graphical Models.
doi: 10.1359/JBMR.0301229

Mullarkey, M. C., Marchetti, I., & Beevers, C. G. (2018). Using Network Analysis to Identify Central Symptoms of Adolescent Depression. *Journal of Clinical Child & Adolescent Psychology*, *0*(0), 1–13. doi: 10.1080/15374416.2018.1437735

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, *69*(1), 511–534. doi: 10.1146/annurev-psych-122216-011836

Neyman, J., & Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical*

Society, *29*(4), 492–510. doi: 10.1017/S030500410001152X

Pratt, J. W. (1977). Decisions as statistical evidence and Birnbaum's 'confidence concept'. *Synthese*, *36*(1), 59–69. doi: 10.1007/BF00485692

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Evanston, Illinois. Retrieved from https://cran.r-project.org/package=psych

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. doi: 10.1037/a0029315

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. doi: 10.1016/j.drugalcdep.2016.02.005

Robert, C. (2007). The Bayesian choice: from decision-theoretic foundations to computational implementation. , 77 - 81.

Rothman, K., & Greenland, S. (2018). Planning study size based on precision rather than power. *Epidemiology*, *29*(5), 599–603.

Thomopoulos, N. T. (2017). Hyper Geometric. In *Statistical distributions* (pp. 149–152). Switzerland: Springer International Publishing. doi: 10.1007/978-3-319-65112-5{_}18

van Borkulo, C. D., Boschloo, L., Kossakowski, J. J., Tio, P., Schoevers, R. A., Borsboom, D., & Waldorp, L. J. (2016). Comparing network structures on three aspects: A permutation test. *Manuscript submitted*(March), 34. doi: 10.13140/RG.2.2.29455.38569

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on psychological science : a journal of the Association for Psychological Science*, *13*(4), 411–417. doi: 10.1177/1745691617751884

Williams, D. R., & Rast, P. (2019). Back to the basics: Rethinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology*. doi: 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*(5), 1–23. doi: 10.1080/00273171.2019.1575716

Wysocki, A. C., & Rhemtulla, M. (2019). On Penalty Parameter Selection for Estimating Network Models. *Multivariate Behavioral Research*, 1–15. doi: 10.1080/00273171.2019.1672516

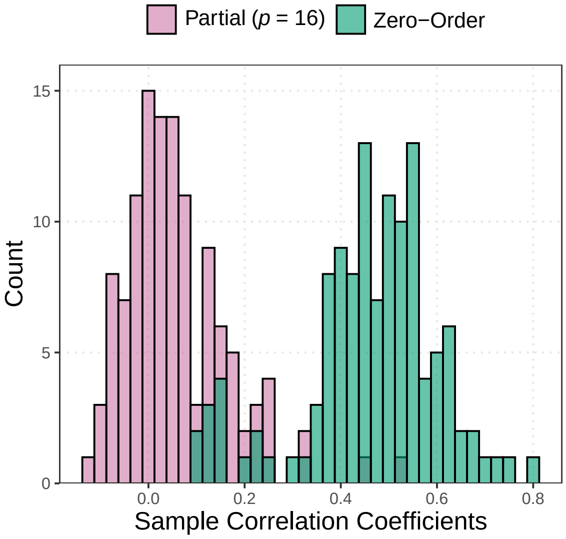

Figure 1. Sample correlation coefficients estimated from 16 post-traumatic stress disorder symptoms (sample 1 from Fried et al., 2018). This juxtaposition highlights an important distinction between partial and zero-order (or bivariate) correlations: the former will typically be much smaller in effect size. This has a direct bearing on partial correlation network replication, in that, when tallying the number of successful replications, mixed results are *expected* because simultaneously detecting hundreds of small effects is no small feat.

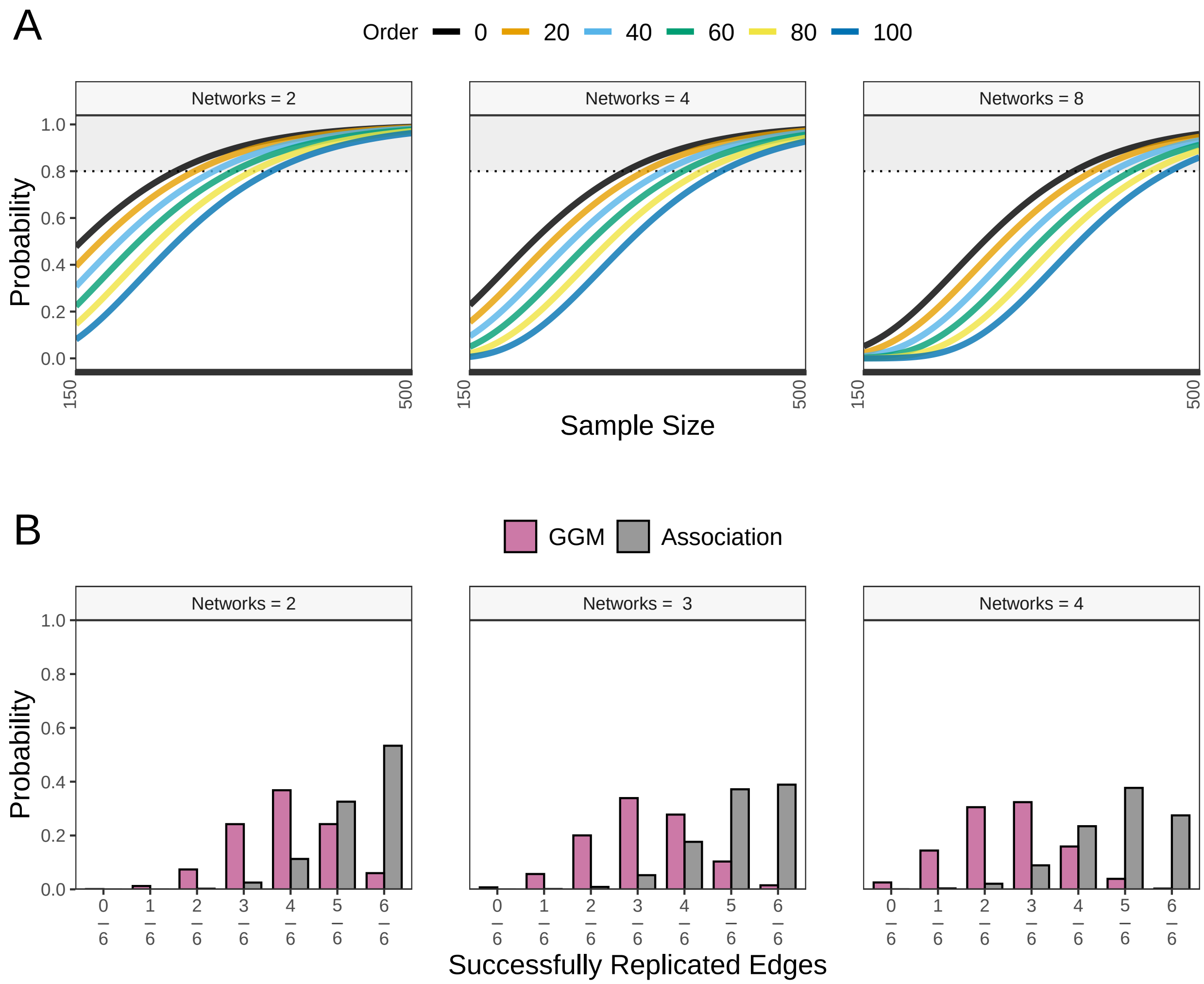

Figure 2. A) The probability of successfully replicating one edge in two, four, and eight networks. The order refers to the number of nodes that are conditioned on. Hence the black line corresponds to a zero-order (or bivariate) correlation. These results reveal that (1) the network replicability decreases with more control variables (i.e., replicating an effect becomes increasingly difficult in larger networks) and (2) replicating the edge in each network becomes increasingly difficult with more replication attempts (i.e., detecting the edge in four out of four networks is less likely than detecting the edge in two out of two networks); B) B) The probability of successfully replicating a given number of edges in a network with four nodes (six relations in total). These results reveal that (1) replicating six edges is improbable, especially in GGMs. For example, even though there was 80 % power to detect each edge, the probability of replicating more than three edges was merely 0.20 in four networks. (2) the probability of replicating most edges is much higher for association than partial correlation networks.

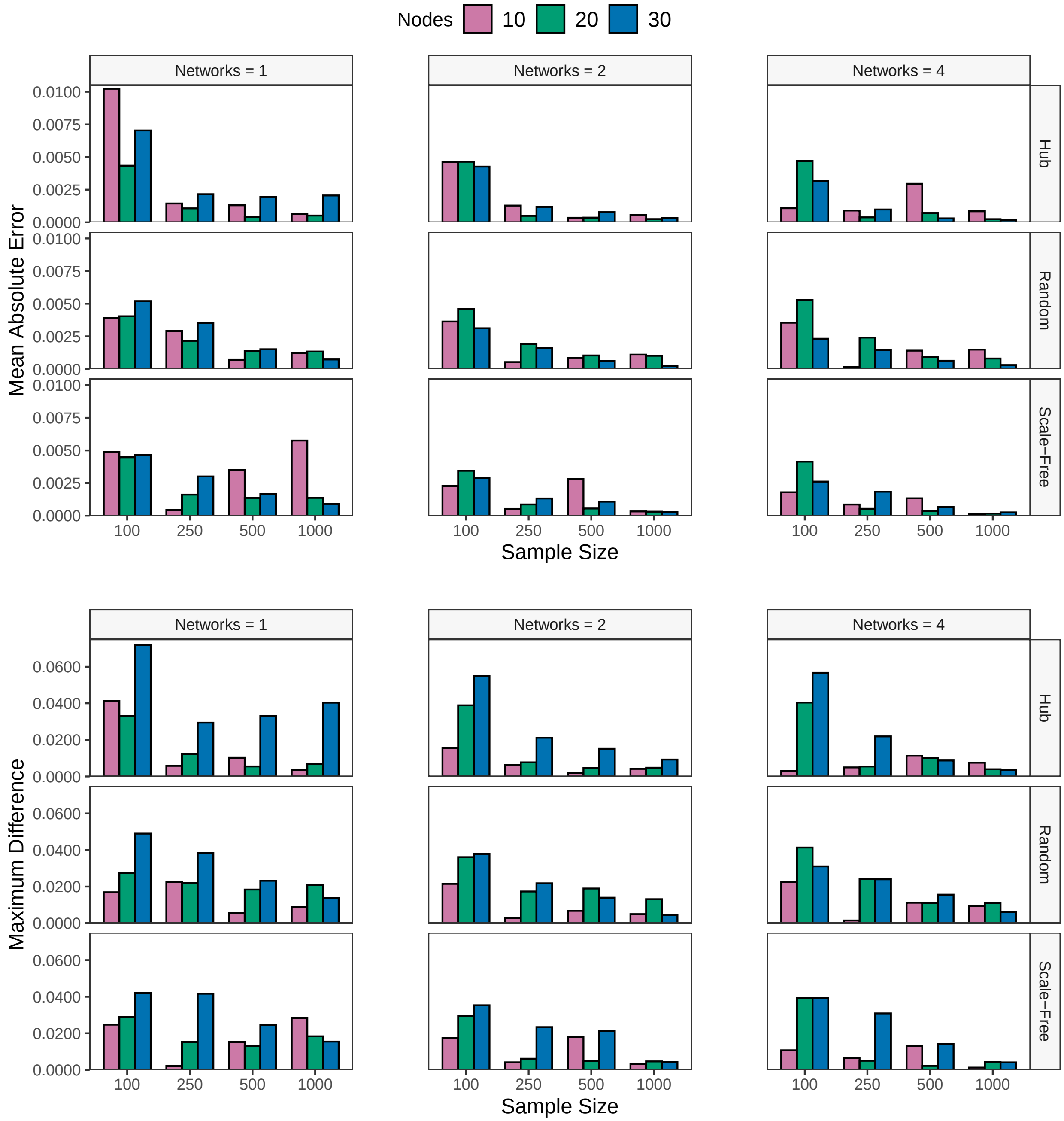

Figure 3. Simulation results (Section Expected Network Replicability: Methodology). The mean absolute error (MAE) is the discrepancy between cumulative density functions (CDFs) estimate with simulation versus the proposed analytic solution. The maximum difference is the *largest* error between the CDFs.

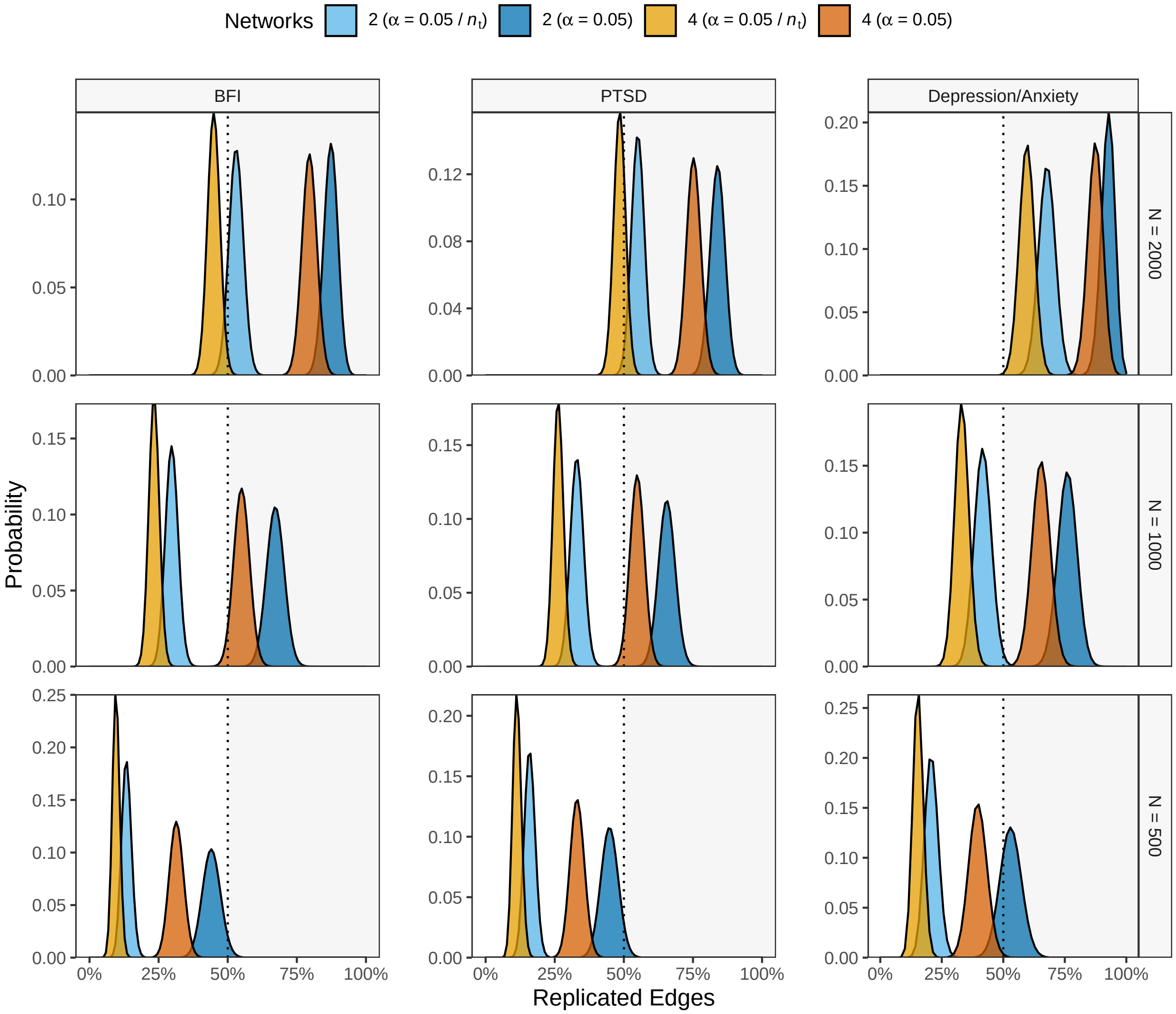

Figure 4. Expected network replicability. The densities are probability mass functions for the Poisson-binomial distribution, given the statistical power $(1 - \beta_e)$ to detect the edges in two and four networks. The y-axis is the probability of replicating a given proportion of edges and the x-axis is the proportion of replicated edges. The shaded area corresponds to more than half of the edges. Hence, if the density is completely in that region, the probability of replicating more than 50 % of the edges is 1.0. Note that the Poisson-binomial distribution is for the number of edges and the proportion was computed by dividing by the number of edges in the respective networks and n_t is the number of tests $(0.05/n_t$ is the corrected alpha level).

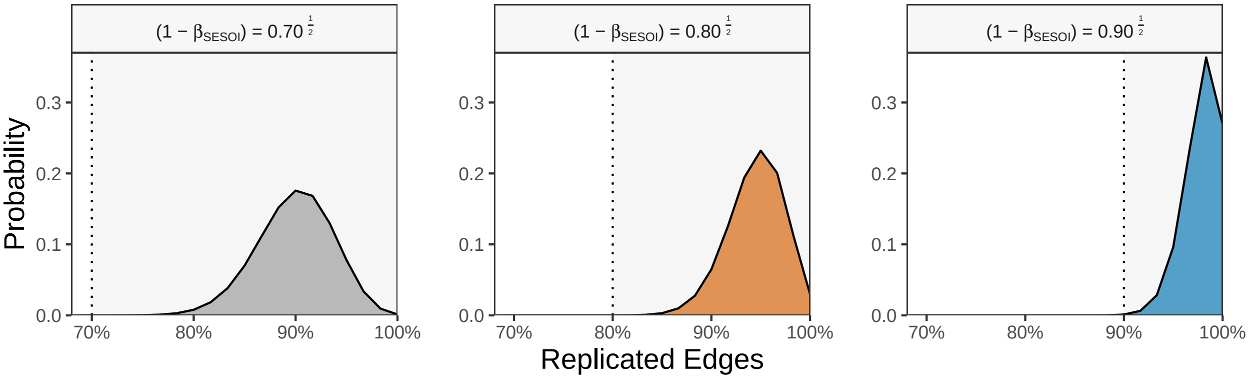

Figure 5. Planning for network replicability by defining the smallest edge size of interest (SESOI). $(1 - \beta_{\text{SESOI}})^{1/2}$ is statistical power to detect the SESOI. The idea is to define power for one network, say, 0.70, and then computing $0.70^{1/2}$ ensures that the probability of detecting the SESOI in two networks is still 0.70. The densities are probability mass functions for the Poisson-binomial distribution for replicating two networks. This reveals that defining statistical power for the SESOI provides a lower bound for replicability.

Appendix

Ordinal Data

It is also important to consider ordinal data which is commonplace in the social-behavioral sciences. Note that, when there are many categories, one strategy would be to assume normality and use the analytic solution based on the Fisher-z transformation (Rhemtulla, Brosseau-Liard, & Savalei, 2012). When there are few categories, however, there is not a simple derivation for computing the standard error of a polychoric partial correlation. Hence, it is necessary to approximate Equation (4) from a bootstrapped sampling distribution, that is,

$$
\begin{aligned}
\beta_s &= \Phi\bigg(Z_{\alpha/2} - \frac{F(\rho_s)}{\sqrt{var(\xi_s^*)}}\bigg) \\
var(\xi_s^*) &= \frac{1}{B-1}\bigg(\sum_{b=1}^{B} \xi_s^{(b)} - \bar{\xi}_s^*\bigg) \\
\bar{\xi}_s^* &= \frac{1}{B}\bigg(\sum_{b=1}^{B} \xi_s^{(b)}\bigg),\ b = 1, ..., B,
\end{aligned}
\tag{10}
$$

where b denotes a given bootstrap sample, B is the total number of bootstrap samples, and ρ_s is the population value for the sth edge. Following Efron and Tibshirani (1986), the Fisher-z transformation is applied to the entire bootstrap distribution. This not only results in approximate normality, which is a key ingredient for computing (10), but it also stabilizes the variance across all values of $F(\rho)$ (i.e., approximately constant). This procedure is applied to simulated multivariate data with a given sample size, number of ordinal categories, and covariance structure.[5] This provides an estimate of the standard error, $\sqrt{var(\xi_s^*)}$, for Fisher-z transformed polychoric partial correlations. I refer interested

[5] The data are generated with an exact covariance structure and not a random sample from that covariance structure. This is possible in the `R` package **MASS**.

readers to Efron and Tibshirani (see Section and Table 2, 1986) for a full discussion of bootstrapping the standard error of correlation coefficients.

Illustrative Example

I characterized ENR in two and four networks. Additionally, ENR was computed for continuous data, as well as ordinal data of decreasing categories (5 and 3). The idea here is to further highlight an additional source of increased sampling variability that presents a challenge for investigating network replicability (over and above going from zero-order to partial correlations). In each replication scenario, the networks had equal sample sizes (250, 500, and 1,000). Note that this is not a requirement for computing ENR, but assuming equal sample sizes does simplify the calculation. Another important consideration is specifying the true values for the edges. This is perhaps never an easy task, but especially in the case of networks, a researcher must assume *true* values for potentially hundreds of effects. Rather than use synthetic data, I estimated partial correlations from symptoms of post-traumatic stress disorder ($N = 221, p = 20$; Armour et al., 2017). I then induced sparsity by setting absolute values less than 0.075 to zero. This was chosen because it resulted in approximately 50% of the nodes sharing a connection. This translates into a rather ambitious attempt to replicate 94 edges.

Results. Figure A1 includes these results. Note that the CDF can be read directly from the densities. For example, with continuous data, $N = 500$, and four networks, the probability of replicating more than 50% of the edges corresponds to the green density in the shaded region. In this case, the probability was nearly 1.0. However, in the same panel, the probability of replicability exceeding 75% was essentially zero. The `enr` function in the `R` package **GGMnonreg** allows for computing these probabilities.

These results highlight a central aspect of this paper: sampling variability present in partial correlations networks can give the appearance of unreliability. In this case, the increases in sampling variability arise when going from continuous data to ordinal data with fewer and fewer categories. The impact on replicating a larger number of edges was non-trivial. With continuous data, $N = 1500$, and four networks, the probability of replicating more than 75% of the edges was 1.0. This suggests that we would expect to consistently estimate three out of four edges in each replication attempt. However, for ordinal data with five categories the probability of replicating more than 75% of the edges was 0.06 and effectively 0 for three categories. This is directly related to the sampling distribution. For continuous data, the standard error is $\frac{1}{\sqrt{1500-18-3}} = 0.026$, whereas the simulated estimates were 0.036 (five category) and 0.046 (three category) for the ordinal data. Note that these are the average standard errors across all edges, and in all individual cases, the same pattern of increasing sampling variability emerged.

These results can also shed light onto the claims of Forbes et al. (2019), where they noted that in four PTSD networks "only [$\approx$] 34% were estimated consistently" (p. 11). Although those ordinal data had four levels and varying sample sizes (with the largest $N = 956$), we can infer that observing that level of replicability *could* be compatible with the results. For example, with three level ordinal data, the probability of replicating more than 34% of the edges was 0.66. Note also their arguments were based on empirical data, and as a result, an exact comparison is not possible. However, these results still show that we would not actually expect to replicate that many edges with ordinal data in particular. And certainly with, say, $N = 403$ (the anxiety networks in Forbes et al., 2019), replicability of even 50% would not be all that common. This is expected.

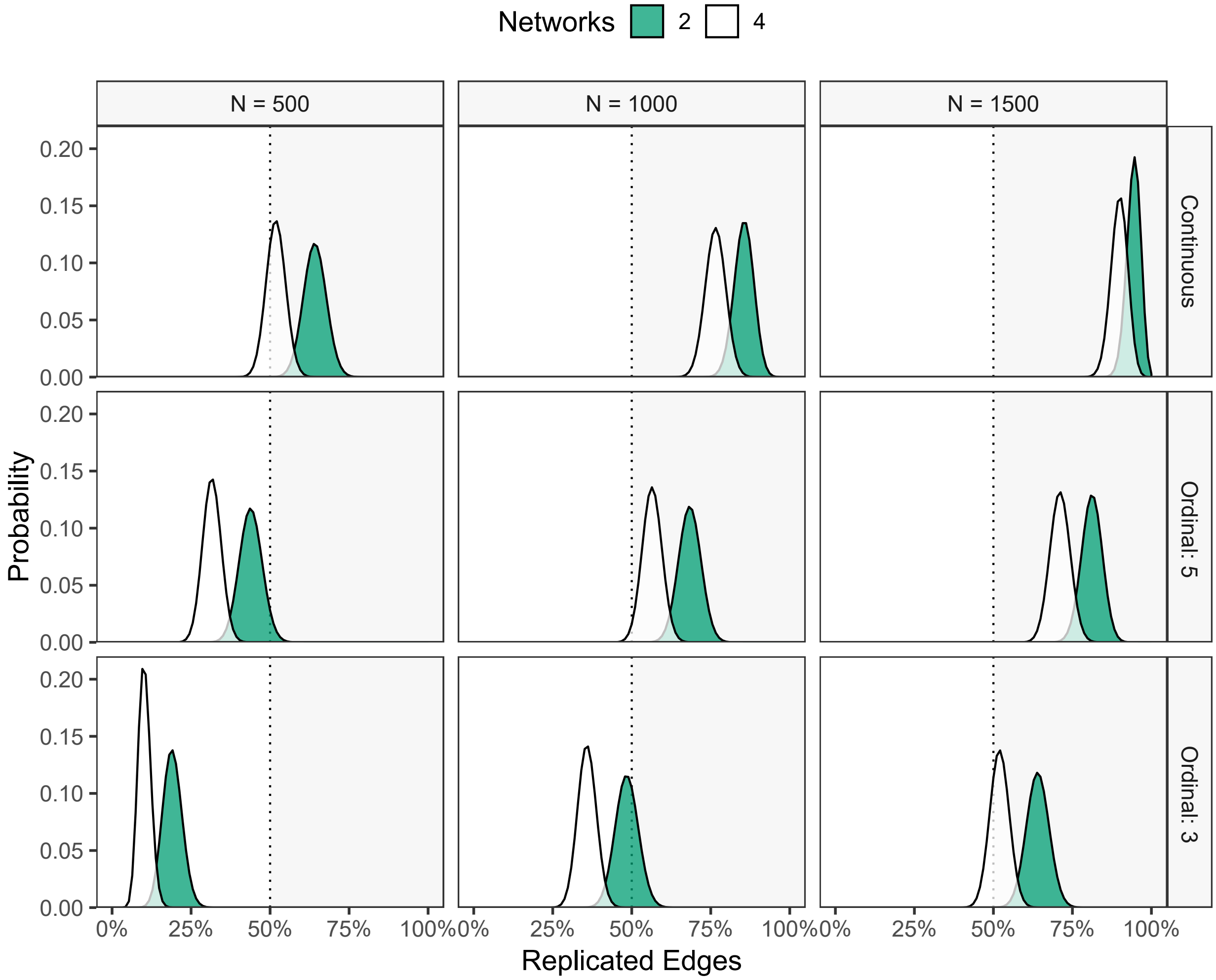

Figure A1. Probability mass function of the Poisson-binomial distribution, given the power to detect each edge in the *true* network. The y-axis is the probability of replicating a given number of edges in two or four networks and the x-axis is the proportion of replicated edges. The shaded area corresponds to more than half of the edges. Hence, if the density is completely in that region, the probability of replicating more than 50% of the edges is 1.0. Furthermore, noting where the distribution is located can also provide useful information. For example, with two networks, $N = 1500$, and three category ordinal data (bottom right panel), the green distribution lies almost completely between 50% and 75%: while more than 50% of the edges are expected to replicate, there is essentially no chance that more than 75% will be detected in all four networks. These results also reveal that the proportion of replicated edges decreases when going from continuous to ordinal data with few categories. This is because there is more sampling variability (i.e., larger standard errors) that translates into reduced power (i.e., type II errors).