Symptom network models in depression research

van Borkulo, Claudia Debora

CHAPTER 5

COMPARING NETWORK STRUCTURES ON THREE ASPECTS: A PERMUTATION TEST

Adapted from:

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

Network approaches to psychometric constructs, in which constructs are modeled in terms of interactions between their constituent factors, have rapidly gained popularity in psychology. Applications of such network approaches to various psychological constructs have recently moved from a descriptive stance, in which the goal is to estimate the network structure that pertains to a construct, to a more comparative stance, in which the goal is to compare network structures across populations. However, the statistical tools to do so are lacking. In this paper, we present the Network Comparison Test (NCT), which uses permutation testing in order to compare network structures from two independent, cross-sectional data sets on invariance of 1) network structure, 2) edge (connection) strength, and 3) global strength. Performance of NCT is evaluated in simulations that show NCT to perform well in various circumstances for all three tests: the Type I error rate is close to the nominal significance level, and power proves sufficiently high if sample size and difference between networks are substantial. We illustrate NCT by comparing depression symptom networks of males and females. Possible extensions of NCT are discussed.

5.1 Introduction

In the past decades, network analysis has rapidly gained popularity as a method of representing complex relations in large datasets, and has been applied in many different fields, from physics and engineering to medicine and biology (Barabási, 2011). Recently, network analysis has also entered the field of psychology, where it has been applied to research on attitudes, intelligence, personality, and psychopathology (Boschloo, Schoevers, Van Borkulo, Borsboom, & Oldehinkel, 2016; Boschloo et al., 2015; Costantini et al., 2015; Cramer et al., 2010; Dalege et al., 2016; Schmittmann et al., 2011). In these applications, network modeling has led to the novel way of representing psychological constructs as complex dynamical systems of interacting variables (Schmittmann et al., 2011). For example, a major depressive disorder may emerge from interactions between depression symptoms, such as depressed mood, fatigue and concentration problems (Borsboom & Cramer, 2013; Cramer, van der Sluis, et al., 2012; Schmittmann et al., 2011). In network approaches, such symptom variables are represented as *nodes* and their interactions as *edges* between nodes.

In the network approach, initial research efforts mainly focused on investigating interaction patterns to reveal potentially important elements in the network (Boschloo, Schoevers, et al., 2016; Boschloo et al., 2015; Costantini et al., 2015; Cramer et al., 2010; Dalege et al., 2016; Fried, Bockting, et al., 2015; Kossakowski et al., 2016; McNally et al., 2015; Robinaugh et al., 2014; Robinaugh & McNally, 2011; Schmittmann et al., 2011). In these studies, the analysis was typically limited to determining a network structure in a single population. More recently, however, the focus has shifted from such single population studies to studies comparing network structures from different populations (Bringmann, Pe, et al., 2016; Bringmann et al., 2013; Koenders et al., 2015; Pe et al., 2015; Van Borkulo et al., 2015; Wigman et al., 2015). A comparative study of our own research group for example showed that the network structure of depression symptoms had a higher level of overall connectivity in a subpopulation of patients with a poor prognosis compared to a subpopulation with a good prognosis (Van Borkulo et al., 2015, see also Chapter 6). Similar comparisons have so far relied mainly on visual inspection of networks structures (Bringmann, Pe, et al., 2016; Bringmann et al., 2013; Koenders et al., 2015; Pe et al., 2015; Wigman et al., 2015), since statistical tests simply have not been available.

Our aim is to fill this gap by developing a statistical testing procedure that allows a direct comparison of two networks as estimated in different subpopulations. This procedure, which we denote the Network Comparison Test (NCT), combines advanced methodology for inferring network structures from large empirical, cross sectional datasets (Epskamp et al., 2012; Van Borkulo et al., 2014) with permutation testing. We focus on tests designed to evaluate three hypotheses that are typically relevant in network analysis: (1) *invariant network structure*, (2) *invariant edge strength*, and *invariant global strength* (3). The first hypothesis, concerns the structure of the network as a whole, and states that this structure is completely identical across subpopulations. The second hypothesis zooms in on the difference in strength of a specific edge of interest. The third hypothesis says that, although networks may differ in structure, the overall level of connectivity is equal across groups.

It should be noted that the present contribution is focused on the comparison of network structures that have to be inferred from data; that is, the network structures involve relations between variables that have to be estimated from the data. This means that the relevant networks should be clearly distinguished from,

e.g., social networks, which pertain to relations between concrete objects (e.g., people) rather than variables, and in which connections (e.g., friendships) are typically treated as observed. In this sense, network approaches in psychometrics are more closely related to graphical models (Lauritzen, 1996) than to social networks. Also note that we focus on the situation where network structures are compared that are inferred from independent, cross-sectional data sets; although extensions to dependent data and even time series networks are possible, these are outside the scope of the present paper.

This paper is structured around three main topics. First, we discuss the general statistical testing framework, including network estimation methods, permutation testing, and an explanation of the test statistics. Second, we present a simulation study to examine the performance of NCT under different circumstances. Third, the utility of the proposed method is illustrated with a real data set. In the discussion, we will propose possible extensions of NCT.

5.2 Network Comparison Test

In this section, we explicate various aspects of NCT. First, we explain the recently developed network estimation methods that are used to construct the networks that form the input for NCT. Second, we elaborate on the test statistics that can be used to test for differences between networks with respect to invariance of network structure, edge strength, and global strength. Third, the statistical testing procedure that underlies NCT is explicated. Finally, we discuss the consistency of the presented test statistics.

5.2.1 Network estimation

Networks relevant to this paper involve connections between variables that are inferred from data. For this purpose, NCT uses recently developed methodology to estimate the network structure from one set of measurements of multiple cases (individuals). The purpose of network modeling in such cases is to determine the network structure most likely to underlie the data. For example, network modeling techniques have been applied to depression symptoms as determined in a community sample (Kessler et al., 2004) or in a sample of depressed patients (Penninx et al., 2008). An example of such a network is given in Figure 5.1.

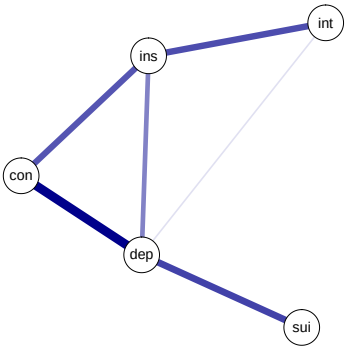

FIGURE 5.1. Hypothetical network, estimated from measurements of depression symptoms of a group of patients. Associations between symptoms are depicted as connections between symptoms with varying width pertaining to the strength of the associations. Associations in this paper are estimated with *eLasso*, a method that is based on ℓ_1-regularized logistic regressions (Van Borkulo et al., 2014). Abbreviations: int - loss of interest, ins - insomnia, con - concentration problems, dep - depressed mood, sui - suicidal ideation.

Although NCT is a general method for all types of data and network estimation methods, it is currently implemented for handling networks derived from continuous and binary data. For continuous data, network estimation can simply be based on partial correlations, where each partial correlation between two variables is computed by conditioning on all other variables in the dataset (Epskamp & Fried, 2016); under the assumption that the data come from a multivariate normal population density, zeros in the matrix of partial correlations (which equals the inverse of the correlation matrix) correspond to conditional independence relations between variables, which in turn translate to missing edges in the network (Koller & Friedman, 2009). For binary data, such computational procedures are not available because partial correlations of zero do not imply conditional independence in binary data. Estimation is, therefore, based on an iterative scheme that combines logistic regression and model fit evaluation (Van Borkulo et al., 2014).

Both estimation methods use ℓ_1-regularization (Tibshirani, 1996) to reduce the number of false positives and elegantly bypasses multiple testing problems that would occur in traditional significance testing (e.g., with only 10 variables, one

would have to perform 45 (10 x 9/2) significance tests — one for each possible edge in the network). This procedure has been shown to converge to the *t*rue network that generated the data if assumptions are met (Van Borkulo et al., 2014); that is, we assume that the data are generated from a network of pairwise, undirected connections with varying intensities (strengths of the connections in the network), in which most of the connections are absent (Ravikumar et al., 2010). The level of sparsity that the method assumes can be adjusted by a so-called hyperparameter (γ) that controls the strength of the penalization involved in the ℓ_1-regularization procedure; in this paper we set γ to zero to obtain networks with the least sparsity.

5.2.2 Test statistics

To assess the difference between networks, we implement three tests that involve hypotheses regarding (1) invariant network structure, (2) invariant edge strength, and (3) invariant global strength.

5.2.2.1 Invariant network structure

The first invariance hypothesis concerns the structure of the network as a whole and states that this structure is completely identical across subpopulations. Formally, the null hypothesis states that all edge weights in A_1 are identical to those in A_2 ($A_1 = A_2$), in which A_1 and A_2 are the connection strength matrices of graphs (networks) G_1 and G_2, respectively. To test this hypothesis, we use a distance measure for symmetric $n \times n$ matrices: the maximum or ℓ_∞-norm. This metric is based on element-wise (absolute) differences and focusses on the largest difference. Let A_{1ij} and A_{2ij} be matrices containing connection strengths between variables i and j of networks G_1 and G_2 respectively, in which A_{1ij} is the connection strength of graph G_1 between nodes i and j. The matrix D with difference scores of all connection strengths contains elements $D_{ij} = |A_{1ij} - A_{2ij}|$. The metric of interest is the largest entry in D and is formally defined as

$$M(G_1, G_2) = max(D_{ij}). \tag{5.1}$$

The test of network structure invariance evaluates the observed value of M in the data against the reference distribution of M that arises from random permutation of group membership across cases to test the hypothesis that $A_1 = A_2$ in the

population from which the sample was drawn. This is explained more extensively in the Procedure section.

5.2.2.2 Invariant edge strength

The second invariance hypothesis zooms in on the difference in strength of a specific edge to evaluate whether that edge is equally strong across subpopulations. Regarding the difference in strength of a specific edge, we simply used the (absolute) difference in edge strength between the focal nodes i and j in both networks:

$$E(\beta_{ij}^{G_1}, \beta_{ij}^{G_2}) = |a_{ij}|. \tag{5.2}$$

Note that this test does not control the family-wise significance level when multiple connections are tested; in this case a Bonferroni-Holm or (local) false discovery rate correction may be applied to counteract the multiple testing problem (Holm, 1979).

5.2.2.3 Invariant global strength

The third invariance hypothesis states that the overall level of connectivity is the same across subpopulations. Overall connectivity can be summarized by global strength and is defined as the weighted absolute sum of all edges in the network (Opsahl et al., 2010). The distance S, based on global strength, between two networks G_1 and G_2 is then formally defined as

$$S(G_1, G_2) = |\sum_{i,j \in V} |A_{1ij}| - \sum_{i,j \in V} |A_{2ij}||. \tag{5.3}$$

Here, V is the set of nodes in networks G_1 and G_2. By randomly permuting the group membership variable across cases to obtain a reference distribution for S, we can evaluate the null hypothesis that $\sum_{i,j \in V} |A_{1ij}| = \sum_{i,j \in V} |A_{2ij}|$ in the population.

5.2.3 Procedure

The procedure that implements NCT consists of three steps. The first step is to estimate the network structure in the different groups using the original, observed (unpermuted) data, which results in a network structure for each group, and the

relevant metric is calculated; this metric will function as the test statistic (see Figure 5.2, step 1). Second, group membership is repeatedly, randomly rearranged across cases, followed by re-estimation of the networks and calculation of the accompanying test statistic (Figure 5.2, step 2). This results in a reference distribution of the test statistic under the relevant null hypothesis. In the third step, the reference distribution can be used to evaluate the significance of the observed test of step 1. The p-value equals the proportion of test statistics that are at least as extreme as the observed test statistic (Figure 5.2, step 3). The method is implemented in R package NetworkComparisonTest (R Development Core Team, 2011; Van Borkulo, Epskamp, & Milner, 2016).

5.2.4 Power of NCT

For comparing ℓ_1-regularized networks, it is difficult to derive a parametric test, since the network parameters (edge weights) can be highly non-normal (Pötscher & Leeb, 2009). In this paper, we deal with this by applying non-parametric permutation testing to circumvent the assumption of normality. Permutation tests have low false positive rates and high true positive rates under many circumstances, whether the data are identically and independently distributed or not (Good, 2006).

A high true positive rate can be achieved asymptotically under two relatively mild conditions (Van der Vaart, 1998). The first condition is that there should be a substantive proportion of edge weights that are independent. Edge weights are dependent when they belong to the same clique (a completely connected subset of nodes). When the network is not one clique (e.g., fully connected in which every node is connected to all other nodes), the true positive rate (power) of our permutation test still converges to 1. However, the more independent edges, the faster the power will converge to 1. With the ℓ_1-regularized network estimation methods that we use, networks will be far from fully connected. Therefore, the first condition is likely to hold. Note that the issue of dependency between edge weights only applies to the test on invariance of network structure and global strength. Concerning the test on edge strength invariance, the test statistic involves only one edge. The second condition is that the distribution of the edge weights is stationary across groups, except for the location. That is, they need to have the same shape, but can have different means. However, the distribution of

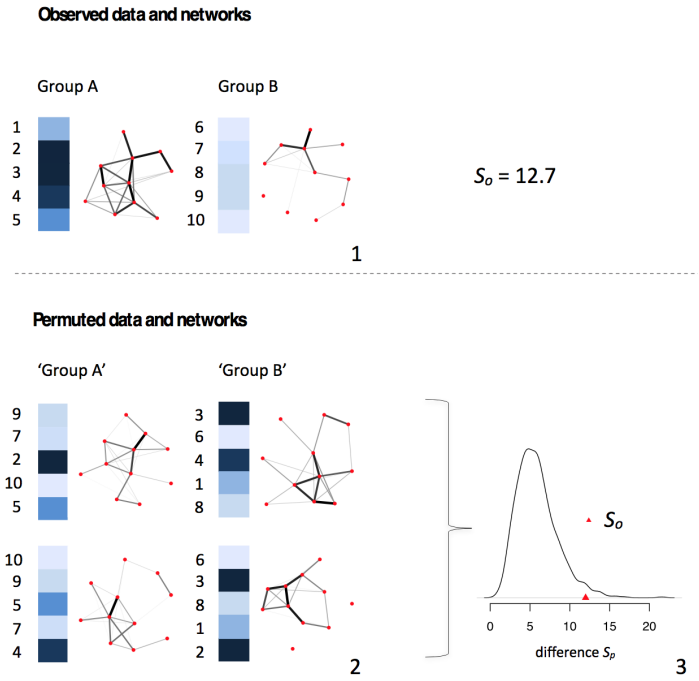

FIGURE 5.2. Schematic representation of the three steps involved in NCT. Step 1: the network structure is estimated of group A and B using the original, observed (unpermuted) data, and the metric of interest S_o is calculated. Step 2: group membership is repeatedly, randomly rearranged; networks are estimated and metrics S_p are calculated based on permuted data ('group A' and 'group B') to create a reference distribution. Step 3: the observed metric S_o is evaluated against the reference distribution under the null hypothesis from step 2, which yields the p-value.

edge weights of ℓ_1-regularized networks is biased by regularization of the parameters that constitute the network (Caner & Kock, 2014; Van de Geer et al., 2014). The strategy of *desparsification* removes the bias and yields approximately normally distributed parameters (Van de Geer et al., 2014). The combination of both conditions implies that NCT can achieve a high true positive rate asymptotically.

5.3 Simulation study

We assessed the performance of NCT using simulations designed to evaluate the three invariance tests on network structure, edge strength, and global strength. In this section, we first explain how the simulation study was set up, followed by the results.

5.3.1 Setup of simulation study

We generated random networks in which nodes are connected by randomly adding edges with varying probabilities, thereby creating networks with varying densities (Erdos & Renyi, 1959). We chose a fixed network size of 36, striking a balance between tractability and representativeness for typical network applications to psychological symptom questionnaires. As the null hypothesis assumes that structures are completely identical across subpopulations (i.e., both groups have the same data-generating mechanism), we simply copied the resulting network to obtain the network for the second group. Weights are assigned to the edges in a realistic range by using squared values from a normal distribution (Van Borkulo et al., 2014). These simulated networks are called the true networks.

To assess performance under the null hypothesis, two binary datasets were generated from identical networks. To assess performance under the alternative hypothesis (i.e., the network structures differ), the network was altered in one of the groups. This was done in two different ways, pertaining to the specific test under investigation in the relevant simulation. For the tests of network structure and edge strength invariance, the edge with the highest strength in one network was changed in the second network by lowering the weight with 50% and 100% (i.e., in the latter condition the relevant edge was set to zero; see Figure 5.3 for examples of these simulated networks). For the test of overall connectivity (global strength), the density was lowered in the copied network by cutting a percentage (25% and 50%) of edges (examples not shown here).

Binary data was simulated with various sample sizes that are realistic in psychology and psychiatry (250, 400, and 700 cases for each group) using the R package IsingSampler (Epskamp, 2013). As sample sizes of groups are not always similar in real data sets, we simulated both equal-sized and unequal-sized groups. In the latter condition, one group has the original sample size (250, 400, or 700

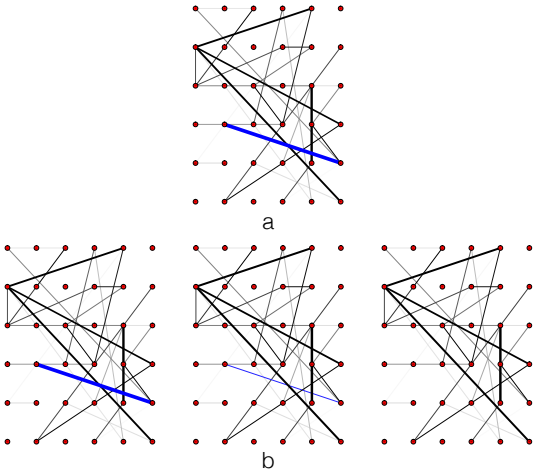

FIGURE 5.3. Examples of networks used in the simulation study to assess performance of the tests on invariance of network structure, an individual edge, and global strength. A random network with 36 nodes was simulated with a probability of an edge of .05 (a). This network was used to simulate data of the first group. For the second group, data was simulated under three conditions: using an exact copy of the network of the first group (b, left panel), using a copy in which the edge with the highest strength (blue edge) was halved (b, middle panel), and using a copy in which the edge with the highest strength (blue edge) was set to 0 (b, right panel). Thickness of the edges represents the weights.

cases) and the other group has 1.5 times that sample size (375, 600, 1050 cases). To investigate whether it matters whether an edge weight is lowered (or a percentage of connections is cut) in the group with the largest or the smallest sample size, we simulated both scenarios.

The resulting setup is a $3\times3\times3\times2\times2$ factorial design, in which the manipulated factors are (a) density (probability of an edge .05, .1, or .2), (b) level of difference (lowering an edge by 0%, 50%, or 100% or by cutting 0%, 25%, or 50% of the edges), (c) sample size (250, 400, 700), (d) equality of sample size (1 or 1.5 times the original sample size), and (e) balancing condition (i.e., whether the network of the smallest or the largest group is cut or lowered). Consequently, the simulation study involved 108 conditions, which were replicated 100 times each. Each condition thus resulted in 100 p-values from which the probability of rejecting the null hypothesis (proportion of $p \leq .05$) was calculated. For conditions under the null

hypothesis (there is no difference), this results in the Type I error, whereas for conditions under the alternative hypothesis (there is a difference) this results in the statistical power of the test.

5.3.2 Results

Performance of NCT was evaluated in terms of Type I error control and statistical power. Results are discussed for each of the three test statistics of NCT.

5.3.2.1 Network structure

NCT adequately retained the null hypothesis in simulations under the null hypothesis (Figure 5.4a, left panel); the Type I error rate (actual alpha) was accurately low (M = .058, SD = .019) across all conditions pertaining to the null hypothesis. When the edge with the highest strength was lowered in one of the identical networks to half of the original strength (Figure 5.4a, middle panel), the average statistical power was moderate across conditions (M = .55, SD = .14). With higher sample size (N = 700), power increased (M = .69, SD = .24). When the strongest edge was lowered to zero in one of the identical networks, inducing a maximal possible difference (Figure 5.4a, right panel), the average statistical power was high across conditions (M = .85, SD = .17).

Zooming in on the specific conditions revealed that, as would be expected, power increased with increasing sample size. In addition, power was highest for less densely connected networks. Moreover, the equality of sample size conditions showed that, when the strongest edge is lowered by 50% (Figure 5.4a, middle panel), it mattered whether groups were equal or unequal-sized. On average, results indicate that the power was more or less similar when groups are equal-sized (M = .50, SD = .25; solid and dotted lines) or when the strongest edge was lowered in the largest group (M = .46, SD = .22; dashed lines). However, when the strongest edge was lowered in the smallest group, average power was higher (M = .60, SD = .21; dotted lines). This effect was also present when the strongest edge was lowered by 100% (Figure 5.4a, right panel). On average, power was similar when groups were equal-sized (M = .84, SD = .18; solid lines) or when the strongest edge was lowered in the largest group (M = .81, SD = .21; dashed lines). But when the strongest edge was lowered in the smallest group, average power was higher (M = .89, SD = .13; dotted lines).

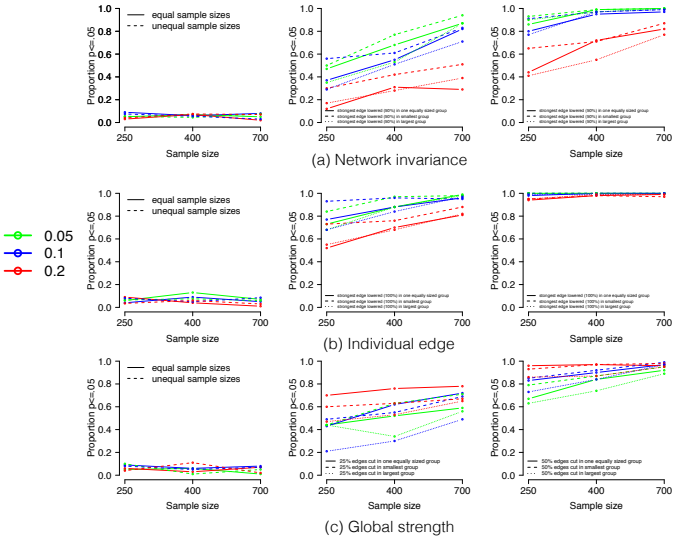

FIGURE 5.4. Proportion of p-values <.05 when performing the tests on invariance of (a) network structure, (b) an individual edge, and (c) global strength test with NCT. The x-axes display sample size, whereas the y-axes displays proportion of p-values $\leq .05$. All three tests were applied on simulated data under the null hypothesis of no difference (left panels) and under the alternative hypotheses that there is a difference to a certain degree (middle and right panels for increasing levels of difference). Data was simulated from networks with different levels of connectivity (probability of a connection .05, .1, and .2; green, blue, and red, respectively) and with equal (solid lines) and unequal sample sizes. Simulations with unequal sample sizes were balanced for simulations under the alternative hypotheses (a dashed line when edges were altered in the smallest group and a dotted line when edges were altered in the largest group); this was not necessary under the null hypothesis, since no edges were altered.

5.3.2.2 Edge strength

For the individual edge strength test, NCT proved slightly too liberal in simulations under the null hypothesis (Figure 5.4b left panel): average Type I error was somewhat increased (M = .062, SD = .028) relative to the network structure invariance test. When the edge with the highest strength was lowered by 50% in one of the identical networks (Figure 5.4b middle panel), average statistical power was high (M = .83, SD = .13). When the strongest edge was lowered to zero in one of the identical networks (Figure 4b right panel), the test on invariance of edge strength almost never failed in any of the simulation scenarios (M = .99, SD = .017), even at the lowest sample size.

Zooming in on the specific conditions revealed that power increased with increasing sample size and that power was highest for less densely connected networks. Moreover, the different sample size conditions showed that, when the strongest edge was lowered by 50% (Figure 5.4b middle panel), it mattered whether groups were equal or unequal-sized. On average, results indicated that the power was similar when groups were equal-sized (M = .80, SD = .15) or when the strongest edge was lowered in the largest group (M = .79, SD = .15; dashed lines). However, when the strongest edge was lowered in the smallest group, average power was higher (M = .89, SD = .09; dotted lines). This effect worn off when the strongest edge was lowered by 100% (Figure 5.4b right panel), since all conditions had very high power.

5.3.2.3 Global strength

NCT adequately controlled Type I errors in simulations pertaining to the null hypothesis (Figure 5.4c); on average, the Type I error (actual alpha) was accurately low (M = .058, SD = .029). For simulations in which the density in one network was lowered by 25% (Figure 5.4c middle panel), average statistical power was moderate (M = .55, SD = .14). When the density was lowered by 50% (Figure 5.4c right panel), average statistical power was high (M = .88, SD = .10).

Zooming in on the specific conditions revealed that power increased with increasing sample size and that power was highest for the most densely connected networks. Note that this is opposite to the other two metrics, in which power was highest for less densely connected networks. The different sample size conditions revealed that, when the density in one network was lowered by 25% (Figure 5.4c

middle panel), power was lowest when density was lowered in the largest group (M = .44, SD = .14). When density was lowered in the smallest group or when groups were equal, power was similar (M = .60, SD = .09, and M = .62, SD = .13, respectively). When the density in one network was lowered by 50% (Figure 5.4c right panel), the average power was high regardless of equal or unequal-sized groups.

Overall, the global density test had more power for more densely connected networks (red lines in Figure 5.4c). Note that this is opposite to the results of the two other metrics, in which power was highest for less densely connected networks.

To conclude, simulations indicated that the three tests in NCT performed well in the scenarios considered in this paper. Tests on invariance of network structure and global strength showed a Type I error close to the nominal level (α = .05) and power increased to high (> .8) when the focal difference between networks increased and/or when sample size was large enough. The test on invariance of an individual edge showed high power, but a slightly elevated Type I error rate; researchers using this test may want to choose a somewhat stricter significance level to accommodate this.

5.3.3 Application to real data

To illustrate the utility of NCT, we used the procedure to evaluate the possible difference in the network structure of depressive symptoms in male versus female depressive patients. It has been shown that, although the prevalence of major depression is higher among women compared to men (Kessler, 2003; Nolen-Hoeksema, 1987), the clinical gender-related differences in depressed patients are limited (Boschloo et al., 2014, 2012; Schuch, Roest, Nolen, Penninx, & de Jonge, 2014). Consequently, with the conception of depression as a network of the symptoms in mind, one could hypothesize that the network of depression symptoms of men and women are overall similar. At a local level, however, one might expect differences in connection strengths. Since men with major depression are known to have a higher suicide risk (Hawton, Casañas i Comabella, Haw, & Saunders, 2013), the symptom of suicidal ideation could be expected to have different connections in the networks of men and women.

5.3.4 Real data

Data were derived from the baseline measurement of the Netherlands Study of Depression and Anxiety (Penninx et al., 2008). For the current analyses, we selected data of men ($N = 351$) and women ($N = 709$) with a past-year major depressive disorder (assessed with the Composite Interview Diagnostic Instrument; Wittchen, 1994). To estimate the network structures, we used scores on 11 DSM-IV criteria pertaining to Major Depressive Disorder (American Psychiatric Association, 2013) as assessed with matching items of the Inventory of Depressive Symptomatology (Rush et al., 1996) and previously described in Van Borkulo et al. (2015).

The criteria, on which the network was based, were originally scored from 0 (not applicable) to 3 (very applicable). Since we focused the simulation study on binary data, these scores were dichotomized. This allowed us to interpret the findings with the real data, and the resulting network, in the light of the simulation study. A score of 0 was interpreted as the absence of a criterion (i.e., a zero in the rescored binary data set), whereas a score of 1 to 3 was interpreted as the presence of a criterion (i.e., a one in the rescored binary data set).

Network structures for male and female patients were estimated with the *eLasso* procedure in which γ was set to 0 and the AND-rule was applied (Van Borkulo et al., 2014). For NCT, 1000 permutations were performed.

5.3.5 Results

From Figure 5.5 it is hard to tell whether the networks of male and female patients differ. Visually, they seem equally densely connected, with some connections stronger in the network for males and some connections stronger in the network for females. The test on network structure invariance revealed that the difference between the network structures is not significant ($M = 1.167$, $p = .251$). When the network structure is found to be invariant, there is no reason to pursue further testing of specific edges. In fact, this can lead to an increased Type I error. Therefore, we did not test edges between suicidal ideation and other symptoms. The test on invariance of global strength also revealed no difference ($S = .618$, $p = .909$). Therefore, as expected, the null hypothesis cannot be rejected; networks of depressed men and women are similar. Repeated subsampling (100 times) from the larger group of women revealed that the difference was significant in only 1 and 2% for network structure invariance and global strength invariance, respectively.

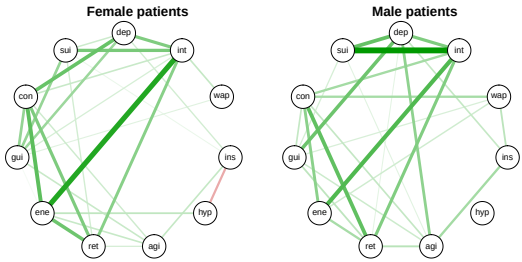

FIGURE 5.5. Networks of female (left panel; $N = 709$) and male patients (right panel; $N = 351$). Connection strengths vary from -.66 (between hyp and ins in women′s network) to 1.97 (between sui and int in men′s network). Abbreviations: dep indicates depressed mood; int, loss of interest or pleasure; wap, weight/appetite change; ins, insomnia; hyp, hypersomnia; agi, psychomotor agitation; ret, psychomotor retardation; ene, fatigue or loss of energy; gui, feeling guilty; con, concentration/decision making; sui, suicidality.

5.4 Discussion

NCT is a novel method to directly test for differences between networks of two independent, cross-sectional data sets. The present study shows that the method performs well under a range of realistic circumstances. Type I error is consistently close to the nominal level ($\alpha = .05$) and power is good ($> .8$) when the differences with respect to the three measures is substantive and/or the sample size is large enough (i.e., relative to the number of variables in the network). Thus, NCT is a viable method to statistically test for several types of differences in various research settings and fills the gap in comparing network structures of psychological constructs.

Simulations indicate three caveats to take into account. First, the edge strength invariance test seems to have a slightly elevated type I error. Researchers may want to choose a somewhat stricter significance level than .05 to deal with this issue. Second, for the network structure invariance and the edge strength invariance test, power is higher for less densely connected networks. For the global strength invariance test, however, this is reversed: power is higher for more densely connected networks. Third, for all metrics, it matters whether the largest (or smallest) group has lowest (or highest) density or connection strength. When the largest group has the lowest density or connection strength, power is lowest. This effect

could be due to the network estimation method as sample size is involved in the penalty of ℓ_1-regularized estimation methods. Researchers that have unequal-sized groups may want to use (repeated) subsampling from the largest group to avoid that sample size differences bias results.

Although NCT is suited for both binary and continuous data, we performed this validation study only with binary data. These results, however, are also applicable to continuous (Gaussian) data; when the number of nodes and sample sizes are equal, performance of NCT with continuous data is at least similar to performance with binary data (Raskutti, Wainwright, & Yu, 2010). The simulation results carry over to networks with other than binary variables, because the test statistic is obtained from the edges; if these are accurately estimated, the NCT will have good properties.

An alternative strategy to compare network structures that are estimated with ℓ_1-regularization, which we did not apply here, involves desparsification. This boils down to removing the bias that is introduced by regularization of the parameters that constitute the network (Caner & Kock, 2014; Van de Geer et al., 2014). This strategy is assumed to yield normally distributed parameters that allows for parametric testing. However, since it is not clear under what circumstances parameters indeed are normally distributed, we chose non-parametric permutation testing for comparing networks, to circumvent the assumption of normality.

The presented methodology may be extended in at least three ways. The first extension involves the incorporation of other measures of difference between networks. Currently, NCT tests the invariance of three different aspects (network structure, edge strength, and global strength), but other aspects could be evaluated. For example, differences in characteristic path length and the global clustering coefficient could be tested; the first measures the average length of all shortest paths between any two nodes (Watts & Strogatz, 1998) and the second measures the proportion of *triplets* (three nodes connected by two connections) which are closed by a third connection (Opsahl, 2013). On a local (node) level, node centrality measures can give an indication of the importance of nodes in a network. It may be interesting to test whether a specific node has a significant higher score on a certain centrality measure in one group compared to the other. An example of a node centrality measure is *betweenness*, which measures the degree to which a node (variable) in the network serves as a bridge between dif-

ferent parts in the network. This measure reflects the degree to which the node can control the information flow through the network (Freeman, 1979).

The second extension is to accommodate NCT to the analysis of dependent data. Often, researchers want to compare a group of participants before and after manipulation of an independent variable (e.g., treatment). This requires a different way of permuting the data. If we take pre- and post-treatment data (measurements of symptoms) as an example, the null hypothesis would be that the network structure (or an individual edge or global strength) before treatment is the same after treatment. If the null hypothesis were true, one would expect that shuffling the label of pre- and post-measurements *within a single person* does not affect results. Related to this extension is one that allows for intensive longitudinal data, gathered according to the Experience Sampling Method (Myin-Germeys et al., 2009) of groups of individuals. Group-level networks that display the temporal dynamics of two groups of individuals could also be compared by, again, randomly shuffling group labels (Klippel et al., 2017). Since group-level networks are similar under the null hypothesis, group membership would not matter. Further research may evaluate whether NCT works in these situations.

Finally, a third extension could be to allow for mixed type variables. Often, data sets are neither strictly binary nor strictly continuous and even may contain categorical variables. One can transform the data to obtain Gaussian or binary variables, but this can lead to unwanted loss of information. Recently, a network estimation method is developed that can handle data with different types of variables that could very well be implemented in NCT (Haslbeck & Waldorp, 2015).

As comparing networks in the field of psychology is becoming more and more popular, NCT seems a valuable tool to do so in a more substantive way. Researchers can now statistically compare networks of two independent groups with a simple but effective permutation test on three different aspects of differences between networks.