Within, between and beyond heterogeneity

Uncovering individual differences using network models

Hoekstra, H.A.

Publication date
2025

Document Version
Final published version

Ria H. A. Hoekstra

within, between, and beyond heterogeneity

uncovering individual differences using network models

WITHIN, BETWEEN, AND BEYOND HETEROGENEITY

UNCOVERING INDIVIDUAL DIFFERENCES USING NETWORK MODELS

RIA H.A. HOEKSTRA

COLOPHON

This document was typeset in LaTeX using `classicthesis` developed by André Miede.

Cover design by Timo Wolf Kamp, persoonlijkproefschrift.nl

The work in this thesis was supported by the Dutch Research Council (NWO; Research Talent Grant no. 406-18-532) and by the Royal Netherlands Academy of Arts & Sciences (KNAW; Van der Gaag Fund).

WITHIN, BETWEEN, AND BEYOND HETEROGENEITY

UNCOVERING INDIVIDUAL DIFFERENCES USING NETWORK MODELS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. P.P.C.C. Verbeek,
ten overstaan van een door het College voor Promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 13 februari 2025, te 10.00 uur

door
Hendrika Aagje Hoekstra
geboren te Noordoostpolder

PROMOTIECOMMISSIE

Promotor:

prof. dr. D. Borsboom	Universiteit van Amsterdam

Copromotor:

dr. S. Epskamp	National University of Singapore

Overige leden:

prof. dr. E.L. Hamaker	Universiteit Utrecht
prof. dr. H.L.J. van der Maas	Universiteit van Amsterdam
prof. dr. R.J. McNally	Harvard University
dr. M. Marsman	Universiteit van Amsterdam
dr. T.F. Blanken	Universiteit van Amsterdam

Faculteit der Maatschappij- en Gedragswetenschappen

CONTENTS

1 INTRODUCTION

Over the last two decades, the network approach to psychopathology has emerged as a novel framework for conceptualizing mental disorders [227, 29]. In this approach, mental disorders are not merely conceptualized as the result of a single underlying common cause, but rather as resulting from complex causal interactions between symptoms through a myriad of biological, psychological, and societal mechanisms [23, 28]. As each individual can be characterized by their own network, this conceptualization fits naturally with the growing recognition of individual differences, i.e., heterogeneity, and the inadequacy of population models to capture these differences. Empirical analysis of such individualized networks has consistently pointed to significant heterogeneity between individuals [e.g., 33, 95, 185, 67, 222]. However, effectively *assessing* and *understanding* the heterogeneity between these individual networks remains a challenge. Overcoming this challenge is vital for improving our theoretical understanding of psychopathology and for determining whether modeling individual networks can help to pave the way for more tailored and effective treatments. This dissertation is an exploration of key questions in this area: How can individual differences be quantified in network models, and how can network modeling techniques contribute to unraveling and understanding these differences?

1.0.1 *From nomothetic to idiographic*

Traditionally, most psychological research has used large samples observed at a single point in time to extrapolate findings to a larger population. This approach, commonly referred to as the *nomothetic* approach, aims to identify common traits and behaviors in the hope that general laws of human psychology can be derived. While the nomothetic approach is effective in uncovering observations at the aggregate level, it faces significant limitations when attempting to apply these findings to individuals, as the insights derived from such studies may not be applicable to all, or even the majority, of individuals within a population [123].

To illustrate this limitation, consider the following example. Suppose we are interested in the relationship between 'drinking coffee' and 'number of words written'.[1] By observing a sample of people at a given point in

1 This example is taken from Epskamp, S., Hoekstra, R. H. A., Burger, J., & Waldorp, L. J. (2022). Chapter 9. Longitudinal design choices: Relating data to analysis. In Isvoranu, A.

time, we might find a positive relationship between 'drinking coffee' and 'number of words written' at the population level. That is, people who, on average, drink a lot of coffee also, on average, write many words. However, if we study the relationship between 'drinking coffee' and 'number of words written' for a particular individual in this population, we might find that on days they drank a lot of coffee, they wrote fewer words, i.e., a negative relation. Taken to the extreme, it would be possible to establish a relationship between variables at the population level, while for each individual in that population, we would find the opposite effect—a discrepancy that is often referred to as Simpson's paradox [169, 246].

As the example illustrates, individuals may not respond or behave in the same way as the average trend in the population. Therefore, applying population-level findings to individual cases should be done with caution. As noted by Molenaar [201], extrapolations from population effects to the individual can only be made with confidence under the assumption of *ergodicity*. Ergodicity implies that there are no long-lasting structural changes over time (i.e., the process is stationary) and that the process is uniform across individuals (i.e., the process is homogeneous). However, this assumption is challenged by the fact that psychological processes are assumed to be inherently dynamic and diverse. As a result, we cannot blindly extrapolate statements derived from standard analyses of large samples to the individual.

This drawback of the nomothetic approach has now been pointed out numerous times [e.g., 33, 95, 126, 169, 201] and has spurred a shift towards *idiographic* approaches in psychological science; where the focus is on understanding unique aspects of individuals rather than seeking to identify general laws that apply to all. This evolution in research approaches reflects a growing recognition of the need for methods that can accurately capture the complexity of human psychology.

1.0.2 *From single cause to network*

Mental disorders have long been viewed through the lens of standard disease models, postulating an underlying "disorder" as the common cause for a set of related symptoms [28, 165]. This model suggests that symptoms of a mental disorder, such as Major Depressive Disorder (MDD), hang together due to a shared underlying disorder causing the symptoms, namely MDD. This conceptualization has spurred efforts to identify the biological, environmental, and psychological "essences" of mental disorders. However, definitive substrates acting as a universal cause have not been identified despite these endeavors. Instead, evidence suggests

M., Epskamp, S., Waldorp, L. J., & Borsboom, D. (Eds.) *Network psychometrics with R: A guide for behavioral and social scientists.* Routledge, Taylor & Francis Group. [86]

mental disorders likely result from a myriad of different causes [26, 167]. Therefore, a shift has been proposed: a shift from seeking the "essence" of mental disorders to investigating the complex and multi-level causal mechanisms underlying mental disorders [55, 164, 167, 22]. This shift is embodied in the network approach to psychopathology.

Unlike the standard disease conceptualization of mental disorders, the network approach to psychopathology has an inherent idiographic interpretation. Instead of conceptualizing mental disorders as a single cause that is shared among every individual diagnosed with a disorder, in the network approach, each individual can be characterized by their own (symptom) network structure. This network structure renders some individuals more susceptible to developing a mental disorder than others [28, 23, 37]; for example, an individual characterized by a highly connected symptom network, may experience cascading effects where the activation of one symptom triggers many others due to the many and strong causal connections within the network, maintaining symptom activation over time. Conversely, a person with a sparse symptom network, characterized by few and weak connections does not exhibit such cascading effects; the likelihood of a single symptom triggering a cascade of many other symptoms is small due to the relatively weak causal interactions in the network [53, 23, 24]. As such, the network approach to psychopathology naturally allows for individual differences in symptom dynamics and aligns with the call for more idiographic modeling within psychology.

While studying such person-specific network models might enable us to uncover person-specific risk factors for developing mental disorders [198, 145, 28], initially, most network models of psychopathology were estimated from large samples measured at a single time point [227], raising questions about the extent to which these inter-individual networks accurately reflect intra-individual processes. As elaborated in the previous section, for such results to hold conceptually, the process should be ergodic in nature, suggesting that the process should be stationary and homogeneous across individuals. Although this is not expected for most psychological processes, it also seems unlikely that associations between symptoms at the individual level would dramatically depart from those at the population level. For example, it seems implausible that the more a person feels worthless, the less likely this person is to experience depressed mood. Thus, associations between symptoms plausibly differ in degree, but perhaps not in kind, such that radical heterogeneity might not be expected for symptom networks [18].

Yet, when the network approach converged with the call for intra-individual modeling as the estimation of person-specific network models become popular [37], the analysis of such empirically estimated individual symptom networks consistently claimed evidence for substantial het-

erogeneity among individuals [e.g., 33, 95, 185, 67, 222], while tools to determine the extent of heterogeneity in idiographic network models are limited. Thus, questions to what extent individual differences in the network structure are present and to what extent inferences from a population level to the individual level are threatened or invalidated by the presence of heterogeneity remain. Moreover, if symptom networks do exhibit radical heterogeneity, understanding and interpreting these differences becomes vital. Such insights are essential for developing and guiding personalized therapeutic interventions, marking a potentially transformative shift in how mental health disorders are approached and treated, potentially leading to more effective interventions and a deeper understanding of the human psyche.

1.1 THIS THESIS

In addressing the need for tools that allow to *assess* individual differences and to *understand* these differences in terms of network models, this thesis is organized into two parts.[2] The first part of this thesis is dedicated to methodologies for *assessing* heterogeneity within idiographic network models. The second part of this thesis focuses on using network methodology to gain a deeper *understanding* of these individual differences. In particular, this thesis evaluates established methodologies and introduces innovative approaches for assessing heterogeneity among idiographic network models. It further presents means to deepen the comprehension of this heterogeneity and discusses challenges and pointers for future directions.

1.1.1 *Part I: Assessing Heterogeneity*

In Chapter 2, simulation study techniques are employed to determine the effectiveness of existing methods to account for individual differences in person-specific network models. Specifically, the extent to which these methods can distinguish illusionary (e.g., sampling variation) from real (i.e., individual differences) heterogeneity is determined. Results show that all investigated methods tend to interpret all sources of variance in the data as individual differences, leading to an overestimation of the amount of heterogeneity between individual network models.

In response to these findings, Chapter 3 introduces a novel methodology, the Individual Network Invariance Test (INIT) to test differences between idiographic network models directly. By utilizing model comparison techniques, INIT compares a model that assumes homogeneity

2 This thesis comprises six prominent works of the author. A full list of publications is available in Appendix K

(i.e., the two network structures under investigation are constrained to be identical) to a model that assumes heterogeneity (i.e., the two network structures under investigation are allowed to vary) and determines which of these competing models describes the data at hand best.

In Chapter 4, a tutorial on the setup and execution of simulation studies within the network modeling framework is presented. Such simulation studies can be used to quantify the extent to which observed model differences can be attributed to estimation accuracy rather than genuine individual differences. By equipping more researchers with the tools to conduct simulation studies, this chapter aims to determine the accuracy of empirical investigations using idiographic network models under various conditions and enhance interpretation.

1.1.2 *Part II: Understanding Heterogeneity*

In Chapter 5, network modeling techniques are used to comprehend heterogeneity within idiographic network structures over time. INIT is employed to determine the stability of idiographic symptom networks over a one-year period, specifically for individuals along the psychosis severity continuum. It explores if network stability over time is associated with psychopathology severity during that same time. The findings revealed that, for the majority of the sample, idiographic symptom networks remained stable over time, while most individuals exhibited changes in psychopathology severity.

In Chapter 6, various techniques to shed light on differences in psychopathology between individuals are used. Of particular interest is the relationship between person-specific network structures and psychopathology stability. The findings demonstrate a strong association between person-specific network connectivity (i.e., the number and strength of connections within the network) and changes (either an aggravation or alleviation) in psychopathology.

In Chapter 7, stability landscapes are presented as a tool to gain insight into the dynamics of statistically estimated idiographic network models. We relate the dynamics of these idiographic network models to both psychopathology severity and stability and provide a theoretical explanation for the results presented in Chapters 5 and 6. Furthermore, the results indicate the technique employed to estimate idiographic networks determines the extent to which the idiographic network can tell us something about symptom stability and/or symptom severity.

This thesis concludes with a general discussion in Chapter 8. A comprehensive summary of the work presented in this thesis is offered and the potential role of the work presented in this thesis is discussed within a broader context.

Part I

ASSESSING HETEROGENEITY

2

HETEROGENEITY: REALITY OR ILLUSION?

ABSTRACT

The use of idiographic research techniques has gained popularity within psychological research and network analysis in particular. Idiographic research has been proposed as a promising avenue for future research, with differences between idiographic results highlighting evidence for radical heterogeneity. However, in the quest to address the individual in psychology, some classic statistical problems, such as those arising from sampling variation and power limitations, should not be overlooked. This chapter aims to determine to what extent current tools to compare idiographic networks are suited to disentangle true from illusory heterogeneity in the presence of sampling error. To this end, we investigate the performance of tools to inspect heterogeneity: (i) visual inspection, (ii) comparison of centrality measures, (iii) investigating standard deviations of random effects, and (iiii) GIMME, through simulations. Results show that power limitations hamper the validity of conclusions regarding heterogeneity and that the power required to assess heterogeneity adequately is often not realized in current research practice. Of the tools investigated, inspecting standard deviations of random effects and GIMME proved the most suited. However, all tools evaluated leave the door wide open to misinterpret *all* observed variability in terms of individual differences. Hence, the current chapter calls for caution in the use and interpretation of new time series techniques when it comes to heterogeneity.

2.1 INTRODUCTION

Ever since Molenaar [201] aimed to bring back the individual into scientific psychology once and for all with his classic manifesto, there has been a rise in idiographic research. This rise is mainly fueled by the realization that inter-individual (nomothetic) and intra-individual (idiographic) levels of analysis do not necessarily yield similar results—a concern that has been pointed out numerous times [e.g., 33, 94, 126, 169, 243, 246]. Pop-

Published as: **Hoekstra, R. H. A.**, Epskamp, S., & Borsboom, D. (2023). Heterogeneity in individual network analysis: Reality or illusion?. *Multivariate Behavioral Research*, *58*(4), 762-786.

ulation heterogeneity is often brought forward as a reason for this lack of overlap: for instance, individuals may differ from each other not only quantitatively but qualitatively, and current research practice struggles to take these differences into account.

Over the years, network analysis has rapidly gained popularity within psychology (for an overview of the literature, see Fried et al. [102] and Robinaugh et al. [227]). In network analysis, psychological constructs are represented by *nodes* and *edges*. Nodes indicate variables that play a role in the psychological construct of interest, e.g., symptoms, where edges represent the statistical relationship between these nodes [23, 28, 54]. This statistical relationship depends on the method used: often, edges represent partial correlations or (logistic) regression coefficients. The most common way to estimate a network is by applying a Gaussian Graphical Model (GGM; [183])—an undirected network model with partial correlations—to cross-sectional data [87, 227]. As such, the edges in these types of networks represent the strength of the statistical association between two nodes while controlling for every other node in the network.[1]

Recently, the network paradigm has converged with the intra- individual modeling tradition, as the estimation of *individual network models* based on time series data has become in favor. In this approach, a single individual is measured frequently over an extended period of time, after which a subject-specific network is estimated [89]. The temporal ordering of time series data adds two challenges to network estimation: on the one hand, because consecutive time points violate independence assumptions, standard GGM estimation techniques for cross-sectional data cannot be used, and on the other hand, the temporal information allows for identifying relationships over time, providing insight into Granger causality [113]. Statistical complications arising from violations of independence can be resolved by estimating a *temporal network*—a network with directed edges that provides information regarding patterns among variables as they unfold over time—in addition to an undirected network containing partial correlations. This latter network is referred to as the *contemporaneous network* and may provide insight into patterns that occur at a time scale different from the one defined by the spacing of the measurement occasions. Especially within clinical practice, a detailed understanding of the individual and their development over time is deemed to be important, and to this end, these types of networks are seen as a promising tool [42].

Intra-individual research using network models has regularly claimed evidence for heterogeneity when comparing individual networks [e.g., 9,

1 Interestingly, the network paradigm was also foreshadowed by Molenaar's work, as he was the first to identify the relation between the Ising model from statistical mechanics—a model very similar to the GGM but applied to dichotomous data—and the Rasch model from Item Response Theory [200, p. 82].

67, 185, 217, 222]. A common way to analyze heterogeneity within network analysis is by estimating intra-individual network models and using tools to compare the individual network models to one another. In accordance, the observation that network models appear to show differences across individuals is often seen as a vindication of the $N = 1$ paradigm, as it seems to support the idea that understanding intra-individual processes requires intra-individual data, so that "[I]f one wants to know what happens in a person, one must study that person" [32, p. 216].

However, this type of research runs the risk of mistaking noise for heterogeneity by directly interpreting *all* observed variability in individual network structures in terms of individual differences. After all, not all variability is due to individual differences; some variability is caused by fluctuations within the data due to sampling error and variance sources unrelated to the constructs of interest. As of yet, it is unclear precisely to what extent current metrics used to detect heterogeneity within network analysis are sensitive to such sources of variance.

Therewith, the aim of this chapter is twofold: we (1) demonstrate how easily noise can be mistaken for heterogeneity, and (2) we shed light on whether popular tools within network analysis can separate heterogeneity from sampling variation. First, we illustrate the influence of sampling variation when comparing individual network models by means of a thought experiment. Second, we investigate the ability of four "heterogeneity metrics" to separate real from illusionary heterogeneity by means of a simulation study. We provide the reader with a measure to calculate the expected overlap between two estimated networks and leave suggestions for future assessments of network heterogeneity.

2.2 NETWORK HETEROGENEITY: A THOUGHT EXPERIMENT

To illustrate the influence of sampling variation within network analysis and its relation to power and heterogeneity, let us conduct a simple thought experiment. Imagine that we have two data-generating network models: One fully connected network containing five nodes and 10 edges and an empty network with the same number of nodes.[2] We refer to these networks as our true underlying networks. For a graphical representation of the two networks, see panel (a) in Figure 1 for the fully connected network case and Figure 2 for the empty network case. Further assume all edges in the network are of equal strength and that our hypothetical sample is entirely homogeneous, i.e., each individual making up our sample has the same true underlying network structure. Now suppose we have a

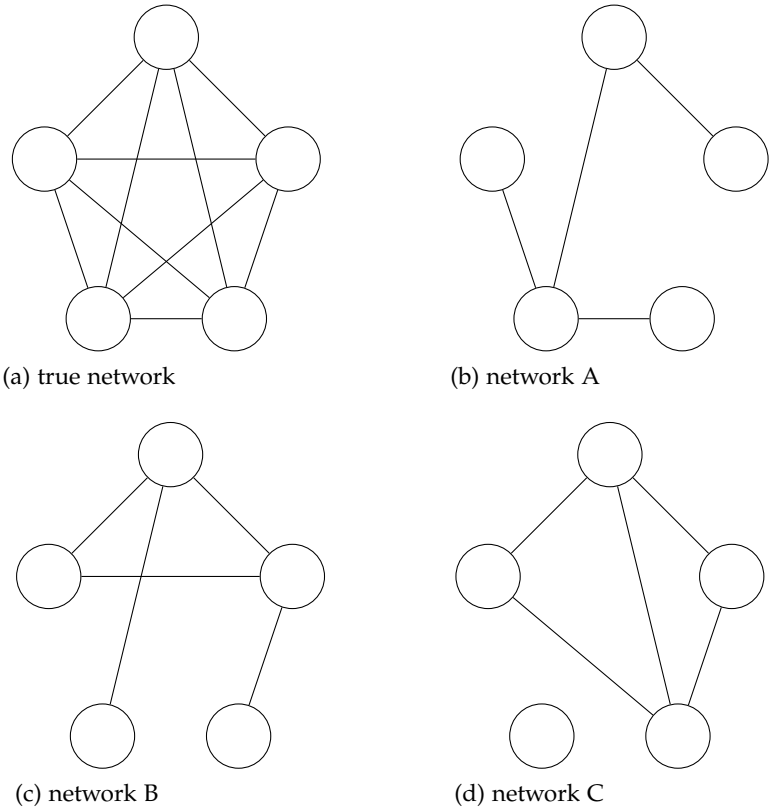
2 For simplicity, we will use undirected contemporaneous networks and ignore temporal effects by assuming temporal networks are empty. The thought experiment presented here could easily be extended to temporal networks as well, as the same logic holds.

sensitivity of 50% and a specificity of 90%. That is, we have a probability of 50% to detect an edge that is truly present within our network, and we have a probability of 90% to reject an edge that is *not* truly present in our network. These values are representative for the sensitivity and specificity when $t = 50$ based on previous simulation studies regarding individual network models [89, 195].

With these assumptions in hand, let us focus on the fully connected network first. Given a sensitivity of 50%, we expect to pick up, on average, five out of the 10 true edges. Assuming all edges are equal in strength, roughly 50% of the edges will show up in our generated network model at random. Now the question is, if we randomly select three individuals from our homogeneous sample, how much overlap between these three individual network models do we expect to see?

When visually inspecting the randomly drawn individual networks based on this scenario, see panel (b), (c) and (d) in Figure 1, one may be tempted to conclude there are plenty of individual differences, as none of the resulting networks look similar to one another. As an indication of the overlap between the networks, we could calculate the ratio of shared edges between two networks, divided by the number of possible edges ($n(n-1)/2$, in which n stands for the number of nodes), and multiply by 100 to convert the ratio into a percentage. Using this as a metric to calculate the resemblance between two networks in Figure 1, the overlap between Network A and B is 20%, the overlap between Network A and C is 10%, and between Network B and C is 40%. If we were to repeat this scenario, on average, we expect to see an overlap of 25% between two randomly selected estimated individual networks, even though the true underlying data generating network structure is exactly the same.

Moving on to the case where our true underlying network is empty, we see even less of an overlap between the generated individual network models, see Figure 2. Remember, we operate with a specificity of 90%. This implies we have 10% chance of finding an edge that is absent in reality. The fact that there are 10 potential edges means we expect, on average, one false edge to be included in each estimated network model. Here again, assuming all edges are equal in strength, this erroneous edge can show up anywhere in our network. Once more, let us generate random networks for three individuals, see panel (b), (c), and (d) in Figure 2. When visually inspecting these networks, again, our initial conclusion would be that there is a great deal of variety in the resulting networks. If we were to calculate the overlap between the networks as we did for our fully connected case, we would now find an overlap between our three networks of 0%. If we were to repeat this scenario, on average, we expect to see an overlap of 10% between any two randomly selected individuals.



(a) true network (b) network A

(c) network B (d) network C

Figure 1: Graphical representation of the fully connected network case in the thought experiment. The left upper panel (a) shows the data-generating network model containing five nodes and 10 edges, i.e., a fully connected network. We refer to this network as the true network. Based on this true underlying network structure, three network structures were randomly generated, yielding a sensitivity of 50%. For simplicity, edge weights are assumed to be of equal strength. The three randomly drawn networks based on these conditions are depicted in panel (b), (c) and (d). Solely based on a visual comparison of the three randomly generated network structures visualized in panel (b), (c) and (d), not knowing the true underlying network, one would be tempted to conclude that these three networks are very different, i.e., there is a great amount of heterogeneity present. While in fact, all of this perceived "heterogeneity" is caused by sampling variation.

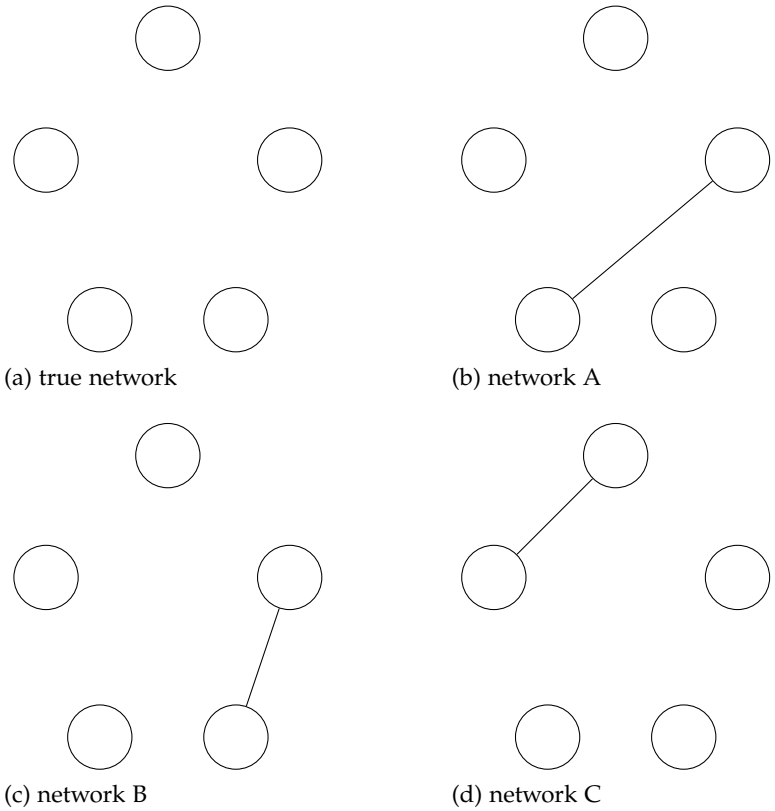


Figure 2: Graphical representation of the empty network case in the thought experiment. The left upper panel (a) shows the data-generating network model containing five nodes and 0 edges, i.e., an empty network. We refer to this network as the true network. Based on this true underlying network structure, three network structures were randomly generated, yielding a specificity of 90%. For simplicity, edge weights are assumed to be of equal strength. The three randomly drawn networks based on these conditions are depicted in panel (b), (c) and (d). Solely based on a visual comparison of the three randomly generated network structures visualized in panel (b), (c) and (d), not knowing the true underlying network, one would be tempted to conclude that these three networks are very different, i.e., there is a great amount of heterogeneity present. While in fact, all of this perceived "heterogeneity" is a direct result of the specificity of the estimation technique.

Analytically, we can derive the probability of obtaining the same network twice in both conditions of our thought experiment (i.e., the true fully connected case and the true empty case) to be:

$$\sum_{m=0}^{k} \binom{k}{m} p^{2m} (1-p)^{\binom{n}{2}-2m+k},$$

in which n is the number of nodes, k is the total number of possible edges (i.e., $n(n-1)/2$), and p represents the sensitivity in the fully connected network case and $1 - specificity$ in the empty network case. Using this expression, we can calculate that the probability of obtaining the exact same network twice in the thought experiment where the true network is a fully connected, with $n = 5$ nodes and sensitivity of $p = 0.5$ to be less than 0.1%. Likewise, we can see that in the case of an empty network with a specificity of 0.9 (i.e., $p = 0.1$) the probability of obtaining two identical networks is 13.7%.

This thought experiment shows that we can find a great deal of variety in individual networks when visually comparing them, despite our true underlying network model being invariant over individuals. This exposes how prone the visual inspection of network models is to error. In addition, often only the number of present edges is taken into account in order to determine overlap between two networks, but if we were to, for example, base the overlap between two networks not only on the estimated edges but also on the absent edges, the overlap between the networks in our fully connected case would rise to 50% and in the empty case to an astonishing 90%. Given our homogeneous sample, another way to increase the estimated overlap between two individual network models could be established by increasing sensitivity. Sensitivity (i.e., the power to detect an existing edge) increases as the number of time points increases. Suppose we increase our sensitivity to 90%. In the case of a fully connected network, this means that if we were to draw two random networks to compare, we would expect an average overlap of 80%, and the chance of two randomly chosen estimated network models to be exactly alike would be 13.7%.

In order to increase power without inflating the Type-I error rate—i.e., increasing sensitivity while keeping specificity high—there is only one solution: we need to collect more data. But how much more data? Much work has been dedicated to identifying the minimum amount of data required when performing ideograph analysis [e.g., 89, 181, 195, 206]. However, a considerable variety of results has been reported, likely as a function of simulation and estimation details as well as variation in the network structure simulated from[3], and simulation results were not al-

3 The work of Epskamp et al. [89] shows $t = 100$ is sufficient when performing a GVAR analysis, while the work of Mansueto et al. [195] shows sufficient sensitivity when $t = 500$.

ways viewed in light of perceived heterogeneity. Therefore, it is unclear how much data is needed in order to achieve the preferred sensitivity to make a valid claim about heterogeneity when comparing individual networks.

The scenario sketched in our thought experiment is merely a hypothetical one. In reality, we deal with various edge strengths and differences in network structures, such as a more sparse or a more densely connected network structure. There is a delicate interplay between edge weights, network structure, properties of the data (such as sample size and effect size), and the estimation technique used. To shed light on this interplay, we turn to a simulation study to determine the effects of sampling variation on detecting heterogeneity in individual networks under different network structures using three popular estimation techniques.

2.3 BEYOND THOUGHT EXPERIMENTS

To further illustrate the main point of this chapter—be careful with interpreting all variability as evidence for individual differences—we conducted a simulation study. To this end, we applied three idiographic network estimation tools (graphical VAR [graphicalVAR], multilevel VAR [mlVAR], and Group Iterative Multiple Model Estimation [GIMME]) to estimate individual network structures from simulated homogeneous data. This simulation study aims to answer the question to what extent current network-based tools to detect heterogeneity yield valid results to determine the amount of heterogeneity present. Throughout this chapter, we have defined the concept of heterogeneity to mean the absence of homogeneity. This is a strong stance on heterogeneity, as the slightest difference is taken as a sign of heterogeneity. Within the network literature, the operationalization of heterogeneity differs. Therefore, it is crucial to have an understanding of different individual network model techniques that can be used to operationalize heterogeneity in order to explore if, and how, these methods account for normal fluctuations within the data. To this end, we provide a brief overview of the most commonly used idiographic network estimation tools.

2.3.1 *An overview of idiographic network estimation tools*

The most common way to estimate individual networks is with some type of Vector Autoregression (VAR) model. There are two modeling frameworks that extend the VAR model to incorporate both temporal and con-

In addition, Lane and Gates [181] shows $t > 60$ is suitable for GIMME to pick up small to moderate effects while the work of Nestler and Humberg [206] shows $t > 100$ is necessary for GIMME to perform well.

temporaneous network structures: the graphical VAR model (GVAR; [89]) and the structural VAR model (sVAR; [46]). The GVAR model represents the contemporaneous network using *undirected* effects, whereas the sVAR model represents the contemporaneous network using *direct* effects. Several methods exist for estimating model parameters (edge weights) coupled with model structure (presence or absence of edges) for both modeling frameworks and for $N = 1$ and $N > 1$ datasets.

In $N = 1$ settings, the GVAR model can be estimated through iterative regularized estimation by using the multivariate regression with covariance estimation algorithm (MRCE; [1, 119], which is implemented in the R packages *graphicalVAR* and *sparseTSCGM*, or the model can be estimated through maximum likelihood estimation as implemented in the *psychonetrics* R package [77, 73]. The sVAR model can be estimated in the $N = 1$ setting through model search using generic structural equation modeling software such as the R package *lavaan* [233]. In the $N > 1$ setting, each of the $N = 1$ methods can be used separately for each individual, a practice investigated in this chapter.

In addition to estimating a network model for each individual separately, methods exist that allow one to borrow information across participants. In particular multi-level estimation is often used with the GVAR model by using the two-step multi-level GVAR algorithm as implemented in the R package *mlVAR* [89, 83], or Bayesian estimation as implemented in *Mplus* version 8 and higher [242]. The sVAR model is often estimated in $N > 1$ settings using GIMME [108], which is implemented in the R package *gimme* [182]. For details on the modeling frameworks used in this simulation study and the estimation techniques used, see Appendix A.

In this chapter, we focus on three methods for GVAR and sVAR estimation: MRCE using *graphicalVAR*, two-step multi-level estimation using *mlVAR*, and Group Iterative Multiple Model Estimation using *gimme*. We focus on these methods because they have been used for the purpose of estimating individual networks and detecting heterogeneity in existing research [e.g., 12, 9, 40, 222, 231]. To simplify the description of results below, we will refer to each of these three methods by referring to their corresponding R packages: *graphicalVAR*[4], *mlVAR* and *GIMME*. We expect results from other methods, such as maximum likelihood estimation of $N = 1$ GVAR models or Bayesian multi-level estimation of $N > 1$ models, to align, as these methods perform similarly in estimating network structures from data (e.g., Mansueto et al. [195] shows a strong overlap between the *graphicalVAR* and *psychonetrics* packages). In the following sections, we will discuss these three methods, the tools used to detect

4 *GraphicalVAR* therefore refers to the MRCE method implemented in the *graphicalVAR* R package, not to the GVAR modeling framework, which can also be estimated using other methods.

heterogeneity, and examples of their use in practice for each method separately.

2.3.1.1 *Regularized estimation using graphicalVAR*

The *graphicalVAR* package uses the MRCE algorithm to estimate the GVAR model [77]. This algorithm makes use of LASSO regularization [252], which iteratively estimates temporal coefficients (regression weights between $t-1$ and t) through regularized regression, and contemporaneous coefficients (partial correlations after controlling for temporal effects) through the graphical LASSO algorithm [106]. The algorithm utilizes two LASSO penalty parameters, which are chosen by optimizing the extended BIC (EBIC; [47, 1, 89].

It is important to note that as of yet there are no techniques available that are developed specifically to detect heterogeneity using *graphicalVAR* estimates. The detection of heterogeneity after estimating a GVAR model through *graphicalVAR* is currently mainly based on (1) differences in network topology, (2) differences in network density, and (3) differences in node connectivity measures (i.e., measures that are often used to assess the relative importance of nodes within the network[5]).

De Vos et al. [67] can be taken as an example of how visual inspection of individual network models is used to detect heterogeneity. De Vos et al. [67] estimated individual network models for people diagnosed with Major Depressive Disorder (MDD) and healthy controls. Participants completed a questionnaire assessing seven positive and negative affect items (e.g., "Feeling cheerful" and "Feeling irritated") three times a day over a period of 30 days. Using visual inspection De Vos et al. [67] concluded that individual networks did not resemble each other in terms of density and topology, taking this as an indication for a very strong level of heterogeneity. Fisher et al. [95] took a somewhat similar approach. The authors estimated individual network models for people with Generalized Anxiety Disorder (GAD) and MDD and for individuals who presented a comorbid clinical picture of both GAD and MDD. Forty participants completed questions on anxiety and depression symptoms (e.g., "Feeling hopeless", "Loss of interest or pleasure"), positive and negative affect, rumination, behavioral avoidance, and reassurance seeking four times a day for at least 30 days. In addition to examining the network topology, Fisher et al. [95] examined strength centrality. Results showed that node centrality metrics differed strongly between individuals. The authors interpret these differences as an indication of heterogeneity among individuals with GAD, MDD, or a comorbid presentation of GAD and MDD.

5 For an overview of connectivity measures often used in psychological networks see Costantini et al. [51] and Robinaugh, Millner, and McNally [229].

The two studies described here should be taken merely as an example. Given the importance of the question of whether individuals differ from one another (and from the between network structure), more studies with similar interests and set-up can easily be found. For more recent examples see [156, 185, 231].

2.3.1.2 *Multilevel estimation using mlVAR*

The second method we discuss is two-step multi-level GVAR estimation using the *mlVAR* package [83], which is based on the multi-level VAR models proposed by Bringmann et al. [40]. This package first estimates temporal coefficients by performing a series of multi-level node-wise regressions between time points $t-1$ and t. Within-person centering of predictors is used in each of the regressions, and person-wise means are included as between-person predictors, such that within- and between-person effects can be separated [127]. This separation also leads to between-person effects, that are gathered in a *between-persons* network, which we do not use in the present chapter (we refer the interested reader to Chapter 6, where we discuss and model such between-person networks). Next, in a second step, another series of node-wise regressions are performed on the residuals of the first series of node-wise regressions, leading to estimates of the contemporaneous networks.

As such, the *mlVAR* package leads to the estimation of three network structures: (a) a contemporaneous network (per subject and fixed-effect structures over all subjects), (b) a temporal network (per subject and fixed-effect structures over all subjects), (c) a between-subjects network (fixed-effects only). In addition, the standard deviations of random effects across the population on the temporal and contemporaneous network parameters are also returned and can be visualized as networks, leading to two more networks: (d) a temporal network of random effects and (e) a network of the standard deviations from the random effects.

In contrast to fully idiographic $N = 1$ estimation, for example, by using the *graphicalVAR* package, multilevel estimation offers a systematic approach to detect heterogeneity. Because of the multilevel structure of the model, one can inspect the network of standard deviations of random effects. These standard deviations show the degree to which network parameters exhibit individual differences. Bringmann et al. [40] recommend a cut-off of one standard deviation for the resulting edge weights for "large individual differences". This means that every edge with a weight above one standard deviation can be seen as a heterogeneous edge. In addition, individual differences can also be observed by visually inspecting personalized networks obtained by adding the fixed and random effects.

Bringmann et al. [40] applied multilevel VAR, on which the *mlVAR* package is based, to ESM data from 129 participants with depressive

symptoms. Participants' mood was assessed by a questionnaire containing six mood variables (e.g., "Cheerful" and "Fearful") 10 times a day over a period of 6 days. Bringmann et al. [40] inferred a network of inter-individual differences by examining standard deviations of the random effects. Results indicated a high level of individual variability on the self-loop for worry, as on the self-loops for cheerful, sad, relaxed, and fearful. Furthermore, variability in the relations between cheerful and relaxed, cheerful and sad, and fearful and worry was found, from which the authors concluded that a fair amount of heterogeneity in mood between the participants with depressive symptoms was present. As opposed to tools used to assess heterogeneity for graphical VAR models, the tool used to assess heterogeneity for multilevelVAR—inspecting standard deviations from the random effects—operationalizes heterogeneity in terms of edge weight differences.[6]

2.3.1.3 *Group Iterative Multiple Model Estimation*

The third and final method we evaluate is GIMME [108, 182], as implemented in the *gimme* package, which is a method for estimating sVAR models in $N > 1$ datasets [11, 181]. GIMME makes use of stepwise model search strategies through structural equation models, for example, by utilizing the *lavaan* package for R. Similar to multilevel modeling, GIMME aims to combine the worlds of nomothetic and idiographic research. However, where multi-level modeling aims to search for quantitative similarity across people (what is the edge-weight?), GIMME aims to search for qualitative similarity across people (which edges are included?). GIMME does this by creating person-specific as well as group-level edges. Group models are estimated by an iterative search procedure that identifies temporal and contemporaneous relationships that would significantly improve model fit for most individuals. After obtaining group-level effects, GIMME will then search for temporal and contemporaneous individual-level effects, continuing this process until an optimal fit is ensured. In addition to individual model output, this leads to a model containing both temporal and contemporaneous group-level edges and temporal and contemporaneous individual-level edges, where the width of the edge corresponds to the number of individuals this edge was estimated for.

The detection of heterogeneity can be executed by inspecting the number of individual paths. Group-level paths reflect homogeneity, as these edges need to be present in at least 75% of the sample, whereas individual paths reflect heterogeneity [12, 11]. This means heterogeneity is operationalized in terms of the number of individual level edges, i.e., edges that did not improve model fit for most individuals but do so for specific

6 Which is indirectly related to network topology and density.

individuals. Beltz et al. [12] applied GIMME to data from 25 individuals with personality pathology. Participants were instructed to complete an online survey on related clinically relevant behaviors for 100 consecutive evenings. Sixteen items from the daily surveys concerning behavioral manifestations of personality disorders were used to take four personality facets into account: negative affect, detachment, disinhibition, and hostility. Results revealed some group-level contemporaneous relations indicating some homogeneity, however, weights for these relations differed across participants. In addition, participants showed different relations on the individual level, reflecting heterogeneity.

In sum, several tools are available to assess the amount of heterogeneity for individual network models. However, the question remains how suitable these metrics are to reveal the prevalence of heterogeneity when sampling variation is taken into account. Our ability to disentangle sampling variation from the true effects in our data, i.e., power, is closely related to our ability to detect heterogeneity. A question that arises is whether, with current tools, we have enough power to separate sampling variation from true heterogeneity. To what degree can we be confident that we are looking at heterogeneity and not just normal fluctuations in our data that arise as a function of sampling variance? We investigate this question in simulation studies reported below.

2.3.2 *Simulation study*

2.3.2.1 *Methods*

NETWORK STRUCTURES Homogeneous data were generated based on three network structures: a synthetic-data network, a more sparse network estimated from data of one clinical patient, and a dense network estimated from data of multiple patients. We will refer to the first network structure as the *synthetic-data network*, the second as the *case-data network* and the third as the *Geschwind-data network*.[7]

The synthetic-data network structure is a sparse chain graph containing eight nodes, with a network density (the number of present edges divided by the number of possible edges) of 29% for the contemporaneous network and 14% for the temporal network. The average absolute edge weights are $M = 0.34$ and $M = 0.32$ for the contemporaneous and temporal effects respectively. For the data generating PDC and PCC matrices see

7 As these data generating matrices are based on GVAR models and GIMME operates under a sVAR model we have included an additional sVAR network structure in our simulation study and performed the simulation set up as described in this section using this sVAR structure. For more details on the sVAR network structure, the simulation set up, and the results for the sVAR network structure, see Appendix A.

Appendix A.[8] We included this synthetic-data network model as previous simulation studies have shown that network estimation works well under this structure [89]. For a graphical representation of the synthetic-data network structure, see Figure 3 panel (a).

In an attempt to create networks that approximate reality, the other two network structures that were used as data-generating structures were estimated from clinical data. The case-data network is estimated from data of one clinical patient ($t = 47$) measured over a period of two weeks. The original network was estimated using graphicalVAR and contains seven nodes regarding the patient's mood such as "Relaxed", "Sad" and "Nervous". For more information on the data, see Epskamp et al. [80][9]. The density of the network is 48% and 10%, with an absolute average edge weight of $M = 0.11$ and $M = 0.16$ for the contemporaneous and temporal networks respectively. For the data generating PDC and PCC matrices, see Appendix A; for a graphical representation of the case-data network, see Figure 3, panel (b).

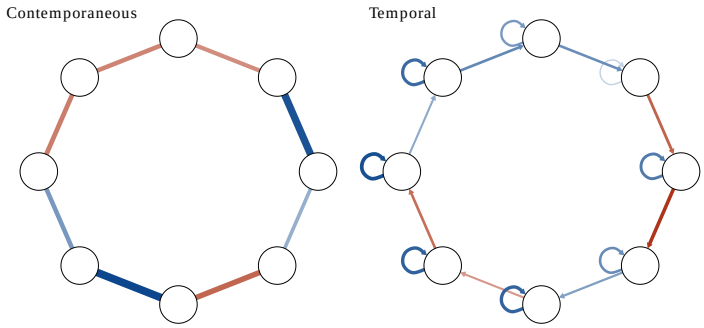
The Geschwind-data network is estimated from data of multiple patients ($n = 129$) measured over a period of six days (mean $t = 60$). For more information on the data set, see Geschwind et al. [111].[10] The original network was estimated using mlVAR and contains six nodes such as "Cheerful", "Sad", and "Relaxed". For this simulation study, the average contemporaneous and temporal networks were taken as the network structure of one subject. The density of the contemporaneous network and temporal network are 62% and 63%, with an average absolute edge weight of $M = 0.16$ and $M = 0.06$ for the contemporaneous and temporal effects respectively. For the data generating PDC and PCC matrices, see Appendix A and for a graphical representation of the Geschwind-data network, see Figure 3, panel (c).

SIMULATION PROCEDURE Taking these three network structures as the true data generating model for each individual, we simulated homogeneous data in which any apparent individual differences are due to sampling variation. Details on the parameter values under which we simulated can be found in Appendix A. In addition to network structure, we varied the number of participants, $n \in \{50, 100, 200\}$, and the number of time points, $t \in \{50, 100, 200, 400\}$. These time points were chosen to represent plausible values, $t \in \{50, 100\}$, and potentially ideal values,

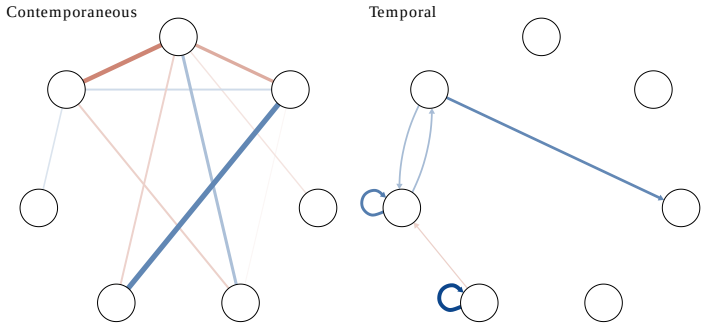
8 In addition to a synthetic-data network structure with 8 nodes, a sparse chain graph containing 16 nodes was generated in a similar fashion to inspect the effect of the size of the network. Results for this larger network structure can be found in Appendix A.

9 The data used for generating the network used in the current simulation study can be found in the supplementary materials of Epskamp et al. [80].

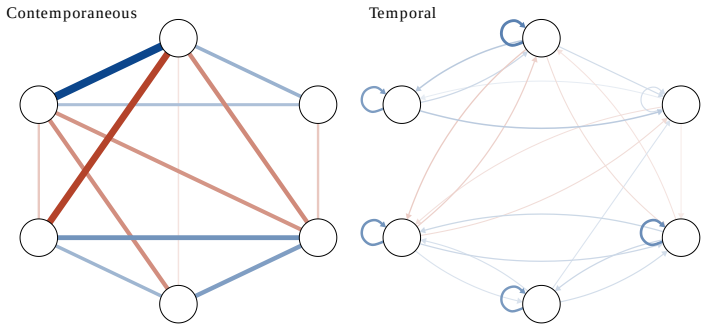
10 The data used for generating the network used in the current simulation study can be found in the supplementary materials of Bringmann et al. [40].



(a) Synthetic-data network model



(b) Case-data network model



(c) Geschwind-data network model

Figure 3: Contemporaneous and temporal data generating network models for the *synthetic-data*, *case-data*, and *Geschwind-data*. The upper panel shows the *synthetic-data network* with eight nodes simulated to be a chain graph, i.e., 1-2, 2-3, etc. The middle panel shows the *case-data network* containing seven nodes estimated from clinical data of one subject measured over time. The lower panel shows the *Geschwind-data network* containing six nodes, estimated from clinical data of multiple subjects. The average contemporaneous and temporal networks are taken as the data-generating network structure for each individual in this study. For each data-generating network, edges are scaled to the maximum edge weight, therewith edges across networks can not be visually compared to one another.

$t \in \{200, 400\}$, to include scenarios where the methods are expected to function well. We used three different estimation routines (grapicalVAR, mlVAR, and GIMME) to estimate the resulting network models, in total, creating a $3 \times 4 \times 3$ design. Each condition was repeated 100 times.

In line with common practice, in order to determine the amount of heterogeneity present in estimated network models through graphicalVAR, we evaluated topology of the resulting networks by visually inspecting three randomly chosen estimated network structures. For these individuals, we compared network density, i.e., the number of actual edges divided by the number of possible edges. Furthermore, across the entire sample, we computed the node centrality measure strength for each individual network. Strength centrality is defined as the sum of the edge weights of a given node (in absolute value). For temporal networks, strength is divided into in-strength, the sum of absolute incoming edge-weights, and out-strength, the sum of absolute outgoing edge weights. Centrality measures have been taken as an indication of the importance of individual nodes in a network [51, 210]. To determine the amount of resemblance in strength measures across individuals, we correlated the centrality measures for each possible pair of individuals. The distribution of these correlations gives us an idea of the spread of the strength of the association between centrality measures. If centrality measures are in fact a suitable measure for separating sampling variation from heterogeneity, we would expect a narrow distribution peaked around a strong positive correlation, because all samples are drawn from completely homogeneous populations.

For mlVAR estimates, we inspected the distribution of standard deviations of estimated random effects, both for contemporaneous and temporal networks. In line with Bringmann et al. [40] we used a cut-off score of one standard deviation. Edges from the standard deviation network with a weight above this cutoff are taken to represent ”large" heterogeneous effects, whereas edges below this cutoff are taken to be sampling error. In order to compare edge weights estimations from the different network structures, edge weights of the standard deviation networks for both contemporaneous and temporal networks were standardized.

To our knowledge, for GIMME, there is no cut-off or rule of thumb that has been proposed in the literature to determine whether a sample is homogeneous. We suggest taking the percentage of homogeneous edges as an indication of the amount of homogeneity. GIMME provides group-level output in which group-level effects are indicated by black edges, and individual effects are indicted with grey edges [181]. Group level edges are seen as homogeneous, whereas individual level edges are seen as heterogeneous [11]. To assess the amount of estimated homogeneity, we propose to take the percentage of estimated group level edges, i.e.,

the number of group level edges divided by the total number of edges estimated in the group level network[11]:

$$\%\text{ homogeneous edges} = \frac{\#\text{ group level edges}}{\#\text{ group level edges} + \#\text{ individual level edges}} \times 100$$

Data was generated based on the three previously described network structures using the `graphicalVARsim` function from the graphicalVAR package in R [77] which simulates data from a graphical VAR model. Network models were estimated from the simulated data using the R packages: graphivalVAR [77], mlVAR [83], and GIMME [182]. The simulation study was performed in R, version 4.0.2 [220]. In addition, a sensitivity analysis was performed for graphical VAR, multilevel VAR, and GIMME to inspect the resemblance between the data generated networks to the estimated networks. Results for the sensitivity analysis can be found in Appendix A for graphical VAR, multilevel VAR, and GIMME, respectively.

2.3.2.2 *Results*

Across all tools evaluated, sensitivity was often too low to make valid claims about heterogeneity (for sensitivity analysis of graphical VAR see Figure 33, for mlVAR see Figure 36, and for GIMME see Figure 37 in Appendix A).[12] In addition, sensitivity was dependent on network structure for the graphical VAR and the ml VAR method, as well as the estimation technique. Sensitivity was highest when the network structure was sparse (i.e., a network with relatively few edges) and the edges present were relatively strong as is the case for the synthetic-data network structure. When the network structure was more dense, as is the case for the Geschwind-data network, or showed relatively weak edges, sensitivity started to decline. This latter effect was particularly strong when estimating temporal networks using the graphical VAR method, and when estimating contemporaneous networks using GIMME.

VISUAL INSPECTION To illustrate the lack of sensitivity and its implications for the validity of claims on heterogeneity, we randomly drew

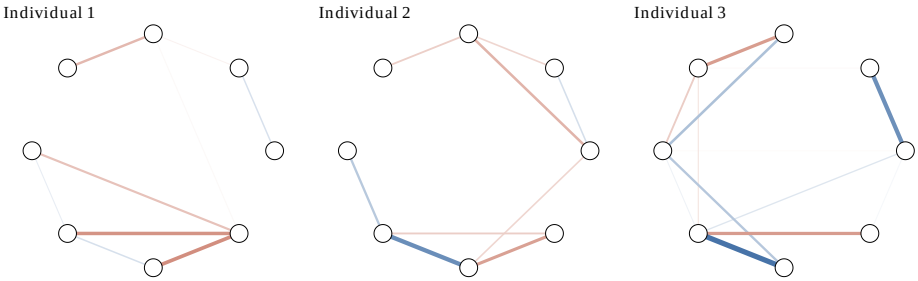
11 It is important to note that individual effects can be both contemporaneous and temporal. If an individual effect was estimated both as contemporaneous for some individuals and temporal for others, this was taken to mean there is some level of heterogeneity, and both edges added up to the total number of individual edges.

12 Sensitivity analysis for GIMME was infeasible for the original three data generating network structures as the data generating model for a GVAR model does not directly correspond to one sVAR model (the modeling framework used by GIMME to estimate networks). However, one sVAR does directly correspond to a unique GVAR model. Therefore, a sVAR data generating network is added to the simulation study, see Appendix A for further details on the data generating network structure and the full simulation study results.

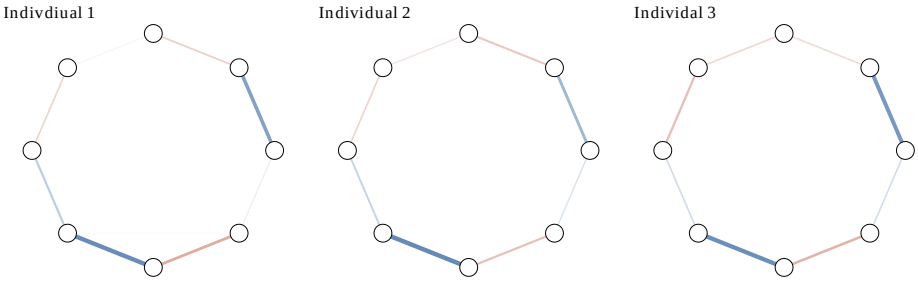
three individual networks estimated using the graphical VAR method from the 5,000 simulated networks for the synthetic-data network condition, see Figure 20 panel (a) and (c) to visually compare the three networks. It is important to note that these networks were chosen to convey how the visual comparison of individual networks can go wrong, especially when sensitivity is low. For $t = 50$ the individual contemporaneous and temporal networks did not differ much with respect to overall edge weight. Average edge weight and standard deviation for the estimated contemporaneous network model of individual 1: $M = 0.11(SD = 0.09)$, individual 2: $M = 0.15(SD = 0.11)$, and individual 3: $M = 0.14(SD = 0.12)$; Temporal network of individual 1: $M = 0.13(SD = 0.09)$, individual 2: $M = 0(SD = 0)$, and individual 3: $M = 0.16(SD = 0.14)$. However, individual networks differed with respect to network density. Network density for estimated contemporaneous networks: 36%, 14%, 46% and temporal network: 44%, 0%, 32%, of individuals 1, 2, and 3 respectively, and in detected edges. Few similarities could be found in terms of detected edges; in addition, the strength of these edges varied across the three exemplary individual networks.

Perceived heterogeneity vanished when visually inspecting the estimated network structures using graphical VAR when $t = 400$, see panels (b) and (d) in Figure 20. When visually inspecting the individual contemporaneous networks, the networks showed significant resemblance. Note that sensitivity is high for the contemporaneous networks; all seven data generating edges were estimated for each of the individual networks. However, in terms of network density, we still found notable differences. For the contemporaneous networks density differed (32%, 43%, 29%, for $t = 400$ for individuals 1, 2, and 3 respectively), as the networks of individual 1 and 2 showed a few weak erroneous edges. These edges can be disregarded in terms of edge weight in comparison to other edges present. In terms of overall edge weight we found slight differences: average edge weight and standard deviation for the estimated contemporaneous network model of individual 1: $M = 0.18(SD = 0.14)$, individual 2: $M = 0.18(SD = 0.14)$, and individual 3: $M = 0.22(SD = 0.12)$.

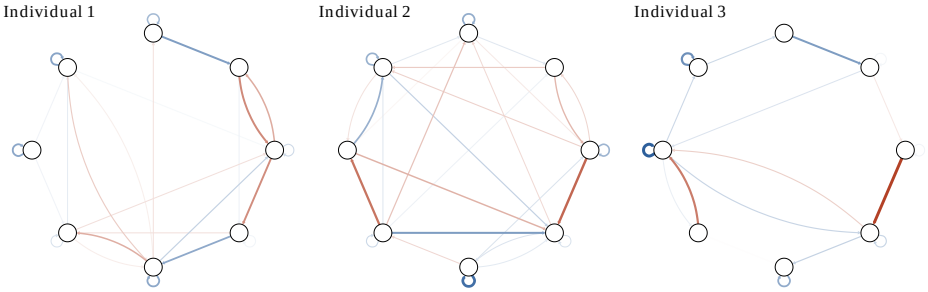
Although sensitivity for temporal networks as estimated with graphical VAR was high at $t = 400$, the exemplary temporal networks of individual 1, 2, and 3 showed a larger range of network density than for their contemporaneous networks (network density of 63%, 57%, 37%, for the estimated temporal networks of individual 1, 2 and 3 respectively). In terms of overall absolute edge weight, we found small differences (temporal network individual 1: $M = 0.15(SD = 0.13)$, individual 2: $M = 0.14(SD = 0.13)$, individual 3: $M = 0.20(SD = 0.13)$). Heterogeneity in terms of estimated edges, network density, and edge weights was more pronounced for the case-data and the Geschwind-data network structures. Examples of three



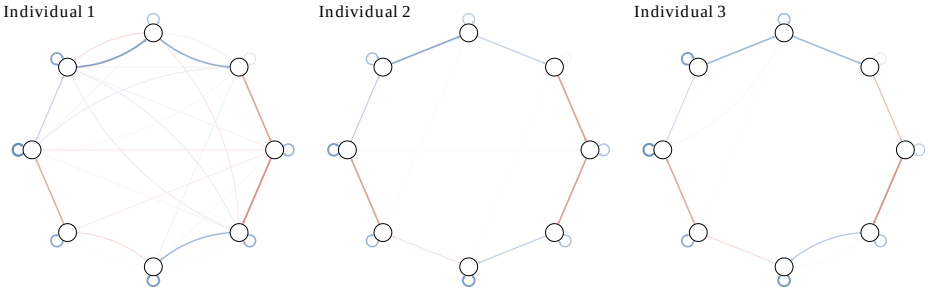
(a) Contemporaneous networks for individual 1, 2, and 3 when t = 50



(b) Contemporaneous networks for individual 1, 2, and 3 when t = 400



(c) Temporal networks for individual 1, 2, and 3 when t = 50



(d) Temporal networks for individual 1, 2, and. 3 when t = 400

Figure 4: Output from graphical VAR. Three randomly selected individual networks estimated using the graphical VAR method, generated under the same synthetic-data network structure (for a visualisation of the synthetic-data network see Figure 3 panel (b)) for two different time points (t = 50 and t = 400). Panel (a) shows the three individual contemporaneous networks when t = 50, panel (b) for t = 400, panel (c) shows their corresponding temporal networks when t = 50, and panel (d) for t = 400.

networks estimated under the case-data and the Geschwind-data network structure can be found in Figure 29 and Figure 30 in Appendix A.

CENTRALITY CORRELATION Moving forward, we investigated the use of centrality measures as an indication of individual differences between networks. For the entire sample, based on each individual network as estimated with the graphical VAR method, node centrality strength was computed. For contemporaneous networks, we computed one measure: strength (the sum of all absolute edge weights connected to the node), for temporal networks, this measure is divided in in-strength (the sum of all absolute incoming edge-weights) and out-strength (the sum of all absolute outgoing edge weights). For all possible pairs of individual networks within the sample, the strength estimates were correlated in order to give an indication of the resemblance in strength centrality. Figure 5 shows the distributions of the correlation of these strength measures for the entire sample. Regardless of network structure, strength measures correlated more with one another for contemporaneous and temporal networks as the number of time points increased. Furthermore, strength measures showed a higher correlation for contemporaneous networks than for temporal networks. For all three network structures, the correlation of strength measures for the contemporaneous network are somewhat similar from roughly $t = 200$ onward.

The median correlation for in- and out-strength for the Geschwind-data network remained low in all cases, in-strength: $r = 0.23$, out-strength: $r < -0.00$. For the case-data network the median correlation for in-strength and out-strength increased from $r = -0.13$ and $r = -0.01$, to $r = 0.86$ and $r = 0.94$ respectively. For the synthetic-data network the median correlation for in- and out-strength increased from $r = 5.14$ and $r = 0.15$ to $r = 0.67$ and $r = 0.63$. Surprisingly, in- and out- strength correlation increased most for the case-data network as opposed to the synthetic-data network.

It is important to note here that the correlation between node strengths, although regularly used in the literature, is a problematic measure of similarity [30, 97]. This is because, in general, weak correlations may be the result of low variance in estimated edge weights (i.e., a situation in which all edge weights are approximately the same). If the variance in individual centrality measures is low this can lead to very weak correlations (around zero) between two individual centrality measures, even when measures are similar. This means a high correlation could be an indication of similarity in strength centrality, while a weak correlation may also be an indication of similarity—making the interpretation of correlations asymmetric and less intuitive. In this particular case, however, weak correlations did not result because of a lack of variance, but rather because of

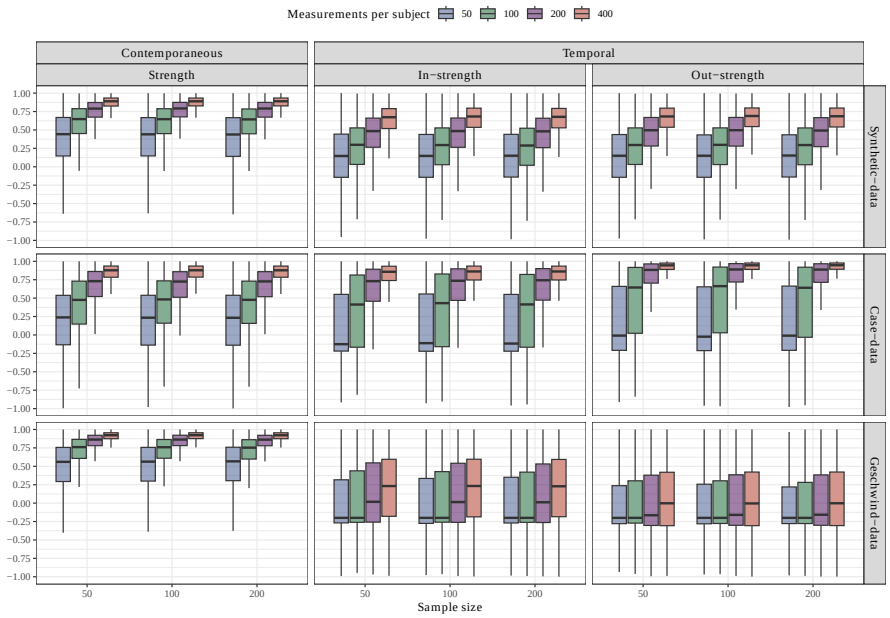


Figure 5: Correlations between node strength centrality calculated from individual networks estimated with graphical VAR. For the contemporaneous network, the correlation between strength centrality is depicted. For temporal networks, the correlations for two strength measures were computed: in-strength and out-strength. Correlations between strength measures of contemporaneous networks are highest for networks estimated under the Geschwind-data structure, while the in- and out-strength correlations are highest for networks generated under the synthetic-data structure.

a lack of a linear relation between individuals' centrality measures. Therefore, here we may safely interpret weak correlations as an indication of a lack of resemblance.

STANDARD DEVIATION OF RANDOM EFFECTS For mlVAR estimation results, we inspected the standard deviation of random effects for the contemporaneous and temporal effects. We took a cut-off of one standard deviation for the edge weights to determine whether population heterogeneity was present. In order to compare the edge weights of the different network structures and their standard deviations, edge weights were standardized. Standardized density distributions of the standard deviation of random effects for the contemporaneous and temporal effects can be found in Figure 6.

For contemporaneous network structures, the estimation of large heterogeneous effects was limited with relatively small sample sizes $n = 50$ and $t = 50$. For each of the three data generating network structures,

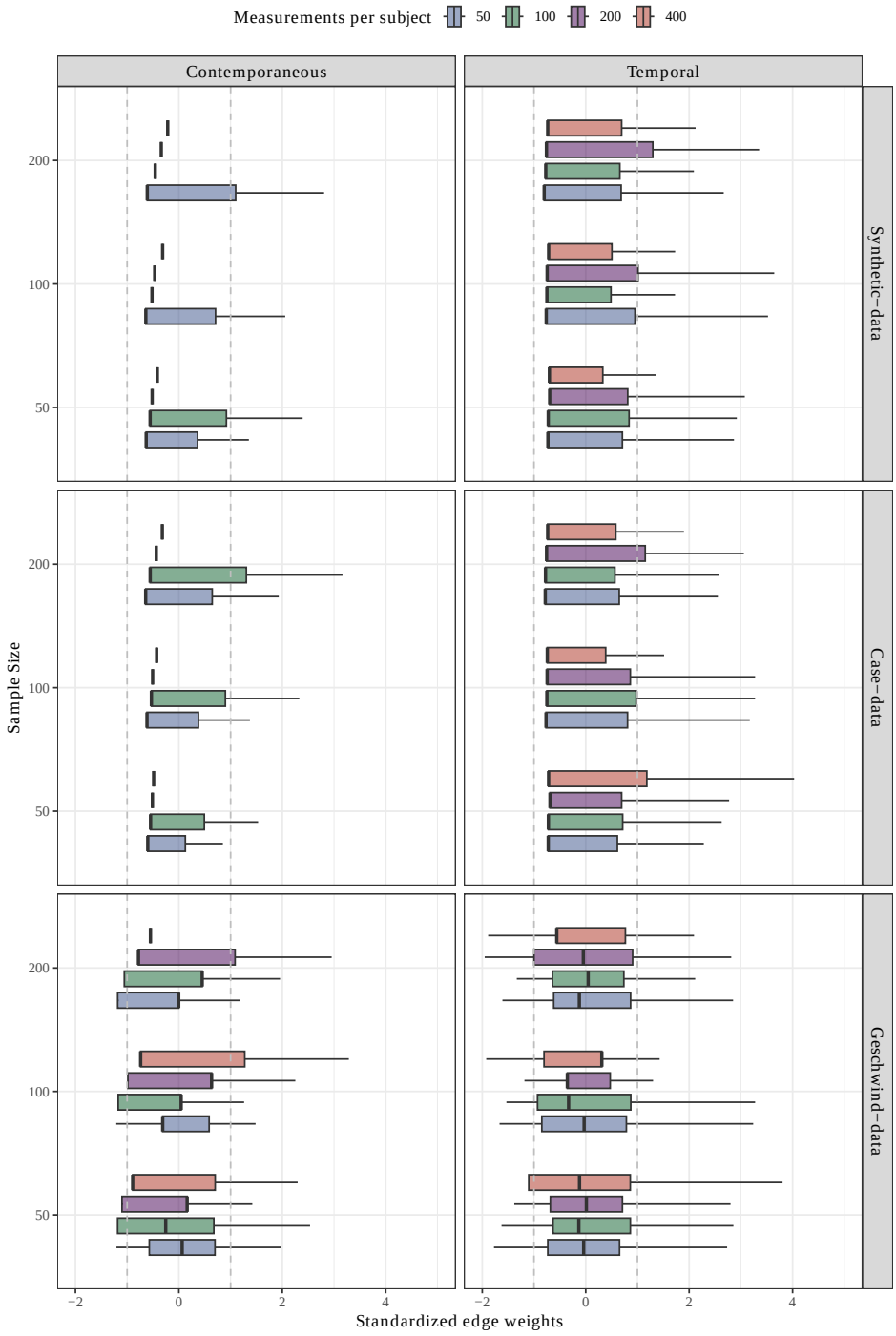


Figure 6: Density distribution of standardized edge weights for standard deviations of random effects for the multilevel VAR model. The vertical black dotted line represents the recommended cut-off of one standard deviation. Edges above or below these cutof'sf are seen as truly heterogeneous.

no heterogeneous edges were detected when the number of individuals increased to $n = 200$ and the number of time points per subject to $t = 400$. For all temporal network structures with $n = 50$, the estimation of heterogeneous edges was limited. Fewer differences were detected for the synthetic-data network structure and the case-data network structure than under a Geschwind-data network structure. When increasing time points to $t = 200$ per individual, even less heterogeneity was detected.

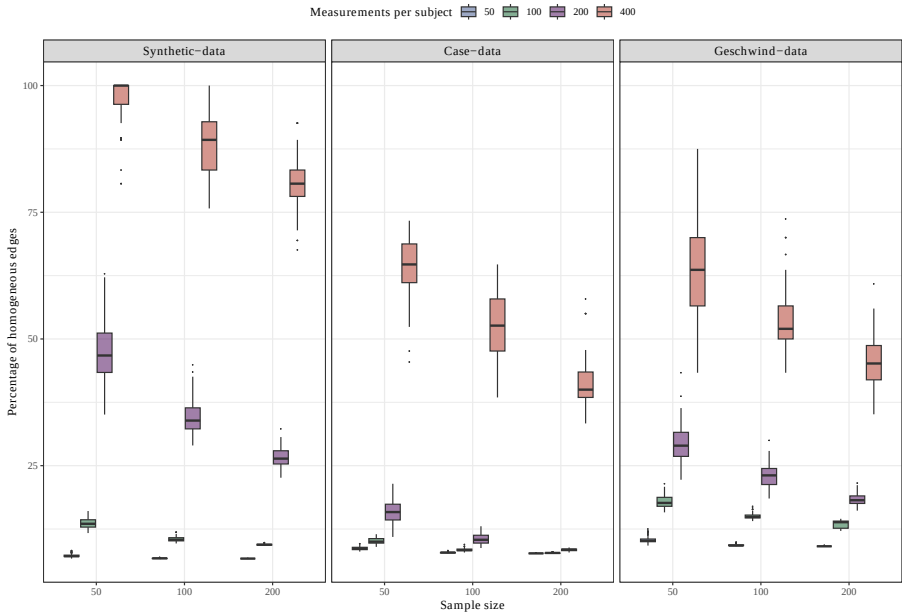


Figure 7: Percentage of homogeneous edges as estimated through GIMME for the synthetic-data, case-data, and Geschwind-data network structures.

PERCENTAGE OF HOMOGENEOUS EDGES In contrast to mlVAR, to the best of our knowledge, for GIMME there has been no cut-off recommended in the literature. We decided to inspect the percentage of homogeneous edges and take this as an indication of the amount of heterogeneity present, see Figure 7. The percentage of homogeneous edges was $7\%(SD = 0.33\%), 9\%(SD = 0.33\%), 10\%(SD = 0.64\%)$ for $n = 50$ and $t = 50$ for the synthetic-data, case-data and Geschwind-data network structures respectively. This percentage increased as the number of time points increased. When $t = 400$ the median percentage of homogeneous edges was $97\%(SD = 4.45\%), 64\%(SD = 5.48\%), 64\%(SD = 8.63\%)$ for synthetic-data, case-data and Geschwind-data network structures. Interestingly, we found that the percentage of homogeneous edges detected decreased as the number of participants increased. The same pattern was

found for a larger synthetic-data network structure, see Figure 23 in Appendix A.

2.4 DISCUSSION

The present chapter aimed to expose effects of sampling variation and power limitations in the investigation of heterogeneity in idiographic network models. We have shown that, even if the underlying network model is invariant across individuals, limitations regarding specificity and sensitivity can impose a great deal of variety in individual network structures estimated from the data, especially when the number of time points is small. In a simulation study, we evaluated four different network tools for assessing heterogeneity: visual inspection, comparing centrality measures, inspecting random effect standard deviations, and applications of GIMME. Results showed that low statistical power places considerable limits on the validity of conclusions regarding heterogeneity for all tools. At low sample sizes, an applied researcher is likely to erroneously conclude that a sample is heterogeneous even if it is in fact, homogeneous. Under the right circumstances (high sample size and favorable generating network structures), inspecting the standard deviation of random effect using multilevel VAR modeling and GIMME proved most suited for detecting heterogeneity, but these circumstances are currently not realized in the great majority of designs in many relevant fields, such as psychopathology research.

2.4.1 *Many lower-powered analyses versus one high-powered analysis*

When comparing idiographic network models, an important pitfall lies in the tendency to interpret *all* variability in resulting networks as heterogeneity between people, while some of this variability is the result of fluctuations in the data created by noise—e.g., sampling variation. Combining sampling variation with overall conservative estimation procedures (high specificity and lower sensitivity at low sample sizes) will inevitably lead to a mismatch in the presence and absence of estimated edges in addition to varying edge weights across individuals. Of course, the most straightforward way to combat these problems is to collect more data in an effort to increase statistical power without inflating the Type-I error (i.e., increasing sensitivity while keeping specificity high).

Importantly, the number of time points needed to estimate robust idiographic network models is dependent on the network structure (e.g., a more sparse or dense network structure) and the estimation technique used. However, we found that even when we simulated under ideal conditions —a sparse network structure with strong edges—at least 200 time

points per individual may be needed to obtain network structures that are sufficiently robust to support comparisons between individuals. The amount of time series data needed quickly adds up if the network structure is less than optimal, such as when one is dealing with a sparse or dense structure with weak edges. In these cases, the number of time points one needs per individual may exceed 400. This amount of data is often not realized in current research practice typical applications in, e.g., psychopathology research. Although some cases have been known to feature sufficiently intense measurement schemes in this respect [177], one rarely encounters time series of this length in, for instance, ESM designs.

In practice, we may see only about 60 observations per individual, [e.g., 9, 156, 231], which might seem like a large amount of data given the difficulty of collecting it, but this may not be enough to estimate a robust individual network. If longer time series cannot be obtained per individual, then the only other solution is to step down from fully idiographic research (estimate one model per person) and instead use an analysis strategy that borrows information from other participants (e.g., multi-level modeling or GIMME). This may then lead to performing one well-powered analysis rather than a sequence of lower-powered analyses at the idiographic level. Such methods performed best in our simulations as well. However, both methods still showed a great deal of illusionary heterogeneity (few common edges in GIMME and relative large random effect sizes in multi-level VAR modeling) and come with their own disadvantages, such as the assumption of heterogeneity, or in the case of multi-level modeling, pulling individual estimations towards the group mean [40].

2.4.2 *Comparing centrality across individuals*

There is an ongoing debate on the use and interpretation of centrality measures in network analysis [39, 64, 122]. Our results add to the discussion by showing that if one wants to use centrality measures as a way to determine heterogeneity, one should proceed with caution. Here we share four considerations to illustrate this. First, as we showed in our simulation study, it is possible to find negative correlations between estimated strength measures for two generated networks with the same true underlying network structure when the number of observations is less than $t = 100$ per individual. Second, although adding data leads to a stronger correlation of strength centrality measures between individuals in most settings, we showed that it is possible that the correlation between centrality coefficients is low even with a large number of data points per individual ($t > 200$). Third, while weak correlations between

strength centrality can be the result of a lack of overlap between two networks, this need not be the case because weak correlations between strength centrality can also be the result of low variance. While strong correlations between centrality measures are an indication of similarities between strength centrality, weak correlations do not necessarily have to be an indication of dissimilarity in strength centrality, making the interpretation of correlations between strength centrality counterintuitive; this warrants even more caution in the interpretation of these correlations. Fourth and last, the supposed difference in centrality measures could be due to sampling variation; hence, in the absence of systematic ways to statistically assess the correlation between centrality measures, it does not provide any way to differentiate between real heterogeneity and random fluctuations in the data.

Importantly, this does not mean that centrality coefficients cannot be used to study network topology or individual differences therein. Instead, it means that, just like with any other statistic, inspecting centrality estimates must always be assisted by assessments of precision and robustness (e.g., confidence or credibility intervals and other functions of sampling distributions). Such assessments have now become standard in cross-sectional network analysis [82]. Future research should investigate extending such methods for application to time series data. Until suitable measures of precision have become available, the use of visual inspection of differences in centrality measures to detect heterogeneity in network analysis must be considered questionable.

2.4.3 *Network heterogeneity and replicability*

The discussion of heterogeneity in idiographic network models both resembles and mirrors the discussion on network replicability in cross-sectional network studies [30, 97]. In cross-sectional network analysis, it has been recognized that there may be a great deal of sampling variation in the estimated network structures, which, combined with conservative estimation methods, may by itself lead to differences in estimated network models [103]. In addition, it has been recognized that fluctuations in centrality measures may be due to sampling variation even at high sample sizes, depending on the generating network structure and the centrality measure chosen [82]. To this end, data-driven sampling methods or more sophisticated statistical methods are routinely used together with presented results to assess accuracy and stability in a sample as well as to statistically compare network structures of different samples [82, 20, 275], as presumably, both *sampling variation* and *heterogeneity* lead to differences in estimated network structures.

In the literature on cross-sectional network models, differences between estimated network structures are sometimes highlighted as evidence for limited replicability of network models [97, 98]. However, these differences should be expected to arise partly due to the expected replicability of a method given its sensitivity and specificity [274]. When studying genuine heterogeneity over different samples used for network analysis, assessments of heterogeneity cannot be assessed without taking sampling variation into account [152]. This also holds true for network models (or any other statistical model) estimated from time series data. Of note, even though we simulated 400 data points as our largest sample size, such a sample size is not deemed very large when estimating a cross-sectional network of, for example, 10 nodes (featuring 45 parameters). Time series data features auto-correlations due to temporal ordering of the data, which both require more parameters to be estimated (145 in total for a graphical VAR model) and leads to effectively less information per observation. Both factors lead to a lower expected 'replicability' of the network structures, which is highlighted in our simulation results.

2.4.4 *Limitations and future directions*

At least two limitations of the current simulations need to be addressed here. First, although GIMME performed well with sufficient amounts of data, there is reason to believe the current simulation setup might put GIMME at a disadvantage. In our simulation study, we generated data based on graphicalVAR models, GIMME, however, operates under a structural VAR (sVAR) model. Running simulation conditions using a sVAR model underlying the data yielded similar results as for the GVAR models. However, our simulation study highlighted the influence of the network structure on the results. As GIMME was evaluated under just one sVAR structure, generalizations must be made cautiously. Thus, the present findings speak to the effectiveness of using GIMME when assessing heterogeneity in data that arise from a given VAR model as specified in this chapter but have less bearing on the general quality of GIMME as a statistical procedure operating within its own data space. Second, although the most common way to estimate individual networks is by using some type of VA R model, it should be noted that many more estimation procedures exist, for example, based on other types of autoregressive (AR) models than the VAR model, such as AR Moving Average (ARMA) or integrated ARMA (ARIMA) [35] (for an overview of these techniques see: Hamaker and Dolan [124]). However, we found that most of the network model papers focusing on the detection of heterogeneity made use of the techniques evaluated in this chapter.

We conclude by highlighting future avenues for the comparison of individual networks regarding heterogeneity. The current chapter has shown that several techniques used to compare individual network models are of limited usefulness in disentangling true heterogeneity from noise. Therefore, there is a pressing need for the development of tools that can do so effectively. A first possible route for future research is to rely more on hierarchical or multi-level models in which heterogeneity across individuals can explicitly be included in the model as random effects and tested to feature non-zero variance. Currently, the *mlVAR* package, which can be used for multi-level graphical VAR estimation, does not allow for this comparison. An alternative is to use the Dynamical Structural Equation Modeling (DSEM) module in *MPlus* [242], which allows for estimating multi-level VAR models, but with marginal correlations rather than partial correlations between innovation terms. This will allow for assessing random effects on temporal structures but not on contemporaneous partial correlation networks. A second option is to extend data-driven methods used to investigate differences in cross-sectional networks to results from time series models [20]. A problematic aspect here, however, is that resampling techniques cannot readily be applied as the temporal ordering of the data plays an important role in the analysis as well. Finally, promising Bayesian methods have been developed to assess not only evidence for differences but also evidence for the similarity between different samples [275]. Such methods could perhaps also be expanded for time series analysis.

2.4.5 *Conclusion*

In the quest to address the individual within psychology, some concerns arising from the use of these individual models are in danger of being overlooked, including common statistical problems that arise due to sampling variation. Although research within the field of individual network analysis is promising, it is vital to recognize the core principle of statistical analysis: accounting for the uncertainty that arises as a result of sampling variation. Sampling variation alone can lead to striking differences (illusory heterogeneity) in estimated models from different subjects, even if these subjects are identical (fully homogeneous). This chapter aimed to function as a wake-up call, addressing some of the concerns regarding analyzing the individual and calls for caution in the use and interpretation of these new time series techniques when making claims about heterogeneity. In the next chapter we provide a solution for these concerns and introduce a novel methodology to directly test for the (in)equality between idiographic network models. Furthermore, in Chapter 4, we provide a tutorial on simulation studies within network psychometrics that

could be used to account for the uncertainties that arises as a result of sampling variation.

3

TESTING FOR INDIVIDUAL DIFFERENCES

ABSTRACT

The comparison of idiographic network structures to determine the presence of heterogeneity is a challenging endeavor in many applied settings. Previously, researchers eyeballed idiographic networks, computed correlations, and used techniques that make use of the multilevel structure of the data (e.g., GIMME and mlVAR) to investigate individual differences. However, these methods do not allow for testing the (in)equality of idiographic network structures directly. In this chapter, we propose the Individual Network Invariance Test (INIT), which we implemented in the R package *INIT*. INIT extends common model comparison practices in Structural Equation Modeling (SEM) to idiographic network structures to test for (in)equality between idiographic networks. In a simulation study, we evaluated the performance of INIT on both saturated and pruned idiographic network structures by inspecting the rejection rate of the χ^2 difference test and model selection criteria, such as the AIC and BIC. Results show INIT performs adequately when $t = 100$ per individual. When applying INIT on saturated networks, the AIC performed best as a model selection criterion, while the BIC showed better results when applying INIT on pruned networks. In an empirical example, we highlight the possibilities of this new technique, illustrating how INIT provides researchers with a means of testing for (in)equality between idiographic network structures and within idiographic network structures over time. To conclude, recommendations for empirical researchers are provided.

3.1 INTRODUCTION

The network approach to psychopathology conceptualizes mental disorders as causal systems of reinforcing symptoms [23, 28, 29]. Studying person-specific dynamics via network modeling techniques is a promising avenue to guide therapeutic interventions and move toward individual process-based therapy [37, 43, 145]. This approach sparked an increase

Published as: **Hoekstra, R. H. A.**, Epskamp, S., Borsboom, D. Nierenberg, A. A., & McNally, R. J. (2024). Testing similarity in longitudinal networks: The Individual Network Invariance Test (INIT). *Psychological Methods*. Advanced Online Publication.

in the number of papers where a type of Vector Autoregressive Regressive (VAR) model is fit to time series data to uncover symptom dynamics at the individual level [e.g., 63, 67, 71, 105, 115, 156, 186]. Although applications of idiographic network analysis utilize different approaches to estimate, inspect, and interpret person-specific networks, they often share one important feature: they aim to compare idiographic network structures in order to identify individual differences in the data-generating mechanisms.

When it comes to comparing these intra-individual networks, researchers have mainly used tools for comparing group-level networks. These tools include visually comparing person-specific network structures, computing correlation coefficients between parameters (e.g., edge weights) that characterize individual network structures, or correlating topological measures of network structure such as strength centrality. A consistent finding emerges from these types of analysis: there is a great amount of heterogeneity among person-specific network structures [e.g., 10, 95, 185, 217, 222, 231]. Consequently, these results vindicate the fundamental reason to pursue personalized modeling [277].

However, tools such as comparing person-specific network structures by eyeballing, computing correlations between individual network structures, or between topological network measures leave the door open to interpret *all* variability in terms of heterogeneity. Yet not all variability is the direct result of individual differences [143], much like other forms of data sampling, some of the variability within individual time series data results from noise and sampling variation. Thus, simple descriptive analyses risk overestimating the amount of heterogeneity. Indeed, evaluating these risks, Hoekstra, Epskamp, and Borsboom [143] showed that when the true data generating network structures were the same, i.e., the sample is homogeneous, in most conditions, the presently available descriptive tools would indicate heterogeneity given standard practices. Accordingly, the risk of over-interpreting the extent of individual differences as well as interpreting them as meaningful, is considerable. To date, few tools have been developed to test for differences within idiographic network analysis.

In this chapter, we propose a straightforward and available solution to test for individual differences within network structures by using model comparison techniques. Model comparison has long been the gold standard in Structural Equation Modeling [SEM; 171], but within the field of network modeling, an avenue yet to be explored.[1] Here, we apply ideas of

1 Although it is important to mention here that previous attempts have been made to incorporate network modeling as part of Structural Equation Modeling by modeling the latent structure as a network model and to compare the latent and network model using model comparison [88]

invariance testing from SEM to propose a systematic model comparison approach, where a model in which equality between individual network structures is assumed is compared to a model in which heterogeneity between individual network structures is assumed. This comprises a rigorous way of testing if individual differences in network structures are plausible given the data at hand. We refer to this practice as the *Individual Network Invariance Test (INIT)*.

In the following sections, we first elaborate on estimation procedures for individual network structures and proceed by explaining how these estimation techniques have enabled comparisons among competing models within the field of network analysis. Second, we introduce INIT as a methodology and its implementation in the R package *INIT*. Next, we evaluate the performance of INIT through simulation studies. Using an illustrative data example, we show how researchers can use INIT to test for (in)equality between individual network structures and to test for (in)equality in the network structure of one individual over time. We end by providing future directions for the use and implementation of INIT.

3.1.1 *Estimating individual networks*

Currently, the most popular way of assessing idiographic network structures is by estimating a graphical Vector Auto-Regressive (GVAR) model [37, 89, 134]. The estimation of a GVAR model results in two estimated network structures that characterize the multivariate structure of time series data for an individual: a *temporal* and a *contemporaneous* network [81, 199]. In the temporal network, the relation from one variable at time point $t-1$ to another variable at t is represented. The resulting network is a directed graph, which represents variables as nodes, and directed edges represent the predictive effects of one variable on another or on itself. In the contemporaneous network, relations among variables within the same measurement occasion that remain after temporal effects have been taken into account are represented as an undirected network. In this network, edges represent partial correlations between residuals after the temporal effects have been regressed out. Typically, the contemporaneous network will include relations between variables that arise from processes that result from a faster time scale than the one embodied in the temporal network.

When considering first-order lagged effects,[2] the lag-1 GVAR can be written as:

$$\mathbf{y}_t = \mathbf{B}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{1}$$

$$\boldsymbol{\varepsilon}_t \sim N(0, \boldsymbol{\Sigma}), \tag{2}$$

in which $\mathbf{y}_t$ represents a vector of centered continuous responses by a given person on p variables at time-point t. The $\mathbf{B}$ matrix encodes temporal effects and is a $p \times p$ matrix containing all lag-1 regression coefficients β_{ij} of variable j on variable i. Its transpose can be used to encode a (personalized) directed weighted network. Moreover, $\boldsymbol{\varepsilon}_t$ represents a vector of normally distributed innovations with variance–covariance structure $\boldsymbol{\Sigma}$. This innovation variance-covariance matrix can be modeled as a Gaussian Graphical Model [GGM; a network of partial correlation coefficients; 89], by taking the inverse of the residual covariance matrix to obtain $\mathbf{K}$, also termed the precision matrix:

$$\boldsymbol{\Sigma} = \mathbf{K}^{-1}. \tag{3}$$

Elements that are modeled to be zero in the precision matrix $\mathbf{K}$ indicate full conditional independence relationships [183]. Off-diagonal elements of $\mathbf{K}$ can subsequently be standardized to obtain model-implied partial correlation coefficients:

$$\omega_{ij} = -\frac{k_{ij}}{\sqrt{k_{ii}k_{jj}}}, \tag{4}$$

in which ω_{ij} represents the partial correlation between variable i and variable j after conditioning on all other variables in $\mathbf{y}$. Standardizing precision matrices to partial correlation matrices is the default practice in network psychometrics. Epskamp, Rhemtulla, and Borsboom [88] proposed combining Equation (3) and Equation (4) into a single psychometric model:

$$\boldsymbol{\Sigma} = \mathbf{K}^{-1} = \boldsymbol{\Delta}(\mathbf{I} - \boldsymbol{\Omega})^{-1}\boldsymbol{\Delta}, \tag{5}$$

in which $\boldsymbol{\Omega}$ is a symmetric matrix that contains zeroes on the diagonal and partial correlation coefficients on the off-diagonal and $\boldsymbol{\Delta}$ is a diagonal scaling matrix that controls the variance of the items.

The benefit of using Equation (5) as a psychometric model rather than Equation (3), is that the partial correlations between two variables can

2 It is worth noting that the lag-1 VAR model does presuppose that auto-regressive and cross-lagged relations extending over a longer time interval (e.g., lag-2) exist, however, that the lag-1 relationships can explain these relationships.

be modeled directly and independently of the variance of the variables. For the purpose of this chapter, this means that we can equate $\boldsymbol{\Omega}$ across individuals (thus testing if two individuals can be modeled appropriately by using the same partial correlation structure) while not requiring the innovation variance (and as a result the variance of $\mathbf{y}$) to be equal across individuals [73].

Estimating the parameters in $\mathbf{B}$ and $\boldsymbol{\Omega}$ per individual—either *saturated* indicating that all parameters are estimated, or *sparse* indicating that only a set of non-zero parameters are estimated—can be done in several ways [41]. Recently, Epskamp [73] proposed to estimate parameters of the GVAR model through a Maximum Likelihood (ML) approach commonly used in SEM estimation of other auto-regressive models: by modeling the Toeplitz variance–covariance matrix of responses at time-point $t-1$ and time-point t [125]. An additional challenge when estimating sparse networks is to choose which parameters should be set to zero and which parameters should be estimated freely [14]. In the ML framework, one straightforward manner in which this can be done is by first estimating a saturated model and subsequently re-moving all parameters that are not significant at a given α level when estimating the sparse model parameters. This process is termed *pruning* [14].

Estimating the GVAR model through ML estimation as implemented in the *psychonetrics* package [74], allows for the application of powerful techniques in network modeling that are already commonly used in SEM, such as handling of missing data through full-information ML estimation [195], the inclusion of latent variables [73], and, most relevant to the current study, invariance and homogeneity testing through multi-group analysis [72].

3.1.2 *The individual network invariance test (INIT)*

3.1.2.1 *General idea*

INIT is a method to test individual network structure similarity. Similar to testing for Measurement Invariance (MI) using multi-group SEM methods to determine if the same measurement model holds across individuals [2], INIT can test if the same network structure applies across different individuals. In this approach, individuals are analogous to groups in the standard MI paradigm, whereas time points are analogous to individuals - the primary difference is that, in contrast to measures of different individuals in the multi-group approach, measures taken at different time points are not independent. Model comparison is subsequently used to compare a model in which all parameters of the GVAR network are estimated freely to a model with equality constraints placed on the parame-

ters of the GVAR network. Therewith, INIT can be interpreted as a null-hypothesis test to compare (strict) homogeneity (i.e., edges between nodes and their parameter values are constant across individuals) to heterogeneity (i.e., edges between nodes and their parameter values are allowed to vary across individuals)[3].

For example, to assess whether both the temporal and contemporaneous network structures are equal for two individuals, we can add subject subscripts to the matrices in Equation (5) and subsequently compare the following models:

$$\begin{aligned} \mathcal{M}_{\text{homogeneous}}: &\quad \boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2, \quad \mathbf{B}_1 = \mathbf{B}_2, \\ \mathcal{M}_{\text{heterogeneous}}: &\quad \boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2, \quad \mathbf{B}_1 \neq \mathbf{B}_2. \end{aligned}$$

These models can be compared using a χ^2 difference test or through information criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The model matrices $\mathbf{B}$ and $\boldsymbol{\Omega}$ can either be saturated (all parameters are estimated to be non-zero) or sparse (some parameters are set to zero). Of note, equality constraints can be placed on both the contemporaneous and temporal network structures of two individual network structures simultaneously ($\mathbf{B}_1 = \mathbf{B}_2, \boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$) or on the different type of network structures separately. For example, equality constraints can be placed on only the contemporaneous network ($\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2; \mathbf{B}_1 \neq \mathbf{B}_2$), or only on the temporal network structure ($\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2; \mathbf{B}_1 = \mathbf{B}_2$).[4] Note this means in all cases $\mathcal{M}_{\text{homogeneous}}$ denotes the model with equality constraints placed on both network structures, while $\mathcal{M}_{\text{heterogeneous}}$ always refers to the less contained model; i.e., this model can entail equality constraints have been placed on $\mathbf{B}$, $\boldsymbol{\Omega}$, or both $\mathbf{B}$ and $\boldsymbol{\Omega}$.

Thus far, we have considered heterogeneity as (in)equality between individuals, but (in)equality may also arise *within* an individual. That is, the network structure of an individual may change over time. Similar to assessing if the same network structure holds for two individuals, INIT can assess whether the same network structure holds for one individual over time, for example, when time series data are collected for one individual over two separate measurement periods.

3.1.2.2 *Methodology*

When estimating saturated networks, INIT applies the following procedure. First, a saturated GVAR model is fit to the time series data of two

3 Definitions for homogeneity and heterogeneity are based on the homogeneity assumptions as defined by Liu, Gates, and Ferrer [188].

4 Likewise, homogeneity in means and variances could be tested by equating $\boldsymbol{\mu}$ and $\boldsymbol{\Delta}$ across groups respectively.

individuals, in which the edges of both networks are estimated freely: each edge weight for each individual network may take on a unique value. In the case where time series data are collected for one individual for two separate periods, a freely estimated saturated GVAR model is fit separately to the data for each measurement period. For the remainder of this chapter, we refer to this model as the *unconstrained model* ($\mathcal{M}_{\text{heterogeneous}}$). Next, a second model is fit to the time series data in which all parameters of the network structure(s) of interest are constrained to be equal across individuals (or, in the case of testing equality over time, across measurement periods). That is, all corresponding edge weights in both individual network structures (or both measurement period networks) are required to take on the same value. We shall refer to this latter model as the *constrained model* ($\mathcal{M}_{\text{homogeneous}}$). Finally, model comparison is used to test which of the two models is preferred.

The procedure above describes a test for equating the network parameters of two individuals while all edges are included. However, often in idiographic research, researchers use model selection procedures to estimate a *sparse* network structure comprising only some edges [81, 14], thereby removing likely spurious edges. To this end, researchers may wish to apply INIT to sparse network structures. The application of INIT to sparse network structures, works for any model selection algorithm—such as regularization [81, 1] or stepwise model search—that returns a sparse idiographic network. In this chapter, we use significance pruning at $\alpha = 0.05$ to obtain estimates of sparse individual network structures [14]. Pruning is a simple form of model selection in which edges are included based on some criterion as a p-value being below a chosen level of statistical significance. By pruning the network structure, we remove edges that do not satisfy the criterion. Next, the network is re-estimated with those edges fixed to zero. This re-estimation ensures that the parameter estimates of the selected edges are based on the final network model. This procedure enhances the accuracy of edge weight estimations relative to, for example, applying a thresholding procedure to remove edges from the saturated network structure [14].

After estimating an individual sparse network per person (e.g., through pruning), we cannot readily test if the non-zero parameters are equal across individuals, because the estimation of sparse network structures may lead to the estimation of a particular edge in one network that might be excluded in another. As it is impossible to place equality constraints on an edge that is estimated in one network but absent in another, applying INIT to sparse networks must occur under the additional constraint that the networks of the individuals (or for the different periods) are similar in structure. To achieve this, we can form a *union network*, which represents all edges that are estimated to be non-zero in at least one of the pruned

network structures. This union network can then be used to perform INIT as described above, such that $\mathcal{M}_{\text{heterogeneous}}$ estimates different parameter values for the non-zero edge weights and $\mathcal{M}_{\text{homogeneous}}$ estimates equal parameter values for the non-zero edge weights for each individual, but for both individuals the same edges that were estimated to be zero and non-zero.[5]

3.1.2.3 *R-package*

INIT has been implemented in the R-package *INIT*.[6] The INIT package functions as a wrapper around the psychonetrics package to estimate idiographic network models using FIML and allows for easy network comparison.

3.2 SIMULATION STUDY

3.2.1 *Simulation design*

To assess the performance of INIT (i.e., the proportion of false positives and its power to detect true differences), a simulation study was conducted.[7] We generated time series data for two individuals from a graphical VAR model under $2 \times 5 \times 11$ conditions. We varied the true underlying data (a) generating network structure: $s \in \{\text{synthetic}, \text{empirical}\}$, (b) the number of time points per individual: $t \in \{50, 100, 250, 500, 1000\}$, and (c) the number of (in)equalities between the two network structures: $r \in \{0\%, 10\%, 20\%, 30\%, \ldots 100\%\}$. For each of the $2 \times 5 \times 11$ data generating conditions, we conducted 100 replications, which led to the evaluation of 11,000 simulated datasets.

The true underlying data-generating network structures consisted of a simulated sparse network and empirical network structures estimated from data collected by McNally and Nierenberg from two outpatients diagnosed with bipolar disorder who were undergoing maintenance treatment at the Dautan Family Center for Bipolar Treatment Innovation at Massachusetts General Hospital (MGH). For the remainder of the chap-

5 Of note, in the sparse setting we compare the $\mathcal{M}_{\text{homogeneous}}$ union model to the $\mathcal{M}_{\text{heterogeneous}}$ union model, as these include the same edges. However, if the $\mathcal{M}_{\text{homogeneous}}$ union model is rejected, we prefer the sparse idiographic network model obtained through pruning over the $\mathcal{M}_{\text{heterogeneous}}$ union model. This is because the $\mathcal{M}_{\text{heterogeneous}}$ union model includes edges for individuals that were not selected in the individual model estimation routines.

6 The *INIT* package can be obtained through the following link: `https://github.com/RiaHoekstra/INIT`.

7 The full simulation code is available on the Open Science Framework (OSF) repository at: `https://osf.io/bqhgc/` [142]).

ter, we shall refer to the first network structure as the *synthetic network* and the second network structure as the *empirical network*.

Synthetic network

In the synthetic network condition, both the temporal and contemporaneous network was characterized as a chain graph containing 6 nodes in which each node is connected to the subsequent node (i.e., 1 – 2, 2 – 3, 3 – 4, etc.). In addition to the chain structure, auto-regressive effects were introduced in the temporal network. Parameter values for the contemporaneous network were drawn from a normal distribution with $m = 0.20$ and $sd = 0.26$. Parameter values for the temporal networks were drawn from a normal distribution with $m = 0.30$ and $sd = 0.10$. As most estimated network structures based on empirical data show a high proportion of positive edges, the proportion of positive edges was set to 80% for both the temporal and contemporaneous network. Previous simulation studies have shown network estimation works well under this structure [e.g., 89]. See Appendix B for an exemplar data generating $\mathbf{B}$ and $\mathbf{\Omega}$ matrices. One of the benefits of this network structure is that it can easily be changed. Therefore, in addition to a 6 node network, we simulated data under a network structure containing 12 nodes.

3.2.1.1 *Empirical network*

In the empirical network condition, the network structures from two outpatients diagnosed with bipolar disorder were used as the data-generating mechanism; see Figure 10 for their idiographic network structures and see Appendix B for the data generating $\mathbf{B}$ and $\mathbf{\Omega}$ matrices. Patients were selected based on the properties of their estimated network structures and the length of their original time series data. Estimated graphical VAR networks based on empirical data often show relatively dense connected contemporaneous and spare temporal networks [e.g., 253, 179, 95, 81, 185]. For both patients, their estimated network structure adheres to these properties. While specificity (i.e., the ability to remove spurious edges) is high, sensitivity (i.e., the ability to estimate a genuine edge) is low when estimating a graphical VAR network with time series ranging from $t = 75 - 100$ [195]. Therefore, we selected patients whose time series data far surpass this number of data points with patient x $t = 282$ and patient y $t = 210$, to ensure both high specificity and sensitivity.

3.2.1.2 *Introducing individual differences*

To introduce individual differences (i.e., heterogeneity) between two network structures, based on the true-underlying data-generating network

structure, a second network was generated by randomly rewiring some of the edges while retaining the same parameter values. Rewiring refers to the process of changing the network structure by permuting edges (e.g., the edge between node 1-2 can be removed and instead assigned to a different node pair such as between node 1-5; see [260]). These (in)equalities were introduced with a probability ranging from 0 to 1 with steps of 0.1 (i.e., $r \in \{0, 0.1, 0.2, \ldots, 1\}$), corresponding to the probability of an edge being rewired. That is, rewiring with a zero probability renders two network structures identical, indicating homogeneity, while rewiring with higher probability generates more different networks The rewiring range corresponds to the extent of heterogeneity: from strict homogeneity ($r = 0$) to heterogeneity ($r = 1$). The degree of rewiring was kept constant for both the temporal and contemporaneous networks within each condition. Importantly, for the temporal network, only cross-lagged edges were rewired. Which edge(s) were rewired was selected at random and thus could vary for each repetition within the simulation.

On the simulated time series data, we estimated individual GVAR network structures by using the *psychonetrics* package (version 0.10; [74]) in R, which uses full information maximum likelihood (FIML) estimation to deal with missing data. Of note, this can be seen as "pseudo" FIML as the Toeplitz covariance structure is modeled rather than the entire time series [203, 125, 96]. This leads to the estimation of a just-identified model. For a detailed description of the model estimation technique used in this chapter, we refer the interested reader to Epskamp [73] for the modeling framework and Epskamp, Isvoranu, and Cheung [72] for a description on the multi-group estimator and imposing equality constraints. We estimated both saturated and sparse network structures (using significance pruning). Equality constraints were either placed on the estimated contemporaneous network structure ($\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$), the temporal network structure ($\mathbf{B}_1 = \mathbf{B}_2$), or on both ($\mathbf{B}_1 = \mathbf{B}_2, \boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$). In all cases, the constrained model ($\mathcal{M}_{\text{homogeneous}}$) was compared to the unconstrained model ($\mathcal{M}_{\text{heterogeneous}}$).

3.2.1.3 *Performance metrics*

To evaluate the performance of INIT, we explored the following outcome variables: χ^2, ΔAIC, and ΔBIC rejection rates (i.e., the rate at which the constrained model [$\mathcal{M}_{\text{homogeneous}}$] is rejected). When heterogeneity is zero, the rejection rate for all three outcome measures should ideally be at the $\alpha = 0.05$ level. With growing heterogeneity, INIT should reject the constrained model based on the χ^2 or fit the data less optimally compared to the unconstrained model as indicated by ΔAIC and ΔBIC.

3.3 RESULTS

We discuss the performance of INIT, using the rejection rates for the χ^2, ΔAIC, and ΔBIC. We will first discuss the rejection rates when estimating saturated networks and second when estimating pruned network structures. For an easier overview of the results, we discuss the results for 0%, 50%, and 100% (in)equalities here, referred to as homogeneity, medium heterogeneity, and high heterogeneity, respectively. Furthermore, as results for the two participants in the empirical condition showed similar patterns, the results for participant x are shown and discussed here. Extensive results for the whole range of (in)equalities and all network conditions appear in Appendix B.

3.3.1 *Performance INIT on saturated networks*

Figure 8 shows the rejection rate for the χ^2, ΔAIC, and ΔBIC when applying INIT to compare saturated networks. Notably, in the case of homogeneity, the rejection rates are Type I error rates. As no noticeable differences were detected between the synthetic network condition and the empirical network condition when comparing saturated networks using INIT, we discuss pattern results for both conditions simultaneously.

The upper panel in Figure 8 shows the results for the χ^2 test. We find the χ^2 test performs unsatisfactorily in both the synthetic network condition and in the empirical network condition: in the case of homogeneity, the χ^2 test has a rejection rate above the preferred $\alpha = 0.05$ level when equality constraints are placed on both the contemporaneous and temporal network structure. With growing heterogeneity and number of time points per individual, the χ^2 test indicates a better model fit for the unconstrained model ($\mathcal{M}_{\text{heterogeneous}}$) in $> 80\%$ of the cases. Compared to the goodness of fit measures discussed below (AIC and BIC), it was more challenging for the χ^2 difference test to maintain Type I error accuracy.

In contrast, the AIC and BIC both correctly favor the constrained model as indicated by a rejection rate of less than 5% shown in the middle and lower panel of Figure 8 when there is homogeneity. In addition, both the AIC and BIC prefer the unconstrained model as a function of increasing heterogeneity and the number of time series data per individual. The BIC, however, has lower power for a given effect size (i.e., degree of heterogeneity) and amount of data (i.e., time series data per individual) in situations where heterogeneity is medium or high. When $t = 500$, the BIC rejects the constrained model in $> 75\%$ of the cases when heterogeneity is medium and 100% when heterogeneity is high. The AIC, on the other hand, reaches an acceptable power when $t = 100$ and heterogeneity is medium, leading INIT to correctly reject the constrained model in 80% of

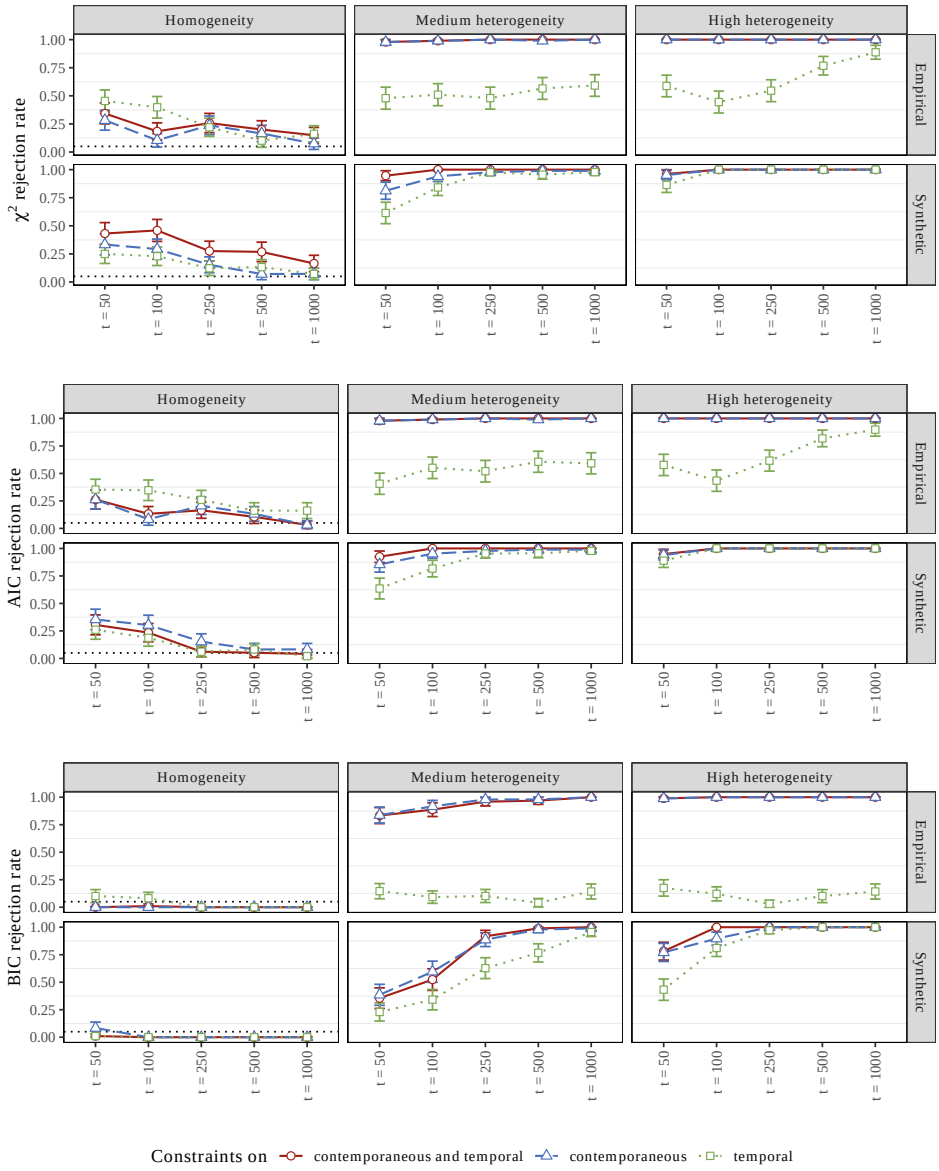


Figure 8: The rejection rate of the χ^2 test (top panel), ΔAIC (middle panel), and ΔBIC (bottom panel) when applying INIT to compare saturated networks. Results are shown for the synthetic and an empirical network condition for 0%, 50%, and 100% (in)equalities between two network structures, referred to as *homogeneity*, *medium heterogeneity*, and *high heterogeneity* respectively. Line color corresponds to the placement of equality constraints. Red indicates equality constraints have been placed on both the contemporaneous and temporal network structure. Blue indicates equality constraints have been placed on the contemporaneous network structure, and green indicates equality constraints have been placed on the temporal network structure. Note, in the case of homogeneity, the empirical rejection rates are Type I error rates; a black dotted line indicates the conventional $\texttt{alpha} = 0.05$ cut-off.

the cases regardless of the placement of equality constraints. When heterogeneity is high, the AIC favors the unconstrained model when $t = 50$ if equality constraints have been placed on both the contemporaneous and temporal network structures. AIC performs best when using INIT to compare (in)equalities in saturated individual network structures.

3.3.1.1 *Performance INIT on pruned networks*

Figure 9 shows the rejection rate for the χ^2, ΔAIC, and ΔBIC when applying INIT to compare pruned networks by utilizing a union network.

When comparing pruned network structures using INIT, both the χ^2 test and the AIC perform unsatisfactorily. When heterogeneity is zero, the χ^2 test, as well as the AIC, has a rejection rate above the $\alpha = 0.05$ level even with larger numbers of time series data ($t = 250$), as is shown in the upper and middle panels of Figure 9 and Figure 9. With growing heterogeneity and number of time-points ($t > 100$), the χ^2 test and the AIC correctly prefer the unconstrained model in most cases ($> 75\%$).

The BIC prefers the constrained model in the homogeneity condition below the desired $\alpha = 0.05$ rate when $t > 50$, shown in the lower panel of Figure 9. With increasing heterogeneity and the number of time series data, the BIC favors the unconstrained model. Interestingly, INIT performs better in the empirical network condition compared to the synthetic network condition when equality constraints are placed on both the contemporaneous and the temporal network or on the contemporaneous network. In the empirical network condition, when heterogeneity is medium and $t = 100$, the BIC favors the constrained model in 88% of the cases when equality constraints have been placed on both the contemporaneous and the temporal network, and 91% when placed on the contemporaneous network, see Figure 9. While in the synthetic network condition, more data are required to reach similar results. When heterogeneity is medium and $t = 250$, the BIC favors the constrained model in 92% of the cases when equality constraints have been placed on both the contemporaneous and temporal network structures and in 89% of the evaluated cases when equality constraints have been placed on the contemporaneous network structure, see Figure 9.

When equality constraints are only placed on the temporal network structure, the unconstrained model is preferred in 63% of the cases in the synthetic network condition; this increases to 76% and 96% when $t = 500$ and $t = 1000$. In the empirical network condition, the unconstrained model is preferred in only 10% of the cases and increases to 14% when $t = 1000$. This is due to the limited number of edges in the temporal network for the empirical condition (the data-generating network model for the participant contained only 2 out of all possible edges). Indicating placing

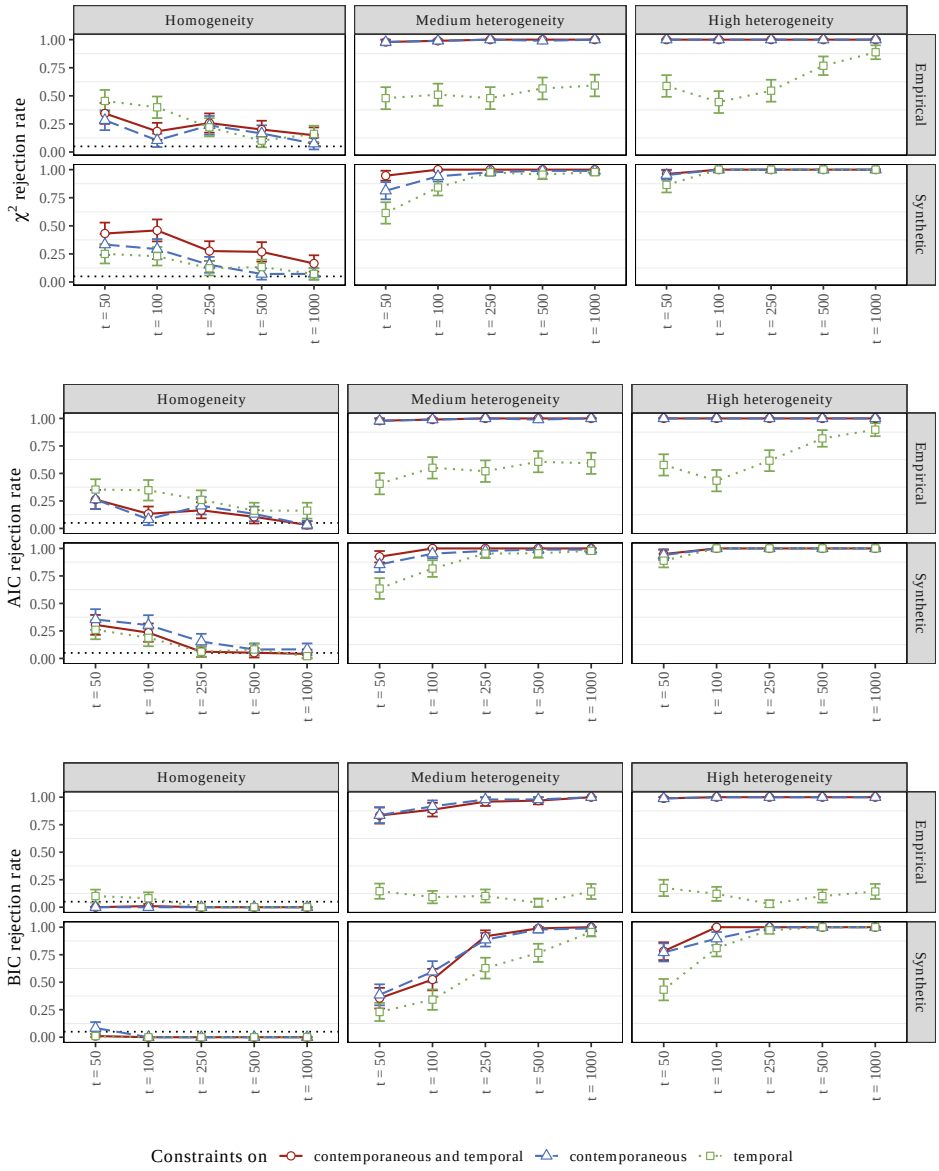


Figure 9: The rejection rate of the χ^2 test (top panel), ΔAIC (middle panel), and ΔBIC (bottom panel) when applying INIT to compare pruned networks. Results are shown for the synthetic and an empirical network condition for 0%, 50%, and 100% (in)equalities between two network structures, referred to as *homogeneity*, *medium heterogeneity*, and *high heterogeneity* respectively. Line color corresponds to the placement of equality constraints. Red indicates equality constraints have been placed on both the contemporaneous and temporal network structure. Blue indicates equality constraints have been placed on the contemporaneous network structure, and green indicates equality constraints have been placed on the temporal network structure. Note, in the case of homogeneity, the empirical rejection rates are Type I error rates, the conventional $\texttt{alpha} = 0.05$ cut-off is indicated by a black dotted line.

equality constraints on a network model with few estimated edges to compare individual differences does not perform well.

It is evident based on Figure 9 that the AIC performs better than BIC in detecting heterogeneity when there is heterogeneity present. However, as our current goal is to introduce a null hypothesis test on an exact difference between idiographic network structures, we especially value the correct identification of a null result. As the BIC outperforms the AIC in case there is no difference between network structures, we conclude the BIC performs best when using INIT to compare (in)equalities in pruned individual network structures.

3.4 ILLUSTRATIVE EMPIRICAL EXAMPLE OF INIT

To illustrate INIT, we analyzed the time series data of two outpatients diagnosed with bipolar disorder who were undergoing maintenance treatment at the Dauten Family Center for Bipolar Treatment Innovation at Massachusetts General Hospital (MGH) and who provided written informed consent to participate in a study on emotion dynamics and symptoms. Participants received 5 pings on their smartphones (09:00 am, 12:00 pm, 3:00 pm, 6:00 pm, and 9:00 pm) daily for six weeks and used the Life data Company's RealLife Exp app to indicate on a zero ("not at all") to 100 ("very much") slider scale how strongly they were currently experiencing six affective states (cheerful, relaxed, anxious, depressed, tired and energetic). For the purpose of illustrating INIT, the two participants with high compliance were chosen as they had the highest number of time series data and a low number of missing data. We will refer to these two participants as participants x and y. This resulted in a sample of $n = 2$ and $t = 492$, where participant x had $t = 282$ with 7% missing data, and participant y $t = 210$ with 1% missing data.

Before assessing individual differences using INIT, the distributional form of the data was inspected and each of the six items was tested for linear trends by regressing the item scores on the time variable. In the case of a significant ($\alpha = 0.05$) regression, the scores were replaced with the residuals. Positive affective states (cheerful, relaxed, and energetic) approximated a normal distribution for participant x, and tended to be positively skewed for participant y, while negative affective states (anxious, depressed, tired) showed a negative skewness for both participants. As the GVAR model assumes multivariate normality, consequently the interpretation of the GVAR parameters is not straightforward [136].

Next, INIT was used to compare the pruned network structures of the two participants. Pruned network structures are obtained by first estimating a saturated network structure and applying a pruning technique in the following step. Saturated networks were pruned at $\alpha = 0.05$, by

removing non-significant edges, and then re-estimated the network, to ensure edge-weight estimates of the selected edges were based on the pruned network structure. α was set to .05 as this level of pruning strikes a balance between discovery (i.e., the sensitivity to detect an effect) and caution (i.e., the specificity to remove truly spurious edges) [14, 149]. Adjusting the alpha level of the pruning, or adding a Bonferroni correction to estimate network structures leads to more conservative networks.

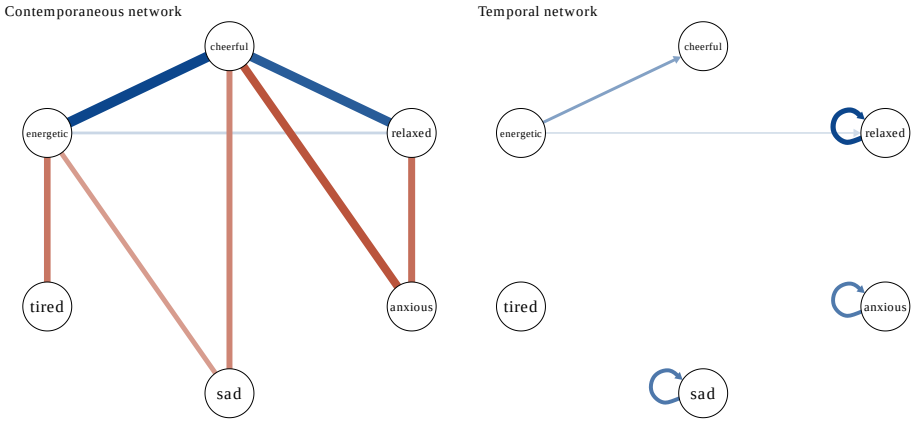
On these pruned network structures, using INIT, we placed equality constraints on both the contemporaneous and temporal networks to compare the (in)equality of the network structures. INIT indicated that the model without equality constraints fit the data best, as indicated by a lower BIC value for the model in which all network parameters were freely estimated. Figure 10 shows the estimated pruned network model for participant x in panel (a), and the estimated pruned network model for participant y in panel (b).

Our analyses indicate that both the contemporaneous and temporal mood networks for participant x and participant y are reliably different from one another (Figure 10, panel (a) and (b), respectively). Despite some structural similarities between the two patients, notable differences are apparent. Consider their contemporaneous networks. For participant x, ratings for cheerful, energetic, and relaxed repeatedly co-occurred in the same measurement period, whereas cheerful was inversely related to anxious and somewhat less so for sad. Relaxed and anxious were inversely associated as were energetic and tired and somewhat less so for energetic and sad.

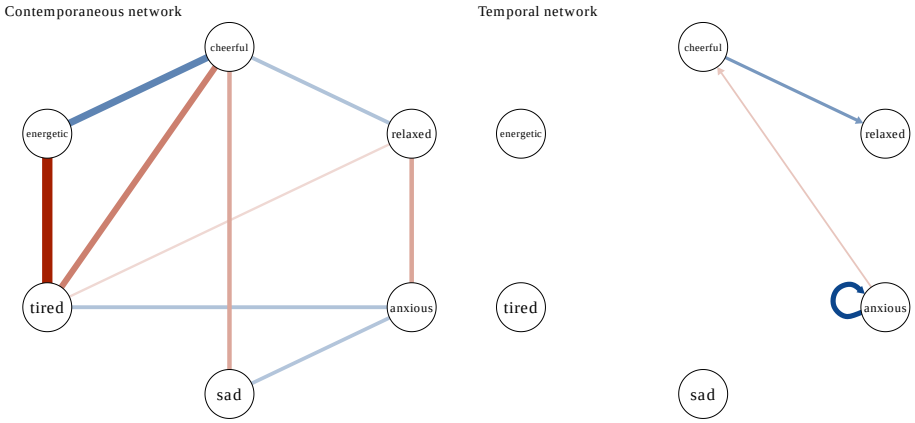
For participant y, tired had very strong inverse associations with energetic and cheerful, and a weak inverse association with relaxed. Energetic was weakly but positively associated with tired. In contrast to participant x, anxious was weak positively associated with tired and sad.

Edges were scarce in the temporal networks for both patients. Underscoring heterogeneity, they signified different predictive relations. For participant x, energetic at one time point predicted cheerfulness and, to a lesser extent, relaxed at the next time point three hours later, whereas for participant y, cheerful at one time point predicted relaxed in the next, and anxious weakly predicted cheerful in the next time point.

To illustrate how individuals can be compared to themselves over time, we have taken the data of the participant with the largest number of time series data, i.e., participant x, and split their time series data into two equal sets of data containing $t = 141$ each. As this split means the last observation of the first section of time series data is dependent on the first observation of the second section of time series data, we have removed the data of the two measurement points in between the sections, creating two equal sets of data containing $t = 140$ observations each. Bear in mind



(a) network participant x



(b) network participant y

Figure 10: Idiographic network structures for participant x (in the upper panel) and participant y (in the lower panel) as estimated with psychonetrics using FIML estimation and model pruning at an $\alpha = 0.05$ level. Using INIT to test for equality constraints on the prunend network structures, results indicated different network structures fit the data best for these two participants by a lower BIC value for the model without equality constraints.

this example is merely illustrative. Ideally, a larger time span would occur between the two measurement periods that figure in the comparison.

Applying INIT to the data, we compared the (in)equality of the pruned contemporaneous and temporal network structure over time. As indicated with a lower BIC value for the constrained model, the model with equality constraints fit the data best, indicating there was no change in network structure over time for participant x. This means the upper panel in Figure 10 shows the estimated saturated network model for participant x for the first part of time series, and the second part of time series. This result is not surprising as measurement took place over the course of $\sim$ 35 days. Ideally, a long(er) break in time series data would surpass between the two time series one wants to compare, and this example should merely be taken as exemplary of the potential use of INIT.

3.5 DISCUSSION

Applied researchers are often interested in the comparison of individual network structures. In this chapter we presented INIT, a novel approach to test differences in idiographic networks. We investigated the performance of INIT by means of a simulation study in which we inspected the rejection rate of the χ^2 difference test, and goodness-of-fit measures such as the AIC, and BIC. Simulation study results indicated INIT performs well when inspecting descriptive goodness-of-fit measures when $t > 100$ per individual. The AIC performed best when placing equality constraints directly on saturated network structures, whereas the BIC performed best when placing equality constraints on pruned network structures. The χ^2 difference test performed unsatisfactorily as the test statistic tended to underreject.

Simulations further elucidate the conditions under which the comparison procedures show adequate power to be used in typical empirical studies. Both goodness-of-fit measures behaved as expected, with the AIC slightly tendency to overfit and the BIC tending to underfit [44]. As the BIC places a stronger penalty on model complexity than the AIC, the BIC has a stronger preference for the simpler model, i.e., the constrained model. Accordingly, it requires stronger evidence, either in terms of effect size or data points per individual, in order to favor the unconstrained model when heterogeneity is present. When comparing saturated network models, the difference between the constrained and unconstrained models is larger than when comparing pruned network models. That is, more parameters need to be constrained to be equal when comparing saturated networks than when comparing pruned networks. These features result in the AIC to overfit when comparing pruned network structures,

whereas the BIC tends to underfit when comparing saturated network structures.

Several reasons can account for the unsatisfactory performance of the χ^2 difference test. One possibility is that the test statistic does not follow a χ^2 distribution. However, when looking at our original results, we did observe that the Type I error rate diminished as the number of time points per individual increased. Simulation studies with a larger number of time points per individual ($t = 1000$) revealed that the χ^2 difference test converges. Unfortunately, convergence emerges too slowly for sample sizes typical of those in psychology.

INIT complements existing methods that allow for the estimation of heterogeneity between (individual) network structures, such as Group Iterative Multiple Model Estimation [GIMME; 108] and multilevel VAR [mlVAR; 40]. Just as INIT, GIMME and ML-VAR can be used to quantify individual differences in a sample. GIMME quantifies individual differences by estimating both group and individual-level edges in an iterative search procedure, where group-level edges are deemed homogeneous, and individual-level edges are deemed heterogeneous [11]. In this iterative search procedure, group-level edges are added to the model based on criteria specified by the researcher (e.g., if including the edge improves model fit for the "majority" of the sample). After the group-level model is fit for each individual in the sample separately, it is determined whether including additional individual edges would improve model fit for the individual. This search-and-add procedure continues until a model with an adequate fit is obtained for each individual separately.

In ML-VAR estimation, individual differences can be quantified as the degree of standard deviations of the random effects [40]. However, multilevel modeling compromises on idiographic research by borrowing information from other individuals to estimate individual-level effects. Therewith, the standard deviations of the random effects often indicate more homogeneity than when one applies idiographic techniques such as INIT or GIMME.

Contrary to GIMME and ML-VAR, INIT can be used to compare idiographic network structures *directly* to one another by means of equality constraints. Therefore, in addition to its use as a stand alone methodology, INIT can be in tandem with GIMME or ML-VAR, for instance as a follow-up test after GIMME or ML-VAR has indicated the presence of heterogeneity. As GIMME and ML-VAR do not allow for a direct test on the invariance of individual network structures, applying INIT after ML-VAR estimation indicated the presence of heterogeneity, for example, would enable investigators to identify which pairs of network structures differ.

As was mentioned previously, heterogeneity is not limited to the comparison *between* individuals, but can also arise *within* an individual across different time windows. To this end INIT, can test for structural differences in the network model of one individual who is measured during two nonconsecutive periods of time, and thus another method for evaluating model change in time series data [e.g., 134, 56, 5], such as the time-varying VAR model [134], for detecting gradual changes in time series data, or the time-varying change point auto-regressive (TVCP-AR) model [5], enabling detection of both abrupt and gradual changes in univariate time series data. INIT can ascertain if a change in network structure have occurred in multivariate data.

3.5.1 *Limitations*

Several limitations should be kept in mind when interpreting our results. If INIT detects a difference between network structures based on the AIC or BIC, this difference does not need to be statistically significant. Unfortunately, the chi-square test for invariance, which could test for significant differences, performed poorly with realistic sample sizes; therefore, we currently advise against its routine use. Accordingly, the question of how to formally test for differences in GVAR networks is still open and provides an important avenue for further research. Relatedly, a difference in goodness of fit measures is not necessarily a *meaningful* difference. That is, although INIT might indicate that the unconstrained model best fits the data, this does not imply that differences in the network structure between two persons are meaningful differences in applied settings, such as psychotherapy. Accordingly, INIT applications should be performed with care.

3.5.2 *Suggestions for applied researchers*

As the χ^2 difference test performed unsatisfactorily, we currently advise against its use when testing (in)equalities between idiographic networks. Additionally, when comparing saturated networks, we recommend using the AIC as a means of model comparison, while recommending inspecting the BIC when estimating pruned network structures. These suggestions are implemented in the R-package *INIT*.[8]

When comparing idiographic network structures on time series data below $t < 100$, we advise applying INIT on saturated networks when pruned network structures show a minimal amount of edges (i.e., only one or two edges were detected). In case where one of the pruned net-

8 The *INIT* package can be obtained through the following link: `https://github.com/RiaHoekstra/INIT`.

work structures (i.e., temporal or contemporaneous) is relatively dense, INIT performs adequately on pruned network structures as well, as the power to detect true differences is higher. In all cases, INIT performs best when applying equality constraints on both contemporaneous and temporal network structures, especially for time series data below $t < 100$.[9]

Finally, when interpreting the results, it is important to bear in mind INIT is a conservative test that errs on the side of caution. This means that INIT tends to *underestimate* the amount of differences present. In contrast to other techniques indicating individual differences between idiographic network structures, INIT takes homogeneity rather than heterogeneity as the null model. The downside of this is that INIT might overlook (small) differences between idiographic network structures.

3.5.3 *Future applications*

Opportunities for future applications of INIT are worth considering. First, the use of equality constraints does not limit itself to testing differences between idiographic network models. As shown in the illustrative empirical example, INIT can be used to test the (in)equality of individual network structures and of one individual over time. For an empirical example, see Chapter 5. In addition, INIT can be used as an omnibus test to test for (in)equality between groups of individuals, for example, to compare the network structure of groups of individuals with the same underlying severity score [70]. In Chapter 8, we discuss this application of INIT in more detail. When used as an omnibus test, a straightforward extension of INIT lies in its use as a post-hoc test to investigate inequalities further.

In addition, INIT could be used to place equality constraints on specific edges to test the similarity of specific edges between individual networks. This highlights a possible expansion of INIT. In its current form, INIT compares an unconstrained model to a constrained model. It would be worthwhile to test (in)equality between individuals based on the whole model space in order to gain more detailed information on the heterogeneity between individuals. This would allow for exploring all possible equality and inequality relations among individuals, ideally while penalizing for multiple comparisons.

9 Of note, pruning is just one of the methods we can use to obtain and compare sparse network structures using INIT. In principle, other methods to obtain sparse network structures could be used, such as LASSO regularization, which could potentially work better with lower sample sizes.

3.5.4 *Conclusion*

Comparing individual network structures to quantify heterogeneity is challenging in many applied settings. In this chapter, we have proposed a possible solution to this problem: INIT. INIT tests the (in)equality between two network structures, allowing researchers to test for (in)equality between and within idiographic network structures. As such, INIT provides a powerful tool for researchers to test heterogeneity between psychological networks.

4

SIMULATIONS: A POWERFUL TOOL

ABSTRACT

Simulation studies are a powerful tool within the field of network psychometrics providing insight into sample size determination and best practices for achieving accurate results. However, the advantages simulation studies provide have thus far been predominately harnessed by psychological researchers with a statistical or methodological background. While user-friendly methods for simulating psychological networks have been developed, there remains a lack of comprehensive guidelines for their effective utilization. In response, this chapter serves as a guide on how to set up simulation studies for the three most widely used network modeling techniques: the Ising model, the Gaussian Graphical Model (GGM), and the Graphical Vector Autoregressive (GVAR) model. We explain the rationale behind simulation studies and present a step-by-step roadmap to initiate one's own simulation studies. This roadmap covers all essential aspects using a practical example to guide the reader. In addition, we provide an overview of open-source software packages available for simulating network models in the statistical programming environment R, and demonstrate their practical application. In doing so, we aim to enable more researchers to explore and resolve statistical questions within network psychometrics.

4.1 INTRODUCTION

In the last decade, network psychometrics has gained substantial attention in psychological sciences [29, 227, 153]. In this field of research, psychological constructs are conceptualized as a complex interplay of components. Psychological networks consist of nodes representing observed variables, connected by edges representing the statistical relationships between these variables. The type of statistical relationship depends on the estimation technique used (e.g., partial correlations when estimating a

Under review as: **Hoekstra, R. H. A.**, Epskamp., S., Finnemann, A. T. K., & Borsboom, D. (2023). Unlocking the potential of simulation studies in network psychometrics: A tutorial.

GGM, regression coefficients when estimating a temporal network structure using GVAR models).

While *estimation* of network models has been described in numerous tutorial papers [e.g., 65, 82, 52, 147], descriptions of how to *simulate* network models is fragmented and a clear overview is missing [65, 82, 93]. As a result, the advantages that simulation studies provide have thus far been predominately harnessed by psychological researchers with a statistical or methodological background [e.g., 151, 84, 195]. Especially within the field of network analysis, more researchers could benefit from the skills to perform simulation studies as, for example, a way to determine the sample size or time points that are necessary to pick up the desired effects is missing. Simulation tools can help researchers determine how much data is required in order to estimate an accurate network model. In addition, several papers have questioned the stability of estimated network models across repeated samples [97, 99, 100, 104]. Simulation studies can be used to determine which factors contribute to the replication of the network structure and to determine the expected reproducibility of a given network structure [29, 82].

The aim of this tutorial is to broaden the accessibility of simulation tools for researchers. We begin by outlining the rationale behind simulation studies in network psychometrics. Using a practical example, we guide readers through specifying a network model, simulating data, estimating a network, and evaluating the simulation results for the three most often used modeling frameworks: Ising model, GGM, and GVAR model. We include an overview of the open-source software to conduct simulation studies in R, and common used evaluation metrics in network psychometrics. R code is available for the reader to replicate our simulations and can be used to perform their own simulations.

4.2 SETTING UP A SIMULATION STUDY

Simulation studies can be thought of as computer-based experiments. One key strength of simulation studies is the ability to understand the behavior of statistical methods under various conditions because the "true" underlying model is known for the data generating process [204]. Therewith, simulation studies are a powerful tool to answer methodological as well as theoretical questions and provide a framework to answer specific questions relevant to one's research.

Similar to empirical studies, simulation studies center around a research question. The research question in a simulation study typically pertains to statistics and involves questions such as understanding the statistical power of a particular method, how a method performs under ideal and non-ideal conditions, or the comparative strengths and limita-

tions of competing methodologies. While some statistical questions may be resolved through mathematical analysis, simulation studies are often better suited for answering more complex questions [121].

With respect to network psychometrics, simulation studies can help us answer questions about the performance of particular estimation techniques under various conditions, i.e., how well does the technique recover the true underlying network structure, [e.g., 17, 88, 89, 276, 133, 20], which network estimation technique works best under which conditions [151], and whether idiographic network model estimation is feasible in clinical practice [195].

In general, simulation studies we follow a systematic process to provide an answer to the research question. Given the research question, one identifies evaluation metrics (e.g., the sensitivity or specificity of an estimation technique), conditions to be varied (e.g., sample size, network structure, number of time series), and specifies a data generating model (e.g., a GGM or GVAR model) . This data generating model serves as the *true* underlying model. With the true model in hand we simulate data, and perform statistical analyses on this simulated data, such as estimating a network model. Since we know the true model for the simulated data, we can compare the estimated parameters to the true parameters and compute our evaluation metrics. This process is repeated multiple times to produce distributions for the evaluation metrics.

Thus, approximately a simulation study can be broken down into eight consecutive steps:

1. Define the aim or research question
2. Define evaluation metric and conditions to be varied based on the aim or research question
3. Specify the true model (e.g., data generating mechanism) based on the aim or research question
4. Generate data based on the true model
5. Perform statistical analyses on the simulated data and retain evaluation metrics
6. Repeat steps 3-4 numerous of times to generate a distribution for the evaluation metrics
7. Analyze the distribution of the evaluation metrics
8. Evaluate the simulation results in light of the research question

In this tutorial, we will provide the tools for readers to walk through all steps involved in conducting a simulation study. As a running example we set up a small simulation study to answer the question: *"To what*

extent can we accurately estimate a GGM for reasonable sample sizes?" (see Tutorial box 1). In order to answer this research question, we varied the sample size condition (i.e., $n \in \{50, 100, 200\}$), and obtained the sensitivity, specificity, and correlation as our evaluation metrics. See Appendix C for a description of the most commonly used evaluation metrics in network psychometrics, and Table 1 for all evaluation metrics implemented in the `evalNet()` function.

In Tutorial box 1, we use the function `parSim()` from the *parSim* package [78] as a means to provide the skeleton for our simulation study. With the specification of just a few arguments this function circumvents the process of combining multiple for loops in order to loop over all conditions in our simulation study. `parSim()` is a parallel simulator that takes a set of conditions and an R expression as arguments and returns a data frame with the simulation results.

We can specify any number of conditions ourselves. Here we specified our sample size condition (i.e., `sampleSize = c(50, 100, 200)`). Within the R expression we can write whatever code we would like to use within our simulation study using the specified conditions as objects. This way, the R expression is performed on each of the specified conditions. With the `reps` argument we control the number of repetitions for each of the conditions. In Tutorial box 1 we set the `reps` argument to 10 by means of a simplification, but ideally this number is much higher such as 100, or 500 repetitions per condition. Within the R expression we load the required R packages, specify the true data generating network, simulate data based upon this network, estimate a network based on the simulated data to ultimately compare the estimated to the true network using our evaluation metrics. The code in Tutorial box 1 can be used as a starting point and adapted as one likes using the techniques described in the subsequent sections of this tutorial.

The remainder of this chapter is structured in accordance with the specification of the R expression within the `parSim()` function in Tutorial box 1. First, we describe how to specify the true data generating network. Second, how to simulate data based upon this true data generating network. And third, how to estimate a network model based on the simulated data. We do so for each of the three most commonly used network models: the Ising model, the GGM, and the GVAR model. Final, we exemplify how to evaluate the simulation results. Annotated R code to follow along with the tutorial boxes presented in this chapter can be found on OSF: `https://osf.io/ygjn7/`.

Evaluation metric	Description
Network accuracy	
Sensitivity	Proportion of existing edges that are correctly identified
Signed sensitivity	Proportion of true same sign edges
Top 50% sensitivity	Proportion of top 50% strongest existing edges that are correctly identified
Top 25% sensitivity	Proportion of top 25% strongest existing edges that are correctly identifieds
Top 10% sensitivity	Proportion of top 10% strongest existing edges that are correctly identified
Specificity	Proportion of non-existent edges that are correctly identified as such
Precision	Proportion of correctly identified edges among all identified edges
Top 50% precision	Proportion of top 50% strongest correctly identified edges among all edges identified
Top 25% precision	Proportion of top 25% strongest correctly identified edges among all edges identified
Top 10% precision	Proportion of top 10% strongest correctly identified edges among all edges identified
Jaccard index	Proportion of shared edges relative to the total number of edges
Edge accuracy	
Edge weight correlation	Correlation between edge weights
Absolute edge weight correlation	Correlation between absolute edge weights
True edge weight correlation	Correlation between true edge weights
Bias for all edges	Average absolute deviation between all edge weights
Bias for true edges	Average absolute deviation between true edge weights
Centrality accuracy[1]	
Pearson correlation centrality measures	Pearson correlation between centrality measures
Kendall correlation centrality measures	Rank order correlation between centrality measures
Top 5 strongest centrality indices	Comparison between top 5 centrality indices
Top 3 strongest centrality indices	Comparison between top 3 centrality indices
Top 1 strongest centrality indices	Comparison between top 1 centrality indices

Table 1: This list of evaluation metrics is based on metrics suggested by Borsboom et al. [30], Forbes et al. [97], and Isvoranu and Epskamp [151].

```
# Load libraries:
library("parSim")

# Run the simulation:
results <- parSim(
  # Specify any number of conditions:
  sampleSize = c(50, 100, 200),

  # Number of repetitions:
  reps = 100,

  # Number of cores:
  nCores = 6, # increase if possible

  # R expression:
  expression = {
    # Load R packages
    library("bootnet")
    source("https://raw.githubusercontent.com/RiaHoekstra/
           simulation_functions/main/evalNet.R")
    source("https://raw.githubusercontent.com/RiaHoekstra/
           simulation_functions/main/GGMsim.R")

    # Specify the true data generating network
    trueGGM <- genGGM(Nvar = 6)

    # Simulate data based on the network structure
    simData <- GGMsim(n = sampleSize, omega = trueGGM)

    # Estimate network using LASSO regularization
    estGGM <- estimateNetwork(simData, default = "ggmModSelect")

    # Obtain evaluation metrics
    metrics <- evalNet(trueGGM, estGGM$graph, metric = c("sensitivity",
                                                         "specificity",
                                                         "correlation"))

    # Return results
    data.frame(metrics)
  }
)

# Inspect results:
View(results)
```

Tutorial box 1: Exemplifying the use of the `parSim()` function from the parSim package. The `parSim()` function is a parallel simulator that takes a set of conditions and an R expression as arguments and returns a data frame with the simulation results. Within the R expression we can write whatever code we would like to use within our simulation study using the specified conditions as objects. This way, the R expression is performed on each of the specified conditions. In the section "Evaluating results", we exemplify how to display the results obtained from this simulation study, see Tutorial box 11.

4.3 SPECIFYING A NETWORK MODEL

A simulation study starts with specifying a research question or aim that leads to the specification of the true, data generating model. However, to specify such a true data generating model within the network modeling framework is not without challenges. In particular, to simulate from a network one needs a) a generating *network structure* and b) a set of *parameters* that specify a probability distribution for the data, based on this structure. However, not all combinations of network structure and network parameters generate data that are useful to evaluate the performance of a model. The choice for a specific generating model thus requires careful consideration.

Broadly speaking, to specify the true data generating network, two routes can be taken: (i) one can use an estimated network structure based on empirical data, or (ii) one can specify a synthetic network structure. We discuss these options in turn.

4.3.1 *Empirical network*

Taking a (previously) estimated network model based on empirical data as the true data generating network has several benefits. One clear benefit is that an estimated network can be chosen that represents the scenario of interest as closely as possible, such that the variables in the empirical network can be a direct representation of the variables of interest and the sample is a representative sample of the sample of interest.

However, limitations need to be taken into account when using an empirically estimated network model as the true data generating network. As the network is based on empirical estimation, the resulting network is as reliable as the data sampling procedure. For example, if the empirical network is based on a relatively large sample, the network may be less likely to contain spurious edges, while if the empirical network is based on a relatively small sample, chances increase that the estimated network contains spurious edges and thus forfeits its representative appeal. Therefore, it is recommend when one wishes to use an empirical estimated network as the true underlying data generating model, one chooses a network that is (i) representative of the situation of interest and (ii) a network that is estimated from a sufficiently large sample.

Regarding the latter, an interesting trade-off should be noted. In general, if the true network structure is a dense network that contains many and weak edges, then the higher the sample size, the more edges can be detected when estimating the network. As a consequence, if we would estimate such a densely connected network based on a $n = 500$ sample compared to a $n = 1000$ sample, the latter data set would allow us to pick up

more effects in the estimated network. However, this leads to a network structure that should be recoverable for data sets with a $n > 1000$ sample, whereas simulating under the network structure we obtained from our $n = 500$ sample, leads to a network structure that should be recoverable for data sets with a $n > 500$ sample. Thus, the higher the sample size we use to determine the true underlying network structure, the higher the sample size that will be required in simulation studies to recover that structure. While this does not render such simulations uninformative, it should be kept in mind in the evaluation of results.

4.3.2 *Synthetic network*

An alternative to using an empirically estimated network model as the true data generating mechanism is to specify the true data generating network oneself. We shall refer to the later as a *synthetic* network. Specifying a synthetic network as the data generating network has some advantages, as it becomes easier to specify a structure with reasonable edge strengths (e.g., no edges with values ambiguously close to zero), and to secure a network that meets desirable conditions (e.g., that results in a positive semi-definite covariance matrix). An additional benefit of specifying a synthetic network structure as the true data generating network, is that it leaves more room to tweak the network. For example, the number of variables in the network can easily be altered[2], as can the number of edges, and the strength of the edge weights.

As is the case for choosing an empirical network as the true data generating network model, specifying a synthetic network model presents us with challenges. We discuss these challenges for the Ising model, GGM, and GVAR model in the following sections. For each of these statistical models, a short description of the model will be given. In addition, we present an overview of open-source software packages and exemplify their accompanying functions to specify a synthetic network for each of these models. For an overview of software packages to specify a synthetic network structure in R and their relevant functions, see Table 2.

2 Note that, when altering the number of nodes in the network, it is important to keep in mind that this has an effect on the edge weights. In particular, changing the number of nodes in the network usually implies that edge weight will need to be smaller if the larger network is to generate association structures that are similar to the smaller network. This means one cannot directly compare a network with 5 nodes to a network with 10 nodes and conclude that the changes found in estimation are a direct result of the change in the number of nodes.

Model	package::function	Description
	set.seed()	random number generator
	ggplot2::ggplot()	visualization
	parSim::parSim()	simulations using multiple cores
	bootnet::replicationSimulator()	simulations to determine reproducibility
	bootnet::netSimulator()	simulations to determine sample size
	evalNet()	computes evaluation metrics for networks
	bootnet::estimateNetwork()	estimates network structure
Ising model		
	genIsing()	generates an Ising model
	IsingSampler::IsingSampler()	generates data from an Ising model
	IsingFit::LinTransform()	transforms between $\{0, 1\}$ and $\{-1, 1\}$ encoding
	IsingFit::IsingFit()	estimates Ising model using eLASSO
GGM		
	bootnet::genGGM()	generates a GGM
	ggmSim()	generates data from a GGM
GVAR		
	genGVAR()	generates a GVAR model
	graphicalVAR::graphicalVARsim()	generates data from a GVAR model
	graphicalVAR::graphicalVAR()	estimates GVAR model using EBICglasso

Table 2: Software packages in R relevant to simulations within network psychometrics. Functions are related to their packages using the '::' notation from R. To access a function, first install the package with `install.package("example")` and then load the package with `library("example")`. The functions `evalNet()`, `genIsing()`, `ggmSim()`, and `genVar()` can be downloaded from github via: `https://github.com/RiaHoekstra/simulation_functions`

4.3.2.1 *Ising model*

The Ising model is a probabilistic model in which variables can have either one of two states. In psychology, the Ising model is frequently used in research on psychopathology to model the absence or presence of symptoms, and in attitude research to study pro versus contra attitudes [17, 66]. The Ising model can be used as a theoretical model to represent and study large systems of interacting elements, and as a statistical model to estimate network structures from binary data. Finnemann et al. [93] discusses both uses of the Ising model in more detail. In this chapter, we only focus on its statistical use.

Formally, the Ising model can be denoted as a probability distribution on a random vector $\mathbf{y}$ in which each element can take one of two states:

$$P(\mathbf{y}) = \frac{1}{Z} \exp\left(\sum_i \tau_i y_i + \sum_{<i,j>} \omega_{ij} y_i y_j\right), \tag{6}$$

in which τ_i indicates a *threshold parameter* (intercept) that is related to the predisposition of a randomly chosen person to respond with one of the two answer categories (either $\{0, 1\}$ or $\{-1, 1\}$ depending on the encoding of the data, see Haslbeck et al. [135]), ω_{ij} is the network edge parameter between node i and node j, and Z is a normalizing constant[3]. Typically, the Ising model also includes a parameter denoting the *inverse temperature* β. However, from a statistical perspective β typically is not identified, and to this end in the expression above we identified the model by specifying $\beta = 1$.

Three aspects are important when specifying a synthetic Ising model: (i) the encoding used (either $\{0, 1\}$ or $\{-1, 1\}$), (ii) the threshold parameters values (τ), and (iii) the network edge parameters (ω). We need to specify these three aspect in such a way that the synthetic Ising model leads to generated data with enough *entropy*. Entropy denotes the average level of information in the data, usually stored in "bits". In essence, entropy answers the question: "how many yes-no questions are on average needed to describe the response pattern of a person in the data set?". Suppose we model a 3-node network, meaning we have 3 dichotomous responses per person. If there is no overlap between the three responses, we would need full information on each of the three responses to describe the outcome. This means that in this case entropy = 3. However, suppose we have strong information on the responses: everyone in the sample always responds with the same answer to each of the questions. For example: the questions are "are you born on Mars?", "are you over 250 years old?", and "can you jump higher than 100 meters?". We would hope that all subjects in our sample would always respond with "no" to each of these questions. In this case, entropy = 0, as we do not need to ask any yes-no question; we already know the answer to each of the responses without information needed from a subject. Ideally, we would like entropy to be somewhere between these extremes, meaning that we want specify the synthetic Ising model in such a way that the data generated from this network is not completely random, while at the same time still having suitable and realistic variation per variable.

To investigate what combination of encoding (either $\{0, 1\}$ or $\{-1, 1\}$), threshold parameters ($\tau \in \{-1, 0, 1\}$), and network edge parameters ($\omega \in$

3 The normalizing constant Z, also termed the *partition function*, is important in parameter *estimation*, but not in data generation as it is not a function of the data and therefore can be ignored when specifying a synthetic Ising model.

$\{0, 0.25, 0.5, 1, 2, 3, 4\}$), would ensure data containing sufficient entropy, we performed a small simulation study. A description of the simulation study and the simulation code is available in Appendix C. Results show that the entropy goes down as a function of the edge weight, meaning that the stronger connected the network, the fewer response patters are generated. Entropy is at a maximum when all threshold parameters (τ) and edge weight parameters (ω) are set to zero, but this is not an interesting network to generate data under as this would just imply completely random (coin-flip) data. Entropy reduces to 0 in most settings, except in symmetric encoded data with all threshold parameters set to zero ($\tau = 0$). In general, when generating symmetric encoded data with all threshold parameters set to 0 edge weight parameters around 0.25 and 0.5 tend to work well. Therefore we recommend when specifying a synthetic Ising model, to use symmetric encoding, set threshold parameters to 0 and sample edge weights around 0.25 and 0.5.

Obliging to the above specified conditions, we can specify a synthetic Ising model in R with the `genIsing()` function. This functions first generates a skeleton for the network structure, i.e., a structure that indicates whether there is an edge present between two nodes. Next, it samples parameter values for the edge weights from a random normal distribution with a specified mean and standard deviation. The `genIsing()` functions has 5 main arguments: `nNode` determines the number of nodes in the network, `rewire` determines the probability of re-wiring random edge weights, `inclusion` determines the probability of edge inclusion, `mean` determines the mean edge weight value, `sd` determines the standard deviation for the edge weight values.

As input the `genIsing()` function simply requires the specification of the number of nodes in `nNode`. In that case a chain graph in which each node is connected to its consecutive node in the form of a chain is generated, i.e., if we consider an undirected network with 3 nodes there is an edge between the nodes 1–2, 2–3, and 3–1, with a mean edge weight value of 0.3 and a standard deviation of 0.1. To move away from this chain-graph structure the `rewire` argument can be specified with a certain probability. This means randomly selected edges in the chain-graph will be rewired. Another way to move away from a chain graph is to specify the `inclusion` argument. The `inclusion` argument can be used to generate a network with a certain probability of edge inclusion. In Tutorial Box 2 we exemplify generating a synthetic Ising model using the `genIsing()` function.

4.3.2.2 *Gaussian graphical model*

The GGM is a model in which the relations between variables are modeled through partial correlation coefficients. In psychology, this model is

To specify an Ising model in R we can use the `genIsing()` function. This function simply requires the specification of the number of nodes:

```
# Load libraries:
source("https://raw.githubusercontent.com/RiaHoekstra/simulation_
       functions/main/genIsing.R")
library("qgraph")

# Generate an Ising model with 6 nodes:
trueIsing <- genIsing(nNode = 6)

# Determine thresholds for each node:
thresholds = 0
thresholds <- rep(thresholds, 6)

# Plot Ising model:
qgraph(trueIsing, layout = "spring", theme = "colorblind")
```

This results in the generating of a chain graph. To move away from this chain structure, a rewiring probability can be specified to randomly rewire a proportion of the edges. Another option is to determine the proportion of edges with a specified probability:

```
# Generate an Ising model with 6 nodes and rewiring probability
# of 50%
trueIsing <- genIsing(6, rewire = 0.5)

# Plot Ising model:
qgraph(trueIsing, layout = "spring", theme = "colorblind")

# Generate an Ising model with 6 nodes and inclusion probability
# of 50%
trueIsing <- genIsing(6, inclusion = 0.5)

# Determine thresholds for each node:
thresholds = 0
thresholds <- rep(thresholds, 6)

# Plot Ising model:
qgraph(trueIsing, layout = "spring", theme = "colorblind")
```

Tutorial box 2: Exemplifying the use of the `genIsing()` function to generate an Ising model. The `genIsing()` function is a developmental function that can be downloaded from github: `https://github.com/RiaHoekstra/simulation_functions/blob/main/genIsing.R`.

often used when we have measured every subject only once and we are interested in the relations between continuous normally distributed variables [e.g., 150, 8, 255]. When we have continuous normally distributed data, we assume that all subjects share the same distribution, that is:

$$\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{7}$$

The covariance matrix $\boldsymbol{\Sigma}$ is our matrix of interest, as the inverse of this matrix, $\mathbf{K}$, the precision matrix, is also known as the GGM [89]:

$$\mathbf{K} = \boldsymbol{\Sigma}^{-1}. \tag{8}$$

The elements in this precision matrix can be standardized to obtain the partial correlation coefficients as follows:

$$\omega_{ij} = -\frac{k_{ij}}{\sqrt{k_{ii}k_{jj}}}, \tag{9}$$

where $i \neq j$. Thus a zero entry in the precision matrix $\mathbf{K}$ implies conditional independence between the connected variables, while a non-zero entry indicates a conditional dependence. This partial correlation is then plotted as a network

The covariance matrix $\boldsymbol{\Sigma}$, precision matrix $\mathbf{K}$, and partial correlation matrix $\boldsymbol{\Omega}$ need to be positive semi-definite. As these matrices have a one-to-one relation, if one matrix is positive semi-definite, all of them are. To determine if the precision matrix $\mathbf{K}$ is positive semi-definite we can check for positive definiteness, which is a stronger condition than positive semi-definiteness. A matrix is positive definite if all of its eigenvalues are positive, and if a matrix is positive definite, it is also positive semi-definite. We can use the `eigen()` function in R to compute the eigenvalues of the precision matrix. If all eigenvalues are strictly positive, then the matrix is positive definite and, by extension, positive semi-definite. Alternatively, we can verify if the inverse of the precision matrix $\mathbf{K}$ leads to a valid covariance matrix $\boldsymbol{\Sigma}$. A covariance matrix is valid when all of its eigenvalues are non-negative. Again, we can use the `eigen()` function in R to compute the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$, to see if all of its eigenvalues are non-negative. If at least one eigenvalue is negative, it implies that the matrix is not positive semi-definite. See Tutorial box 3 for an example how to check if the covariance matrix as obtained from the `genGGM()` function is positive semi-definite.

A way to specify a GGM in R obliging to the above mentioned condition can be done using the `genGGM()` function from the *bootnet* package [85]. This function generates a GGM as described in Yin and Li [279] using the Watts and Strogatz [260] algorithm for generating a graph structure. This creates the skeleton for our network. Generating a GGM with this function can be done by specifying the number of nodes one wants to include

To specify a GGM in R we can use the `genGGM()` function from the `bootnet` package. This function simply requires the specification of the number of nodes:

```
# Load libraries:
library("bootnet")
library("qgraph")

# Generate a GGM with 6 nodes:
trueGGM <- genGGM(Nvar = 6)

# Plot GGM:
qgraph(trueGGM, layout = "spring", theme = "colorblind")
```

This results in the generating of a chain graph. To move away from this chain structure, a rewiring probability `p` can be specified to randomly rewire a proportion of the edges. The probability to estimate a positive edge can be specified through the `propPositive` argument:

```
# Generate a GGM with 6 nodes:
trueGGM <- genGGM(Nvar = 6,
                  p = 0.5,
                  propPositive = 0.8)

# Plot GGM:
qgraph(trueGGM, layout = "spring", theme = "colorblind")
```

To check if the covariance matrix is a positive semi definite matrix, we can check if all eigenvalues are non-negative:

```
# Check if covariance matrix is positive semi definite:
all(eigen(diag(ncol(trueGGM)) - trueGGM)$values >= 0)
```

Tutorial box 3: Exemplifying the use of the `genGGM()` function from the bootnet package in R to specify a GGM.

in the network. This results in the generation of a chain or ring graph. In a chain graph each node is connected to its consecutive node in the form of a chain, i.e., if we consider an undirected network with 3 nodes there is an edge between the nodes 1–2, 2–3, and 3–1. To move away from this chain structure, a rewiring probability `p` can be specified to randomly rewire a proportion of the edges. Depending on the proportion of edges rewired, we obtain a random graph. Between the chain graph and random graph, we find a small-world network. To ensure small-world properties of the network, we can specify the `graph` argument as `"smallworld"` to create a small world network. Clusters in the network can be created by specifying `"clusters"` in the `graph` argument. The probability to estimate a positive edge can be specified through the `propPositive` argument, and

the range of the parameter values can be specified through the argument `parRange`. For exemplary use of the `genGGM` function to specify a synthetic network structure, see Tutorial box 3.

4.3.2.3 *Graphical vector autoregressive model*

The Graphical VAR (GVAR) model is an implementation of the VAR that is often used to model time series data of one individual [e.g., 10, 81]. The lag-1 GVAR model can be depicted as:

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{B}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_t, \tag{10}$$

$$\boldsymbol{\varepsilon}_t \sim \mathrm{N}(0, \boldsymbol{\Sigma}), \tag{11}$$

where $\mathbf{y}_t$ represents a vector of continuous responses by a given person on time-point t and $\boldsymbol{\mu}$ is the intercept. The $\mathbf{B}$ matrix contains all lag-1 regression coefficients β_{ij} of variable j on variable i encoding temporal effects, and its transpose is used to display a personalized directed weighted network structure referred to as the *temporal* network. Moreover, $\boldsymbol{\varepsilon}_t$ represents a vector of normally distributed innovations with variance–covariance structure $\boldsymbol{\Sigma}$. This innovation variance-covariance matrix $\boldsymbol{\Sigma}$ matrix can further be modeled as a GGM [89]. The inverse of the residual matrix $\mathbf{K} = \boldsymbol{\Sigma}^{-1}$ is used to create a *contemporaneous* network structure per individual. While the temporal network gives insight in how deviations from the person-wise mean in one variable at a certain measurement occasion predict deviations from the person-wise mean in the next measurement occasion, the contemporaneous network gives insight how variables within the same measurement occasion co-occur after controlling for temporal effects [41].

Thus, when specifying a GVAR model in R we need to specify *two* true data generating networks: the contemporaneous network, and the temporal network. As the contemporaneous network can be modelled as a GGM, we can simply use the `genGGM()` function in R as discussed in the previous section to generate the contemporaneous network structure. We can specify a temporal network by first creating a skeleton for the network structure. This means a structure that indicates whether there is an edge present between two nodes and in which direction. Next, we can sample the parameters, i.e., edge weights, from a random normal distribution with a certain mean and standard deviation. This procedure has been implemented in the `genGVAR()` function.

`genGVAR()` requires the specification of the number of nodes, and the mean (`mean` and standard deviation `sd` for the sample distribution determining the edge weights of the temporal network. As `genGVAR()` functions as a wrapper around `genGGM()` to generate the contemporaneous network, all arguments for `genGGM()` can be directly specified in the `genGVAR()`

To specify a synthetic GVAR model in R we can use the `genGVAR()` function. This function requires the number of nodes as input and a mean and standard deviation for the temporal edge weight parameters:

```
# Load in libraries:
source("https://raw.githubusercontent.com/RiaHoekstra/simulation_
       functions/main/genGVAR.R")
library("qgraph")

# Generate a GVAR with 6 nodes:
trueGVAR <- genGVAR(nNode = 6,
                    mean = 0.3,
                    sd = 0.1)

# Plot contemporaneous network:
qgraph(trueGVAR$PCC, layout = "spring", theme = "colorblind")

# Plot temporal network:
qgraph(trueGVAR$PDC, layout = "spring", theme = "colorblind")
```

Tutorial box 4: Exemplifying the use of the `genGVAR()` function to specify a GVAR model. The `genGVAR()` function is a developmental function that can be downloaded from github: `https://github.com/RiaHoekstra/simulation_functions/blob/main/genGVAR.R`.

function to generate the contemporaneous network. In addition, the proportion of positive edge weights the contemporaneous can be specified using the argument `propPositiveCon`. As the edge weights for the temporal network are sampled from a random normal distribution, a mean and standard deviation need to be specified to determine the shape of this distribution (see Tutorial box 4 for the use of `genVar()`). As output, `genVar()` returns the partial correlation network, `PCC`, that can be used to plot the contemporaneous network and the directed partial correlation network, `PDC`, that can be used to plot the temporal network in addition to the $\mathbf{B}$ and $\mathbf{K}$ matrix.

4.4 SIMULATING DATA BASED ON A NETWORK MODEL

The open-source software packages to specify psychological network structures discussed in the previous section, often contain methods to generate *data* from these structures (for an overview, see Table 2). In this section we illustrate the use of the `IsingSampler()`, `GGMsim()`, `GVARsim()`, to simulate data from an Ising model, a GGM, and a GVAR model respectively.

4.4.1 *Ising model*

To simulating data from an Ising model one can use the `IsingSampler()` function from the *IsingSampler* package [76]. The `IsingSampler()` function employs three methods to generate data from the Ising model. Data can be generated using a Metropolis-Hastings algorithm [48], using the Metropolis-Hastings algorithm together with Coupling from the Past [205], and simply computing the probability for every possible state and sample directly from the distribution of states using these probabilities, see Epskamp [76]. The latter method quickly becomes intractable (roughly above 10 nodes).

In Tutorial Box 5, we exemplify the use of the `IsingSampler()` function using the default Metropolis-Hastings algorithm to sample data from our specified synthetic Ising model in Tutorial box 2. `IsingSampler()` has four main arguments: `n` specifies the number of generated observations, `graph` specifies the underlying true network structure, `thresholds` controls the node specific external fields, and `responses` controls the encoding. See Tutorial box 5, how to generate data for 500 responses within the $\{-1, 1\}$ domain.

We can generate data from an Ising model in R by using the `Isingsampler()` function from the *IsingSampler* package.

```
# Load libraries:
library("IsingSampler")

# Generate data:
dataIsing <- IsingSampler(n = 500,
                          graph = trueIsing,
                          thresholds = thresholds,
                          responses = c(-1, 1))
```

Tutorial box 5: Exemplifying the use of the `IsingSampler()` function from the Isingsampler package to generate binary data from an Ising model.

4.4.2 *Gaussian graphical model*

To simulate data based on a GGM we can use the `GGMsim()` function. `GGMsim()` provides a way to simulate data from a multivariate normal distribution. As input, it requires the number of individuals for one wishes to generate data in the `n` argument, and the partial correlation matrix to generate random numbers from a multivariate normal distribution with a

mean of zero and covariance matrix $\boldsymbol{\Sigma}$ in the `omega` argument, see Tutorial Box 6.[4]

To generate multivariate normal data for $n = 100$ based on a GGM we can use the `GGMsim()` function in R.

```
# Load libraries:
source("https://raw.githubusercontent.com/RiaHoekstra/simulation_
       functions/main/GGMsim.R")

# Generate data:
dataGGM <- GGMsim(n =100,
                  omega = trueGGM)
```

Tutorial box 6: Exemplifying the use of the `GGMsim()` function from the xxx package in R to generate multivariate normal data from a GGM.

4.4.3 *Graphical vector autoregressive model*

To generate data from a GVAR model, we can use the `graphicalVARsim()` function from the *graphicalVAR* package [77]. This function provides a way to simulate stationary multivariate time series data based on a GVAR model using the `rmvnorm()` function[5]. In essence, `graphicalVARsim()` generates data following equation 10 for every time point using a specified $\mathbf{B}$ and $\mathbf{K}$ matrix. Therefore we need to provide the $\mathbf{B}$ and $\mathbf{K}$ matrix as input in the `graphicalVARsim()` using the `beta` and `kappa` arguments. Either we have obtained these matrices by estimating a GVAR model from data, or we obtained these matrices by generating a GVAR model using the `genVar()` function. In addition, we want to specify the number of time series data we want to generate, using the `nTime` argument. Optionally, a vector $\boldsymbol{\mu}$, containing person means can be specified in the `mean` argument, see tutorial box 7 for the use of `graphicalVARsim()`.

4 `GGMsim()` in essence is a wrapper around the `rmvnorm()` function from the `mvtnorm` package, that generates random values from a multivariate normal distribution.

5 Stationarity is a core assumption of the GVAR model and refers to a time series in which no changes over time are indicated by its defining characteristics, such as no change in means, variances, and network parameters [41]. Violations of this assumption arise if there are trends in the time series data, such as linear or seasonal trends. It is important to note that deviations from the stationarity assumption are not implausible in psychological time series. For example, we might observe shifts in the means of symptoms after certain life events or following treatment. However, as stationarity is one of the model assumptions, we simulate stationary time series data from a GVAR model.

To generate stationary time series data based on the GVAR model we can use the `graphicalVARsim()` function from the graphicalVAR package in R. The function requires the specification of the number of time series, the **B** and the **K** matrix:

```
# Load in libraries:
library("graphicalVAR")

# Generate time series data from GVAR model:
dataGVAR <- graphicalVARsim(nTime = 200,
                            beta = trueGVAR$beta,
                            kappa = trueGVAR$kappa)
```

Tutorial box 7: Exemplifying the use of the `graphicalVARsim()` function from the graphicalVAR package in R to simulate multivariate normal data from a GVAR.

4.5 ESTIMATING A NETWORK MODEL

After defining the data generating network and simulating data from this network, depending on the defined research question or aim of the simulation study, we perform some form of statistical analysis on the data. Often this involves the estimation of a network model based on the simulated data. What estimation technique to use however is entirely up to the researcher and dependend upon the research question. This part of the simulation study leaves a lot of room for the researcher as, for example, this is also the part where newly developed methods can be applied to the data to ultimately determine their performance.

To estimate network models based on data many estimation techniques exist [e.g., 153, 17, 82, 148, 276, 151], and it goes beyond of the scope if this chapter to give an overview of all network estimation techniques to estimate an Ising model, a GGM, and a GVAR within network psychometrics. Numerous tutorial papers describe these estimation procedures in detail, [e.g., 65, 82, 52, 148, 153, 93], and we refer the interested reader to those. In this section we briefly touch upon the estimation of network models as implemented in the `estimateNetwork()` function from the *bootnet* package as this function allows for the estimation of all three models discussed in this chapter.

4.5.1 *Ising model*

Here, we exemplify how to estimate an Ising model using the *bootnet* package function `estimateNetwork()` [85]. This function takes the data as input, and the `default` argument needs to be specified as `"IsingFit"` in

order to estimate an Ising model.[6] Most network estimation techniques fit the Ising model in a binary domain (i.e., $\{0, 1\}$), however, we simulated symmetric encoded data (i.e., $\{-1, 1\}$).[7] Therefore we need to transform the symmetric data to binary data before we can estimate the network model using the `estimateNetwork()` function. We can do so by encoding our -1 variables as 0, see Tutorial Box 8.

We can estimate an Ising model using the `estimateNetwork()` function from the *bootnet* package. As many statistical models to estimate the Ising model require the data to be binary, we first transform the data from $\{-1, 1\}$ encoding to $\{0, 1\}$ encoding:

```
# Load libraries:
library("bootnet")

# Transform encoded data to binary data:
dataIsing[dataIsing == -1] <- 0

# Estimate Ising model from simulated data in binary domain:
estIsing <- estimateNetwork(data = dataIsing,
                            default = "IsingFit")
```

Next we can transform the estimated network parameters to the $\{-1, 1\}$ domain in order to compare the estimated network parameters to the true network parameters:

```
# Load libraries:
library("IsingSampler")

# Transform estimated parameters from in binary domain to symmetric
domain:
estIsing <- LinTransform(graph = estIsing$graph,
                         thresholds = estIsing$results$thresholds,
                         from = c(0, 1),
                         to = c(-1, 1))
```

Next we can plot the estimated Ising model:

```
# Load libraries:
library("qgraph")

# Plot estimated network:
qgraph(estIsing$graph, layout = "spring", theme = "colorblind")
```

Tutorial box 8: Exemplifying the use of the `estimateNetwork()` function from the bootnet package to estimate an Ising model, and the `LinTransform()` function from the IsingSampler package to transform the network parameters from the $\{0, 1\}$ to $\{-1, 1\}$ domain.

6 The `estimateNetwork()` function in essence a wrapper around the `IsingFit()` function from the *Isingfit* package [21, 85].

7 There are some packages that allow for the estimation of Ising models within the $\{-1, 1\}$ domain such as *psychonetrics* [79].

In simulation studies we often like to say something about how well a certain technique managed to estimate the network parameters. As we now have a true network structure based on the $\{-1, 1\}$ domain, and an estimated network based on the $\{0, 1\}$ domain, we cannot simply compare the obtained network parameters from our estimated network to our true data generating network. We first need to transform our estimated network parameters back to the $\{-1, 1\}$ domain. This can be done using the `LinTransform()` function from the *IsingSampler* package [76]. The `LinTransform()` function requires the estimated network graph and threshold values as input, the specification from which domain the network is estimated (can be specified in `from` the argument), and to which domain we would like to transform the network parameters (can be specified in the `to` argument), see Tutorial Box 8. This way we can compare the estimated network parameters to the true network parameters.

4.5.2 *Gaussian graphical model*

To estimate a GGM from continuous normal data several estimation techniques exist. We refer the interested reader to Isvoranu and Epskamp [151] for an in dept discussion of the most common estimation techniques and recommendations for when to use which. Many of the model estimation procedures for a GGM are implemented in the `estimateNetwork()` function from the *bootnet* package. We shall illustrate how to estimate a GGM using an unregularized estimation technique (i.e., *ggmModSelect* algorithm).

To estimate a GGM we can use the `estimateNetwork` function from the bootnet package in R. This function requires the data and the type of model selection procedure as input:

```
# Load libraries:
library("bootnet")
library("qgraph")

# Estimate GGM from simulated data:
estGGM <- estimateNetwork(dataGGM,
                          default = "ggmModSelect")

# Plot estimated GGM:
qgraph(estGGM$graph, layout = "spring", theme = "colorblind")
```

Tutorial box 9: Exemplifying the use of the `estimateNetwork()` function from the bootnet package to estimate a GGM.

This method has been implemented in the `estimateNetwork()` function from the *bootnet* package [85] and can be used by setting the `default` argu-

ment to `"EBICglasso"`. The function simply requires the data as input, as well as the specification of which estimation technique to use by setting the `default` argument, see Tutorial Box 9.

To estimate a GVAR model we can use the `estimateNetwork()` function from the *bootnet* package. The functions requires the timeseries data of one individual as input as well as the specification of which estimation technique to use.

Estimating a GVAR model using the `estimateNetwork()` function can be done as follows:

```
# Load libraries:
library("bootnet")
library("qgraph")

# Estimate GVAR model from simulated data:
estGVAR <- estimateNetwork(data = dataGVAR,
                           default = "graphicalVAR")
```

We can plot the estimated GVAR as follows:

```
# Plot results:
layout(t(1:2))

# Plot temporal network:
qgraph(estGVAR$graph$temporal, layout = "spring",
       theme = "colorblind")

# Plot contemporaneous network:
qgraph(estGVAR$graph$contemporaneous, layout = "spring",
       theme = "colorblind")
```

Tutorial box 10: Exemplifying the use of the `graphicalVAR()` function from the graphicalVAR package and `gvar()` function from the psychonetrics package in R to estimate a GVAR.

4.5.3 *Graphical vector autoregressive model*

The most commonly used technique to estimate a GVAR model by finding a sparse solution for both **B** and **K** is by estimating a regularized temporal and contemporaneous network using LASSO regularization. This algorithm requires two tuning parameters, one for the temporal coefficients and one for the contemporaneous coefficients, that are selected via Extended Bayesian Information Criterion [EBIC; 89, 41]. This procedure of estimating individual network models has been implemented in

the `estimateNetwork()` function from the *bootnet* pacakge.[8] The function `estimateNetwork()` simply requires the generated time series data as input, and the specification of the `default` argument to `"graphicalVAR"`, see Tutorial Box 10.

4.6 EVALUATING RESULTS

In the previous sections, we discussed how to determine the data generating network model, how to simulate data based upon this true data generating network, and how to estimate a network based on the simulated data for each of the three most popular modeling frameworks (i.e., Ising model, GGM, GVAR). Let us return to our original example from Tutorial Box 1 to exemplify the evaluation of the simulation results.

After running our simulation study, we want to analyze the results. How, and which results to display is highly dependent upon the research question. In general, we would like to summarize our results in the form of a table or plot showing the distribution of evaluation metrics obtained for our various conditions. We can use the packages *dplyr* and *tidyr* for data manipulation, and *ggplot2* for plotting the results [270, 268, 269, 267].

Recall that the example research question in this chapter asks to what extent we can accurately estimate a GGM for reasonable sample sizes. In order to answer this research question, we varied the sample size condition (i.e., $n \in \{50, 100, 200\}$), and obtained the sensitivity, specificity, and correlation as our evaluation metrics. In order to plot our results with the `ggplot()` function, some data modifications need to be made. First we want to format our results in such a way that can be used within `ggplot()`. That means, we need to go from a wide data format to a long data format. We can use the `gather()` function from the `tidyr` package to do so, see Tutorial box 11. In addition, we want to group the metrics according to their condition. With the `mutate()` function from the `dplyr` package we can add an additional column to the results to change the name of the metric variable, see Tutorial box 11.

After formatting the data, we can use `ggplot()` to display the simulated results and start interpreting the results (see Tutorial box 11). Our results indicate that the median correlation between the true and estimated network increases as a function of sample size. So does the median sensitivity, and specificity, meaning we find increasingly more true effects, and include less false effects with increasing sample size, see Figure 11.

Appendix C contains a simulation set up like the one from Tutorial Box 1 in combination with how to analyze the results as exemplified in Tutorial box 11 for the Ising model and the GVAR model.

8 In essence this function makes use of the `graphicalVAR()` function from the *graphicalVAR* package [77]

```
# Load libraries:
library("dplyr")
library("tidyr")
library("ggplot2")

# Inspect results:
View(results)     # These are the results obtained from Tutorial box 1

# Put data frame in long format for use of ggplot:
long <- gather(results, metric, value,
               sensitivity, specificity, correlation)

# Make factors with labels:
long$nSample <- factor(long$nSample,
                       levels = c(50, 100, 200),
                       labels = c("50", "100", "200"))

# Let's plot:
ggplot(long, aes(x = factor(nSample), y = value, fill = metric)) +
  geom_boxplot() +
  facet_grid(~ metric) +
  theme_bw() +
  scale_fill_manual(values = c("#4C71BB","#CC79A7","#A22016")) +
  xlab("sample size") +
  ylab("") +
  guides(fill = guide_legend(title = ""))
```

Tutorial box 11: Tutorial on how to analyze simulation results as obtained from Tutorial box 1.

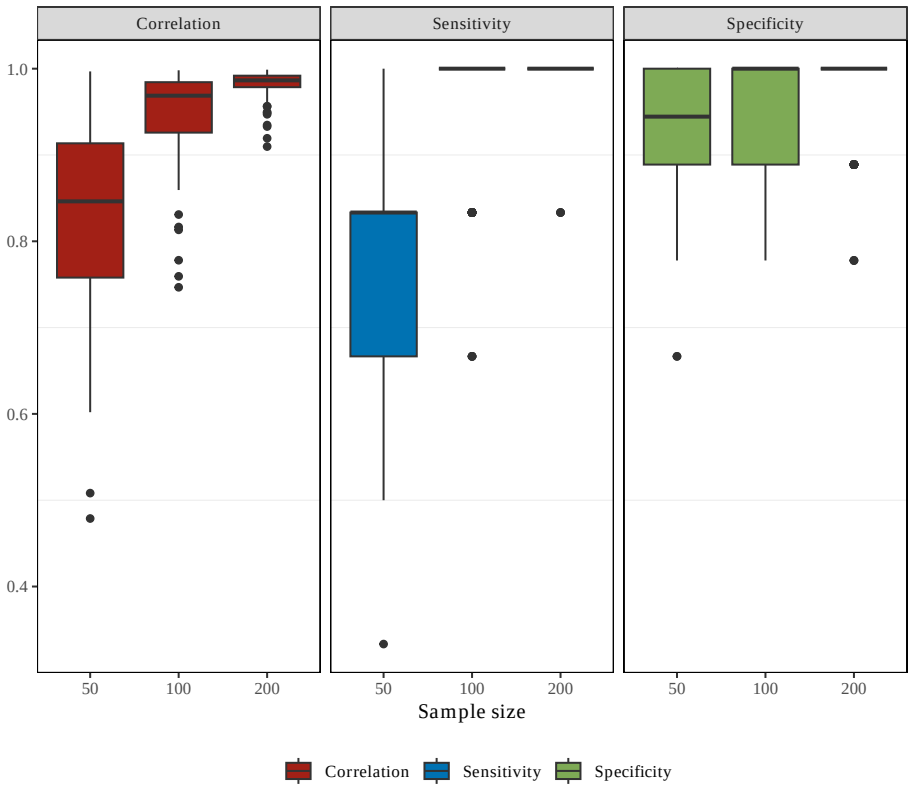


Figure 11: Results from the Gaussian graphical model simulation tutorial, see Tutorial Box 11.

4.7 CONCLUDING COMMENTS

In this chapter, we showed how to use contemporary open-source software to set up a simulation study within the field of network psychometrics. Our aim was to make simulation tools available to a broader audience of researches by providing a step-by-step guide in setting up a simulation study for the three most often used models: the Ising model, GGM, and the GVAR model. As a running example, we highlighted how these simulations tools can be used to determine the estimation accuracy of model parameters for various sample sizes. The general idea we highlighted with this example can be extended to the modeling frameworks discussed in this chapter (see Appendix C for a tutorial), to different estimation techniques, and to asses the expected replicability for a given network structure, by, for example, adding an additional independent sample to the simulation procedure.

Being able to performing simulation studies has many advantages as for example methods for sample size determination are limited in general, and especially within the network modeling framework where estimation

quality may depend on the network structure relevant in one's application. Some simulation based tools have been proposed in the context of cross-sectional network estimation to determine the expected accuracy of a given network structure. For example, `netSimulator()` from the *bootnet* package provides a framework for simple simulation studies regarding the performance of network estimations techniques as a function of the sample size [85]. In addition, recently a new R-package has been introduced to compute sample sizes: *powerly* [50]. *Powerly* makes use of an automated Monte Carlo algorithm to find an optimal sample size given a network model structure, a performance measure of interest (e.g., a sensitivity of 0.7) and a corresponding target value (e.g., reaching a sensitivity of 0.7 with a probability of 0.8). One of the benefits of *powerly* its hat the implemented stratified bootstrapping procedure allows to quantify uncertainty around the recommendations provided by the package. While the methods exploited in both packages can be extended to other types of network models, currently the packages only support the determination of sample sizes for cross-sectional network estimation techniques such as GGM.

Thus, currently it not possible to apply these simulations based tools to asses the number of time-points needed for network estimated from intensive longitudinal data using GVAR estimation as implemented in the *GVAR* package or multilevel network estimation as implemented in the *mlVAR* package [77, 83]. In addition, it is not possible to asses the required sample size for more recently developed cross-sectional estimation techniques such as (pseudo) ML estimation as implemented in the *psychonerics* package [73, 79] or Bayesian estimation of network models as implemented in the *easyBGM* package [147, 148]. The simulation tools provided in this tutorial will enable researchers to do so, and therefore lift these limitations.

Assessing expected replicability is another benefit of performing simulation studies. Given sampling variability, we should not expect networks to reproduce completely, and to what extent we do expect networks to replicate depends on several factors such as the network structure. Therefore, it is important to determine the replicability and generalizability of estimated psychological network models across samples. For this purpose, tools have been developed to asses the expected replication in the context of cross-sectional models such as the `replicationSimulator()` function from the *bootnet* package [85]. The `replicationSimulator()` function can be used to asses how well a network based on a second independent sample would replicate based on the first independent sample. As for the `netSimulator()`, and `powerly()` the `replicationSimulator()` is only applicable for cross-sectional network estimation techniques.

Armed with the simulation tools described in this chapter at hand, more researchers can extend the ideas behind functions as `netSimulator()` and `replicationSimulator()` to any type of network estimation technique to wish to employ.

Part II

UNDERSTANDING HETEROGENEITY

5

RELATING NETWORK STABILITY TO SYMPTOM SEVERITY

ABSTRACT

A hypothesis flowing from the network theory of psychopathology is that symptom network structure is associated with psychopathology severity. In turn, one may expect that individual network structure changes along the levels of psychopathology severity. However, this expectation has rarely been addressed directly. In this chapter we aim to examine (1) the stability of individual contemporaneous symptom networks over a one-year period and (2) whether network stability is associated with a change in psychopathology. We used daily diary data of $n = 66$ individuals located along the psychosis severity continuum, from two separate 90-day periods, one year apart ($t = 180$). Based on the outcome of the Individual Network Invariance Test (INIT) presented in Chapter 3 to determine symptom network stability, participants were divided into two groups: stable and unstable. We tested whether these groups who differed in their network stability over the study period, differed in their absolute change in psychopathology severity during that same period. The majority of the sample ($n = 51, 77.3\%$) showed a stable network over time while most individuals showed a decrease in psychopathological severity. We found no significant association between a change in psychopathology severity and individual network stability. The results presented in this chapter call for further evaluation of the association between idiographic networks as estimated from data and psychopathology to optimize the implementation of current methods in clinical applications.

5.1 INTRODUCTION

The development of psychopathology involves complex interactions between processes at different levels and time scales [90, 194]. Understanding this complex development might be advanced by viewing psychopathol-

Published as: van der Tuin, S.*, **Hoekstra, R. H. A.***, Booij, S. H., Oldehinkel, A. J., Wardenaar, K.J., van den Berg, D., Borsboom, D. & Wigman, J. T. W. (2023). Relating stability of individual dynamical networks to change in psychopathology. *PloS ONE*, *18*(11), e0293200.

*authors contributed equally

ogy from a network perspective. The network perspective conceptualizes mental disorders as the result of the direct interplay between symptoms that can trigger and maintain each other [24]. This interplay can be visualized in a network, in which symptoms are represented as nodes, and statistical relationships between these symptoms are represented as edges [82]. The network approach has rapidly gained substantial influence in the field of psychiatry [24, 38, 227, 13]. Networks of psychopathological symptoms can be modeled in many ways and a plethora of studies have discussed the principles, applications, merits and drawbacks of a network approach to psychopathology [24, 38, 227, 84, e.g.,]. Symptom networks have been estimated from cross-sectional and, increasingly, longitudinal and time series data. An important advantage of this latter data type is that symptom networks can be constructed per individual by modeling the immediate associations between symptoms using multiple time points—creating the opportunity to gain insight into the dynamic interplay of symptoms at an individual level [28].

One of the main hypotheses flowing from the network theory of psychopathology is that the network structure is associated with the level of psychopathology severity. The symptom structure in a strongly connected network is thought to be self-sustaining: one small perturbation to the network can activate all symptoms in the network, posing a vulnerability to (more) severe psychopathology. On the other hand, when connections between symptoms in such a network are weak, perturbations to the network may have a limited effect, and the system is thought to return to its initial healthy state quickly [28]. We refer the interested reader to Chapter 7, where we discuss and investigate the most important postulates from the network theory of psychopathology in more detail.

Several studies have investigated the connection between the structure of symptom networks as estimated from data and psychopathology, and found a an apposing effect: several studies found that more strongly connected symptom networks are related to more severe and persisting psychopathology [172, 266, 271]. Using a multilevel-based estimation technique, De Vos et al. [67] found an overall higher network density within individuals suffering from major depressive disorder (MDD) compared to healthy individuals. However, when taking an individual-level approach using sparse vector autoregression techniques on that same data, they found more and stronger associations in the population network of healthy individuals compared to the network of individuals suffering from MDD. In addition, they found no relation between individual network density and psychopathology as measured at baseline. This demonstrates that the relation between a network and psychopathology in longitudinal settings is complex and sensitive to the type of analysis

and underlines the necessity of further research in order to elucidate the relation between individual network structure and psychopathology.

To date, research examining the association between network structure and psychopathology has mainly focused on the association between network structures estimated from longitudinal data and and the cross-sectional assessment of psychopathological status or severity. However, given the hypothesized association between symptom network structure and psychopathology, one may expect variations in psychopathological severity may parallel alterations in the network's structure. For example, a diminution in the density of the symptom network can be hypothesized to correlate with an improvement in mental health, whereas an enhancement in the density of the network structure is posited to lead to a deterioration in mental health upon activation of the symptom network. This hypothesis has frequently been implicitly adopted in studies employing network models for clinical applications, particularly in the development of interventions aimed at optimizing treatment targets based on network structure [224, 173]. Despite this, to the best of our knowledge, the stability of individual symptom networks over time and the potential association between alterations in individual symptom network structures and psychopathological changes remain unexplored. Understanding this association could contribute to the refinement and development of targeted interventions using network models.

In this chapter we aim to examine (1) the stability of individual symptom networks over a period of 1 year and (2) whether symptom network stability was associated with psychopathology during this time period. To that end, we compared two separate periods of diary assessments one year apart, each spanning 90 days of daily symptom reports. Two thing are important to note here. First, there is no clear consensus of what defines "stability". The way we define stability in the current chapter is depended on our statistical approach. Concretely, we employed the Individual Network Invariance Test (INIT), as was introduced in Chapter 3, to determine network stability. This means we consider networks stable over time when a single network structure fits the two periods of time series data better in terms of model fit, than two different network structures. Second, the study was performed on a convenience sample of young adult individuals located along the psychosis severity continuum who were considered at increased psychometric or clinical risk for psychosis [16].

5.2 MATERIALS AND METHODS

5.2.1 *Participants and study design*

Data came from the Mapping Individual Routes of Risk and Resilience (Mirorr) study [16, 272]. Participants were individuals located along the psychosis severity continuum who were considered at increased psychometric or clinical risk for psychosis (N = 96, age 18–35 years). Participants were assessed during two daily diary periods of 90 days each (one at baseline and one at 1-year follow-up), as well as with cross-sectional questionnaires at baseline and three-yearly follow-up (T0-T3) measurements on mental health and functioning. The sample consisted of N = 25 individuals from the general population who were at increased psychometric risk for psychosis. These were the 25% highest scoring individuals on a questionnaire assessing psychotic experiences (Community Assessment for Psychic Experiences; CAPE) [175]. The other N = 71 individuals were recruited from mental health care institutions in the Netherlands and thus were in clinical care at the moment of inclusion. As the sample consisted of individuals who are at risk for psychosis, individuals with a history of, or current psychotic episode, were excluded. The participants who were recruited from mental health care institutions presented with a broad range of mental disorders, but the majority was diagnosed with depression and/or anxiety. This high level of comorbidity and specifically of mood disorders is reflective of the nature of individuals at UHR for psychosis [141].

Inclusion criteria were: (1) age between 18 and 35 years, (2) ability to read and speak Dutch fluently, (3) being capable of following the research procedures, and (4) providing informed consent. Exclusion criteria were: (1) history of or current psychotic episode according to the Diagnostic and Statistical Manual of Mental Disorders 4 (DSM-4) criteria, (2) significant hearing or visual problem impairments, and (3) pregnancy. We refer the interested reader to Booij et al. [16] for a more detailed description of the design and procedure of the Mirorr study.

The study was approved by the medical ethical committee of the University Medical Centre Groningen, Groningen, the Netherlands (registration number MEC no. 2015/159, ABR no. NL52974.042.15). The study was conducted in accordance with the Helsinki Declaration. All participants provided written informed consent.

For this chapter, data from two assessment waves were used: baseline (T0) and follow-up after one year (T1). Both waves contained a 90-day diary study in which several psychopathological symptoms, thoughts, emotions, and behaviors were assessed daily via the participants' smartphones. To assess symptom network stability over time, we only included

participants who completed diary data at both waves, resulting in a sample of $n = 66$ individuals. Individuals in this final sample were slightly older than the dropouts (mean age completes: $m = 25.5$, $sd = 4.49$, mean age dropouts: $m = 23.1$, $sd = 3.07$, $p = .03$), and did not differ on gender, education and psychopathology.

5.2.2 *Instruments*

DIARY ITEMS The diary questionnaire contained 80 items on a broad range of feelings and emotions, functioning, and behaviors. All items were scored on a 100-point Visual Analogue Scale (VAS) ranging from 'not at all' to 'very much', and were completed once a day in the evening. Procedures were identical at T0 and T1. We limited the number of variables in the analysis to ensure adequate estimation accuracy and power by computing domain scores through averaging the scores of the items related to the same domain [195]. In line with previous work on these data [254], five domains of psychopathology: irritation, stress, depression, psychosis, anxiety, and one confidence domain were chosen for the current study, which were based on 16 individual items. We focused specifically on transdiagnostic domains as the early clinical stages for psychosis are characterized by a transdiagnostic and diffuse nature [141]. Domain structure was chosen based on theoretical grounds, i.e., based on which items typically cluster together, and then assessed psychometrically by checking the composite reliability scores. Thus, for all domains, we calculated composite reliability scores taking the multilevel structure into account [110] through the R-package 'multilevelTools' [273], resulting in within-person omega's ranging between .60 and .86 and between-person omega's ranging between .83 and .99 based on our final sample of $n = 66$ individuals.

PSYCHOPATHOLOGY QUESTIONNAIRE As a measure of psychopathology, the Dutch version of the Symptom Checklist Revised [SCL-90-R; 6] was used. The SCL-90-R measures general psycho-pathology over the past week, with 90 items scored on a 5-point Likert scale. The SCL-90-R has high reliability [$\omega = .98$; 247]

5.2.3 *Statistical analyses*

All analyses were performed in R version 4.1.0 [220] and the conventional alpha $p < .05$ was used as the inference criterion.

DESCRIPTIVE ANALYSES Descriptive statistics on age, gender, education, general psychopathology, and the five diary domains for the total

sample were calculated. Mean psychopathology severity (SCL-90-R) was compared between T0 and T1 by means of a paired t-test. For each of the diary domains, the within-person median over the 90 days was calculated, and then averaged across the total sample. As the data was nested within persons, the scores on the diary domains between T0 and T1 were compared using multilevel models (one per domain) with the R package 'nlme' [218]. All multilevel models had time-point (T0 versus T1) as the fixed effect, random intercepts for subjects, and random slopes for time (the 90 measurements) and time-point.

PREPROCESSING STEPS To adhere to the model assumptions of the graphical Vector Autoregressive model (GVAR) used to estimate individual network models, before constructing these symptom networks, (1) missing data were imputed for each individual using exponential moving average, (2) data were transformed with a nonparanormal transformation [91], and (3) detrended per individual (for more information, see Appendix D). Detrending led to a situation where only model deviations around the individual's trend were analyzed, so that dynamics in the networks reflect the relation between the deviations from the trend of each domain score. Thus, networks should not be interpreted as reflecting the degree to which stable increases in one symptom domain are associated with stable increases in another domain, but rather as reflecting the degree to which changes in one variable are associated with changes in another variable.

SYMPTOM NETWORK CONSTRUCTION Individual symptom networks were estimated for each individual at both T0 and T1 as GVAR models in R using version 0.10 of the *psychonetrics* package [79]. In *psychonetrics*, first a fully connected network model is estimated. After this, several model selection techniques can be applied to obtain sparse network structures in which only some edges are included. In this chapter, we use significance pruning at $\alpha = 0.05$ to obtain sparse individual network structures. By pruning the fully connected network model, edges are excluded based on the set significance level. Next, the network structure is re-estimated with those edges fixed to zero. This re-estimation ensures that the edge weight estimates of the included edges are based on the final model. This leads to more accurate edge weight estimations [14].

Estimating GVAR networks based on longitudinal individual data results in the estimation of a directed, temporal network (based on VAR coefficients) and a contemporaneous network (based on VAR residuals). The temporal network shows the relationship between two variables from the previous measurement occasion onto the next measurement occasion while controlling for the temporal effect of all other variables. The contem-

poraneous network reflects the unique bidirectional associations among variables that occur after time effects have been taking into account.[1]

In previous work using the same data, we found that most symptom associations in this dataset occur within days (i.e., contemporaneous) rather than between days (i.e., temporal) [254, 253]. This is in line with other research on diary studies that assess mental states once a day [241]. In addition, assessing the stability of temporal networks with little to no edges has proven to yield limited results [144]. Therefore, for this study, we only interpreted the contemporaneous networks.

INDIVIDUAL SYMPTOM NETWORK INSPECTION To gain more insight into symptom network structures at both time points, we assessed which edges between nodes were estimated in the pruned network structures at T0 and T1, both between- and within-individuals, by calculating the percentage of edges estimated at T0 and T1 across individuals, and calculating the percentage of individuals for whom an edge was estimated similar at both T0 and T1.

INDIVIDUAL SYMPTOM NETWORK STABILITY To test the stability of symptom networks of individuals over time, INIT was used.[2] INIT makes use of idiographic network estimation techniques as implemented in the *psychonetrics* R package [79]. In the psychonetrics R package, individual network models are estimated as GVAR models using (pseudo) Full Information Likelihood (FIML) estimation. Using psychonetrics, first a saturated network model is estimated (i.e., a network in which all edges are estimated to be non-zero), after which several methodologies can be applied to remove edges from the saturated network structure. To make the comparison between network structures and stability outcomes not dependent on the methodology to induce sparse network structures, we performed INIT on the saturated (i.e., fully connected) network structures.

INIT compares a model in which the edge weights of both network structures at T0 and T1 for one individual are estimated freely, (i.e., no constraints are placed between an individual's network at T0 and T1), to a model in which the edge weights for both networks at T0 and T1 of the individual are constrained to be equal, (i.e., equality constraints are placed between an individual's network structure at T0 and T1). INIT determines which of these two models fits best given the data by computing and comparing a model fit index for both models. Model fit is evaluated according to the Akaike Information Criterion (AIC), as simulation stud-

1 We refer the interested reader to Epskamp et al. [81] for an in-depth discussion on the differences between temporal and contemporaneous networks and their potential in clinical research.

2 The package *INIT* can be installed via https://github.com/RiaHoekstra/INIT.

ies showed the AIC fit index performs best compared to the χ^2 difference test and the Bayesian Information Criteria (BIC) when equality constraints are placed on saturated individual network structures, see Chapter 3, and Appendix D. The model with the lowest AIC is the best fitting model.[3]

As a sensitivity check, we also performed INIT on pruned network structures and calculated correlation coefficients between individual network structures at T0 and T1, see Appendix D.

Using the above-described approach, we determined for each person whether their symptom networks at T0 and T1 were similar or different over time. Those who had similar symptom networks at T0 and T1 were assigned to the "stable" group and those who had different symptom networks at T0 and T1 were assigned to the "unstable" group. Group membership resulting from INIT was used for further analyses.

DESCRIPTIVE COMPARISON OF INIT GROUPS The stable and unstable groups were compared on age (t-test), gender (Fisher's exact test), and education (Fisher's exact test). Psychopathology severity (t-test) and the scores on the six diary domains (multilevel models) were compared between the groups at T0 and T1. As both psychopathology and the diary domains were measured at T0 and T1, we performed additional tests to assess changes from T0 to T1. For a difference in psychopathology, we assessed whether there was a significant change between T0 and T1 for the stable and unstable groups separately by means of a paired t-tests. For differences in diary domains, we assessed whether there was a significant change from T0 to T1 per group separately using multilevel models.

THE ASSOCIATION BETWEEN NETWORK STABILITY AND CHANGE IN PSYCHOPATHOLOGY SEVERITY To assess whether a change in network structure was associated with a change in psychopathology, we tested whether absolute change scores of psychopathology severity differed per group through an independent samples t-test.

POWER Previous simulation studies regarding estimating idiographic network structures showed that we have sufficient power to reliably estimate idiographic networks with 90 measurements per person [195]. We conducted an additional simulation study to test the power to detect instability in individual network models using INIT, see Appendix D. Results of this simulation study indicated with six variables and 90 time points per individual, INIT is able to pick up moderate to high instability between network structures.

3 For a more detailed description of the INIT methodology, see Chapter 3.

5.3 RESULTS

5.3.1 *Demographics*

For information on the full sample demographics and differences in means on psychopathology as measured with the SCL-90, as well as mean difference for diary domains at T0 and T1, see Table 3.

5.3.2 *Symptom network structure*

The number of times an edge was detected across individuals at T0 and T1, as well as the number of individuals for whom an edge was estimated similar at both T0 and T1 was relatively stable, see Table 4.

Demographics		
Age *mean, (SD)*	25.5 (4.49)	
Gender *frequency, (%)*		
Female	53 (80.3%)	
Male	13 (19.7%)	
Education *frequency, (%)*		
Low	10 (15%)	
Middle	36 (54%)	
High	20 (30%)	
	Measurement Time	
	T0	**T1**
Psychopathology *mean, (SD)*		
SCL-90*	189.24 (57.66)	160.30 (47.40)
Diary Domains *media, (IQR)*		
Irritation*	19.3 (28.7)	14.5 (21.1)
Stress*	31.1 (25.8)	25.8 (31.9)
Depression*	25.6 (28.1)	22.3 (20.8)
Psychosis*	7.6 (9.9)	6.4 (9.3)
Anxiety*	17.9 (28.3)	12.7 (23.3)
Confidence*	47.9 (22.3)	50.1 (18.8)

Note: *indicates significant difference between measures on T0 and T1, $p < .05$

Table 3: Information on full sample demographics ($n = 66$), difference in means on psychopathology as measured with the SCL-90, and dairy domains for T0 and T1.

	T0		T1		T0 & T1	
Edge	Frequency	Percent	Frequency	Percent	Frequency	Percent
Depression-Confidence	59	89.4%	60	90.9%	55	83.3%
Stress-Anxiety	44	66.7%	39	59.1%	43	65.0%
Irritation-Depression	31	47.0%	30	45.5%	35	53.0%
Depression-Anxiety	25	37.9%	29	43.9%	54	81.8%
Irritation-Stress	20	30.3%	27	40.9%	33	50.0%
Irritation-Psychosis	20	30.3%	26	39.4%	40	60.6%
Psychosis-Anxiety	19	28.8%	28	42.4%	43	65.0%
Stress-Depression	15	22.7%	10	15.2%	43	65.0%
Psychosis-Confidence	15	22.7%	9	13.6%	52	78.8%
Depression-Psychosis	14	21.2%	15	22.7%	43	65.2%
Stress-Confidence	13	19.7%	18	27.3%	39	59.1%
Irritation-Confidence	12	18.2%	4	6.1%	50	75.8%
Stress-Psychosis	11	16.7%	19	28.8%	44	66.7%
Irritation-Anxiety	10	15.2%	13	19.7%	45	68.2%
Anxiety-Confidence	10	15.2%	8	12.2%	52	78.8%

Table 4: Frequency and percentage of estimated edges at T0, T1, and at T0 & T1, for the full sample ($n = 66$).

5.3.3 *Symptom network stability*

Application of INIT revealed, that for 51 out of the 66 individuals the model with equality constraints showed the best model fit. Thus, the symptom networks of these individuals were considered similar in structure (i.e., edge weights) at both time points. These 51 individuals were labeled as the 'stable group'. For the other 15 individuals, the model without equality constraints fit the data better. Thus, the symptom networks of these individuals were considered different in structure at both time points. These individuals were labeled the 'unstable group'.

5.3.4 *Comparison between stable and unstable group*

The stable ($n = 51$) and unstable group ($n = 15$) did not differ significantly in age, gender, and education level. For more details on demographics per group, see Table 6. In addition, we found no significant group differences in psychopathology severity, nor in diary domain scores. While there is no statistical significant difference between absolute change scores of psychopathology between the stable and unstable group, some observations can be made. It appears that the stable group shows more stability in the median scores on the diary domains than the unsta-

ble group, especially for Stress, Depression, and Confidence, see Table 6 for more details on psychopathology severity and diary domain scores.

	stable group	**unstable group**
Demographics	$n = 51$	$n = 15$
Age *mean, (SD)*	25.7 (4.4)	25.4 (4.6)
Gender *frequency, (%)*		
Female	41 (80%)	12 (80%)
Male	10 (20%)	3 (20%)
Education *frequency, (%)*		
Low	7 (14%)	3 (20%)
Middle	26 (51%)	10 (66%)
High	18 (35%)	2 (13%)

Table 5: Demographics for the stable and unstable group.

The difference in psychopathology severity between T0 and T1 was significant for both the stable group ($t = 4.6, p < .01$) and the unstable group ($t = 3.9, p < .01$). Multilevel models showed that out of the five domain scores, Depression, Psychosis, Anxiety, and Confidence, differed significantly between T0 and T1 for the stable group, and all five domain scores differed significantly between T0 and T1 for the unstable group.

	stable group $n = 51$		**unstable group** $n = 15$	
	T0	T1	T0	T1
Psychopathology *mean, (SD)*				
SCL-90	185.2 (56.0)	159.9 (44.9)	203.1 (63.1)	161.8 (56.9)
Diary Domains *median, (IQR)*				
Irritation	19.2 (30.2)	14.3 (23.2)	19.5 (24.0)	14.7 (16.2)
Stress	29.9 (28.1)	25.6 (32.2)	45.5 (20.8)	28.2 (22.1)
Depression	21.4 (27.2)	21.3 (23.1)	37.4 (26.8)	25.3 (10.7)
Psychosis	7.5 (8.2)	6.4 (9.2)	7.9 (10.2)	6.4 (8.9)
Anxiety	18.3 (24.8)	12.8 (23.3)	14.8 (46.1)	9.2 (13.4)
Confidence	49.3 (20.6)	49.5 (16.9)	41.4 (29.6)	52.2 (28.3)

Table 6: Psychopathology and diary domains for the stable and unstable group.

5.3.5 *The association between network stability and change in psychopathology*

A two-sample t-test showed that the stable and the unstable group did not differ significantly in their absolute psychopathology-severity change scores (stable = $38.96(25.45)$ [mean(sd)], unstable = $42.87(39.13)$ [mean(sd)], $t = .39, p = .70$).

5.4 DISCUSSION

This study aimed to explore the association between the stability of individual symptom networks over time and changes in psychopathology severity in a convenience sample of individuals situated along the psychosis severity continuum, identified as being at increased psychometric or clinical risk for psychosis. Inspecting the stability of individual contemporaneous networks, we found that for a substantial portion of the participants (51 out of 66 participants), the networks were consistent across two measurement points separated by a one-year interval. Furthermore, the majority of participants exhibited a reduction in psychopathology severity, as well as a change in domain scores after one year, regardless of the stability of the symptom network. We did not find evidence linking the stability of individual contemporaneous symptom networks to modifications in psychopathology severity. Consequently, our findings challenge the notion to link psychopathological severity directly to contemporaneous symptom network structures.

Our findings are consistent with those of Snippe et al. [248], who found that treatment did not change the dynamic structure of symptom networks for individuals diagnosed with depression, despite the observable impact of treatment on average symptom levels. Their analysis, however, was rooted in group-based multilevel methodologies, potentially concealing the individual variances related to temporal shifts in symptom networks [33, 95, 123, 202]. Our individual-centered approach led to similar conclusions. In our sample, participants also showed a significant decrease in mean levels of psychopathology and symptom domains, with most individuals showing no changes in symptom network structure. Our results force us to reflect critically on the usefulness of network structure evaluations as a means to explain psychopathology severity. In the development of intervention programs to improve mental health that work with feedback based on individual network structures ([e.g., 224, 173, 232]) it is crucial to acknowledge the yet-to-be-fully-elucidated link between the network structure (or elements thereof) and psychopathology severity or clinical state. To facilitate this understanding, Chapter 7 introduces a tool to potentially bridge the gap between individual network structure dynamics and psychopathology severity and stability.

Upon inspecting the individual symptom network structures, we found that the edge between the Confidence domain and Depression domain persisted across the majority of participants both at T0 as well as T1. This is in line with earlier work on this sample[254], and highlights the usefulness of individual symptom networks construction to gain more insight into how different domains of symptoms are connected, which in turn could provide useful input for research aimed at understanding the devel-

opment of psychopathology. Especially in the case of confidence and depression, this highlights the potential relevance to incorporate constructs from positive psychology, e.g., increasing confidence as a protection mechanism to prevent the emergence or worsening of depressive symptoms [154].

Some limitations need to be taken into account when interpreting our results. First, several factors have limited the chance of finding a difference in psychopathology changes between the stable and unstable group. Group membership was based on INIT. As INIT tests similarity between individual network structures by comparing a model with equality constraints to a model in which all parameters have been estimated freely, this results in a large increase in the number of parameters estimated for the model without equality constraints. Hence, when it comes to detecting individual differences, INIT errs on the side of caution. Further applications could look into comparing multiple models to test intermediate forms of inequality in individual network structures over time. Second, the application of INIT resulted in a small number of individuals being allocated to the unstable group, which decreased the power to detect a statistical association between the stable and unstable group and changes in psychopathology. Moreover, as a statistical difference is not necessarily a meaningful difference, hypothetically, the difference between the stable and unstable group may have been smaller than anticipated, resulting in less discriminating groups than expected.

Furthermore, it is possible that important changes in individual network structures occur on a different level than modeled in this study. We used the most popular method to estimate individual network structures: the GVAR model. The GVAR model assumes stationarity and so estimates deviations around an individual's trend, rather than stable increases. We want to mention two points related to this in the current chapter. First, stationarity is a strong assumption that might not be tenable, especially in longer time series. Second, idiographic network structures as estimated using GVAR show how temporal deviations from one variable's mean affect another variable's temporal deviations from its means. However, relations of the mean trends themselves are not part of the model, i.e., the model assumes one stable mean over time. Therefore, the extent to which mean changes in one variable influence mean changes in another variable, are currently not modelled, even though this change is arguably the most meaningful from a clinical (and experiential) perspective. Incorporating changes in the mean of symptom (node) scores into the estimation of network structures is an important future step that potentially shed light on the relationship between psychopathology and idiographic network structure. To the extent that network theory describes patterns between means levels, our results are less informative for the network theory of

psychopathology, but highlights the importance of when and under what conditions the network theory is expected to hold giving our current modeling tools. We refer the interested reader to Chapter 7 for a more in dept discussion on the link between individual network models and the network theory of psychopathology.

A consideration to take into account when interpreting our results is that in the network structures, we included five negative emotion nodes and one positive emotion node. We deliberately did not interpret whether network stability is positive or negative for each individual, rather, we investigate the association between network stability and psychopathology changes. In addition, our domain structure was chosen based on theoretical reasoning (i.e., which items typically cluster together clinically) rather than using statistical techniques (e.g., factor analysis). A second consideration to keep in mind is that symptoms were measured only once a day, while the processes of symptom development may unfold within a shorter time frame. In previous studies on the same dataset [253, 254], we observed that most relationships between variables appeared within days rather than over days. Therefore, we looked exclusively at contemporaneous relationships between variables, which due to the design of our study reflect the associations within a day. If the variables had been assessed on a different timescale, e.g., multiple measurements within a day, we might have found more temporal associations and we cannot exclude that this would have shown a different pattern of stability over time. One important difference between contemporaneous and temporal networks is that in the latter, the direction of associations can be estimated, and taking this directionality into account might lead to different results. For example, it is possible that the direction of the association changed (e.g., from depression –> anxiety to anxiety –> depression) between measurement occasions. This is something that temporal networks could detect, but contemporaneous networks could not. Thus, in our study, which is based on contemporaneous networks, we cannot make any conclusions about the stability of the directionality of associations. Therefore, this study should be replicated with measurements on different timescales to better understand potential temporal relations. In addition, it might be fruitful to explore alternative methodologies that explicitly account for the absence of temporal relationships.

A final limitation concerns the generalization to other samples and research settings. Replication in other samples that reflect other populations (e.g., the general population, individuals with depression), apply other designs (e.g., six-month diary periods, three assessments per day) and different operationalization of stability (e.g., density, centrality) is necessary to determine the extent to which our findings are specific to our study.

Nevertheless, our study is valuable for its unique design. Our convenience sample was designed as a diary study to predict course and outcome in individuals at different levels of risk for psychosis. In this sample, a change in psychopathology was expected, as, about 75% of participants in our sample received mental health care, which likely alleviates psychopathological severity over time. Therefore, it is a highly relevant sample to investigate the stability of symptom networks specifically in this group of individuals. It does seem likely to expect that, as these individuals received mental health care for some form of mental distress, that these interventions influenced levels of psychopathological severity. Clinically, our sample is quite diverse and although this is representative of those at UHR for psychosis, the heterogeneity in received care is also high. Unfortunately, we did not have detailed information on treatment and therefore could not take this into account in the analyses. In addition, data were collected over two separate diary periods spanning 90 days each, creating the unique opportunity to compare individual network models over time. With 90 days for each diary period, we were able to assess symptoms over a relatively long period, as many diary studies last only 5–7 days. This is an important advantage as more variation could be captured, and in turn, results are less dependent on the specific moment of the measurement period. In addition, with 180 measurements per individual for 66 individuals, we had a total of 11.880 data-points divided over two diary periods, creating the novel opportunity to relate network structures to changes in psychopathology within individuals along the early stages of the psychosis continuum.

5.4.1 *Conclusion*

The quest to unravel the evolution of psychopathology within individuals over time is of critical importance. Early detection plays a pivotal role as this may prevent progression to more severe stages of mental illness. To this end, the application of individual network models presents promising avenues for both clinical research and practice, although this approach is still in its early days. Our study is the first in their effort to explicitly explore the link between changes in individual network models and changes in psychopathology severity. The absence of a clear connection between alterations in individual network structure and psychopathology severity, necessitates a thorough evaluation of the clinical utility of these models. This work serves as an initial step towards a deeper inquiry into the relation between individual network dynamics and psychopathology, a journey that we continue in the Chapters 6 and 7.

6

DISENTANGLING HETEROGENEITY IN SYMPTOM DYNAMICS

ABSTRACT

Many studies have found that depressive complaints are associated with the regulation of affect while facing stress. Individuals inclined towards the experience of negative affect are more vulnerable to developing depressive complaints, while frequent experiences of positive affect buffer the development of such complaints. To better understand the dynamic mechanisms between affect and depression in detail, this chapter investigates how different evaluations of depressive complaints over a prolonged period of stress relate to fluctuations in affect. We included assessments of affect (Positive and Negative Affect Scale) and depressive complaints (Patient Health Questionnaire) in 228 participants who completed at least 20 assessments spanning between $9-14$ weeks. We (i) explored affect trajectories for different evolutions of depressive complaints, (ii) estimated longitudinal multilevel network models to examine the direct interplay between affect and depressive complaints in detail, and (iii) investigated how person-specific network density relates to changes in depressive complaints over time. When separating affect trajectories based on depressive complaints, we identified that individuals consistently experiencing depressive complaints ($\mathrm{PHQ} \geqslant 4$) report higher negative affect levels than positive affect. Contrary, individuals consistently reporting no depressive complaints ($\mathrm{PHQ} \geqslant 4$) showed the opposite pattern. Furthermore, the longitudinal networks included many and strong relations between the affects and depressive complaints variables. Lastly, we found a strong correlation between the density of person-specific networks and their change (aggravation or alleviation) in depressive complaints. We conclude that affect fluctuations and evolution of depressive complaints are directly related both within- and across individuals over time.

Published as: Lunansky, G.*, **Hoekstra, R. H. A.***, & Blanken, T. F. (2023). Disentangling the role of affect in the evolution of depressive complaints using complex dynamical networks. *Collabra: Psychology*, *9*(1), 74841.

*authors contributed equally

6.1 INTRODUCTION

Major Depression (MD) is ranked as the most significant contributor to non-fatal health loss globally [212]. The illness places a significant burden on the suffering individual, negatively impacting their quality of life and daily activities [219]. In addition, MD can have a long-lasting and serious impact on the inflicted person's social environment and society as a whole [114]. Despite decades of research, it remains difficult to understand how depressive complaints arise and develop which seriously hampers the quest for effective treatment interventions [164]. One complicating factor in understanding the etiology of MD is that people can receive the same MD diagnosis while reporting very different combinations of depressive complaints [101].

One potential explanation for this heterogeneity of depressive complaints is that differences in affect regulation in response to stressful situations play an important role in determining whether complaints are developed, and if so, which complaints arise [67, 117, 157, 158]. Typically, affect is divided into positive affect (PA), for example, feeling inspired or enthusiastic, and negative affect (NA), such as feeling afraid or upset [259]. Experiencing negative affect in the face of stress helps activate necessary behaviors, such as running away from a threat when feeling afraid [176]. However, potentially traumatic events or prolonged stress, such as childhood abuse, may deregulate affect responses to novel stressors in a long-lasting way [131]. Individuals who have faced severe adversity may show overly aroused nervous system responses (e.g., fight, flight, or freeze responses), leading to perpetual perceptions of threat. Because negative feelings such as distress and sadness have appeared in severe ways or over longer periods of time in a stressful context, these feelings could persevere in other stressful contexts as well, even if these novel situations cause no real harm [112]. As such, individuals with deregulated affect responses, inclined towards the experience of negative affect, may be more vulnerable to develop mental health complaints [67].

In turn, frequent experiences of positive affect may buffer the development of such mental health complaints [249]. It has been suggested that positive affect, such as laughter, can dampen stress responses in cardiovascular, metabolic, and inflammatory systems [238, 280]. Various hypotheses exist on how exactly positive affect buffers the negative effects from stressful experiences (van Steenbergen, Sauter, et al., 2021). For example, through neurological systems that influence our 'wanting' and 'liking' [207], by balancing the hypothalamic-pituitary-adrenal (HPA) axis [250], or through directing behavior towards situations that offer rewards and pleasure [223]. Thus, frequent PA is generally associated with a reduced

risk of developing depressive complaints [168, 261], while persistent NA is related to an increased risk of developing depressive complaints [264].

While it is clear that both PA and NA influence the evolution of depressive complaints, PA and NA are not independent of each other [7]. It is possible that the concerted influence of PA and NA on depressive complaints (i.e., when both PA and NA are taken into account simultaneously) is different than when only considering either the effect of PA or NA. Therefore, it is also of interest to understand how the interplay between PA and NA is associated with the development of depressive complaints. Research suggests that zooming in on day-to-day experiences in a detailed level, such as smaller components of positive affect (feeling 'happy' or 'excited') and negative affect (feeling 'afraid' or 'upset'), can generate insights that are easily overlooked in more macro-level measurements of affect and psychopathology [40, 262]. Specifically, looking into the interrelations of smaller components of PA, NA, and depressive complaints as one integrated system can contribute to our mechanistic knowledge of the development of depressive complaints.

The possible dynamic mechanisms underlying the relation between positive and negative affect and the development of depressive complaints can be investigated using network analysis [146]. Network models estimate the conditional associations between variables on a detailed level [40, 262]. This means that all smaller components of PA, NA and distinct depressive complaints are both dependent and independent variables simultaneously [235]. With longitudinal networks that are estimated from data including assessments of PA, NA, and depressive complaints over a longer stressful period from multiple participants, we can look into various directed pathways of how fluctuations in negative and positive affect are related to evolutions of specific mental health complaints. In this way, we aim to deconstruct the relationship between affect and depressive complaints in the face of stressful times and take a step further in understanding the heterogeneity of depression.

Network models consist of nodes, which represent variables (in the current study: affect and depressive complaints), and edges, which represent the direct conditional associations between the nodes [23, 53, 84]. Edges can be weak or strong, representing the magnitude of the relations between the variables. The strength of the associations between the detailed elements in the network is summarized as the density of the network. It has been suggested that this density indicates the system's vulnerability to evolve towards an unhealthy situation [23]. A simulation study with symptom networks found that the density of the network is related to the severity of the depression [53]. As such, the density of a network with PA, NA and depressive complaints could suggest potential causal mech-

anisms between affect and depressive complaints in which more strongly connected networks signal vulnerability [23].

To capture these possible dynamic mechanisms, we need to investigate the relationship between affect and depressive complaints over a longer period. Longitudinal studies of affect mostly collect data within an intensive but brief timespan [e.g., five times a day for two weeks; 239], capturing relations between momentary affect and current depressive complaints. A longer timespan needs to be taken into account to ensure that depressive complaints could increase or decrease, while taking frequent assessments to capture affect fluctuations. Additionally, we are interested in investigating whether differences in affect regulation in times of prolonged stress may be related to the evolution of depressive complaints. In this study we aim to investigate how affect fluctuations and evolutions of depressive complaints are associated within and across people over longer periods of time ($9-14$ weeks, 12 ± 1 [mean$\pm$SD]), during a prolonged period of stress as induced by the COVID-19 pandemic.

The current chapter uses data that are part of a longitudinal investigation from the Boston College, in which the repercussions of the COVID-19 pandemic on mental health were investigated [62]. The studied period (March 20th 2020 until June 26th 2020) commenced a day after the first"stay-at-home" order was issued in California, which covers the moments leading up to the first large COVID-19 wave in the US. Clearly, this was a period of great uncertainty, and many effects on mental health problems during this time have been reported across all levels of society [116, 162, 215, 180]. Interestingly, previous work indicated that while some people experienced an increase in mental health complaints, others reported that their mental health improved [174]. Improvements in mental health may be related to an increase in PA, which is why it is likely that the studied period of this study captures significant fluctuations of both PA and NA. The reported period thus yields an unprecedented opportunity to investigate how fluctuations in positive and negative affect are related to the evolution of depressive complaints during a prolonged period of stress.

To elucidate the working mechanisms behind the interplay between affect and depressive complaints we first aim to explore whether the evolution of depressive complaints in the beginning of the COVID-19 pandemic (i.e., worsening or alleviation of complaints) is accompanied by general trends in affect fluctuations over that same period. Since we have scores on depressive complaints throughout the studied period, we divide the sample into sub-groups of people who experienced a meaningful change during the studied period, either an aggravation or alleviation of complaints, and sub-groups whose scores can consistently be classified as no complaints, mild complaints, or severe complaints. We then visually

inspect the trajectories of positive and negative affect during that same period. We expect that people who experienced an increase in depressive complaints would show high levels of NA, whereas people who reported a decline in complaints would show high levels of PA. Following our hypotheses that the evolution of depressive complaints may be linked to differences in affect regulation, we then explore the affect trajectories for people who experience a meaningful change in their depressive complaints over the studied period.

Second, to investigate the direct interplay between depressive complaints and affect over the studied period at a more detailed level, we model their relations more directly using longitudinal multilevel network models [28, 81, 82]. These network models allow us to estimate the direct relations between affect and depressive complaints across individuals (i.e., between-subjects and fixed-effects network) and for every individual (i.e., random-effects) in a multilevel framework. For the between-subjects and fixed-effect networks, we expect to find positive associations within the PA and NA elements, but negative associations between these elements. Additionally, we expect to find positive associations between the NA elements and depressive complaints, and negative associations between the PA elements and depressive complaints. For the person-specific networks (i.e., random-effect networks) we aim to investigate whether the structure of these person-specific networks is related to a clinically meaningful change (increase or decrease) of depressive complaints. We expect person-specific network structures to be associated with a clinically meaningful change in depressive complaints, such that higher person-specific densities correspond to a larger increase in depressive complaints.[1]

6.2 METHODS

6.2.1 *Participants*

The data were obtained through the Boston College daily sleep and well-being survey [62]. The study was set-up during the first wave of COVID-19 (March 20th 2020 until August 5th 2020) and participants were recruited online. All English-speaking individuals older than 18 were eligible to participate in the study, resulting in N=1,518 enrolled participants (mean $\pm$ SD age 35.2 ± 15.1 years old, range 18-90 years old). The participants provided informed consent, and the study received ethical approval from the Institutional Review Board at Boston College. More details on the study and recruitment can be found in Cunningham, Fields, and Kensinger [62].

1 In Chapter 7, we critically reflect upon this initial expectation.

6.2.2 *Procedure*

The study started with a demographic survey, and upon completion, participants received daily surveys on their sleep and well-being. The daily surveys were divided into a short and a full version, where the full version included additional questions containing validated assessments of mood [Positive and Negative Affect Schedule (PANAS); 259] and depressive complaints [Patient Health Questionnaire-9 (PHQ-9); 178]. The full version was sent on the first three days of the study enrollment. After enrollment, the long survey was sent on two randomly selected days of the week, and the short version was sent on the five remaining days. For a more detailed description of the assessments, we refer to [62].

6.2.3 *Materials*

For the current study, we include the demographics survey, PANAS items, and PHQ-9 assessments from the full version of the questionnaire.

PANAS The PANAS is a 20-item questionnaire on the experience of positive (e.g., enthusiastic) and negative (e.g., scared) affect rated on a five-point Likert scale ranging from 1 ("very slightly/ not at all") to 5 ("extremely") [259]. To reduce the number of variables for more power within the conducted statistical analyses (i.e., the estimated network models described in the "Statistical Analyses" section of the current chapter), we selected the ten items of the PANAS that have been validated in the short-form [193]. The reliability of both PA and NA short-form scales is high, with a Cronbach alpha of .78 for the PA scale and .87 for the NA scale. For the assessment of PA, items include: inspired, alert, excited, enthusiastic, and determined, and for NA, items include: afraid, upset, nervous, scared and distressed. In the survey, participants were explicitly asked to rate how they felt in the current moment for each of the attributes ("For each of the following attributes, indicate which description best describes how you currently feel, right now in the moment").

PHQ-9 The PHQ-9 is a 9-item questionnaire to measure depression severity by assessing each of the 9 DSM-IV criteria for depression on a four-point Likert scale ranging from 0 ("not at all") to 3 ("nearly every day"). The internal reliability of the PHQ-9 is high, with a Cronbach's alpha ranging between $.86 - .89$ [178]. In the current implementation, the item on suicidal thoughts was omitted. Participants were asked to rate the severity of complaints over the last several days ("In the last several days, how often have you been bothered by any of the following prob-

lems: not at all, some of the time, more than half of the time, almost all of the time").

6.2.4 *Data selection and pre-processing*

STUDY PERIOD We selected the study period from March 20th 2020 until June 26th 2020, ending three days after the final full survey was sent out. Of the total number of participants enrolled in this study, $n = 1355$ (89.3%) completed at least one assessment during the selected study period.

PRE-PROCESSING OF ASSESSMENTS The full version of the questionnaire was sent out twice a week on random selected days. Sometimes participants completed the full survey multiple times on a single day: on 79 occasions the survey was completed twice, and on two occasions the survey was completed three times. For these 81 assessments we chose the survey that was completed first. Inspecting the response rate of the full survey over the study period, a clear three-day interval pattern can be seen (see Appendix E, Figure 49). Therefore, we chose to group the days into 'measurement occasions', defined by a three-day window (e.g., March 20th-22nd 2020 is measurement occasion 1). In this way, the entire study period is grouped into 33 measurement occasions of three days each. The advantage of this grouping is twofold. First, it circumvents the problem of large differences in the number of completed surveys per assessment. Second, given that the full surveys were sent out randomly twice a week, dividing the assessments into three-day intervals makes the time between two completed surveys more equidistant. In case participants completed multiple surveys within one measurement occasion, we averaged their responses. Following the recommendations for estimating a multilevel VAR model [159], we selected participants who completed surveys for at least 20 measurement occasions, resulting in a final sample size of $n = 228$ participants (16.8%). Thus, for each included participant, we have completed data for a minimum of 20 and a maximum of 33 measurement occasions that span $9 - 14$ weeks.

6.2.5 *Statistical analyses*

VISUAL INSPECTION OF TRAJECTORIES To explore whether the evolution of depressive complaints (i.e., aggravation or alleviation of complaints) is accompanied by general trends in affect fluctuations over the same period we defined different groups of depressive evolution. First, we defined two groups of participants that experienced a clinically meaningful change in their depressive complaints, defined by a 5-point difference

in their PHQ-9 total score [189, 234], and indicative of an aggravation or alleviation of their complaints. Second, we defined groups based on the overall level of experienced complaints throughout the studied time period: participants who are consistently without depressive complaints, defined as a PHQ-9 score of four or lower on all assessments; participants who experience occasional depressive complaints, defined as a PHQ-9 score higher than four on at least one assessment (and lower than four on at least one other assessment); and participants who experience consistent depressive complaints, defined as a PHQ-9 score consistently higher than four on all assessments [178].

To visually explore whether general trends can be observed, we plotted the smoothed means of each affect over time using locally estimated scatterplot smoothing (i.e., a loess curve). We did this for all participants together as well as for each of the defined groups to explore whether trends in affect trajectories accompany differences in evolution of depressive complaints. We used loess regression to facilitate the exploration of general trends. Loess regression is a non-parametric method that fits least squares regressions in localized subsets of the data [49]. The amount of smoothing that is applied depends on the number of data points that are used in each local regression (i.e., the neighborhood) and is controlled by setting the smoothing parameter α between 0 and 1. The larger the values for α, the more data points are being selected in the neighborhood (i.e., $n\nu \times \alpha$ data points are selected, where n represents the total number of datapoints). More datapoints in the local regression results in smoother functions, that are more robust to fluctuations in the data. We set the smoothing parameter to $\alpha = 0.2$ in order to aid the visualization of patterns in the data, without losing sensitivity of fluctuations in the data.

LONGITUDINAL NETWORK MODEL To directly investigate the interplay between depressive complaints and affect we estimated a multi-level network model including both affect and depressive complaints. To estimate the relations while taking the longitudinal structure of the data into account, we estimated a two-step multi-level GVAR model as implemented in the mlVAR package [83]. A multi-level GVAR has two major benefits: (1) a single model can be estimated, leading to an adequately powered analysis, and (2) the within-person effects can be separated from between-person effects [81].

The edges in the multi-level GVAR network are computed from partial correlations, meaning they portray the unique association among two variables after controlling for all other variables in the network. Edges can be positive or negative, indicating the corresponding nature of the associations between the nodes [84]. The mlVAR package estimates three types of network structures: (a) a temporal network (both a fixed-effect

structure over all persons and a random-effect structure per person), (b) a contemporaneous network (both a fixed-effect structure over all persons and a random-effect structure per person), and (c) a between-persons network[81]. In line with Epskamp, Borsboom, and Fried [82], for the remainder of this chapter, the term within-person will refer to the fixed-effect within-person network (either temporal or contemporaneous) and the term person-specific will refer to random-effects within-person networks (either temporal or contemporaneous).

The temporal network indicates how well a variable predicts another variable at the next time point while controlling for all variables at the current time point. For example, a direct association from the depressive complaint 'trouble concentrating' to 'feeling distressed' indicates that having 'trouble concentrating' now predicts 'feeling distressed' at the next time point, taking into account all other current affects and depressive complaints. In the contemporaneous network we controlled for all these temporal effects, which shows the unique association among variables within the same time window. For example, a direct and positive edge between 'feeling distressed' and 'feeling afraid' indicates that these two negative affects are positively associated after removing the lagged effects. Finally, the between-persons network shows the relationships among the means of persons in the data. For example, a positive edge between 'trouble concentrating' and 'feeling distressed' would indicate that persons who have, on average, more trouble concentrating, also, on average, feel more distressed.

NETWORK DENSITY Based on the theorized role of affect dynamics in the course of depressive complaints, we expected the network structure to differ between people who experienced a meaningful change in their depressive complaints (either aggravation or alleviation) and people who did not experience such a meaningful change. Specifically, we expected that more densely connected networks result in more depressive complaints over time. To explore this hypothesis we computed, for each participant, the average absolute strength of their temporal associations [i.e., their density; 211]. Subsequently, we correlated their density to the (absolute) maximum change in their PHQ-9 score over the study period: the difference between an individual's maximum PHQ-9 score, and their minimum PHQ-9 score over the course of 4 months. To compute the maximum difference, we took the order in which the scores occur into account, to capture whether an aggravation or alleviation was experienced. For example, two individuals A (PHQ-9 scores: $3, 7, 6, 8, 6$) and B ($7, 8, 4, 4, 3$) who both have a minimum PHQ-9 score of 3 and a maximum PHQ-9 score of 8, differ in their PHQ-9 difference score reflecting their aggravation (A: maximum difference $3 - 8 = -5$) or alleviation (B:

maximum difference $8 - 3 = 5$). The PHQ-9 difference score indicates the maximum change in depressive complaints this individual reported over the study period. In addition, we inspected the relation between network density and change in PHQ-9 regardless of an aggravation or alleviation in depressive complaints by correlating person-specific network density to an individual's absolute maximum change in PHQ-9.

SENSITIVITY ANALYSES To examine the extent to which the observed correlation between person-specific temporal network density and their change in PHQ-9 score could be dependent on methodological choices we conducted two sensitivity checks. First, we determined whether potential trends in the data could have affected our observed correlation. We checked for linear trends in the variables and detrended any variables with significant (linear) trends. Then we followed the same procedures as described above and computed the correlation. Second, since the network includes the PHQ-9 items we examined if the correlation between network density and maximum change in PHQ-9 score might be driven by this overlap. Given this overlap, the temporal density is in part based on the same information (i.e., the PHQ-9 items) as the change score in PHQ-9 items. To that end, we re-estimated a mlVAR network including only the affect states as nodes and correlated the person-specific temporal network density of the affect states with participants' change in PHQ-9 scores.

STABILITY ANALYSES To examine the stability of the correlation between person-specific temporal network density and change in PHQ-9 score we followed a similar procedure as described in [156]. We estimated 100 mlVAR networks that included a random selection of 80% of the original data (i.e., using the data of $\sim$ 182 participants per estimation). For each of these re-estimated networks we followed the same procedure as described above, (i.e., we calculated person-specific network density and correlated the network density to change in PHQ-9 score).

All analyses were performed in R (version 4.0.5) [220] using the packages 'ggplot2' (version 3.4.0) [267], 'mlVAR' (version 0.5) [83], 'qgraph' (version 1.9.3) [75], 'dplyr' (version 1.0.10) [268], 'nnet' (version 7.3-18) [225], and 'ggpubr' (version 0.5.0) [161]. The derived data used for analyses and corresponding code can be found on the Open Science Framework (OSF): `https://osf.io/2zh4f/`.

6.3 RESULTS

6.3.1 *Sample characterization*

We included 228 participants that completed assessments on at least 20 measurement occasions within March 20th 2020 and June 26th 2020. Over the course of four months, on average, participants completed 23 ± 3 ($mean \pm SD$) assessments. On average, 9 ± 3 ($mean \pm SD$) assessments were missing, thus, on average, 28% waves of data are missing per individual. As responses for all variables of interest where forced, there is no within wave missingness. The majority of the participants were female ($n = 186$, 81.6%) and, on average, 45 ± 19 ($mean \pm SD$) years old. For more details on sample characterization, see Table 7. Table 8 shows the means, standard deviations, and range for each variable used in this study. Our sample ($n = 228$) differed from the original full sample ($n = 1355$) in their mean PHQ-9 score at baseline; the original full sample had a higher PHQ-9 mean at baseline. Our sample did not differ in their PANAS scores at baseline from the original full sample ($n = 1355$), see Appendix E, Table 11 for more details on sample characterization, and Table 12 for means standard deviations, and range of the PHQ-9 and PANAS for the original full sample.

6.3.1.1 *Visual inspection of trajectories*

Figure 12 (a) shows the affect trajectories across the 33 measurement occasions for all participants concertedly. Averaging over all participants shows that PA is generally rated somewhat higher than NA, which is consistent over the entire study period.

We inspected the affect trajectories for participants based on (i) their change in depressive complaints over time, and (ii) the severity of the experienced depressive complaints. Over the studied time period, 116 participants experienced a meaningful change in their depressive complaints: 56/116 participants (48.3%) experienced an aggravation in their depressive complaints (i.e., an increase in PHQ-9 score of at least 5), and for 60/116 participants (51.7%) their depressive complaints alleviated (i.e., a decrease in PHQ-9 score of at least 5). Splitting the affect trajectories for each of these individuals, shown in Figure 12 (b), indicates that both for participants who experienced an aggravation and an alleviation of their depressive complaints over time, on average, their affects are more intertwined. Interestingly, no marked differences are seen in participants who experienced an aggravation of their complaints compared with participants who experienced an alleviation of their complaints.

Inspecting the consistency of participants depressive complaints, we found that 50/228 participants (21.9%) were consistently without depres-

Characteristics	Mean	SD
Age	44.8	19.2
	n	**percentage**
Gender		
Female	186	81.6%
Male	42	18.4%
Race/ethnicity		
African-American	1	0.4%
Asian	21	9.2%
White	207	90.8%
Hispanic/Latinx	5	2.2%
More than one race	4	1.8%
Cultural Background		
North-America	206	90.4%
South-America	2	0.9%
Africa	1	0.4%
Asia	4	1.8%
Europe	10	4.4%
Oceania	5	2.2%
Annual household income		
$0-$25,000	21	9.2%
$25,001-$50,000	36	15.8%
$50,001-$75,000	42	18.4%
$75,001-$100,000	40	17.5%
$100,001-$150,000	32	14%
$150,001-$250,000	31	13.6%
$250,000+	26	11.4%
Education		
High School Diploma	5	2.2%
Some college	19	8.3%
College degree	56	24.6%
Some post-bacc education	29	12.7%
Graduate, medical or professional degree	119	52.2%

Table 7: Sample characteristics ($n = 228$).

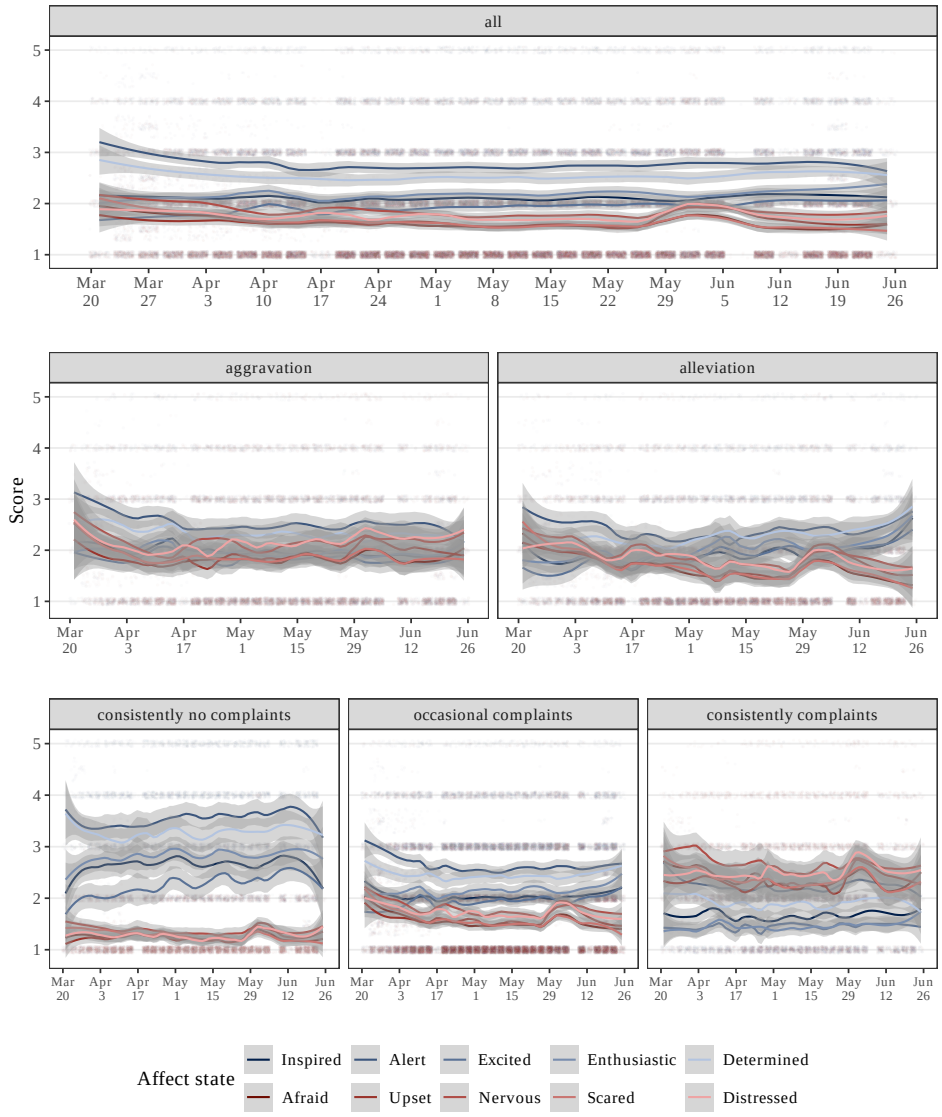


Figure 12: Affect trajectories from March 20th 2020 (measurement occasion 1) until June 26th 2020 (measurement occasion 33). Panel (a) shows the smoothed conditional mean trajectories of all participants for each of the affect states, together with its 95% confidence interval (shaded area). Panel (b) on the left shows the trajectories for participants whose depressive complaints aggravated during the study period, and on right shows the trajectories for participants whose depressive complaints alleviated during the study period. The (c) panels show the trajectories for participants who consistently experienced no depressive complaints (left), those that occasionally experienced at least mild depressive complaints (middle), and those that consistently experienced depressive complaints (right). Blue lines correspond to the smoothed conditional means of the positive affect states 'inspired', 'alert', 'excited', 'enthusiastic', and 'determined'; and red lines correspond to the smoothed conditional means of the negative affect states 'afraid', 'upset', 'nervous', 'scared', and 'distressed'. Affect states are scored on a Likert scale from 1 to 5. Decimal scores were obtained when participants completed multiple assessments within one measurement occasion.

Variable	Mean	SD
PHQ9	5.90	4.14
PANAS		
Inspired	2.07	1.03
Alert	2.95	1.07
Excited	1.8	0.88
Enthusiastic	2.04	1.03
Determined	2.61	1.15
Afraid	1.84	1
Upset	1.67	0.86
Nervous	2.12	1.02
Scared	1.97	0.97
Distressed	1.92	0.89

Table 8: Mean and standard deviation per variable ($n = 228$).

sive complaints, 137/228 participants (60.1%) experienced depressive complaints at least occasionally, and 41/228 participants (18.0%) experienced depressive complaints consistently. Figure 12(c) shows the affect trajectories for each of these participants. Here, we see clear differences in affect trajectories across participants: in people without depressive complaints there is, on average, a clear distinction between the PA scores, which are rated relatively high, and the NA scores, which are all rated consistently low. In the people with occasional depressive complaints, the PA scores are, on average, still rated higher than the NA scores, but the distinction is less clear. Finally, in people who consistently experience depressive complaints, the ratings of positive and NA have flipped, as NA is, on average, rated higher than PA.

6.3.1.2 *Longitudinal network model*

We investigated the dynamical relations among affect and depressive complaints by estimating a multilevel network model. Figure 13 shows the three estimated network structures: (a) the average temporal associations for all participants; (b) the average contemporaneous associations for all participants; and (c) the between-persons network structure. For the interpretation of the relations within each network structure, it is important to note that questions regarding affect and depressive complaints were prompted differently. Affect was prompted: "Indicate which description best describes how you currently feel, right now in the moment" while questions regarding depressive complaints were prompted: "In the last several days, how often have you been bothered by any of the following problems". This difference in phrasing, as well as the temporal scaling of our study (roughly 3 days between assessments) influences the interpre-

tation of the estimated networks. The relations in the average contemporaneous network reflect how depressive complaints over the days prior to the assessment and affect at the moment of assessment co-occur (e.g., how depressive complaints over a period of 3 days, let's say Monday through Wednesday, co-occurs with affect on Wednesday). Relations in the average temporal network reflect how depressive complaints over the days prior to the assessment predicts affect at the next assessment occasion (e.g., how depressive complaints over a period of 3 days, for example, Monday through Wednesday, predicts affect on Friday) and vice versa. The relations in the between-person network reflect how, on average, depressive complaints relate to affect.

The average temporal network shows strong autoregressive effects for depressive complaints, indicating depressive complaints have a (relatively) strong reinforcing tendency, see the top panel of Figure 13. We found positive relations within the three domains (e.g., the positive relations between the NA variables "afraid" and "nervous", or the positive relations between the PA variables "determined" and "enthusiastic"), while negative and positive associations can be found between the three domains (e.g., a negative relation from the PA variable "inspired" to the depressive complaint "loss of interest"). Interestingly, the temporal network shows many relations between affect states and depressive complaints (e.g., the negative relation from "afraid" to "depressed mood" or the negative relation from "excited" to "depressed mood"), whereas only one relation between PA and NA can be found (i.e., from "determined" to "afraid").

Compared to the temporal network, the average contemporaneous network shows clearer demarcations between PA, NA, and depressive complaints: edges are comparatively stronger within than between the three domains. This can be seen in the middle panel of Figure 13, and is confirmed by inspecting the edge weights of the contemporaneous network. For example, strong positive associations can be found between NA items such as "distressed" and "upset", and "afraid" and "scared", as well as between PA items such as "enthusiastic" and "excited", and between depressive complaints such as "loss of interest" and "depressed mood". Weak associations can be found between the three domains (e.g., the negative association between "alert" and "depressed mood").

The between-persons network portrays stronger relations between the different domains compared to the contemporaneous network, indicating that the average affect people experience is related to their average depressive complaints. As expected, we found positive associations within PA, NA, and depressive complaints, see the bottom panel Figure 13. In addition, we found positive associations between the NA items and depressive complaints (e.g., the association between "afraid" and "feelings of worthlessness"), and negative associations between the PA items and depressive

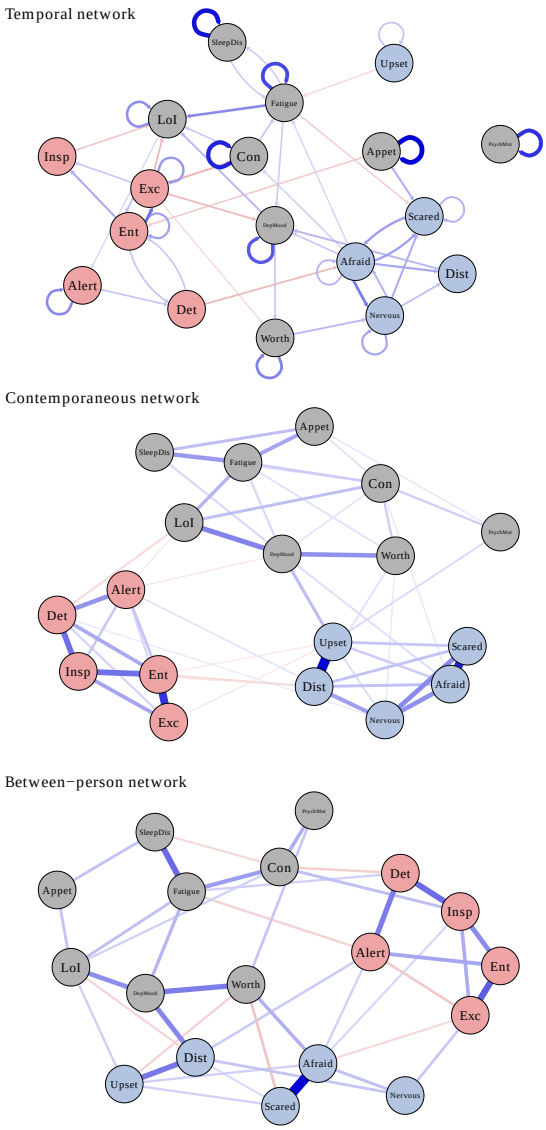


Figure 13: Output from mlVAR for $n = 228$ and $t \geqslant 20$. In the top panel, the temporal network model is displayed, portraying the average within-person relations from one measurement occasion onto the next. The center displays the contemporaneous network model, portraying the average within-person effects in the same measurement occasion, after controlling for the temporal effects. At the bottom, the between-persons network model is displayed, indicating the average effects between persons. Blue edges indicate positive relations, whereas red edges indicate negative relations. Node colors correspond to PA (light blue), NA (salmon pink), and depressive complaints (grey). Abbreviations: Insp = inspired; Alt = alert; Exc = excited; Ent = enthusiastic; Det = determined; Afr = afraid; Ups = upset; Ner = nervous; Scar = scared; Dist = distressed; LoI = loss of interest; DepMood = depressed mood; SleepDis = sleep disturbances; Fatigue = fatigue; Appet = loss of appetite; Worth = feelings of worthlessness; Con = concentration problems; PsychMot = psychomotor agitation or retardation.

complaints (e.g., the association between "alert" and "fatigue"). However, contradictory to our expectations, we also found negative associations between the NA items and depressive complaints (e.g., the association between "distressed" and "loss of interest") and positive associations between PA and depressive complaints (e.g., the association between "alert" and "fatigue").

6.3.1.3 *Network density*

To further investigate the relations between affects and depressive complaints, we correlated person-specific network density of the temporal network to their (absolute) maximum change in PHQ-9 score. As shown from the correlation plot in Figure 14 (right panel), there is a strong correlation ($r = 0.77$) between person-specific network density and absolute change in PHQ-9 score. Interestingly, and contrary to our initial expectations, stronger network densities relate to both a more substantial aggravation and to a more substantial alleviation in depressive complaints, as can be seen in Figure 14 (left panel) illustrating the correlation between person-specific network density and maximum change in PHQ-9 score. Network density is thus related to clinically meaningful change in both directions, a finding that we elaborate on in Chapter 7.

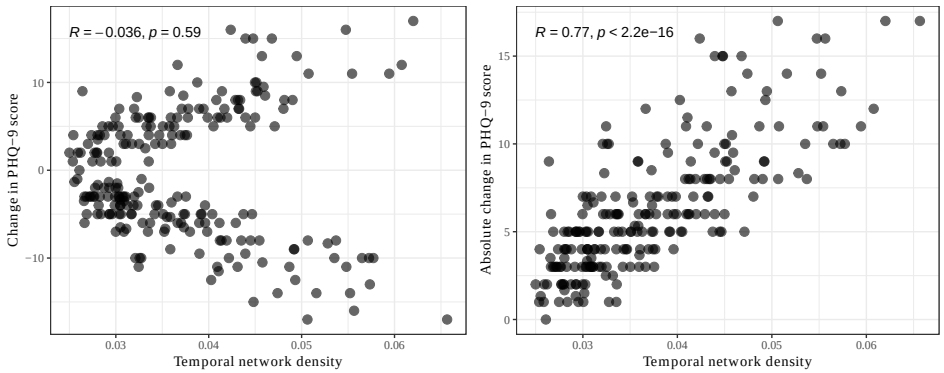


Figure 14: In the left panel the correlation between person-specific network density and maximum change in PHQ-9 score is shown. In the right panel the correlation between person-specific network density and the absolute maximum change in PHQ-9 score is shown.

6.3.1.4 *Sensitivity analyses*

Sensitivity checks showed the strong correlation between absolute change in PHQ-9 and person-specific temporal network density was still present when data was detrended ($r = 0.79$). When only affect states were included in the network structure, the strength of the observed correlation

decreased ($r = 0.4$), however we still observed a positive association between person-specific network density and absolute maximum change in PHQ-9 score.

In addition, the relation between network density to both a more substantial aggravation and a more substantial alleviation in depressive complaints was still present when data was detrended, or when only affect states were included in the network structure. See correlation of $r = -.04$, ranging from $-0.15 - 0.07$. See Appendix E for more details on the performed sensitivity checks.

6.3.1.5 *Stability analyses*

Stability checks indicated the originally found correlation between person-specific network density and maximum change in PHQ-9 score was stable. The strong correlation between absolute maximum change in PHQ-9 and persons-specific temporal network density was still present when recursively re-estimating networks using 80% of the data at a time. The median correlation between person-specific network density and their maximum change in PHQ-9 score using 80% of the data was $r = 0.77$ and ranged from $0.72 - 0.81$. In addition, we found stable results for the relation between person-specific network density and the aggravation and alleviation of depressive complaints; when looking at the correlation between maximum change in PHQ-9 and person-specific network density, we found a median correlation of $r = -.04$, ranging from $-0.15 - 0.07$. (See Figure 53 in Appendix E).

6.4 DISCUSSION

Depressive complaints have often found to be associated with the regulation of affect while facing stress. To better understand the mechanisms between affect and depression, in the current chapter, we set-out to investigate the dynamic interplay between affect and depressive complaints during a prolonged and eventful time imposed by the start of the COVID-19 pandemic. First, we explored whether the evolution of depressive complaints is accompanied by general trends in affect fluctuations using a longitudinal dataset from the early phase of the COVID-19 pandemic. We divided the sample into groups of people who experienced a clinically meaningful change during the studied period (i.e., a substantive aggravation or alleviation of complaints), and groups whose scores can consistently be classified as no complaints, mild complaints, or severe complaints. We then visually inspected affect fluctuations of these groups. As expected, we found differences in the affect trajectories among people who experienced consistently no depressive complaints (showing higher

PA than NA) compared with consistent depressive complaints (showing the reversed pattern). Crucially, these differences pertained to both PA and NA trajectories, showing that there is a clear link between consistent severity levels of depressive complaints and both positive and negative affect. Contrary to our expectations, affect trajectories were similar for people experiencing either an aggravation or alleviation of depressive complaints, meaning that we reject our hypothesis that people who experienced an increase in depressive complaints would show high levels of NA, whereas people who reported a decline in complaints would show high levels of PA.

Second, we investigated the link between affect and depressive complaints in more detail by estimating a multilevel longitudinal network model. The network showed many and strong relations between affect and depressive complaints, indicating that these elements are indeed directly linked, both at the average intra-individual level, as indicated by the contemporaneous and temporal networks (i.e., fixed effects), and at the inter-individual level as indicated by the between-person network. At the inter-individual level, we found mostly positive associations within the PA and NA elements, and within the depressive complaints, as expected. We hypothesized to find positive associations between the NA elements and depressive complaints and negative associations between the PA elements and depressive complaints in the between-subjects network. This was true for some (e.g., positive association between the NA 'feeling distressed' and the depressive complaint 'depressed mood'), but not all associations in the network (e.g., positive association between the PA 'feeling inspired' and the depressive complaint 'concentration problems'). The average temporal and contemporaneous networks did show the expected associations, with negative (positive) associations between PA (NA) and depressive complaints.

Furthermore, we investigated the relation between the strength of the associations in the temporal person-specific networks. As expected, we found a strong relation between the density of the person-specific networks and the individuals' absolute change in PHQ-9 score. Interestingly, similar network densities are related to both a worsening and improvement in depressive complaints. This indicates that, contrary to our initial expectations, strong associations between affect and depressive complaints are related to change in any direction: both to an improvement of complaints as well as to a worsening of complaints. The second part of our hypothesis should, therefore, be rejected, as the expectation was that a higher density would specifically be related to an aggravation in PHQ-9.

The strong relation between network density and change in depressive complaints may have important implications for the clinical interpretation

of networks, as network density has generally been related to more severe psychopathology [e.g., 45, 19]. However, our study shows an alternative situation in which a larger density of networks indicates more fluctuations and potential for flexibility [138]. One possible explanation for our finding is that phase transitions in a wide variety of systems (e.g., transitioning from mild depressive complaints to severe complaints) are often characterized by a period of instability, in which the behavior of a system shows many fluctuations [209, 263]. This increase in fluctuations before a clinically meaningful alleviation or aggravation could be reflected in the increased network density. This explanation possibly corroborates findings in other studies that found larger densities in networks to be related with a decrease in psychopathology symptoms over time [197].

While the relation between network density and change in depressive complaints may signal relevant clinical importance, it is important to note that this strong relation could also merely reflect a well-known property of test reliability, namely that the variance of a total score (in our case the change in depressive complaints) consists of the sum over the variance in all items (in our case the individual affects and individual PHQ-9 items) and the sum over their covariances [57]. Since denser networks reflect stronger covariances, it is a statistical necessity that denser networks are accompanied by larger variations in the total score (i.e., the variation in PHQ-9 score). However, our finding does not merely reflect a methodological artefact, as our sensitivity analysis indicate that only including the affect states in the network structure (thus, removing the PHQ-9 items from the network) still rendered a positive association between network density and absolute change in PHQ-9 score (see Appendix E for more details on the performed sensitivity checks). Additionally, it is important to note that one should be cautious to compare network densities of different types of network models. Thus, our finding here cannot simply be generalized to different types of network models [e.g., non-linear models such as Ising network models; 53]. In the next chapter, we compare the dynamics of different types of idiographic network models and relate their dynamics to psychopathology).

Some limitations in the current study warrant attention. First, the questionnaires were sent out twice a week at random intervals, thereby violating the assumption of equidistant measures for longitudinal analyses. We partly addressed this problem by defining measurement occasions as three-day periods. Second, due to the high amount of missing data in the original full sample, we selected participants that answered enough measurement occasions. The selected sample of participants who completed at least 20 measurement occasions had a lower PHQ-9 mean at baseline than the original full sample. Therefore, it can be that we may have missed more severe depression cases. At the same time, our selected sample still

included 129 (56.6%) participants who scored above the cut-off of severe depressive complaints at baseline and 41 participants (18.0%) who consistently scored above the cut-off for severe depressive complaints.

Third, affect and depressive complaints may operate on different time-scales. To the best of our knowledge, there are no current techniques available that can account for these possible time-scale differences in processes in the field of network analysis [38]. However, the time-scale differences in affect and depressive complaints were reflected in the phrasing of the questionnaire. Questions regarding affect were prompted: "Indicate which description best describes how you currently feel, right now in the moment" while questions regarding depressive complaints were prompted: "In the last several days, how often have you been bothered by any of the following problems". This difference in phrasing captures the idea that affect operates on a faster time-scale (e.g., fluctuates within a day), while depression operates on a slower time-scale (e.g., fluctuates from day-to-day).

Fourth, it should be noted that while we are interested in mechanisms of change, the current available network estimation techniques assume that the mean and variance of the time series data remains the same (i.e., stationarity) [159]. Unfortunately, alternative time-varying network models require many more datapoints than present in the current dataset [e.g., see 134]. Therefore, there is a mismatch between our data, our interest in change, and the available statistical models. In addition, although multilevel VAR estimation allows for the estimation of person-specific networks, these networks are not purely idiographic. With multilevel estimation, a shared underlying distribution for all model parameters is assumed. This means person-specific edge weights are shrunken to follow the same underlying group distribution. As a result, individual differences are smoothed out in the estimation process. Therefore, future research should indicate whether the result found here hold on a purely idiographic level.

To conclude, we found that affect fluctuations and the evolution of depressive complaints are strongly related during stressful periods. We showed this relation in three ways. First, we visually explored the trajectories of PA, NA, and depressive complaints at the beginning of the COVID-19 pandemic. Individuals who consistently reported no or mild depressive complaints showed higher levels of PA, while people with consistently severe depressive complaints showed higher levels of NA. Second, we zoomed in on these relations by estimating longitudinal networks, both at the individual and group level. These network models showed many direct relations between PA, NA, and depressive complaints. Third, we found that individuals with an alleviation or aggravation of depressive complaints have more densely connected networks. This means that

the stronger affect and depressive complaints are connected over time, as indicated by the network model, the larger the change is in depressive complaints. Together, these findings shed light on the potential underlying mechanisms of change and the development of mental disorders.

7

MAPPING MODEL DYNAMICS TO PSYCHOPATHOLOGY

ABSTRACT

The network theory of psychopathology posits that mental disorders are stable states of symptom activation that arise from networks of causally interrelated symptoms. This implies that, ceteris paribus, a person with a more strongly connected symptom network (i.e., a network characterized by more and stronger causal relationships), will have an increased risk of developing a mental disorder. However, empirical studies relating idiographic network connectivity to mental disorders show mixed results, leaving unclear how the statistical model relates to the theoretical framework. In this paper we map the dynamics of idiographic network models to the network theory of psychopathology. In particular, we use stability landscapes to map the dynamics implied by the Ising model with a $\{0, 1\}$ and $\{-1, 1\}$ encoding, and the graphical vector autoregressive (GVAR) model onto symptom severity and the variability in symptom activation. Only the Ising model with the $\{0, 1\}$ encoding shows behavior consistent with the network theory. The Ising model with the $\{-1, 1\}$ encoding partially aligns with the network theory. Contrary to popular belief, we show that, in the GVAR model, network connectivity is independent of symptom severity. Instead, in the GVAR model connectivity strongly relates to variability in symptom activation: more strongly connected networks are associated with more variance around the stable state. Our results suggest that more systematic investigations should precede linking psychological theory to statistical network models. To this end, we present stability landscapes as a thinking tool to gain insight into the dynamics of idiographic network models and as a means to link model-implied behaviour to theory-implied behaviour.

In preparation as: **Hoekstra, R. H. A.***, de Ron, J.*, Epskamp, S., Robinaugh, D. J., & Borsboom, D. (2024). Mapping the dynamics of idiographic network models to psychopathology using stability landscapes.

*authors contributed equally

7.1 INTRODUCTION

The way we conceptualize mental disorders and the statistical models we use to study them are intimately connected. For example, the introduction of the network theory of psychopathology [24] has gone hand in hand with the increasing popularity of statistical network models in psychology [e.g., Ising model, GGM, GVAR model; 227]. However, theories and statistical models should not be conflated: One could estimate a latent variable model without assuming an underlying common cause [40] just as one could estimate a network model without assuming underlying causal interactions between variables [25]. Because statistical models are not necessarily coupled to psychological theories, it is important to consider how our theories map onto the dynamics of our statistical models [137, 68]. In this paper, we aim to relate the dynamics of statistical idiographic network models to the most fundamental principle of the network theory of psychopathology, namely that an individual's symptom network reflects an individual's propensity to develop a mental disorder.

The network theory of psychopathology characterizes mental disorders as a causal network of symptoms, where the presence of a mental disorder is defined as a stable state of symptom activation [28, 24]. External shocks or perturbations caused by factors outside the symptom network, such as losing one's partner, can activate symptoms in the network, such as a depressed mood. Two characteristics of the network determine how the network responds to these external shocks. First, the *thresholds* of the symptoms determine the tendency of a symptom to be activated. Low (high) thresholds, correspond to more (less) likely symptom activation.[1] According to the network theory, the thresholds for most of the symptoms should be high, since most symptoms are rare and symptom distributions are typically skewed to the right [251, 24]. Second, the *connectivity* of the network (i.e., the strength and structure of the relations between symptoms calculated by summing all the weighted connections) determines how easily a symptom activates other symptoms. Symptom activation in a strongly and fully connected network triggers the activation of many other symptoms, resulting in cascading feedback effects that maintain symptom activation over time.[2] Conversely, symptom activation within a sparsely connected network does not necessarily exhibit this effect; the likelihood of a single symptom triggering a cascade of many other symptoms is small due to the relatively weak causal relations [53].

On the basis of the network theory of psychopathology, many researchers have examined whether individuals who have (or are at risk of devel-

1 The thresholds are equivalent to the difficulty parameters in a Rasch model [196]

2 This is shown in a simulation by Cramer et al. [53] based on the dynamics of the Ising model in the $\{0, 1\}$ domain.

oping) a mental disorder have higher network connectivity than those who do not. The empirical literature examining this relationship between connectivity and psychopathology has shown decidedly mixed results. Network studies have reported that increased network connectivity is associated with *increased* levels of psychopathology [e.g., 19, 197, 214, 271, 192, 245, 244], with *decreased* levels of psychopathology [e.g., 34, 107, 67], and both to increased and decreased levels of psychopathology [e.g., 191, 163].

These mixed results illustrate that the translation from the network theory of psychopathology to statistical networks is not straightforward. Whether the *statistical* network model aligns with the network *theory* minimally depends on (a) whether the variables assessed correspond to the symptoms that are hypothesized to play the role of network nodes in the theory, (b) whether the data level corresponds to the theory-implied dynamics, and (c) whether the statistical network is capable of representing dynamics in a way that is consistent with the theory. If we consider the alignment between network theory and network methods we observe several inconsistencies. For instance, nodes in network theory correspond to symptoms as encoded in diagnostic manuals such as the DSM-5 [24]. In contrast, in many ESM studies using network modeling, nodes would correspond to variables that assess the fluctuation in momentary mood states rather than symptoms.[3] Oftentimes empirical studies use between-person data, whereas the network theory is described at the within-person level. Furthermore, although the network theory posits that both thresholds and connectivity play a role in determining if the system remains in a stable state of symptom activation after external shocks, not all statistical network models (e.g., GVAR, GGM) have a parameter that directly relates to thresholds. In addition, the GVAR model is linear and exhibits stationarity, and as a result, it cannot exhibit the type of nonlinear behavior that would be expected from the network theory [137, 237].

Neither of the minimal requirements are trivial, however, (a) and (b) are a matter of design choice, while (c) requires an intricate understanding of the model- and theory-implied dynamics. Therefore it is important that researchers have tools to examine the connection between the network theory and statistical idiographic network models[4] so that they can assess what they *would* expect to see in the statistical model *if* the connectivity or thresholds of the model change. To that end, stability landscapes are highly useful to represent model behavior, because they offer an easy-to-understand visual tool to study the long-term dynamics of a model

3 There is no logical reason to assume that a person with stronger causal connections between symptoms should also have higher (partial) correlations between momentary mood states.

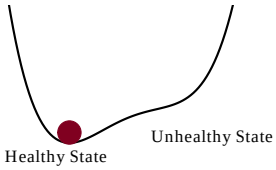
4 With idiographic network models we refer to any person-specific network model, so both GVAR and Ising model.

based on all parameters simultaneously: in a single visualization, a stability landscape gives insight into the effect of external shocks, thresholds, and connectivity.

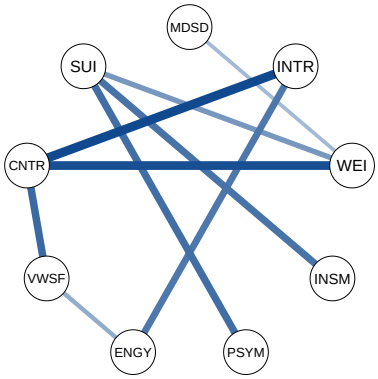
In this chapter, we build upon the work by Cui et al. [60, 61] and Cui, Hasselman, and Lichtwarck-Aschoff [59], and develop stability landscapes to examine the connection between idiographic network models and theory implied network dynamics of psychopathology. In particular, we systematically vary the two network characteristics (i.e., thresholds and connectivity) and investigate the impact of these characteristics on the dynamics of two idiographic network models, namely the Ising model (both within the $\{0, 1\}$ domain and the $\{-1, 1\}$ domain) and the GVAR model. We show that only in the Ising model within the $\{0, 1\}$ domain, the dynamics can directly be translated to the network theory. In the Ising model within the $\{-1, 1\}$ domain the state with no symptom activation also becomes more stable with increasing connectivity. Furthermore, we show that in the GVAR model, network connectivity is not associated with symptom severity but with the variance in symptom activation.

7.2 STABILITY LANDSCAPES

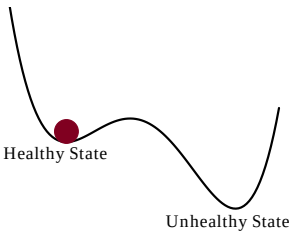
Stability landscapes summarize the dynamics of a system by visualizing the tendency of the system to be in each possible state given the initial conditions of its variables. In psychology, stability landscapes have mainly been used as a conceptual illustration—the so-called *ball-in-cup* metaphor [208, 160, 36, 265, 139]. Figure 15, Panel (a) and (c) shows two conceptual stability landscapes with hills and valleys in which the valley's depth represents the system's stability (where the deeper the valley, the more stable the state). The red ball represents the system's current state: The ball can roll over the landscape due to perturbations that represent the influence of external shocks (e.g., adverse life events). How the system responds to these perturbations depends on the shape of the landscape. In an unstable state (hill), a small perturbation pushes the ball towards a stable state (valley). In a stable state (valley), a perturbation temporarily disturbs the system, but the ball moves back to its original state over time. The steeper the valley, the quicker the system returns to the stable state, and the more likely it is that the system gets trapped in that state as it is more difficult to transition from one state to another (e.g., only a significant perturbation can push the ball up the hill). The point where the system transitions from one stable state to the other, as represented in the stability landscape by the hill, is called a tipping point.



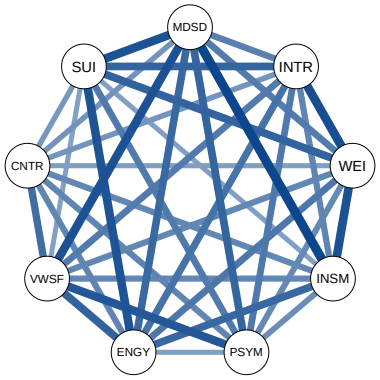
(a) Resilient stability landscape



(b) Weakly connected network



(c) Vulnerable stability landscape



(d) Strongly connected network

Figure 15: The network theory of psychopathology posits that individuals vulnerable to developing a mental disorder have a more strongly connected symptom network and lower thresholds than resilient individuals, and therefore their stability landscapes look different. Panel (a) shows a resilient stability landscape against developing a mental disorder and Panel (b) shows the corresponding symptom network with low connectivity. If an external shock activates one of the symptoms (i.e., the ball is perturbed to the right side away from the stable state), due to the weak connections in the network other symptoms will not maintain activated in response, and the system will recover rapidly to its original state. Panel (c) shows a stability landscape that is vulnerable to end up in the unhealthy phase and Panel (d) shows the corresponding symptom network with high connectivity. In a densely connected network, the activation of one symptom easily cascades to the activation of other symptoms, and is more likely to cause the system to move the system's state over the so called tipping point into the unhealthy phase.

7.2.1 *Linking the network theory to stability landscapes*

In the network theory of psychopathology, mental disorders are hypothesized to be stable states of symptom activation. Stability landscapes offer a useful illustration of the network theory of psychopathology as they give a visual representation of the symptom activation (i.e., position of the stable state) and the stability of the symptom activation [i.e, the variability in symptom activation over time as indicated by the steepness of the valley around the stable state; e.g, 24, 184, 190, 145].

Two postulates of the network theory of psychopathology can be illustrated by coupling the two network characteristics (i.e, thresholds and connectivity) to stability landscapes [23, 53, 24]. The first postulate is that the *stability* of different psychological states (i.e., healthy and unhealthy) is in part governed by an individual's network connectivity and thresholds. A network with low connectivity and high symptom thresholds, see Panel (b) of Figure 15, gives rise to a stability landscape that is characterized by a single stable state, corresponding to a healthy state of low symptom activation, see Panel (a) of Figure 15. In contrast, a network with high connectivity and low symptom thresholds, see Panel (d) of Figure 15, gives rise to a stability landscape that is characterized by two alternative stable states, a healthy and an unhealthy state, and features a tipping point, see Panel (c) of Figure 15. For such individuals, the healthy state is a shallow basin and the unhealthy state is characterized by a deep and steep basin; *ipso facto*, states of persistent symptom activation, i.e., mental disorders, are both accessible and stable.

The second postulate is that with increased connectivity, the system's behavior becomes increasingly discontinuous [53, 227]. For a resilient person with a low connected symptom network, the system responds proportionally to external shocks: The ball can roll to an unhealthy state but once the external shock is removed, the system returns back to its original, healthy state. However, for a vulnerable person with a highly connected symptom network, the system displays non-linearity and responds disproportionately to external shocks. This means that, if the person is near a tipping point, small perturbations can have large effects. Once the tipping point between the two stable states is reached, the system rapidly moves to the unhealthy state. Since the unhealthy state itself is stable, the disappearance of the external shocks will not cause the system to return to its original healthy state. Mathematically, these dynamics can be described using a cusp catastrophe [92].

In summary, the network theory predicts that in networks with high thresholds [24] (1) a more strongly connected idiographic network leads to stability landscapes where the unhealthy state is more severe (i.e., higher symptom severity) and more stable than the healthy state (i.e.,

the unhealthy state has a deeper and steeper basin than the healthy state). And (2) more strongly connected networks lead to stability landscapes where the behavior of the system becomes increasingly discontinuous (i.e., the tipping point between the stable states becomes more prevalent). That is, within a stable state there is *less* variability in symptom activation but within the whole system there is *more* variability in symptom activation.

Recently proposed stability landscape methodology [60, 61, 59] allows us to move beyond the metaphorical use of the stability landscapes and instead construct landscapes based on the parameters of statistical network models. In the next section, we map the theoretical principles of the network theory to idiographic network model dynamics by constructing stability landscapes. Specifically, we present a simulation study where we vary both the thresholds and the connectivity for the idiographic network models to systematically investigate the effect of these different parameterizations on symptom *severity* and *variability* in symptom activation.

7.3 METHODS

In our simulation study, we focus on two idiographic network models that can be estimated for an individual on longitudinal, multivariate data: The Ising model for binary data and the GVAR model for continuous data.

7.3.1 *Ising Model*

The Ising model defines the probability of a node x_i to be "on" (encoded as 1) or "off" (encoded as 0 or -1) given all other nodes in the network. In psychological network models, nodes correspond to symptoms, and a node being "on" or "off" corresponds to a symptom being present or absent respectively. The probability for each node to be "on" is based on two parameters, namely the thresholds (τ) and the strength of the interactions between nodes, which is represented by edge weights (ω), following the equation:

$$P(x_i \mid x_{-i}) = \frac{1}{1 + e^{\tau_i + \sum \omega_{i,j} x_i}}, \tag{12}$$

where x_i represents the state of node i, x_{-i} represents the state of all other nodes in the network, and τ_i is the threshold of node i, which represents the tendency of node i to be activated independently of the other nodes in the network. A threshold of zero is neutral, a negative threshold corresponds to the tendency for a symptom to be off and a positive threshold corresponds to the tendency for a symptoms to be on. Note that

the thresholds in the Ising model are mirrored relative to the representation of thresholds used in the original formulation of the network theory of psychopathology, in which a low (high) threshold correspond to the tendency for a symptom to be on (off) [24, 4].

The edge weights are represented by matrix of size $n \times n$ nodes, where each edge weight $\omega_{i,j}$ represent the strength of alignment between two nodes. However, the exact interpretation of the edge weights is dependent upon the encoding of the nodes, either as $\{0, 1\}$ or $\{-1, 1\}$ [135]. When nodes are coded as $\{-1, 1\}$ a weight parameter above zero indicates that nodes align with each other in either direction, either $\{1, 1\}$ or $\{-1, -1\}$, and a weight parameter below zero indicates that opposite labels align with each other, $\{-1, 1\}$. In contrast, when nodes are coded as $\{0, 1\}$ the strength of the positive weight parameter indicates the increased probability of two nodes to be in state $\{1, 1\}$ relative to the probability of all other states $\{0, 1\}$, $\{1, 0\}$, and $\{0, 0\}$. As the dynamics of the Ising model are dependent upon the encoding of the nodes, we investigate the dynamics of both the Ising model in the $\{-1, 1\}$ and the Ising model in the $\{0, 1\}$ domain.

7.3.2 *GVAR model*

The GVAR model defines the value of each variable at time t as a linear combination of itself (so-called auto-regressive or lagged effects) and all other variables in the model (so-called cross-lagged effects) on the previous time point $t - 1$ as follows [130, 89]:

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{B}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_t, \tag{13}$$

$$\boldsymbol{\varepsilon}_t \sim \mathrm{N}(0, \boldsymbol{\Sigma}), \tag{14}$$

where $\mathbf{y}_t$ denotes a vector of responses of a given person at time t, μ is the mean of the distribution of each variable, $\mathbf{B}$ is a matrix containing the lagged and cross-lagged relations also known as the temporal network, $\mathbf{y}_{t-1}$ is the vector of responses of person at a previous time point $t - 1$ and $\boldsymbol{\varepsilon}_t$ is a Gaussian noise process with a time-invariant positive definite covariance matrix $\boldsymbol{\Sigma}$ [89, 81].

Note that the GVAR model does not contain a parameter that directly translates to the thresholds in the network theory of psychopathology. In principle, continuous variables as used in the GVAR model do not have thresholds in the sense that categorical variables do, as the nature of change is fundamentally different (gradual change versus switching states). In contrast to the threshold parameter in the Ising model, the GVAR model features a mean parameter μ_i that controls the location of the distribution of variable i. If one formulated a network theory of psychopathology using continuous symptoms, this parameter could play

Theory	Ising model	GVAR model
Thresholds	τ	μ
high = *lower* symptom activation	positive = *higher* symptom activation	high = *higher* symptom activation
low = *higher* symptom activation	negative = *lower* symptom activation	low = *lower* symptom activation
Connectivity	$\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \lvert\omega_{ij}\rvert$	$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lvert\beta_{ij}\rvert$

Table 9: Connecting network characteristics to statistical models. Thresholds in the network theory of psychopathology refers to the amount of external force a symptom needs in order to become active (switch from being absent to being present). According to the network theory, a *high* threshold relates to *low* symptom activation and vice versa. Thresholds in the Ising model can be interpret as the probability of symptom activation. Meaning symptoms with a *positive* threshold, have a *higher* probability of being active. For the GVAR model, the mean of the symptom refers to symptom activation, where a *higher* mean is indicative of *high* symptom activation. Connectivity is defined as the weighted average strength of the estimated network model for both the Ising model and the GVAR model.

an analogous role to the threshold in the Ising model, as the higher the mean for a given symptom, the higher its intensity, and thus the more the symptom is "activated". See Table 9 for an overview of the different interpretations of the network characteristics in the network theory, the Ising model, and the GVAR model. Connectivity (i.e., the strength and structure of the network) in the network theory could potentially be operationalized in the Ising model or the GVAR model as the network density, i.e., the average absolute edge weight [211]. However, as we will show later in this chapter, these operationalizations are not on par in representing the theory.

7.3.3 *Simulation Set Up*

To evaluate whether the alternative representations in Table 9 match the theory, one can use simulation approaches.[5] In the simulation study reported here, we vary both the thresholds (i.e., the τ parameters for the Ising model and the μ for the GVAR model) and the network connectivity and we investigate their effect on the stability landscape. In order to achieve this goal, our simulation study follows three steps. In the first step, we construct a baseline network for the Ising model in the $\{0, 1\}$ and $\{-1, 1\}$ domain and the GVAR model. Next, we vary the thresholds and

5 Note that for these models one could also derive the impact of changing parameters on means/variances. For the GVAR model see Appendix F for such a derivation.

connectivity of these baseline networks. In the second step, we construct a stability landscape for each of the different network parameterizations. In the third step, we inspect the stability landscapes by looking at the number of stable states, their location and stability. We review each step of the simulation process in further detail in the next section.

STEP 1: PARAMETERIZE NETWORKS To construct the baseline Ising networks, we use the data from the Virginia Adult Twin Study of Psychiatric and Substance Use Disorders [VATSPSUD; 166]. This data set includes binary variables for nine depressive symptoms of 8,973 twins from the Mid-Atlantic Twin Registry. To estimate the Ising model we used the function `estimateNetwork` from the *bootnet* package with `IsingFit` as default argument. In the original data set the variables are encoded as 1 (symptom present) and 0 (symptom absent). When estimating the baseline Ising network within the $\{-1, 1\}$ domain we first recode absent symptoms as -1.

Although the baseline Ising network is derived from cross-sectional data whereas the postulates of the network theory are based on idiographic networks, the parameterization is in line with the network theory. That is, according to the network theory, the thresholds for most of the symptoms should be high (and thus negative in the Ising model), since most symptoms are rare and symptom distributions are typically skewed to the right [251, 24]. In the baseline Ising network within the $\{0, 1\}$ domain we have thresholds between -6 and -2 [which is comparable to the simulations in 53]. Furthermore, according to the network theory, relations between symptoms are predominantly positive, as symptoms are more likely to exacerbate rather than alleviate each other [24]. The baseline Ising network is almost fully connected with all strongly positive connections (only one edge between the nodes fatigue and suicidal ideation is missing).

We construct a baseline temporal GVAR network ($\mathbf{B}$) with nine variables. We sample the weights for the cross-lagged relations from a uniform distribution between 0.1 and 0.2, and the auto-regressive relations from a uniform distribution between 0.2 and 0.4. This implies that the temporal network has stronger auto-regressive effects than cross-lagged effects. In line with the network theory, all connections between nodes in the network are positive. We construct the contemporaneous network ($\mathbf{\Sigma}$) using the `genGGM()` function from the *bootnet* package [85] with 80% positive edges (setting argument `propPositive` to 0.8).

For both type of network models, we multiply the baseline network with a constant, which we will refer to as *multiplier*. The multiplier determines the connectivity of the network, where we have three connectivity conditions: For the Ising model we multiple the network with 0.9, 1, or

1.3. For the GVAR model, we multiply the network either by .5, 1 or 2. This results in a less strongly connected network, the baseline network, and a more strongly connected network. In addition, we alter the threshold parameters for the Ising model (τ) and the means (μ) for the GVAR model. For the Ising model, we study three different threshold conditions, namely baseline thresholds (as these are negative they represent the tendency of all symptoms to be off), the baseline thresholds multiplied by 2 (leading to even more negative thresholds) and positive counterpart of the baseline thresholds (by multiplying the thresholds by -1; representing the tendency of all symptoms to be on). For the GVAR model, we study three different mean conditions, namely $\mu = 0$, $\mu = 3$ and $\mu = 6$.

For our simulation, we end up with three different network models: Ising model in the $\{-1, 1\}$ domain, the Ising model in the $\{0, 1\}$ domain and the GVAR model. Thus, in total, we construct 27 stability landscapes: 3 (type of network model) $\times$ 3 (thresholds or mu's) $\times$ 3 (connectivity multipliers).

STEP 2: CONSTRUCTING STABILITY LANDSCAPES Stability landscapes visualize a potential energy function [60, 61, 59]. A potential energy function defines the probability of system to be in a certain configuration (e.g., the observed values of the nodes in the network). For unidimensional linear systems, the potential energy function can be obtained by taking the integral. However, when we have a multidimensional (non-)linear system, e.g., an Ising model or a GVAR model with multiple variables, taking the integral quickly becomes intractable. In such situations, we can approximate the generalized potential energy function (U) of each state of the system (X) based on the steady-state distribution (Pss). To get from the steady-state distribution to the potential energy function, we take the negative logarithm of the steady-state distribution of each configuration [257, 60, 59, 61]. As we take the negative logarithm, states with a lower density in the steady state distribution have a higher potential energy function, as is given by:

$$U(X) = -\ln P_{SS}(X), \tag{15}$$

where $\mathbf{X}$ is a specific configuration of node values in the network, so-called *microstate* of the network. For instance, if we have two binary variables, x_1 and x_2, than there are four possible microstates: ($x_1 = 0$, $x_2 = 0$), ($x_1 = 1$, $x_2 = 0$), ($x_1 = 0$, $x_2 = 1$), ($x_1 = 0$, $x_2 = 0$). However, here we more interested in the *macrostate*, i.e., the number of active symptoms, as it represents general symptom severity. In that case, a healthy state in the stability landscape corresponds to low number of active symptoms and an unhealthy state corresponds to a high number of active symptoms.

Cui et al. [61] define a generalized potential function for the macrostate by summing over the potential function for all microstates with a similar number of active symptoms in the network. For instance, if we want to find the potential energy function of a macrostate that corresponds to one symptom being present within the $\{0, 1\}$ domain, we must sum the potential energy function of the two microstates, namely $(x_1 = 1, x_2 = 0)$ and $(x_1 = 0, x_2 = 1)$, following:

$$\begin{aligned} U(X) &= -\ln P_{SS}(X), \\ &= -\ln \sum_{(\sum x_i)=X} P(x_i). \end{aligned} \tag{16}$$

For the Ising model, we know the potential energy function. This function is typically denoted as the Hamiltonian represented as $H(x)$:

$$H(x) = -\sum_{\langle i,j \rangle} w_{i,j} x_i x_j - \sum_i \tau x_i. \tag{17}$$

The steady-state distribution needed to construct our stability landscape $P(x)$, is then proportional to the exponent of $H(x)$[6]:

$$P(x) \propto e^{-H(x)}. \tag{18}$$

Following Equation 16, we can then obtain the generalized potential function for the macrostate (i.e., number of nodes in state 1) of the Ising model as follows:

$$\begin{aligned} U(X) &= -\ln P_{SS}(X), \\ &= -\ln \sum_{(\sum x_i)=X} P(x_i), \\ &= -\ln \sum_{(\sum x_i)=X} e^{-H(x)}. \end{aligned} \tag{19}$$

Similarly to the Ising model, to construct the stability landscape we need to derive the steady state distribution of the GVAR model. The GVAR model assumes all parameters are stationary across the measured time-period [41]. The only exogenous random variable in the GVAR model is $\boldsymbol{\varepsilon}$, which is Gaussian. As the model assumes stationarity, it follows that the response variable y is also Gaussian. This is a convenient property of the GVAR model, as this allows us to derive that the steady state distribution of a variable in the GVAR model is a Gaussian distribution with a mean of μ and variance of [73]:

6 We ignore the β parameter, so-called inverse temperature in physic literature, by setting it to one. We do so as the β parameter is unidentifiable (that is, it is unobtainable when estimating the Ising model based on data).

$$\text{vec}\left(\text{var}(\mathbf{y})\right) = (\mathbf{I} - \mathbf{B} \otimes \mathbf{B})^{-1} \text{vec}\left(\mathbf{\Sigma}\right), \tag{20}$$

where $\mathbf{I}$ an identity matrix containing one's on the diagonal and zero's on the off-diagonal elements.

Again, we are not interested in the microstate of the system, (i.e., the exact values of the nodes in the network), but in the general macrostate of the system. However, in contrast to the Ising model where the macrostate is the number of present symptoms, we work with continuous variables and thus the macrostate refers to the sum score of the variables in the network. To obtain the potential landscape of the macrostate for the GVAR model, we can sum over all the μ's, and the square root of the sum over all implied variances (see Appendix F for the complete derivation).

STEP 3. INSPECT STABILITY LANDSCAPE In the third and final step, we investigate the stability landscapes for each of the constructed networks. We inspect two properties of the stability landscapes: (1) the location of the stable state(s) relating to symptom severity and (2) the steepness of the valley around the stable state(s), relating to variability of symptom activation (for Ising model) or the sum score of the variables (for GVAR model). Steeper valleys correspond to smaller response to perturbations to the network. As the GVAR is a linear model, it has only one stable state [130]. If there are two stable states in the in the landscape, as to be expected for the Ising model, we investigated the steepness of the valley for all stable states.

7.4 RESULTS

7.4.1 *Ising model dynamics*

$\{0, 1\}$ DOMAIN Figure 16 shows nine stability landscapes with varying parameter values of the Ising model within the $\{0, 1\}$ domain. The x-axis shows the number of active symptoms and the y-axis shows the value of the energy function, U. The columns indicate the difference in threshold parameters (i.e., mean $\tau \in \{-7, -4, 4\}$), while the rows indicate the difference in the multiplier that was used to either make the baseline network sparser (i.e., $\mathtt{multiplier} = 0.9$) or denser (i.e., $\mathtt{multiplier} = 1.08$).

The value of the threshold determines the location of the stable state: When the thresholds are positive (mean $\tau = 4$, see right column of Figure 16) the system is most likely to be in an unhealthy state (i.e., all symptoms are active) whereas when the thresholds are strongly negative (mean $\tau = -7$, see left column of Figure 16) the system is most likely to be in a healthy state (i.e., all the symptoms are absent). For the positive

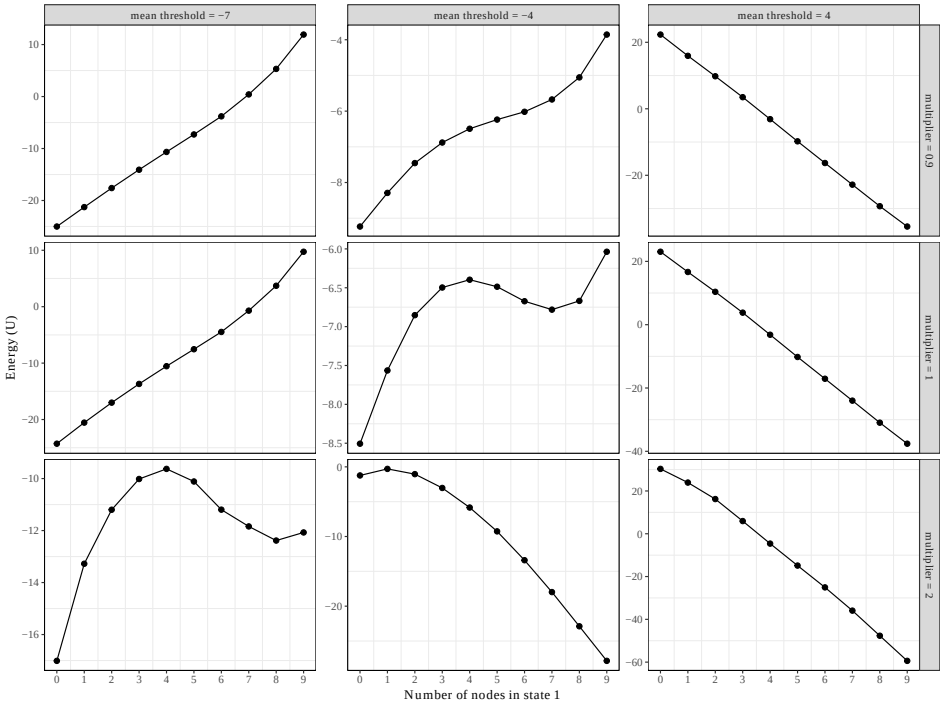


Figure 16: The potential stability landscapes for Ising model, where absent symptoms are coded as 0 and present symptoms as 1. We varied the density of the network and the thresholds. The x-axis represents the state of the system, namely the sum score of the variables, and the y-axis the potential energy (U) of that state. The lower the potential, the more stable the state is.

threshold conditions (mean $\tau = 4$), the Ising model show only one stable state in which all symptoms are active.

For the threshold condition based on the original data with a mean threshold of -4 there are four consequences of increasing connectivity. First, as the connectivity increases, the number of stable states changes. There is one stable state when network connectivity is low (multiplier = 0.9) but there appear two stable state in the conditions with increased network connectivity (although the in the condition with the highest network connectivity the healthy stable state almost disappears). Second, increasing the network connectivity results in a more deep and steep unhealthy stable state and a more shallow healthy stable state. Third, as the connectivity increases, the unhealthy stable state becomes more severe (more symptoms are present). Fourth, initially when the connectivity increases, a pattern of increasing discontinuity arises. That is, at lower levels of connectivity, the Ising model may have only one stable state and exhibit a relatively smooth transition between different levels of symptom activation. However, as the connectivity increases a tipping point (as indicated in the landscape by a hill) emerges. This is the classic dynamical behaviour of

the cusp catastrophe [92], where connectivity plays the role of splitting variable and external perturbations act as the normal variable. However, when the connectivity increases even further, only one unhealthy stable state emerges.

In summary, the results within the $\{0, 1\}$ domain, at least within subset in the range of possible parameters (e.g., when connectivity becomes too high the discontinuity disappears), are in line with the postulates of the network theory. Furthermore, there is an interaction effect between thresholds and connectivity. The more negative the thresholds, the more connectivity is needed in order to obtain two instead of one stable state(s).

$\{-1, 1\}$ DOMAIN Figure 17 shows nine stability landscapes with varying parameter values of the Ising model within the {-1,1} domain. The x-axis shows the number of active symptoms and the y-axis shows the value of the energy function, U. The columns indicate the difference in mean thresholds (i.e., mean $\tau \in \{-1, 0, 1\}$), while the rows indicate the difference in the multiplier that was used to either make the baseline network sparser (i.e., $multiplier = 0.9$), or denser (i.e., $multiplier = 1.3$).

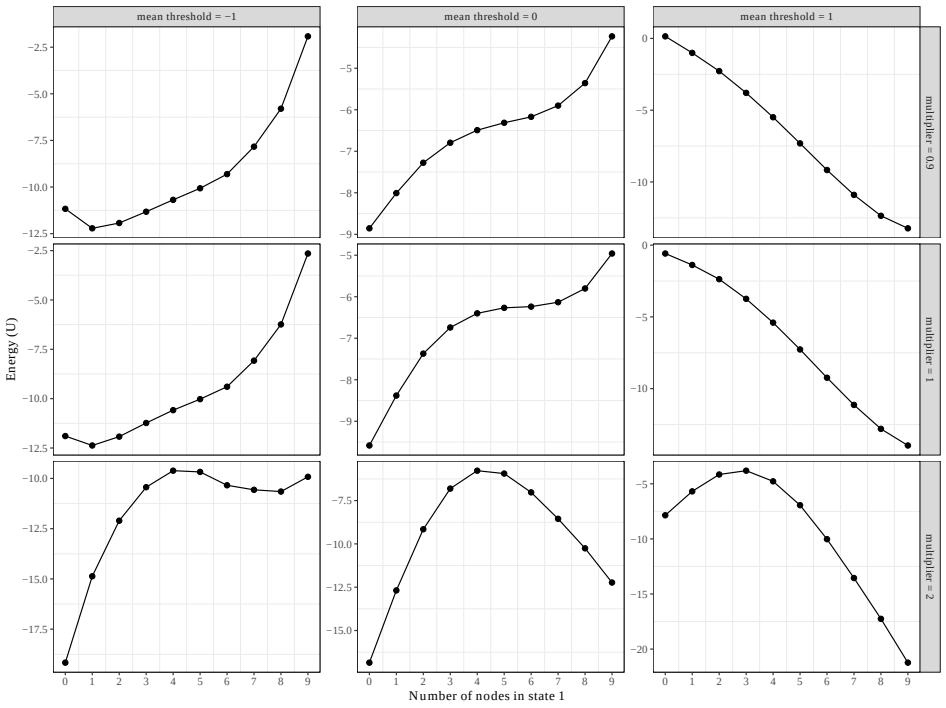


Figure 17: The potential stability landscapes for Ising model, where absent symptoms are coded as -1 and present symptoms as 1. We varied the connectivity of the network and the thresholds. The x-axis represents the state of the system, namely the sum score of the variables, and the y-axis the potential (U) of that state. The lower the potential, the more stable the state is.

The threshold condition with a mean $\tau = 1$, shows a landscape with only one stable state: The unhealthy state where all symptoms are active (i.e., the number of active symptoms is 9), see right column in Figure 17. The threshold condition with a mean $\tau = -1$, shows a landscape with either one or two stable states, depending on the connectivity. When connectivity is low (multiplier = 0.9) or at baseline (multiplier = 1), the healthy state (i.e., number of active symptoms is 1) is the most likely to occur, see left column in Figure 17. However, with increased connectivity (multiplier = 2) a new unhealthy stable state emerges (i.e., number of active symptoms is 7), see lower right panel in Figure 17. Similar to the Ising model within the $\{0, 1\}$ domain, the unhealthy state becomes more stable with increasing connectivity, however, in contrast to the Ising model within the $\{0, 1\}$ domain, the healthy stable state also becomes more stable. That is, due to the encoding increasing connectivity not only render two nodes in a $\{1, 1\}$ state more likely but also in the $\{-1, -1\}$ state [135].

When the mean thresholds are zero and there is network connectivity is high (multiplier = 1.3), the Ising model shows two stable states: one basin appears around 0 active symptoms and one basin appears around 9 active symptoms, see the plot at the bottom of the middle column in Figure 17. Given mean thresholds of zero, with increasing connectivity *both* stable states become more resilient to perturbations. In addition, when the system is not near a tipping point, higher connectivity implies the the system will return to its original stable state more quickly after a perturbation (see the bottom row in Figure 17). With decreasing connectivity the system responds more strongly to perturbations, as indicated by a more shallow basin, and the system would more slowly return to its current stable state after a perturbation (see the top row in Figure 17). When the mean thresholds are positive ($\tau = 1$), increasing connectivity causes a second (healthy) state to appear. Thus, contrary to the network theory, stronger connectivity leads to a system that is more likely to be in a healthy state than a less connected network.

7.4.2 *GVAR model dynamics*

Figure 18 shows nine stability landscapes with varying parameter values of the GVAR model. The columns indicate the difference in mean parameters (i.e., $\mu \in \{0, 3, 6\}$), while the rows indicate the difference in the multiplier that was used to either make the baseline network more sparse (i.e., $\mathtt{multiplier} = 0.5$), or more dense (i.e., $\mathtt{multiplier} = 1.2$).

As the GVAR model is a linear model, the stability landscapes show only one stable state. The mean, μ, determines the position of the stable state. When $\mu = 0$, we find a stable state at the minimum sum score of 0.

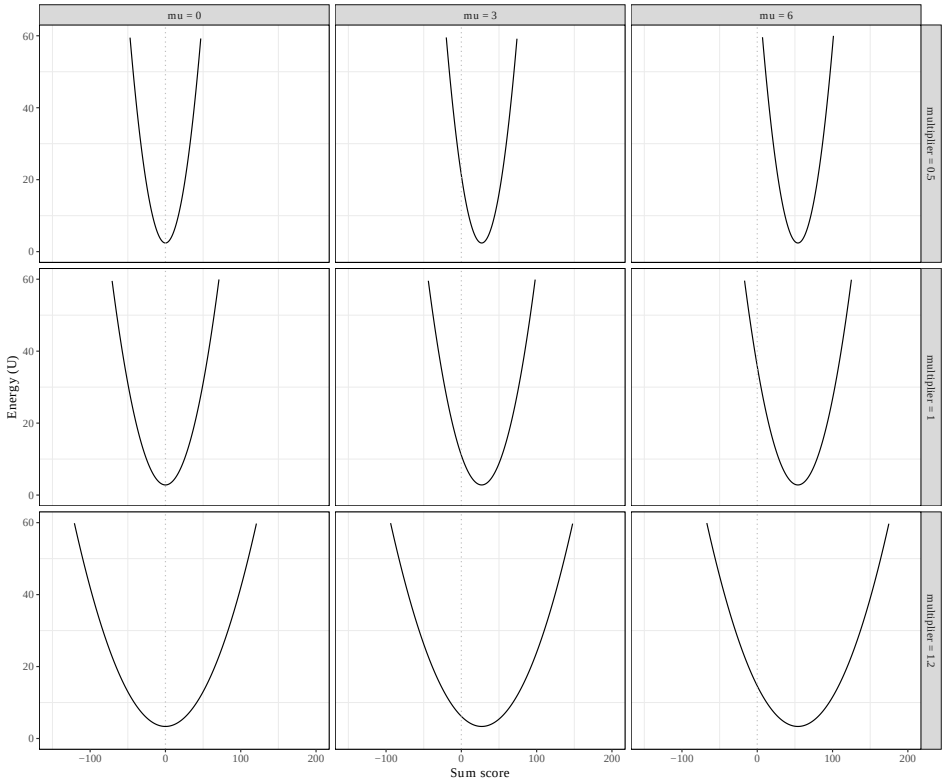


Figure 18: The potential stability landscapes of the GVAR model, where we varied the density of the temporal network β and the $\mu's$. The x-axis represents the state of the system, namely the sum score of the variables, and the y-axis the potential (U) of that state. The grey dotted line indicates a sum score of 0. The lower the potential, the more stable the state is.

If we increase the mean, the position of the basin shifts to the right and thus we increase the value of the stable state, regardless of the network connectivity. The network connectivity determines the steepness of the valley in the landscape, i.e., the amount of variation possible around the stable state. When connectivity is low ($multiplier = 0.5$), there is less variation around the stable state as indicated by a steep valley. When connectivity is high ($multiplier = 1.2$), there is more variation around the stable state, as indicated by a more shallow stable state.

The connectivity of the GVAR model determines the shape of the landscape, whereas the mean values of the variables determine the location of the landscape. When mean is low ($\mu = 0$) and connectivity is high ($multiplier = 1.2$), the systems shows more variation around a stable state of sum score 0. In contrast, when the mean is high ($\mu = 6$) and connectivity is low ($multiplier = 0.5$), the systems shows less variation around a sum score of 50, see the right upper panel and left lower panel in Figure 18 respectively.

7.4.3 *Discussion simulation results*

Table 10 gives an overview of the simulation results. Regarding the dynamics of the Ising model, we see that the qualitative behavior of the model depends on the encoding of the variables: We found that for the conditions we investigated, only a subset of the parameter space of the Ising model within the $\{0, 1\}$ domain corresponds to the dynamics postulated by the network theory of psychopathology. The Ising model within the $\{-1, 1\}$ domain only partly aligns with the network theory: Increasing network connectivity leads to increasingly discontinuous behavior. However, as network connectivity increases, the unhealthy stable states becomes more stable but contrary to the network theory, the healthy state also becomes more stable–this is to be expected as positive connections within the $\{-1, 1\}$ domain indicate two variables tend to take on the same value, without a preference for -1 or 1. As the GVAR model is a linear model it only has one stable state and both postulates of the network theory cannot be directly translated to the GVAR model.

Postulates of the network theory of psychopathology	**Ising model $\{0, 1\}$ domain**	**Ising model $\{-1, 1\}$ domain**	**GVAR model**
When connectivity is increased, the healthy state becomes less stable and the unhealthy state becomes more severe and stable	✓	×	×
When connectivity is increased, the system becomes more discontinuous	±	✓	×

Table 10: Recovery of the dynamics for the Ising model and the GVAR model as expected by the network theory of psychopathology.

7.5 DISCUSSION

Ever since the introduction of the network approach, methodological advances to estimate idiographic network models have developed in parallel [227]. In this chapter we systematically investigated the link between the network theory of psychopathology and the parameterization of different statistical idiographic network models using stability landscapes. In particular, we constructed stability landscapes to map the two postulates of the network theory of psychopathology to the dynamics of the Ising model and the GVAR model. First, the network theory postulates when thresholds are high (corresponding to negative thresholds in the Ising model), more strongly connected idiographic networks lead to stability landscapes where the healthy state becomes less stable and unhealthy

state becomes more severe (i.e., high symptom activation) and stable (i.e., steeper basin). Second, more strongly connected networks lead to stability landscapes where the behavior of the system becomes increasingly discontinuous (i.e., the tipping point between the stable states becomes more prevalent) so that the overall behavior becomes more variable. Our results indicate that the translation of the network theory of psychopathology onto statistically estimated models depends heavily on the chosen model (e.g., Ising model or GVAR) and its parameterization.

Using stability landscapes we have shown that only for a subset of the parameters the Ising model within the $\{0, 1\}$ domain corresponds to the two postulates stated by the network theory. Thus, the characterization of networks that are highly connected as networks of individuals that are vulnerable to develop a mental disorder depends on choosing the Ising model in the $\{0, 1\}$ domain. This result is expected given that this model is equivalent to the model used by [53] on which the network theory of psychopathology is based. The dynamics of the Ising model within the $\{-1, 1\}$ domain only partly aligns with the network theory: The network theory states that increased connectivity should only make the unhealthy state more stable, however, in the Ising model within the $\{-1, 1\}$ domain, both the healthy (low symptom activation) and unhealthy (high symptom activation) state become more stable with increasing network connectivity. Such dynamics seem more likely for attitudes than for psychopathology [135, 66].

Furthermore, we have shown that the dynamics of the GVAR model do not correspond to the network theory of psychopathology. The fundamental assumption of the GVAR model is that variables fluctuate around a stable mean. This inherently limits its capacity to represent multiple stable states, such as the dichotomy between healthy and unhealthy states depicted in the network theory. This limitation means that while the GVAR model can provide insights into behaviors within a stable state, it cannot provide insights into transitions between different stable states. For example, in a highly connected temporal network increased variability and a slower return rate to the stable state are observed (i.e., visualized in a stability landscape by a shallower basin). This increased variability could be an indication of critical slowing down [53], and therefore an early warning signal that the system is close to a tipping point moving from an healthy to an unhealthy state or vice versa [184]. These results are in line with research that found that an increased connectivity of temporal GVAR networks relates to more symptom fluctuations and not to symptom severity per se [163, 191].

In Chapters 5 and 6, we aimed to elucidate the relation between network connectivity, as estimated using a VAR model (either GVAR or mlVAR), and symptom severity (explored in Chapter 5) and symptom sta-

bility (explored in Chapter 6). The results presented in this chapter provide a theoretical underpinning for understanding the dynamics of the VAR model and the implications of the results found in Chapter 5 and Chapter 6. Specifically, Chapter 5 revealed an absence of correlation between changes in idiographic network structure and symptom severity. Chapter 6 demonstrated a strong correlation between the connectivity of person-specific networks and their variation (either aggravation or alleviation) in depressive complaints. The results presented in the current chapter refine these findings by illustrating that connectivity of idiographic network models as estimated using a VAR model are more indicative of psychopathology stability rather than psychopathology severity. The next chapter will further dissect these findings, by offering a comprehensive discussion on the congruence of these results and the subsequent ramifications for the application of idiographic network models within psychology.

7.5.1 *Practical considerations*

In clinical context, attempts have been made to link the connectivity of statistically estimated network models to psychopathology. Where it is often hypothesized, in line with the network theory, that high network connectivity means bad news in terms of psychopathology [e.g., 191, 214, 271, 192, 245, 67, 107]. In this chapter we have shown we should be careful with interpreting high idiographic network connectivity as indicating something “bad". Only within a subset of negative threshold parameters when estimating an Ising model in the $\{0, 1\}$ domain, could we interpret higher network connectivity as an indication of higher symptom activation. While for the other parameterizations and idiographic network models, higher network connectivity could either indicate “bad" or “good" news depending on the threshold parameters in the Ising model, or the mean parameter in the GVAR model. Therefore, our results highlight it is valuable to report both network connectivity and thresholds when relating an Ising model, or the mean when relating a GVAR model, to psychopathology. Furthermore, we encourage researchers to carefully consider the relation between the statistical model and their research question. For example, if the research questions regards critical slowing down, estimating an idiographic GVAR model might be better suited to estimating an idiographic Ising model.

7.5.2 *Limitations and future outlook*

The primary limitation of our study lies in the examination of a limited range of network models and parameter values. Our findings underscore

the significant impact of the network model parameterization on the resulting dynamics within the stability landscape. Consequently, it becomes vital to determine the dynamics of diverse model parameterizations. The stability landscape methodology presented here provides a valuable tool to explore the effects of such diverse model parameterizationson on the stability landscape. With the code presented online researchers can explore the stability landscapes of the parameterization of their estimated Ising model or GVAR model. Furthermore, future researchers could develop methods to investigate the stability landscape of different types of idiographic network models, such as the Time Varying-VAR model [TV-VAR; 134], threshold VAR model [T-VAR; 118], multilevel integrated VAR model [ML-VARI; 140], mean-switching Hidden Markov Model [HMM; 128, 221]. This approach can enhance our understanding of the intricate dynamics of idiographic network models.

Second, we simulate under a GVAR model where we assume stationarity and normally distributed variables. However, GVAR models are often estimated based on affect variables (i.e., negative and positive affect), which do not always follow a normal distribution. Negative affect variables are often highly skewed [136]. Furthermore, both negative and positive affect variables are bounded to zero–affect states can be absent but not minus. Such constraints on the data could affect the interpretation of network connectivity with regards to psychopathology. An increased connectivity in a network with highly skewed negative affect variables could mean that negative affect was more often prevalent, possibly corresponding to psychopathology. Therefore, future studies could focus on examining how these distributional characteristics of variables influence the dynamics displayed by the GVAR model and their theoretical interpretation.

7.5.3 *Conclusion*

The current chapter warrants caution in translating concepts rooted in the network theory of psychopathology directly onto the dynamics of statistically estimated idiographic network models. To that end, stability landscapes can function as a thinking tool to examine the connection between substantive theory and statistical network models.

8

GENERAL DISCUSSION

In the last two decades, we have witnessed two paradigm shifts in psychological research: (i) a transition from a focus on the *nomothetic* approach to a focus on the *idiographic* approach, and (ii) a transition from viewing mental disorders as having a *single* underlying cause to understanding them as causal *networks* of mutually reinforcing symptoms. The work discussed in this thesis straddles both paradigm shifts by examining how we can model heterogeneity within the context of idiographic network models to better understand individual differences in symptom dynamics.

8.1 THIS THESIS

8.1.1 *Part I: Assessing Heterogeneity*

The first part of this thesis focused on assessing heterogeneity in idiographic network models. Research using idiographic network modeling techniques has regularly claimed evidence for heterogeneity when comparing the network structures of individuals [e.g., 33, 95, 185, 67, 222]. In Chapter 2, we identified an important pitfall when comparing such estimated idiographic network models: a tendency to interpret *all* variability in the data as an indication of individual differences, while some of this variability could be the result of fluctuations created by noise—e.g., sampling variation. In a thought experiment, we showed that even if the true data-generating network structure is the same for two individuals, the network estimated from their data might differ substantially, as statistical methods are subject to the reliability of their estimation. Combining sampling variation with overall conservative estimation procedures will inevitably lead to a mismatch in the presence and absence of estimated edges in the network, in addition to varying edge weights across individuals. This entails high sensitivity and specificity are needed to separate illusionary from true heterogeneity when estimating idiographic network models.

We conducted a simulation study to determine the extent to which popular idiographic network modeling techniques are capable of separating such illusionary heterogeneity (i.e., sampling variation) from real heterogeneity (i.e., individual differences). Results showed that low statistical power places considerable limits on the validity of conclusions regarding heterogeneity for all tools investigated. At low sample sizes, a researcher

is likely to erroneously conclude that a sample is heterogeneous even if, in fact, the sample is homogeneous.

In response to these results, in Chapter 3, we proposed a rigorous way of testing if heterogeneity between idiographic network structures is plausible given the data at hand. We employed ideas of invariance testing from SEM to propose a systematic model comparison approach, where a model in which equality between idiographic network structures is assumed is compared to a model in which inequality between idiographic network structures is assumed. We deemed this practice the *Individual Network Invariance Test (INIT)*. INIT can be interpreted as a null hypothesis test to compare (strict) homogeneity (i.e., edges between nodes and their parameter values are constant across individuals) to heterogeneity (i.e., edges between nodes and their parameter values are allowed to vary across individuals). Contrary to the tools evaluated to compare idiographic network models in Chapter 2, INIT takes homogeneity rather than heterogeneity as the null model. As a result, INIT is a more conservative practice that errs on the side of caution, only deeming the variance in the data as individual differences when the data provide enough evidence to do so.

In Chapter 4, we presented simulation studies as a general framework to determine the sensitivity and specificity of various (idiographic) network estimation techniques. We exemplified the setup and execution of such simulation studies within network psychometrics in a tutorial. The aim of this chapter was to enable a broader range of researchers to examine the accuracy of network estimation techniques in order to determine to what extent model differences can be attributed to estimation accuracy rather than genuine individual differences.

8.1.2 *Part II: Understanding Heterogeneity*

In the second part of this thesis, we focused on understanding heterogeneity in idiographic network models. In particular, we tried to unravel heterogeneity by investigating the relationship between idiographic network structures and psychopathology severity and stability. In Chapter 5, we examined the relationship between idiographic network stability as estimated with a GVAR model over time and change in psychopathology severity for individuals located along the psychosis severity continuum. As the network theory of psychopathology links psychopathology severity to individual network structures, such that dense symptom networks could be associated with increased psychopathology severity, one might expect that individual networks change alongside psychopathology severity. To examine this expectation, we determined the stability of idiographic symptom networks over one year and identified whether network stability was associated with psychopathology severity. Using

the INIT methodology presented in Chapter 3, we identified individuals with a stable network over time and those with an unstable network over time. We tested whether these groups differed in their absolute change in psychopathology severity over that same time period. Most of our sample showed a stable symptom network over time, while most individuals showed a decrease in psychopathological severity. Thus, based on the results in Chapter 5, we found no relationship between a change in psychopathology severity and a change in idiographic networks over time.

In Chapter 6, we aimed to disentangle heterogeneity in depressive complaints and affect by investigating both psychopathology severity and stability using several tools, i.e., descriptive tools and by estimating a multilevel VAR model. We used descriptive tools such as visual inspection of the time series data to investigate how the evaluation of depressive complaints relates to affect fluctuations. We revealed that affect fluctuations show different patterns for individuals with an aggravation or alleviation in depressive complaints, as well as for individuals with consistent no depressive complaints, occasional depressive complaints, and consistent depressive complaints. Furthermore, the examination of person-specific network models revealed that the connectivity of idiographic networks was strongly related to a change (aggravation or alleviation) in depressive complaints. Thus, taking the results of Chapter 5 and Chapter 6 together, heterogeneity in idiographic network models when estimated using a VAR modeling approach seems to relate to psychopathology stability rather than psychopathology severity.

Chapters 5 and 6 demonstrate a clear understanding of model-implied behavior is essential in order to use statistical models to unravel the underlying mechanisms of mental disorders and the potential role individual differences play with respect to these dynamics. In Chapter 7, we used stability landscapes as a novel tool to systematically investigate the model implied behavior of several idiographic network estimation techniques: the Ising model and the VAR model. By mapping the two postulates of the network theory to idiographic network models, we revealed that both the estimation technique employed and the model parameters chosen matter a great deal for the dynamics idiographic network models as estimated from data can display. We provided a potential explanation for the results found in Chapter 6, and Chapter 5 namely that for the most commonly used idiographic network modeling technique, i.e., the VAR model, idiographic network structures relate to variance, and thus rather to psychopathology stability than psychopathology severity.

The work discussed in this thesis represents an initial endeavor to assess and understand heterogeneity using idiographic network models. As the initial surge of interest in estimating idiographic network models wanes, shortcomings of commonly employed methodologies have

emerged. These shortcomings underscore the challenges we are confronted with in order to adequately incorporate individual differences into psychological science. Addressing these challenges is crucial for deepening our understanding of the ways in which individuals differ and the ways in which they are similar.

8.2 BEYOND THIS THESIS

Most of the work discussed in this thesis primarily utilizes the VAR model to estimate individual network models in order to assess and understand heterogeneity. Although the VAR model is currently the most popular way to estimate individual network models within psychology, the VAR model has assumptions which are not always ideal for modeling psychological data [41, 37]. Two assumptions of the VAR model have direct implications for our ability to assess and understand differences among individuals. First, the VAR model assumes the time series data needed in order to estimate an idiographic network are stationary, and second, time points between measures are equally spaced in time. These assumptions of the VAR model and aspects of the data, limit our ability to capture the full spectrum of individual differences. I shall discuss each of these limitations and their consequences for modeling individual differences in more detail in the following sections.

8.2.1 *Modeling change*

The stationarity assumption posits that key statistical properties of the time series modeled—such as mean, variance, and network parameters—remain invariant over time. This assumption may be tenable for certain psychological processes observed over specific time scales. For example, personality might exhibit stationarity when measured over days, but not over decades. Similarly, while mood fluctuations are expected throughout the day, they might not be apparent on a second-to-second basis. The assumption of stationarity could also be tenable when the measurement time frame is relatively brief (e.g., two weeks vs. two months). However, conceiving a fully stationary psychological process is challenging in practice and requires researchers to consider whether the psychological process being studied is plausibly stationary. Furthermore, it is often non-stationary processes that we are interested in when modeling clinical data, for example when we want to model a change from a healthy state to a disorder state when comparing pre-treatment to post-treatment data [199].

When encountering non-stationary time series data, a common approach is to detrend the data to align with the stationarity assumption of the VAR

model [41]. This approach involves regressing out linear trends from the data and modeling the residuals instead [41]. Simulation studies showed that detrending improves the specificity and sensitivity of idiographic network models [81]. However, the clinical appropriateness of detrending remains debatable, especially given the interpretative challenges of detrended data and the clinical relevance of the trends themselves. We might expect reasonably stable means and variances during a pre-treatment baseline period, as well as after treatment, but the very aim of treatment itself is to reduce the variances and means for symptoms throughout therapy [199]; removing these trends from the data through detrending might be detrimental [258].

Despite ongoing methodological deliberations, no clear consensus has emerged on managing non-stationary time series. As such, this highlights an important avenue for future research. Simulation studies could aid in understanding the interpretative challenges posed by detrended data. Such studies could help identify conditions under which detrending obscures or preserves clinically relevant information and help determine the clinical implications of detrending the data by examining how the removal of trends affects the identification of clinically meaningful patterns in symptom dynamics. Alternatively, researchers could plan their study in such a way that the psychological process studied is plausibly stationary during the measurement time frame. For example, by collecting data once symptoms have stabilized [216].

Since the stationarity assumption requires constant statistical properties over time, the VAR models as employed in this thesis are not well-suited to capture time-varying characteristics in the data. This limitation directly affects the detectability of heterogeneity in idiographic network models. For example, VAR models may fail to capture changes in individual characteristics over time, such as evolving trends, thereby restricting the identification of inter-individual differences stemming from non-stationary behaviors. Furthermore, the impact of trend *removal* on discerning heterogeneity in idiographic networks is yet to be fully understood and the impact of trends in the data on the behavior of heterogeneity measures currently employed for idiographic network models (such as visual inspection of idiographic network models when estimated with a VAR model, the inspection of the standard deviations of random effects in the multilevel VAR model or INIT) has not been systematically reviewed.

Examining how the removal of trends affects the identification of heterogeneous patterns in the data, and the impact of trends in the data on heterogeneity measures could be explored using simulation studies. For example, the INIT methodology presented in Chapter 3 in this thesis could be employed to determine the effect of trend removal on heterogeneity by comparing idiographic network models estimated from data

containing trends, to (G)VAR models estimated from detrended data, and stationary data under various levels of heterogeneity. This would allow us to see how robust INIT is to model misspecification[1] when simulated data is homogeneous, and how INIT behaves under various levels of heterogeneity when trends are present in the data. Second, the simulation tutorial presented in Chapter 4 makes it possible for researchers to set up a similar simulation study as presented in Chapter 2 to determine the effect of trends in the data on various other tools used to detect heterogeneity in idiographic network models. Exploring these avenues could lead to more nuanced insights into the effects of trend removal techniques on the detectability of heterogeneity in idiographic networks.

Alternatively, we can turn to models that allow for change. To take changes in temporal dynamics into account in idiographic network models, methodologies such as INIT and time-varying (TV-VAR) network models can be employed [137, 134]. The INIT methodology, as detailed in Chapter 3 of this thesis, implicitly allows modeling change in idiographic network models over time. While INIT does not specify the nature of parameter value changes over time, INIT can be used to determine whether a change is necessary to accurately describe the data. This is particularly useful when comparing pre-treatment and post-treatment periods in a clinical context, see also Chapter 5. In such instances, INIT can be used to determine if the pre-treatment and post-treatment data can be modeled using one or two network models. If INIT determines one network model can describe the pre-treatment and post-treatment data better than two network models, this implies there is no change in the network structure for the pre-treatment and post-treatment data. While if INIT determines two network models can describe the data better than one network model, this implies a change in the network structure to describe the pre-treatment and post-treatment data has occurred. However, as INIT employs the VAR model to estimate idiographic network models, this methodology still requires the data to be stationary during the pre-treatment data collection period and the post-treatment data collection period.

There are situations where it might be favorable to entirely abandon the assumption of stationarity and directly model change in order to understand how model parameters evolve over time. For instance, when we are interested in analyzing the effect of treatment, such that we do not merely want to establish that there *is* an effect, but aim to explicitly model the process by which this effect unfolds. In that case, we can model changes over time explicitly using time-varying network models [137, 134]. These models are particularly useful for exploring transitions from healthy to disordered states or the impact of life events or treatment

1 We can view trends in the data as a form of model misspecification

interventions on individuals. However, the estimation of these models requires assumptions about the temporal nature of the process (e.g., do we believe the process continuously changes as a function of time, or does the process display discrete changes?) and specification of the duration over which changes occur. This means that to utilize these models effectively, we should not only have an understanding of the nature of the change we aim to model, but also its temporal scale. Often it is this precise knowledge we are lacking, and the very thing we set out to unravel using these time-varying models; a challenge I will deliberate on in the following section.

8.2.2 *Continuous modeling*

The second assumption of the VAR model is that all measurements are equidistant in time. As a consequence, results obtained from the VAR model heavily rely on the chosen time interval between measurements. Thus, to capture the dynamics of interest in a VAR model, it is essential that the measurement interval aligns with the inherent temporal scale of the process being modeled for an accurate representation of the dynamics of interest [41]. However, theoretical reasons for choosing a time interval in longitudinal study designs are a rare exception, and it is often challenging to ascertain the true temporal scale of psychological processes. Furthermore, practical constraints might make it infeasible to measure at the ideal frequency (e.g., people sleep roughly 7 hours a night). Therefore, the reasons for the chosen time interval in longitudinal study designs are mostly practical in nature [69]. The consequences of using inappropriate time intervals are quite detrimental: we might completely miss detecting important relationships or misestimate effects [3].

Discrepancies between the true temporal dynamics and the measurement interval occur in at least two scenarios. First, when the process we are interested in unfolds more rapidly than the frequency at which we measure. For example, consider modeling indicators of a panic attack such as perceived threat, rapid heartbeat, and shortness of breath or hyperventilation, which are likely to occur within seconds or minutes but are measured in hours. In such instances, we fail to capture the dynamics between these indicators in the temporal predictions of the VAR model. In this case, we might be able to pick up these effects in the contemporaneous network of the VAR model [82]. Although these contemporaneous results might be informative, the key temporal dependencies we are interested in are not modeled, leaving the temporal dependencies between the variables unclear.

Second, the true dynamics might evolve more slowly than the frequency at which we measure. For example, if we were to assess the dynamics

between depression symptoms and personality every hour. In such instances, oversampling might lead to increased noise in the data, as we are capturing more of the random fluctuations that don't necessarily reflect the underlying process. This noise can obscure the true relationship between our variables of interest and induce autocorrelations in the time series. A practical solution, if possible, would be to employ strategies that involve examining different sampling rates in order to identify the optimal sampling frequency. Additionally, slower effects might be better understood using panel designs, which allow for greater temporal spacing between assessments [73, 278].

An inability to accurately capture the process of interest directly also impacts the heterogeneity observed in idiographic network models. Misalignment between the timing of the dynamics and the measurement interval can lead to erroneous interpretations of individual differences in network parameters. Since the strength of temporal relationships depends on the interval between the observations, researchers studying the temporal relationship between, for example, stress and depression, may reach very different conclusions about the nature of these dynamics depending on the timing of these dynamics for each individual and the measurement interval chosen by the researcher. This highlights the possibility that measurements obtained within a certain time frame do not necessarily reflect the true dynamics occurring only within that interval, nor do they occur within the same time frame for each individual [129]. Thus, it is possible that the dynamics of interest are homogeneous for two individuals in terms of their parameters, but do not play out within the same time frame. In such cases, we might incorrectly assume individuals are heterogeneous with respect to their VAR parameters instead of the timing of the process. This emphasizes the necessity of carefully considering the temporal aspects one expects to find in the data based on theory to prevent misinterpretation of individual differences in dynamic processes.

As an alternative to discrete-time VAR models as employed in this thesis, continuous-time (CT) models or differential equations offer a promising avenue worth exploring. CT-VAR models accommodate unequal time intervals between data points, facilitating a deeper understanding of the underlying process by exploring how lagged relationships are expected to vary and evolve as a function of the time interval [236, 120, 256]. Computational models employ differential equations to model the rate of change rather than observed values. One such computational model, the panic model proposed by Robinaugh et al. [230], exemplifies the utility of modeling change over time. Using simulations, they specified computational, testable models of panic disorder with agoraphobia. By specifying a (network) theory of panic disorder and formalizing that theory as a computational model, they explain how individual differences in the propensity

to experience a panic attack occur, the onset of a panic disorder, and the efficacy of treatment, highlighting the fruitful combination of theory and computational models in order to understand and explain individual differences.

Such computational or formal models can combat the aforementioned difficulties in at least two ways. Relatively few theories in psychology explicitly specify the role of time in the conceptualization of psychological phenomena [240]. Computational models can aid in theorizing about the time scale over which the phenomena of interest play out. Explicitly theorizing about the role of time—whether we believe our variables to be relatively stable across a certain time interval, or whether we believe our variables to change across a certain time interval—might also advance our knowledge by creating new research questions that might otherwise have gone unnoticed [226, 31]. Future research should focus on the role of time in psychological processes itself, theorizing on questions such as when relationships between phenomena occur, over what time span they unfold, and what these relationships look like. Second, theorizing about the role of time not only helps us combat some of the challenges of using more complex VAR models such as the TV-VAR, but also provides guidance for determining the appropriate sampling rate for our phenomena of interest.

8.2.3 *Heterogeneity All the Way Down*

A so far understated possible application of the methodology outlined in Chapter 3, is highlighted in the work of Ebrahimi et al. [70]. Ebrahimi et al. [70], employed INIT to investigate heterogeneity in idiographic depression symptom networks among individuals with comparable overall symptom severity. This study uncovered distinct symptom dynamics in 63% of individuals, despite similar severity levels. These results highlight the nuanced and individualized nature of psychopathological experiences beyond mere severity quantification. Expanding upon these findings, it is imperative to further explore the implications of heterogeneity in clinical practice and psychological theory. For instance, acknowledging and understanding individual-specific symptom dynamics by tailoring interventions to the unique psychological patterns of each individual could revolutionize personalized therapeutic approaches. Additionally, these insights challenge existing theoretical frameworks that often generalize psychological processes across individuals, calling for a more nuanced and individualized understanding of psychological phenomena.

The introduction of more sophisticated models, such as the CT-VAR, TV-VAR models, and computational models mentioned in the previous sections, presents both opportunities and challenges in understanding in-

dividual differences. As these advanced models accommodate a greater degree of heterogeneity, they increase the possibility of assuming "heterogeneity all the way down." Yet, as a scientific discipline, our goal is to discern some level of universality or generalizable laws governing these processes. The biggest challenges, therefore, lie in discerning when and how heterogeneity is meaningful and significant to the psychological construct under investigation. This necessitates a more deliberate approach in defining the type of individual differences we are interested in and clarifying our objectives in the context of this heterogeneity.

Computational models can aid in explicitly theorizing about the role of heterogeneity in psychological processes, as they leave room for individual differences and provide opportunities to explore the consequences of heterogeneity. Allowing heterogeneity in our theories (and our models) means individual differences are no longer ignored but modeled explicitly, and it forces us to carefully consider what aspects of the phenomena we determine are important to differ from individual to individual. Second, utilizing such models might help us determine the importance of these individual differences on the phenomena of interest, e.g., why do some individuals develop a psychological disorder when faced with adversity, while others do not? Furthermore, adopting more theory-driven approaches might make us able to determine how individuals differ in terms of what intervention will have the strongest impact on them and underscores the need to understand how individuals differ in their treatment needs [15].

Moreover, the advent of advanced modeling techniques also invites a reconsideration of the type of statistical methodologies employed in psychological research and the type of research questions we wish to ask. The presumption that psychological dynamics are linear, occur at a singular temporal scale, and remain stable over time has thus far dominated the modeling of individuals in psychological research. However, these assumptions likely oversimplify the complexity and variability inherent in psychological processes. Current statistical methods may not be sufficiently equipped to model the complexity of psychological processes unfolding within individuals over time. Consequently, there is a growing need for the development and application of novel techniques that can effectively capture, analyze, and interpret the rich heterogeneity present in psychological data. Such an evolution in methodology, together with a stronger emphasis on theory, and how these two align, not only enhances our ability to understand individual differences but also pushes the boundaries of psychological research, fostering a deeper and more comprehensive understanding of the human psyche.

8.3 CONCLUDING WORDS

In this thesis, I have aimed to assess and understand heterogeneity using idiographic network models. It has become evident to me that we are merely at the beginning of unlocking the potential the source of information time series data brings in order to model individual trajectories. As we continue to unravel the intricacies of psychological dynamics, it becomes increasingly clear that the path to understanding these dynamics lies in embracing their inherent complexity and variability.

Part III

APPENDIX

A

SUPPLEMENTS TO CHAPTER 2

In this appendix, we provide detailed information regarding the simulation study conducted in Chapter 2, present results from additional simulation studies, and perform sensitivity analyses.

A.1 DETAILS SIMULATION STUDY

In this section, we provide more details regarding the simulation study conducted in Chapter 2. The goal of our simulation study was to identify how well equipped different modeling techniques for longitudinal individual data analyses are in separating noise from heterogeneity. We sought to address this objective by simulating illusory heterogeneity, e.g., noise, and aimed to assess the effect of noise on the conclusions about heterogeneity one would draw based on three different estimation approaches. We evaluated three commonly used estimation techniques: graphical VAR (GVAR), multilevel GVAR and Group Iterative Multiple Model Estimation (GIMME). These three estimation techniques fall within the realm of two modeling frameworks. In the following section, we discuss each modeling framework and their occupying estimation techniques in more detail, after which we provide details on parameter values used within the simulation study.

A.1.1 *Modeling frameworks*

Two frameworks have been used within this simulation study: graphical VAR (GVAR) and structural VAR (sVAR). The GVAR framework is used to estimate network models from temporally ordered data of single subjects (GVAR), as well as for temporally ordered data of multiple subjects using multilevel techniques (ml-VAR). The sVAR framework is only used to estimate network models from temporally ordered data of multiple subjects (GIMME).

Both modeling frameworks are reformulations of the VAR model. In the following subsections, these modeling frameworks are explained in more detail. Throughout these explanations three assumptions are important to keep in mind: (1) all models explained here assume the time

series is centered around the within-subject means, (2) is generated from a stationary process, and we assume a lag of order 1[1].

All formulae depicted in this appendix are subject-specific unless noted otherwise.

VAR The VAR is a model used to quantify lagged associations between variables across time and can be denoted as a regression model on the previous measurement occasion. The standard lag-1 VAR can be depicted as:

$$\mathbf{y}_{(t)} = \mathbf{B}\mathbf{y}_{(1-t)} + \boldsymbol{\varepsilon}_{(t)} \tag{21}$$

where $\mathbf{y}$ represents a p-variate vector of scores on p measures for an individual at time t, $\mathbf{B}$ here is the $p \times p$ matrix containing the lagged regression coefficients, and $\boldsymbol{\varepsilon}$ is a p-variate vector of residuals with a mean of zero and a time-invariant $p \times p$ dimensional covariance, here denoted as $\boldsymbol{\Theta}$:

$$\boldsymbol{\varepsilon}_{(t)} \sim N(0, \boldsymbol{\Theta}) \tag{22}$$

GRAPHICAL VAR In the GVAR modeling framework estimation is done by utilizing the inverse of the sample covariance matrix, e.g., the *precision matrix* [89]. The precision matrix is defined as the inverse of the residual covariance matrix of the VAR model:

$$\mathbf{K} = \boldsymbol{\Theta}^{-1} \tag{23}$$

Estimating sparse solutions for both $\mathbf{B}$ and $\mathbf{K}$ per individual can be done through LASSO regularization as implemented in the *graphicalVAR* package [77]. In short, multivariate regression with the covariance estimation (MRCE) algorithm can be used to jointly estimate the contemporaneous ($\mathbf{K}$) and temporal effects ($\mathbf{B}$). Optimizing $\mathbf{B}$ can be done by iterative estimation of temporal coefficients using cyclical-coordinate descent, and optimizing $\mathbf{K}$ can be done by using graphical LASSO [1, 106]. This optimization routine requires two tuning parameters that are governed by one hyperparameter: γ. The best fitting model can be selected via Extended Bayesian Information Criterion (EBIC) [47, 89].

Using the technique described above, from the estimation of the $\mathbf{B}$ and the $\mathbf{K}$ matrix, the PDC's and the PCC's for a given individual can be computed. The standardized lagged associations between variables are denoted as:

$$\mathbf{PDC}_{(y_i, y_j)} = \frac{\beta_{ij}}{\sqrt{\theta_{ii} k_{jj} + \beta_{ij}^2}} \tag{24}$$

1 Note that a lag is a backward shift in time, where a lag of 1 represents an association between time point t and $t-1$, and other lagged associations across time between variables are possible such as a lag of 2 or lag of 3.

where $\mathbf{PDC}_{(y_i,y_j)}$ denotes the standardized partial directed correlations between the variables y_i and y_j at time point t and $\mathbf{B}$ and $\boldsymbol{\Theta}$ are defined as in equation 21 and 2. The PCC's are denoted as:

$$\mathbf{PCC}_{(y_i,y_j)} = -\frac{k_{ij}}{\sqrt{k_{ii}k_{jj}}} \tag{25}$$

where $\mathbf{PCC}_{(y_i,y_j)}$ denotes the PCC between the variables y_i and y_j at time point t.

MULTILEVEL VAR When multiple subjects are measured and one is interested in the dynamics over multiple subjects, we can estimate average effects using multilevel modeling [40, 89]. As with the single case, we can model the data with an individual GVAR model:

$$\mathbf{y}_{(t)} = \boldsymbol{\mu} + \mathbf{B}(\mathbf{y}_{(t-1)} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_{(t)} \tag{26}$$

$$\boldsymbol{\varepsilon}_{(t)} \sim N(0, \boldsymbol{\Theta}) \tag{27}$$

$$\boldsymbol{\Theta}^{-1} = \mathbf{K} \tag{28}$$

where a stationary mean vector for a subject is entered in the model denoted by $\boldsymbol{\mu}$. As with the single subject GVAR, the $\mathbf{B}$ denotes the temporal network containing the lagged effects and $\mathbf{K}$ is the contemporaneous network containing the contemporaneous effects for a given individual. We can now add a second level to the model to estimate a general network structure: *the fixed effects*. Deviations from these fixed effects are called *random effects*. Going into the details of these equations goes beyond the scope of this supplement and an excellent depiction of the multilevel VAR model and the different estimation techniques can be found in [89].

STRUCTURAL VAR Similar to the GVAR modeling framework, the sVAR modeling framework allows for the estimation of contemporaneous relations. Contrary to the GVAR modeling framework, contemporaneous effects are modeled using a directed network instead of an undirected network. The sVAR can be denoted as:

$$\mathbf{y}_{(t)} = \boldsymbol{\Gamma} y_{(t)} + \mathbf{B}_{(1-t)} + \boldsymbol{\varepsilon}_{(t)} \tag{29}$$

where $\boldsymbol{\Gamma}$ is a $p \times p$ matrix containing the contemporaneous associations, y_t represents the p-variate vector of scores on p-variables ate time t, and $\mathbf{B}$, in correspondence with the VAR and GVAR modeling framework, is a $p \times p$ matrix containing the lagged associations. The contemporaneous effects are extracted from the VAR model by forcing the innovations to be uncorrelated [109]

The estimation of sparse sVAR models can be done through the use of unified Structural Equation Model (uSEM) techniques. In short, uSEM

takes an iterative approach where in the first step null values for all matrices in the sVAR model ($\boldsymbol{\Gamma}$, $\mathbf{B}$, $\boldsymbol{\Theta}$) are assigned. Next, parameters that would maximally improve model fit for an individual are freed for estimation. This process is repeated until no additional parameter significantly improves model fit for an individual given all modification indices [109, 170].

GROUP ITERATIVE MULTIPLE MODEL ESTIMATION (GIMME) When multiple subjects are measured and one is interested in the group dynamics as well as in the individual dynamics, one can estimate sVAR models using GIMME. GIMME can be denoted as [11, 213]:

$$\mathbf{y}_{(i,t)} = \left(\boldsymbol{\Gamma}_{(i)} + \boldsymbol{\Gamma}_{(g)}\right) y_{(i,t)} + \left(\mathbf{B}_{(i)} + \mathbf{B}_{(g)}\right) y_{(i,1-t)} + \boldsymbol{\varepsilon}_{(i,t)} \tag{30}$$

where $\mathbf{B}$ and $\boldsymbol{\Gamma}$ have now been split into an estimate for the individual: $\mathbf{B}_{(i)}$ and $\boldsymbol{\Gamma}_{(i)}$, and the group: $\mathbf{B}_{(g)}$ and $\boldsymbol{\Gamma}_{(g)}$.

Like uSEM, GIMME estimates null models for all individuals. However, while uSEM adds parameters that improve model fit for a single individual, GIMME uses information from all individuals to assess if adding parameters will improve model fit. By default, a parameter is added to the group model if it will improve model fit for 75% of the sample.[2] After estimating group-level effects, GIMME will then continue to estimate contemporaneous and temporal individual-level effects in an iterative manner until an optimal fit is ensured. This leads to both group and individual model output.

STRUCTURAL EQUIVALENCE A model equivalence exists between GVAR and sVAR models where one sVAR model corresponds to a single GVAR model. However, one GVAR model cannot be transformed to a single sVAR model, e.g., multiple equivalent sVAR models correspond to a single GVAR model [89, 213]. A sVAR model can be transformed into a GVAR model by first transforming the sVAR model to a VAR model. Once this is done, one can invert the $\boldsymbol{\Theta}$ matrix to derive the precision matrix $\mathbf{K}$. From the precision matrix, a contemporaneous network of partial correlations (PCC) can be computed for each individual, and from the $\mathbf{B}$ the standardized partial directed correlations can be computed. Any direction in the contemporaneous network model is diminished after this procedure, and the initial directionality of the contemporaneous sVAR effects has now manifested itself in the PDC's of the GVAR model [213].

2 This default can be altered according to user preferences.

A.1.2 *Data generating structures*

In this section, we will display parameter values for each of the data-generating structures in the simulation study as described in Chapter 2. In the simulation study, we used three data-generating network structures to generate homogeneous data, e.g., all data within one condition was generated from a single data-generating structure. In addition to network structure, we varied the number of participants, $n \in \{50, 100, 200\}$ and the length of the time series, $t \in \{50, 100, 200, 400\}$; in total, creating a $3 \times 3 \times 4$ design. time series values were chosen to represent plausible time series values, $t \in \{50, 100\}$, and potentially ideal values, $t \in \{200, 400\}$. Each condition was repeated 100 times. Following are the parameter values for each of the three data generating network structures: the synthetic-data, case-data, and Geshwind-data network structures respectively.

SYNTHETIC-DATA NETWORK The synthetic-data network structure is a sparse chain graph containing eight nodes in which each node is connected to its consecutive node, e.g., 1–2, 2–3, etc. In addition to a synthetic-data network structure with eight nodes, a sparse chain graph containing 16 nodes was generated. For the synthetic-data network structure containing 8 nodes, the PDC matrix contains the lagged relationships, and the PCC matrix contains the contemporaneous parameters:

$$\text{PDC} = \begin{bmatrix} .29 & .32 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & .13 & -.36 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .37 & -.47 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .32 & .27 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .44 & -.24 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .44 & -.33 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .50 & .24 \\ .34 & 0 & 0 & 0 & 0 & 0 & 0 & .42 \end{bmatrix},$$

$$\text{PCC} = \begin{bmatrix} 0 & -.27 & 0 & 0 & 0 & 0 & 0 & -.29 \\ -.27 & 0 & .49 & 0 & 0 & 0 & 0 & 0 \\ 0 & .49 & 0 & .22 & 0 & 0 & 0 & 0 \\ 0 & 0 & .22 & 0 & -.36 & 0 & 0 & 0 \\ 0 & 0 & 0 & -.36 & 0 & .52 & 0 & 0 \\ 0 & 0 & 0 & 0 & .52 & 0 & .29 & 0 \\ 0 & 0 & 0 & 0 & 0 & .29 & 0 & -.30 \\ -.29 & 0 & 0 & 0 & 0 & 0 & -.30 & 0 \end{bmatrix}.$$

CASE-DATA NETWORK The case-data network, containing 7 nodes, was estimated from data of one clinical patient measured over a period of two weeks ($t = 47$) using graphical VAR. For more information on the dataset used, see Epskamp et al. [80][3]. The model parameters used under which data was simulated are:

$$\text{PDC} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .41 & -.08 & 0 \\ 0 & 0 & 0 & 0 & 0 & .29 & .15 \\ 0 & 0 & .26 & 0 & 0 & .15 & 0 \end{bmatrix},$$

$$\text{PCC} = \begin{bmatrix} 0 & -.15 & -.05 & .14 & -.08 & 0 & -.22 \\ -.15 & 0 & 0 & -.02 & .26 & 0 & .08 \\ -.05 & 0 & 0 & 0 & 0 & 0 & 0 \\ .14 & -.02 & 0 & 0 & 0 & 0 & -.08 \\ -.08 & .26 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .06 \\ -.22 & .08 & 0 & -.08 & 0 & .06 & 0 \end{bmatrix}.$$

For the PDC matrix and the PCC matrix respectively.

GESCHWIND-DATA NETWORK The Geschwind-data network is originally estimated from data of multiple patients ($n = 129$) measured over a period of 6 days (mean $t = 60$) containing 6 nodes using mlVAR. For more information on the data set, see Geschwind et al. 2011 [111]. The average fixed-effects temporal and contemporaneous network were taken as the network of one individual in our simulation study.Model parameters for the PDC matrix and the PCC matrix are:

3 The dataset used for generating the network used in the current simulation study can be found in the supplementary materials of Epskamp et al. [80].

$$\text{PDC} = \begin{bmatrix} .23 & .06 & -.05 & 0 & -.07 & .09 \\ 0 & .06 & -.03 & 0 & -.05 & .03 \\ -.04 & 0 & .20 & .07 & .07 & 0 \\ 0 & .05 & .06 & .18 & .06 & 0 \\ -.06 & -.05 & .07 & .06 & .19 & 0 \\ .07 & .09 & 0 & 0 & 0 & .18 \end{bmatrix},$$

$$\text{PCC} = \begin{bmatrix} 0 & .14 & -.17 & -.04 & -.28 & .33 \\ .14 & 0 & -.08 & 0 & 0 & .11 \\ -.17 & -.08 & 0 & .17 & .19 & -.16 \\ -.04 & 0 & .17 & 0 & .13 & -.17 \\ -.28 & 0 & .19 & .13 & 0 & -.08 \\ .33 & .11 & -.16 & -.17 & -.08 & 0 \end{bmatrix}.$$

A.2 ADDITIONAL SIMULATION RESULTS

In this section, we present the results from two additional simulation studies to assess the performance of heterogeneity metrics for graphicalVAR, mlVAR, and GIMME. We extend the simulation procedure as discussed in Chapter 2 to (1) a larger network structure containing 16 nodes, and (2) for a network structure generated as a structural VAR (SVAR) (as opposed to a GVAR). The goal of our simulation study was to identify how well equipped different modeling techniques for longitudinal individual data analyses are in separating noise from heterogeneity. With these additional simulation studies we aimed to identify the role of the size of the network, and if modeling techniques such as GIMME perform better when data is generated under the same model structure (a SVAR). This latter addition to our simulation also created the opportunity to conduct sensitivity analyses for GIMME.

A.2.1 *Simulation study: Large synthetic network*

A.2.1.1 *Methods*

In Chapter 2, simulation results are provided for a sparse chain graph containing eight nodes. We included this *synthetic-data* network model as previous simulation studies have shown that network estimation works well under this structure [89]. To assess the influence of the number of

nodes on the performance of heterogeneity metrics used within the different estimation techniques, we constructed a data-generating network in a similar fashion as we did for the synthetic-data network model, now increasing the number of nodes to 16. As for the original synthetic-data network, both the contemporaneous network and the temporal network were simulated to be a chain graph (e.g.,$1-2$, $2-3$, $3-4$, etc.). For a graphical representation of the large resulting network structure see Figure 19. The contemporaneous network has a density (the number of edges divided by the number of possible edges) of 0.13 and an average absolute edge weight of $M = 0.28$. The temporal network has a density of 0.07 with an absolute average edge weight of $M = 0.29$. We refer to this network as the *large-synthetic* network.

The PDC and PCC matrix for the large-synthetic network are:

$$\text{PDC} = \begin{bmatrix}
.37 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.36 \\
.18 & .34 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & .36 & .50 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & .26 & .42 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -.33 & .24 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -.21 & .45 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & .49 & .28 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .37 & .26 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -.28 & .21 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.19 & .07 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .13 & .32 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.33 & .42 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.25 & .23 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.37 & .10 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .4 & .37 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.16 & .39
\end{bmatrix},$$

$$
\text{PCC} = \begin{bmatrix}
0 & .17 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .12 \\
.17 & 0 & -.28 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -.28 & 0 & -.30 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -.30 & 0 & .31 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & .31 & 0 & .40 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & .40 & 0 & -.29 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -.29 & 0 & -.34 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -.34 & 0 & .32 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & .32 & 0 & -.56 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.56 & 0 & -.17 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.17 & 0 & -.29 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.29 & 0 & -.26 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.26 & 0 & -.26 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.26 & 0 & -.18 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -.18 & 0 & .26 \\
.12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .26 & 0
\end{bmatrix}.
$$

This network structure was used as the data-generating network structure for each individual. This means we generated homogeneous data based on the same network structure for each individual. As a result, fluctuations in the *estimated* network model are a combination of sampling variation and the power of the estimation technique. Homogeneous data was generated using the *graphicalVARsim* function from the graphicalVAR package in R [77], which simulates data from a graphical VAR model. Network models were estimated from the simulated data using the R packages: graphivalVAR [77], mlVAR [83], and GIMME [182]. The simulation study was performed in R, version 4.0.5 [220].

As in the original simulation study, we varied the following conditions:

- Sample size: The sample size was varied, $n \in \{50, 100, 200\}$. These values were chosen to represent plausible values in psychological studies.

- Number of observations: The number of observations per subject was varied between $t \in \{50, 100, 200, 400\}$. These values were chosen to represent plausible values, $t \in \{50, 100\}$, and potential ideal values, $t \in \{200, 400\}$.

- Estimation techniques: graphicalVAR, mlVAR, and GIMME. These estimation techniques were chosen as they are used to assess heterogeneity in network analysis.

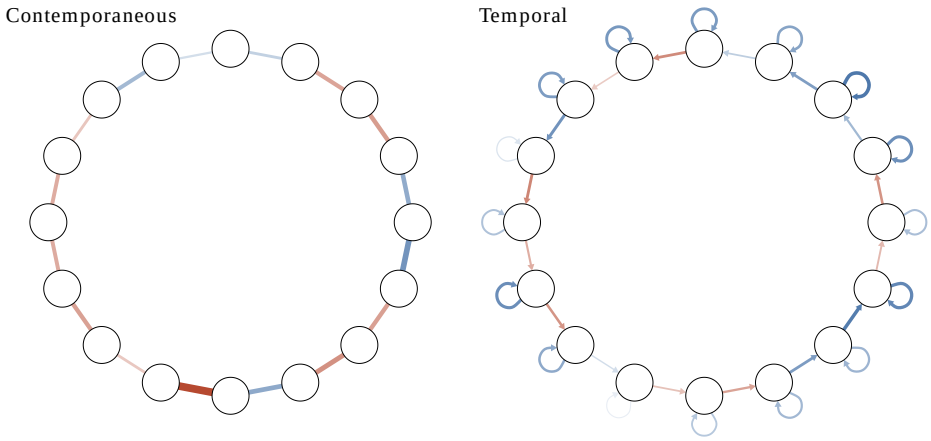
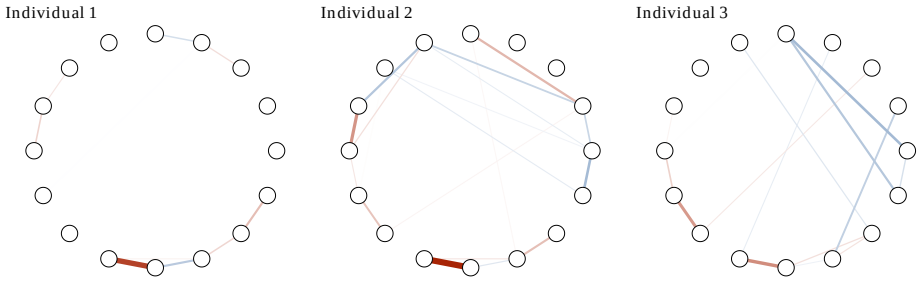


Figure 19: Contemporaneous and temporal data generating network models for each individual. The network model was constructed to be a chain graph, i.e., 1-2, 2-3, etc, containing 16 nodes.

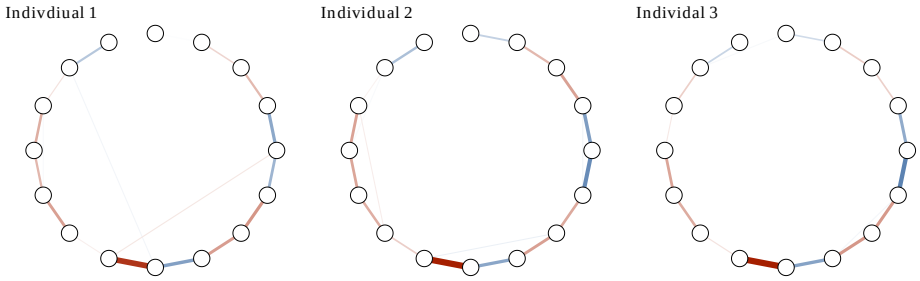
A.2.1.2 *Results*

VISUAL INSPECTION To illustrate the lack of sensitivity and the possible implications for the validity of claims on heterogeneity, we randomly drew 3 individual networks estimated using the graphical VAR method from all simulated networks, see Figure 20. It is important to note the aim of this procedure is to portray how the visual comparison of individual networks can go wrong, specifically when sensitivity is low. When sensitivity is low, which is the case for $t = 50$, the contemporaneous and temporal networks of individual 1, 2, and 3 show dissimilarities, which in practice, can be taken for heterogeneity. When sensitivity increases ($t = 400$), the individual networks of individual 1, 2 and 3 show more resemblances. The at first instance perceived heterogeneity, was in fact illusionary heterogeneity, created by a lack of sensitivity. For a sensitivity analysis for the graphicalVAR method under the large-synthetic network structure see Figure 33.

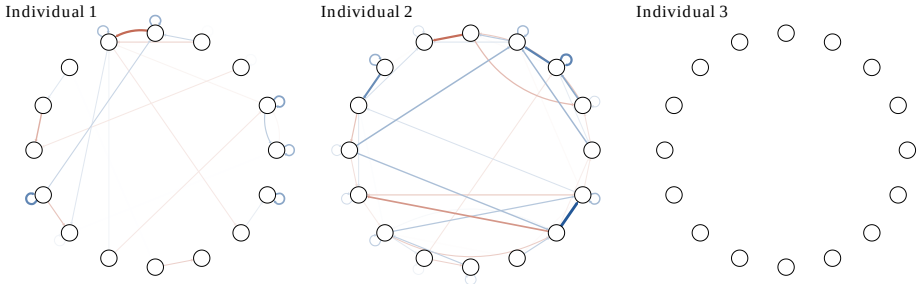
CENTRALITY CORRELATION Correlations between strength centrality showed a more clear-cut result for the large-synthetic generating network, see Figure 21. Results for correlations between strength centrality were ambiguous for the synthetic-data network structure as discussed in Chapter 2 (correlations could be weak due to little variation in strength centrality, meaning low correlation could not be taken as a sign of heterogeneity). However, results for the large-synthetic network showed an increase in correlation when power increases. Correlations between strength centrality measures are moderately strong from $t = 200$ to strong when $t = 400$ onward.



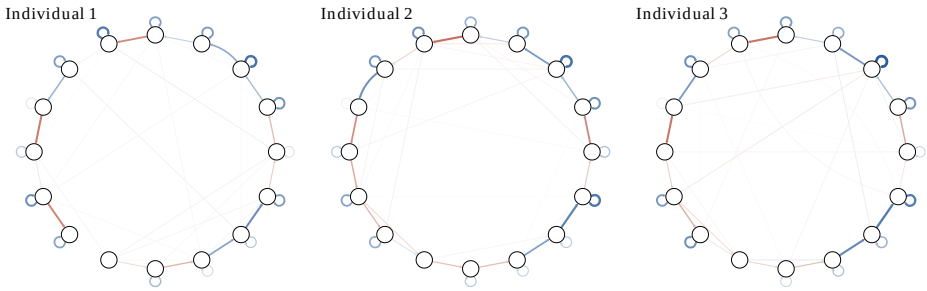
(a) Contemporaneous networks for individual 1, 2, and 3 when $t = 50$



(b) Contemporaneous networks for individual 1, 2, and 3 when $t = 400$



(c) Temporal networks for individual 1, 2, and 3 when $t = 50$



(d) Temporal networks for individual 1, 2, and. 3 when $t = 400$

Figure 20: Three randomly selected individual networks estimated using the graphical VAR method, generated under the same large-synthetic network structure (for a visualization of the network see Figure 19 panel (b)) for two different time points ($t = 50$ and $t = 400$). Panel (a) shows the three individual contemporaneous networks when $t = 50$, panel (b) for $t = 400$, panel (c) shows their corresponding temporal networks when $t = 50$, and panel (d) for $t = 400$.

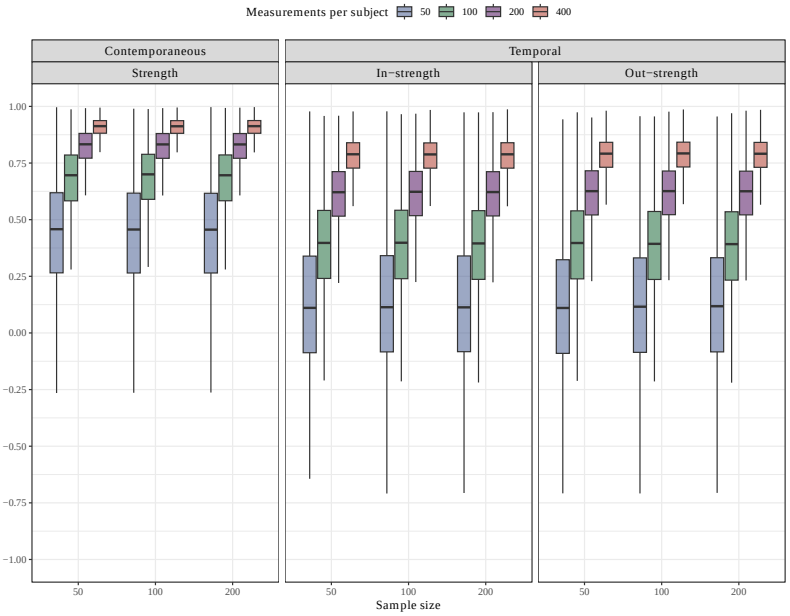


Figure 21: Correlation between strength centrality as computed on estimated network models with graphical VAR.

STANDARD DEVIATIONS OF RANDOM EFFECTS For mlVAR, we inspected the standard deviations of random effects for the contemporaneous and temporal effects. We take a cut-off of one standard deviation for the edges to determine whether large differences were present. See Figure 22 for a graphical representation of the density distributions of standardized edge weights. As is apparent from Figure 22, most edge weights fall within the range of one standard deviation. As the number of time points increases, the number of edge weights that fall outside this range become less. More heterogeneity is detected for temporal networks than for contemporaneous networks. This may be due to the number of parameters; more parameters need to be estimated for the temporal networks, which requires more power. For a sensitivity analysis for multilevel VAR see Figure 36.

PERCENTAGE OF HOMOGENEOUS EDGES To the best of our knowledge, for GIMME no cut-off to determine the amount of heterogeneity has been recommended in the literature. In Chapter2, we proposed to inspect the percentage of homogeneous edges as a number of the total estimated edges in the network as an indication of the amount of homogeneity estimated. It is important to keep in mind that individuals were simulated to be homogeneous. Therewith, if a *high* percentage of homogeneous edges were detected this can be taken as an indication that the

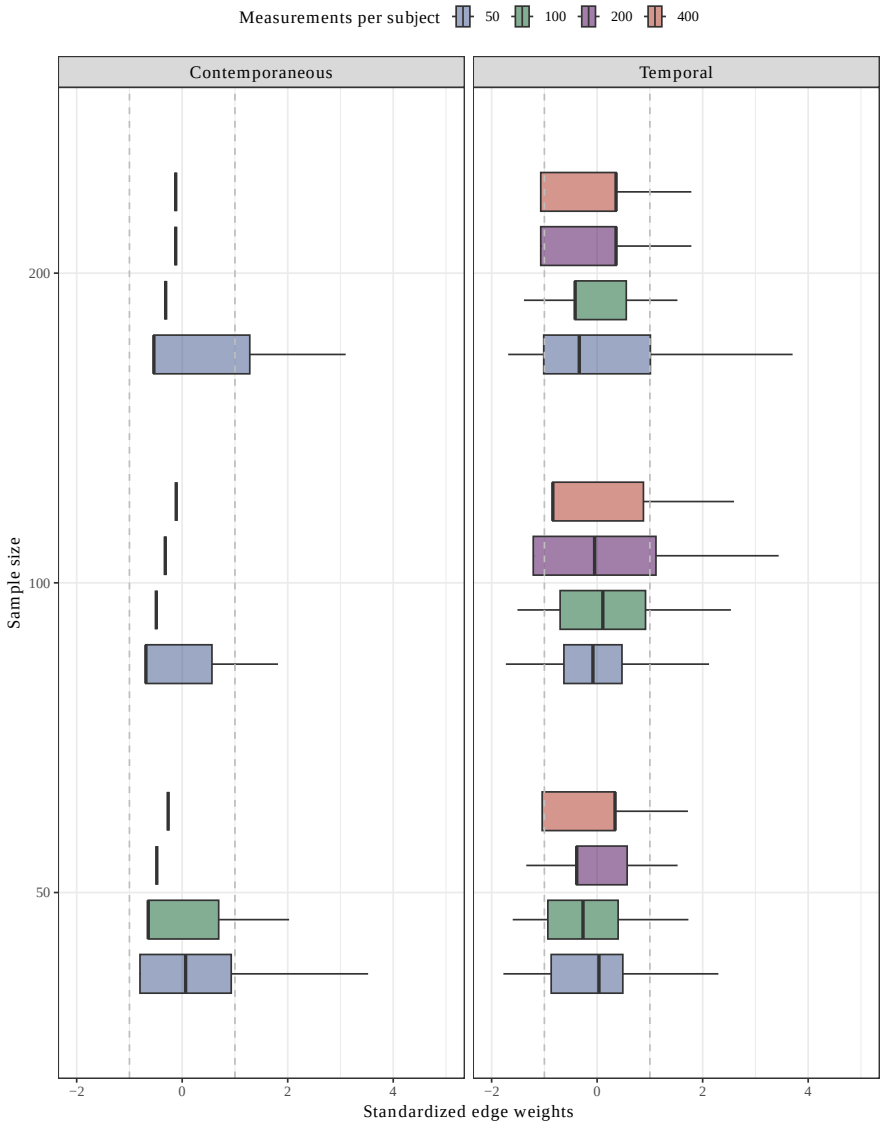


Figure 22: Standardized standard deviation edge weights of random effect as estimated with multilevel VAR. Dotted line represents the cut-off of one standard deviation. Edges above or below this cut-off are seen as truly heterogeneous.

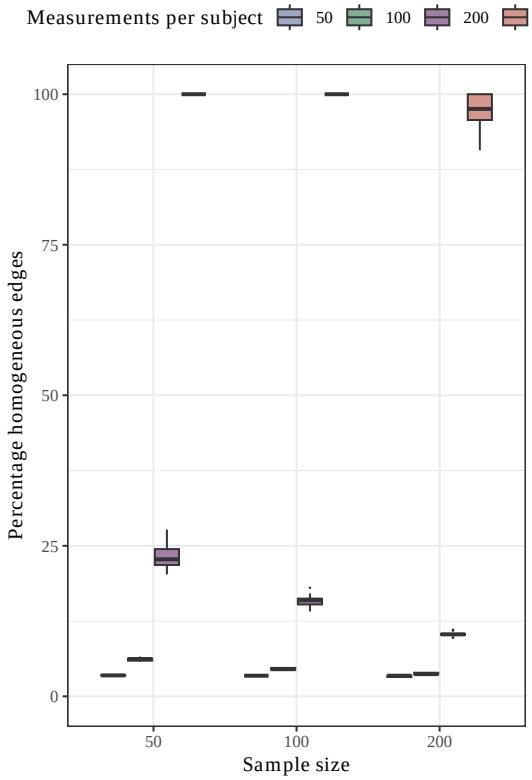


Figure 23: Percentage of homogeneous edges for GIMME when data was simulated under the large-synthetic-network condition.

technique performs well. GIMME performed well when $t = 400$. When $n = 50$ or $n = 100$, and $t = 400$ GIMME's output indicates complete sample homogeneity, see Figure 23.

A.2.2 *Simulation study: SVAR network*

To make a better comparison between the estimation techniques used in our simulation study, as well as to inspect the performance of GIMME in terms of sensitivity and precision, we included a simulation study using a SVAR model as the data generating model.

A.2.2.1 *Methods*

We used the following parameter values to generate data from a SVAR model for a 4-node network:

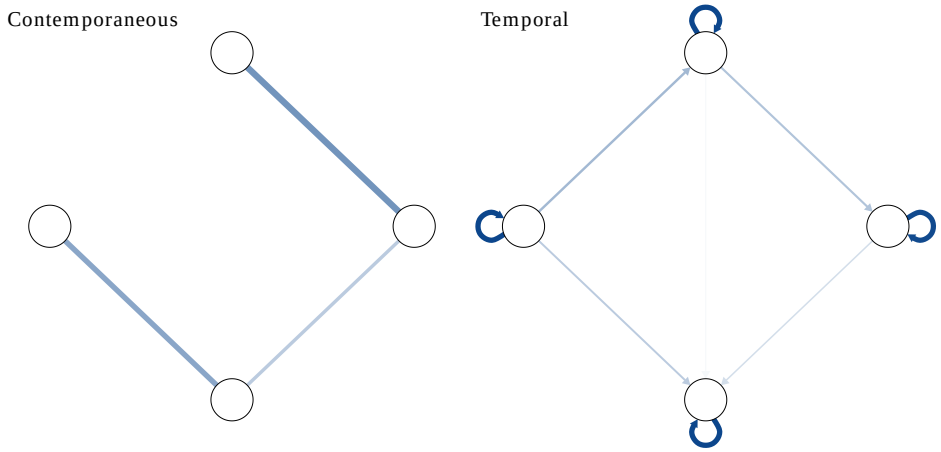


Figure 24: Contemporaneous and temporal data generating network models for each individual. Network model was constructed in accordance with previous simulation studies (see Park et al. [213] and is referred to as the synthetic-structural-VAR network.

$$\Gamma = \begin{bmatrix} 0 & .30 & 0 & 0 \\ 0 & 0 & .15 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & .25 & 0 \end{bmatrix}, B = \begin{bmatrix} .60 & 0 & 0 & 0 \\ 0 & .60 & 0 & 0 \\ 0 & 0 & .60 & 0 \\ .20 & 0 & 0 & .60 \end{bmatrix}, I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

These parameter values were chosen as they resemble parameter values found in empirical studies and as these values have been used in previous simulation studies to evaluate the performance of GIMME [12, 109, 213].

As a SVAR model corresponds to a single GVAR model, we can transform the SVAR model to a GVAR model [213]. From the SVAR matrix we derived the following coefficient matrices for the GVAR metric:

$$PDC = \begin{bmatrix} .50 & .16 & .02 & 0 \\ 0 & .49 & .09 & 0 \\ 0 & 0 & .50 & 0 \\ .19 & 0 & .14 & .50 \end{bmatrix}, PCC = \begin{bmatrix} 0 & .29 & 0 & 0 \\ .29 & 0 & .14 & 0 \\ 0 & .14 & 0 & .24 \\ 0 & 0 & .24 & 0 \end{bmatrix}.$$

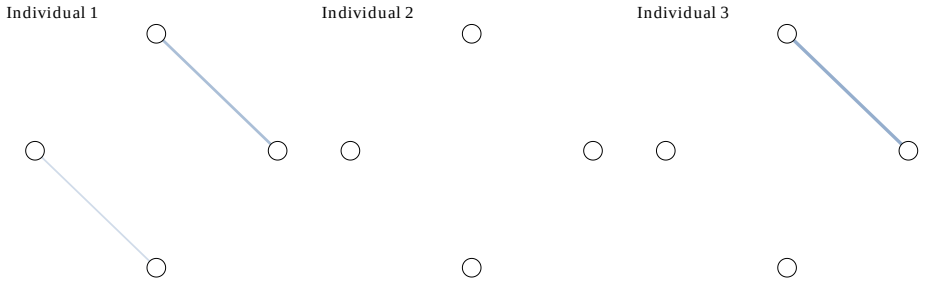
For a graphical representation of the graphical VAR network structure see Figure 24. The contemporaneous network has a density of 0.5 and an average absolute edge weight of $M = 0.22$ and the temporal network has a density of 0.42 with an average absolute edge weight of $M = 0.12$. We refer to this network as the *synthetic-structural-VAR* network.

We followed the same procedure as previously described: we used the synthetic-structural-VAR network as the data generating network structure for each individual. This means we generated homogeneous data based on the same network structure for each individual and fluctuations in the *estimated* network model are the result of a combination of sampling variation and power of the estimation technique. Homogeneous data was generated using the *graphicalVARsim* function from the graphicalVAR package in R [77]. Network models were estimated using the R packages: graphivalVAR [77], mlVAR [83], and GIMME [182]. As data were generated from a structural VAR model, sensitivity metrics (i.e., correlation, sensitivity, specificity) could be calculated for each of the estimation techniques used. The simulation study was performed in R, version 4.0.5 [220].

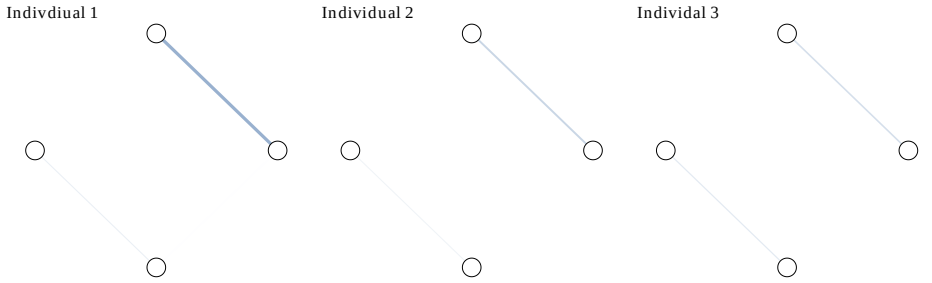
A.2.2.2 *Results*

VISUAL INSPECTION To illustrate the lack of sensitivity and its possible implications for the validity of claims regarding heterogeneity between individuals, we randomly drew 3 individual networks estimated using the graphical VAR method, see Figure 25. The aim of this procedure is to convey how visual comparison of individual networks is prone to error, specifically when sensitivity is low. When sensitivity is low, which is the case when $t = 50$, the contemporaneous and temporal networks of individual 1, 2, and 3 show a large portion of dissimilarities. When sensitivity increases (as the number op time points per individual increases $t = 400$), the individual networks of individual 1, 2 and 3 are identical in terms of estimated edges; heterogeneity is only perceived in terms of edge weights. For temporal networks, there are still some dissimilarities in terms of estimated edges even when $t = 400$. However the differences diminished in comparison to when $t = 50$.

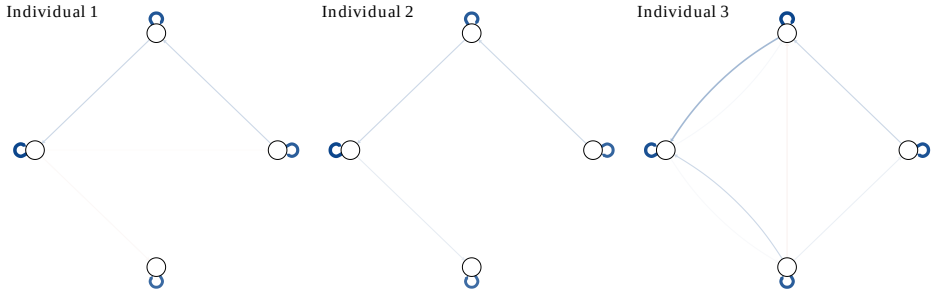
CENTRALITY CORRELATION Correlations between strength centrality showed a great deal of variance under the structural-VAR-data generating network structure, see Figure 26. For both the contemporaneous as well as the temporal network structures, the correlation between strength values for the individual network models increases when t increases. However, for the contemporaneous network structure, while the median correlation increases, the spread for the correlation is large. Even when $t = 400$ correlations still range -0.10—0.88. Meaning, it is still possible to find a negative, or small correlation between individuals networks strength centrality when power is high. For the temporal network, the spread of correlations becomes smaller as the number of time points increases. Interestingly, this effect is stronger for the in-strength correlation, than for the out-strength



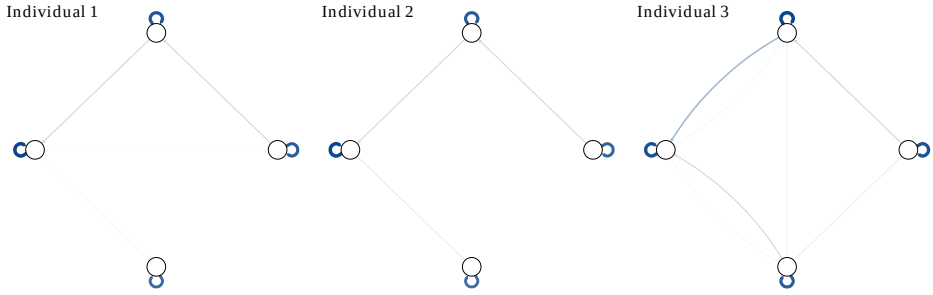
(a) Contemporaneous networks for individual 1, 2, and 3 when $t = 50$



(b) Contemporaneous networks for individual 1, 2, and 3 when $t = 400$



(c) Temporal networks for individual 1, 2, and 3 when $t = 50$



(d) Temporal networks for individual 1, 2, and. 3 when $t = 400$

Figure 25: Three randomly selected individual networks estimated using the graphical VAR method, generated under the same structural-VAR-data network structure (for a visualisation of the network see Figure 25 panel (b)) for two different time points ($t = 50$ and $t = 400$). Panel (a) shows the three individual contemporaneous networks as estimated using graphical VAR when $t = 50$, panel (b) for $t = 400$, panel (c) shows their corresponding temporal networks when $t = 50$, and panel (d) for $t = 400$.

correlation. For a sensitivity analysis for the graphicalVAR method under the synthetic-structural-VAR network see Figure 31.

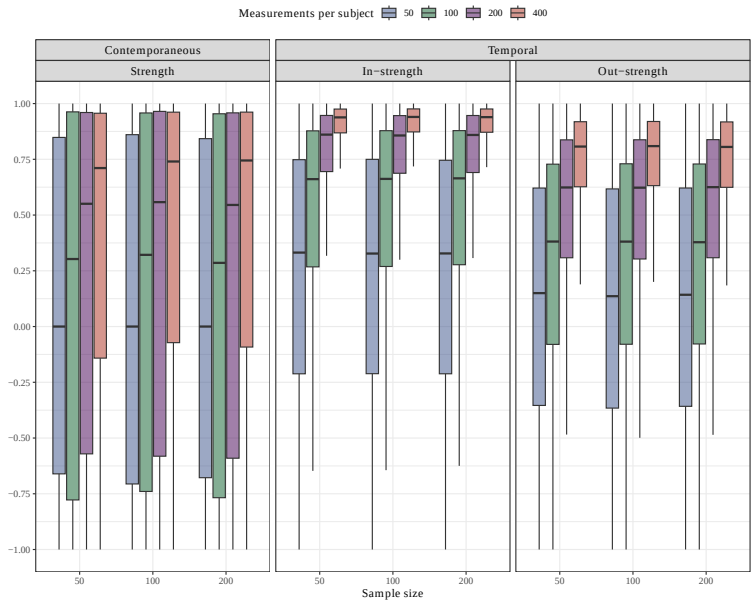


Figure 26: Correlation between strength centrality as computed on estimated network models with graphical VAR.

STANDARD DEVIATION OF RANDOM EFFECTS For mlVAR, we inspected the standard deviations (SD) of random effects for contemporaneous and temporal effects. We took a cut-off of one standard deviation for the edges to determine whether large differences were present [40]. See Figure 27 for a graphical representation of the density distributions of standardized edge weights for the SD network.

Although, some edge weights fall outside the range of the one standard deviation cut-off, most edge weights fall within the range of one standard deviation. Larger heterogeneous effects are detected for temporal networks than for contemporaneous networks. In the previous section we provided a possible reason for this result: it may be due to the number of parameters; more parameters need to be estimated for the temporal networks, which requires more power. For a sensitivity analysis for multilevel VAR see Figure 34.

PERCENTAGE OF HOMOGENEOUS EDGES For GIMME, to the best of our knowledge, no cut-off has been recommended in the literature. We proposed to inspect the percentage of homogeneous edges as a number of the total estimated edges in the group network as an indication of the amount of homogeneity estimated. GIMME performed well when

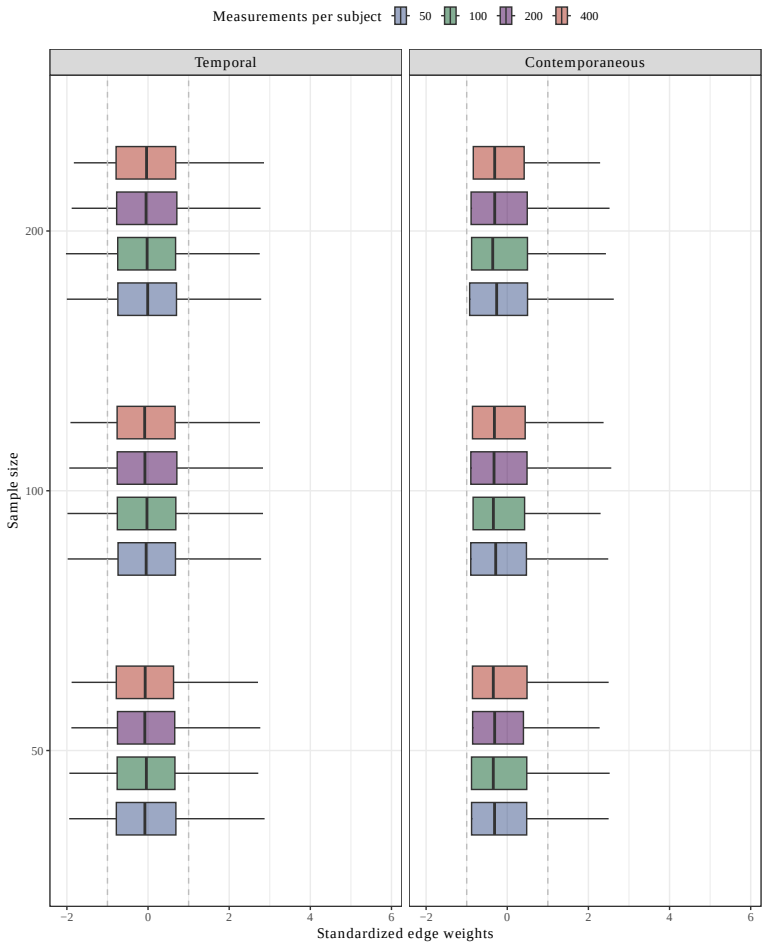


Figure 27: Standardized standard deviation edge weights of random effect as estimated with multilevel VAR. The dotted line represents the cut-off of one standard deviation. Edges above or below this cut-off are seen as large heterogeneous effects.

$t = 400$, see Figure 28. When $n = 50$ and $t = 400$, GIMME's output indicates complete sample homogeneity. Interestingly, GIMME performs worse in terms of percentage of homogeneous edges when n increases, while power increases when n and t increase, see Figure 37 for a sensitivity analysis for GIMME.

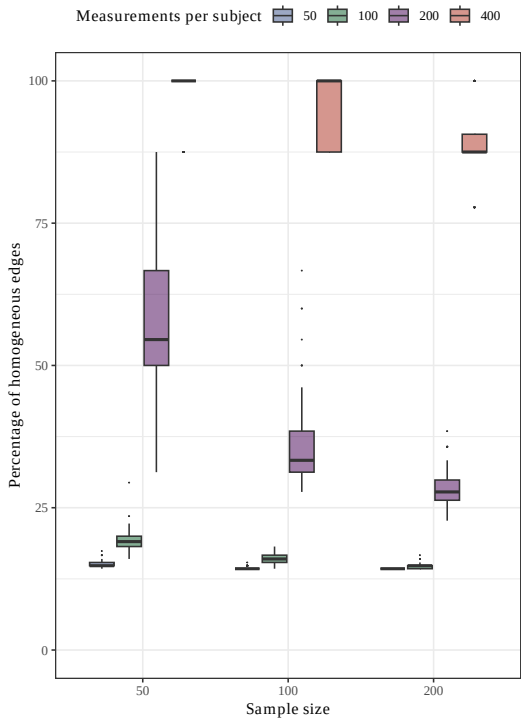
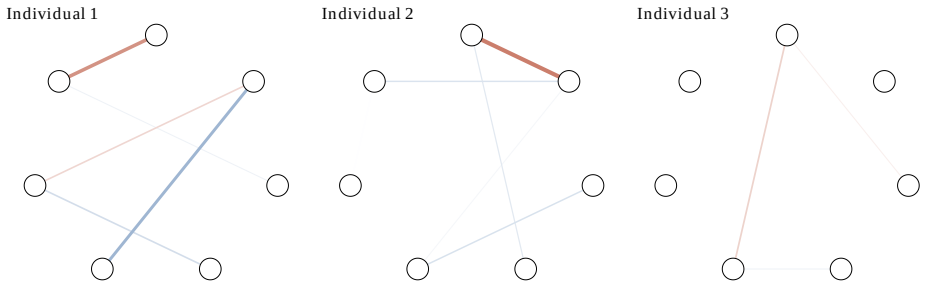


Figure 28: Percentage of homogeneous edges for GIMME.

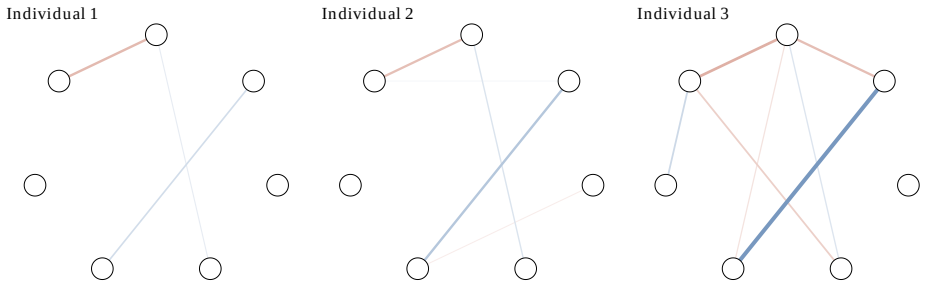
A.3 VISUAL INSPECTION

In Chapter 2, we illustrated the consequences for the lack of sensitivity for the validity regarding claims on heterogeneity in for the synthetic data set by randomly drawing three individual networks from the 5,000 simulated networks to visually compare. In this section, we further illustrate the lack of sensitivity and its implications for the validity of claims on heterogeneity, for the case-data and Geschwind-data network conditions.

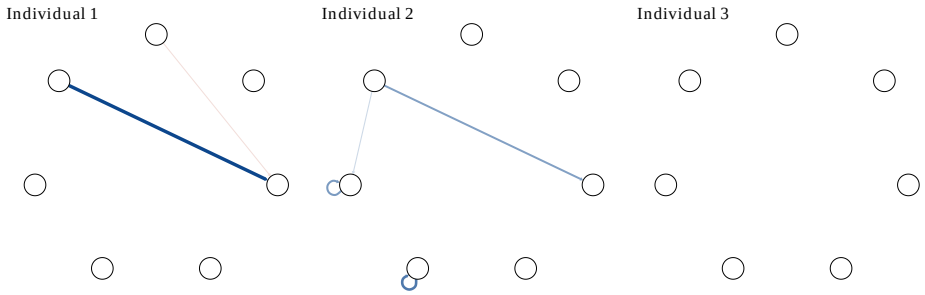
Where in the synthetic-data condition, the perceived heterogeneity vanished when visually inspecting the estimated network structures for $t = 400$ (see panel (b) and (d) Figure 20), there still seemed to be considerable heterogeneity present when visually inspecting the estimated network structures for the case-data and Geschwind-data network structures when $t = 400$ (see Figure 29 and Figure 30 panel b and d).



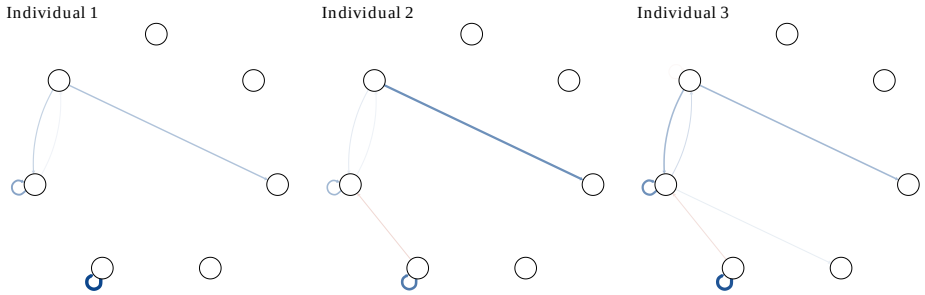
(a) Contemporaneous networks for individual 1, 2, and 3 when $t = 50$



(b) Contemporaneous networks for individual 1, 2, and 3 when $t = 400$

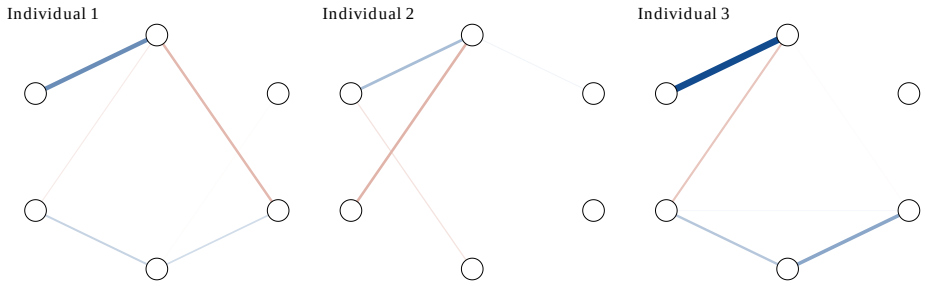


(c) Temporal networks for individual 1, 2, and 3 when $t = 50$

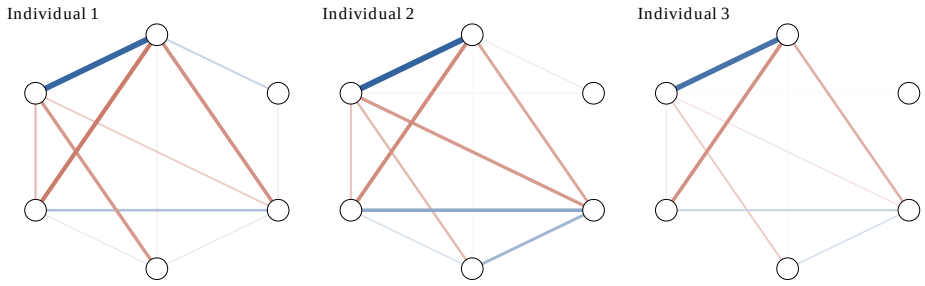


(d) Temporal networks for individual 1, 2, and. 3 when $t = 400$

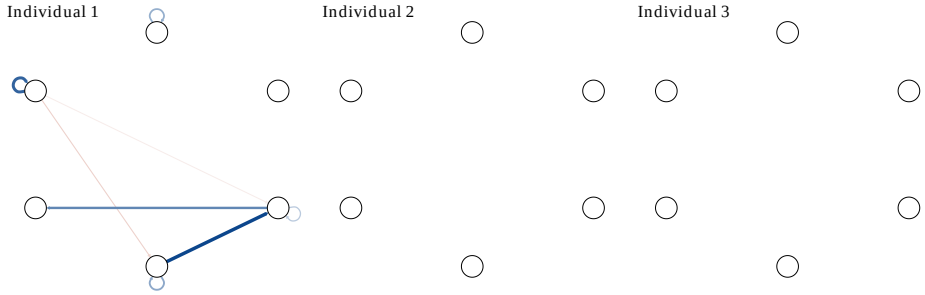
Figure 29: Output from graphical VAR. Three individual networks (contemporaneous and temporal) were generated under the same case-data network structure. Panel (a) shows the contemporaneous networks for $t = 50$, (b) for $t = 400$, (c) their corresponding temporal networks with $t = 50$, and (d) for $t = 400$.



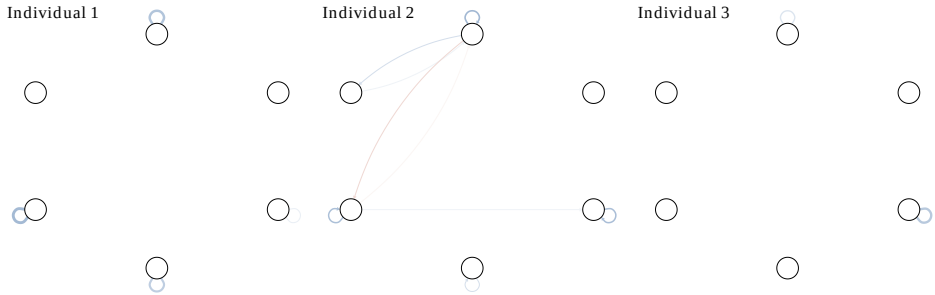
(a) Contemporaneous networks for individual 1, 2, and 3 when $t = 50$



(b) Contemporaneous networks for individual 1, 2, and 3 when $t = 400$



(c) Temporal networks for individual 1, 2, and 3 when $t = 50$



(d) Temporal networks for individual 1, 2, and. 3 when $t = 400$

Figure 30: Output from graphical VAR. Three individual networks (contemporaneous and temporal) were generated under the same case-data network structure. Panel (a) shows the contemporaneous networks for $t = 50$, (b) for $t = 400$, (c) their corresponding temporal networks with $t = 50$, and (d) for $t = 400$.

A.4 SENSITIVITY ANALYSIS

A.4.1 *Graphical VAR*

Sensitivity result for the graphical VAR model.

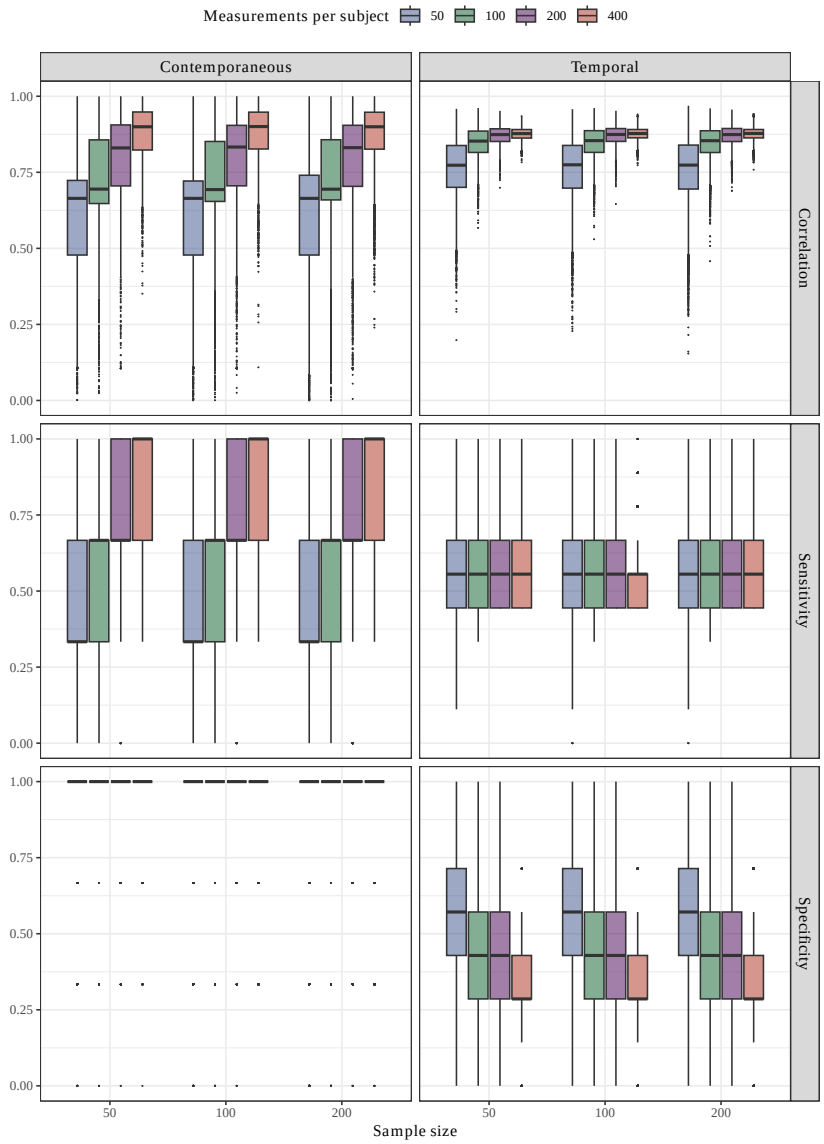


Figure 31: Sensitivity results for graphical VAR. Correlation, sensitivity, and specificity for contemporaneous and temporal individual networks.

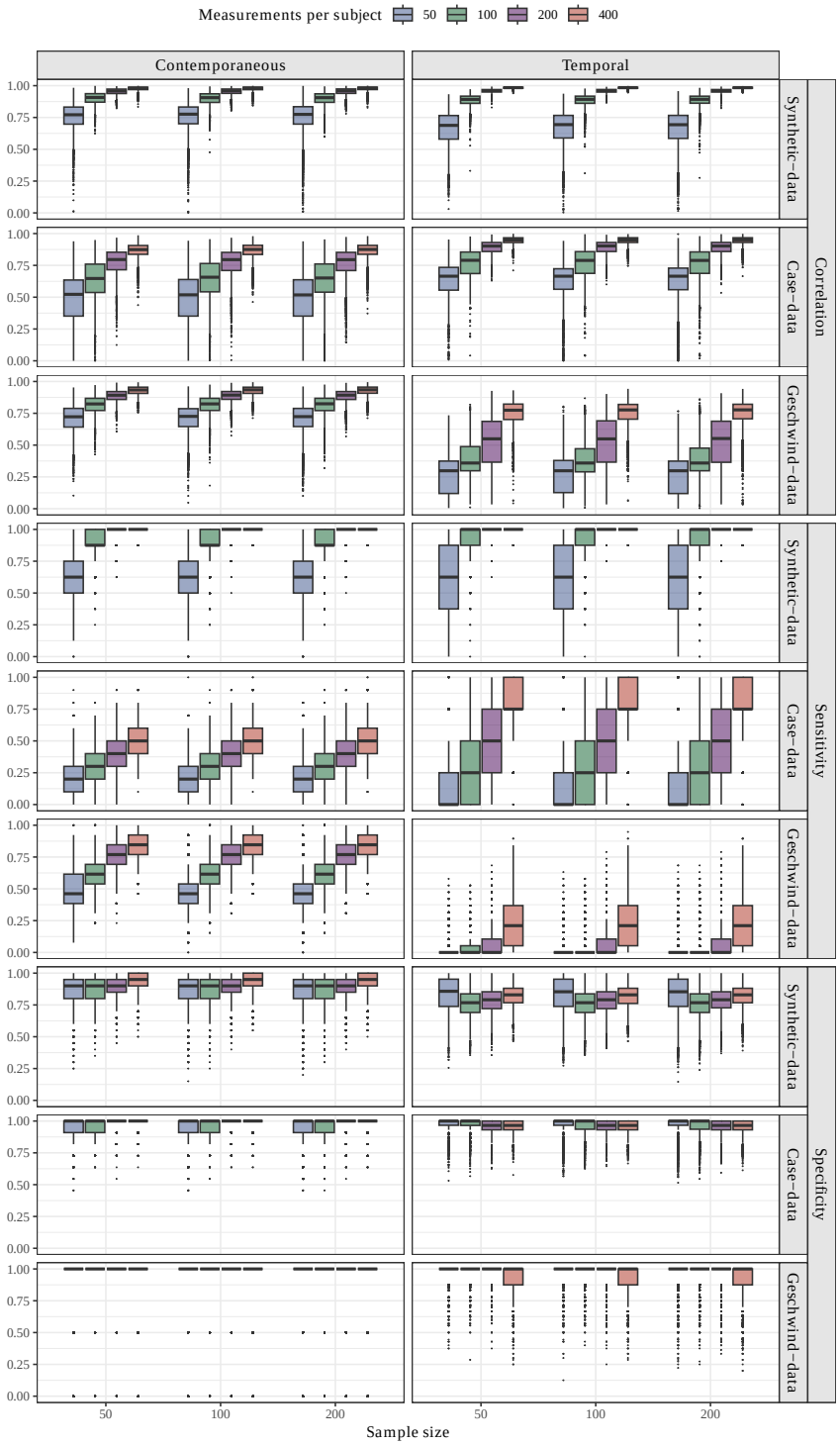


Figure 32: Sensitivity results for graphical VAR. Correlation, sensitivity, and specificity for contemporaneous and temporal individual networks for synthetic-data, case-data, and Geschwind-data network structures.

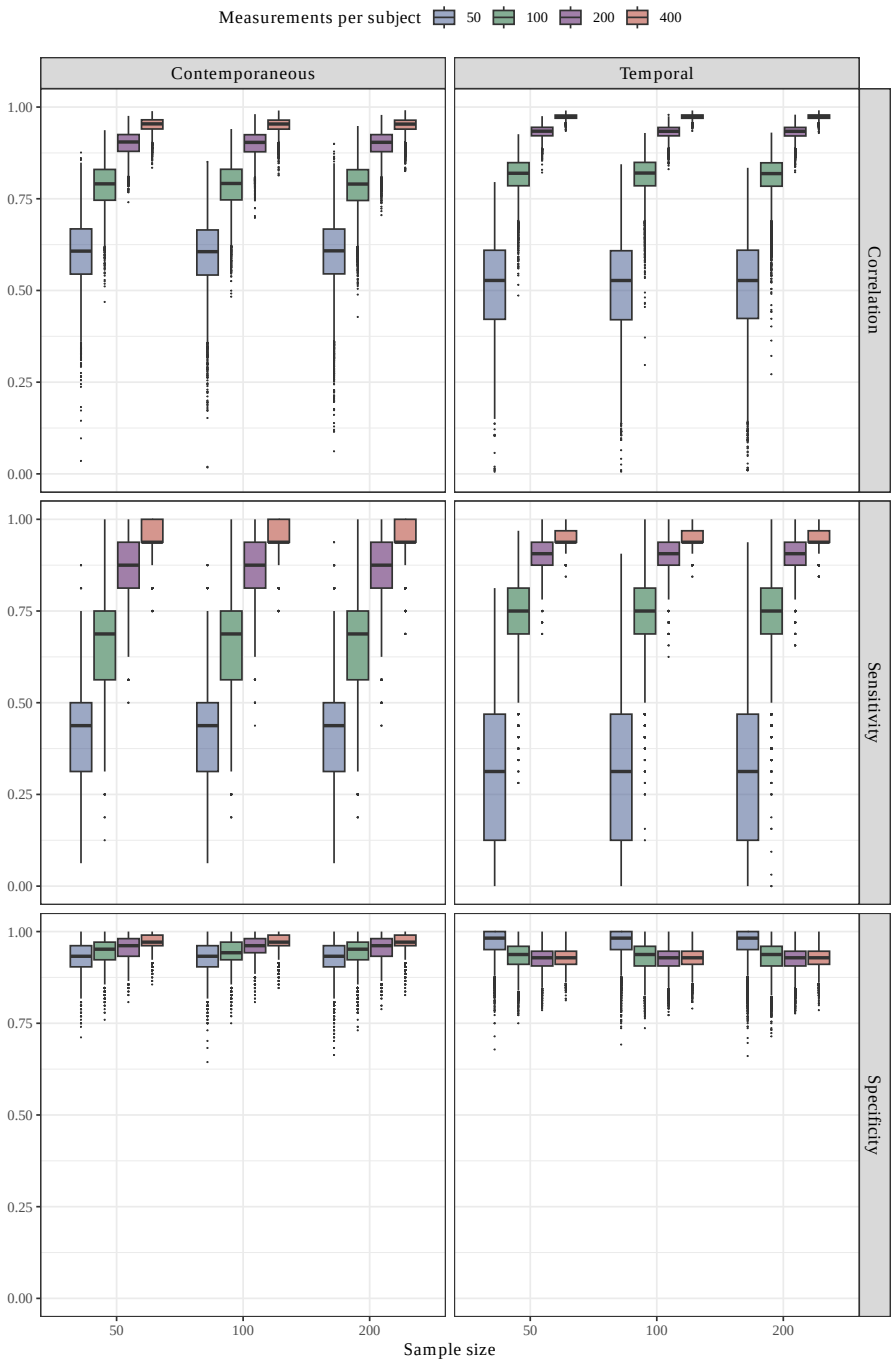


Figure 33: Sensitivity results for graphical VAR when data was simulated under the large-synthetic-network condition. Portrayed are the correlation, sensitivity, and specificity for contemporaneous and temporal individual networks.

A.4.2 *Multilevel VAR*

Sensitivity result for the multilevel VAR model.

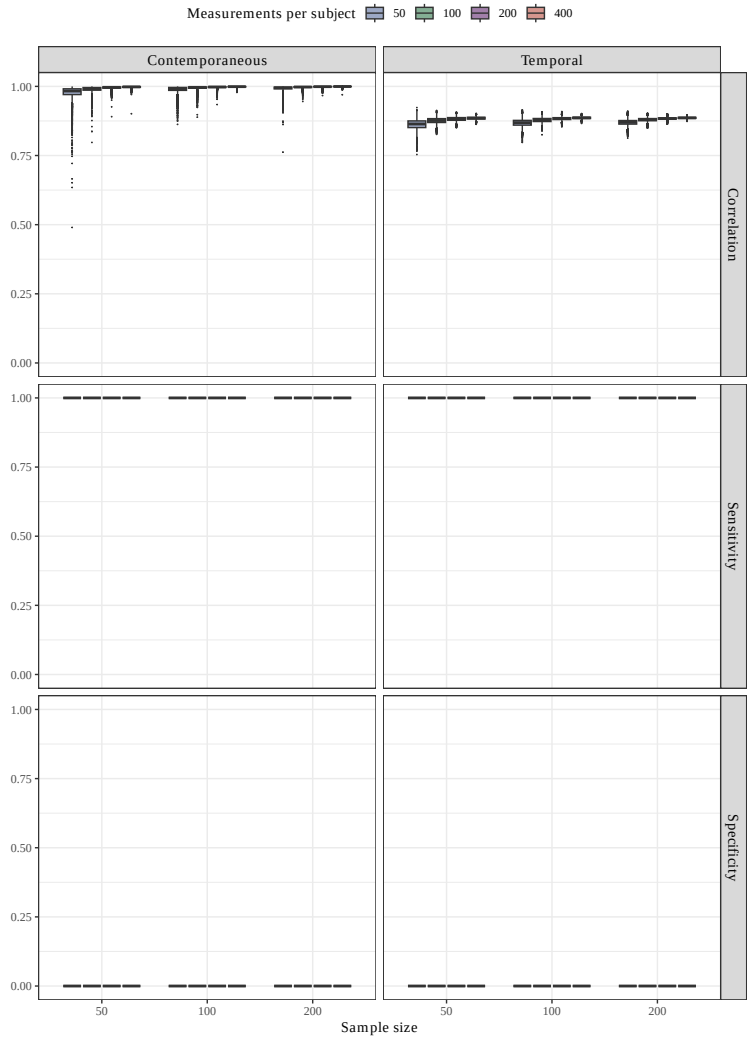


Figure 34: Sensitivity results for mlVAR. Correlation, sensitivity, and specificity for contemporaneous and temporal individual networks.

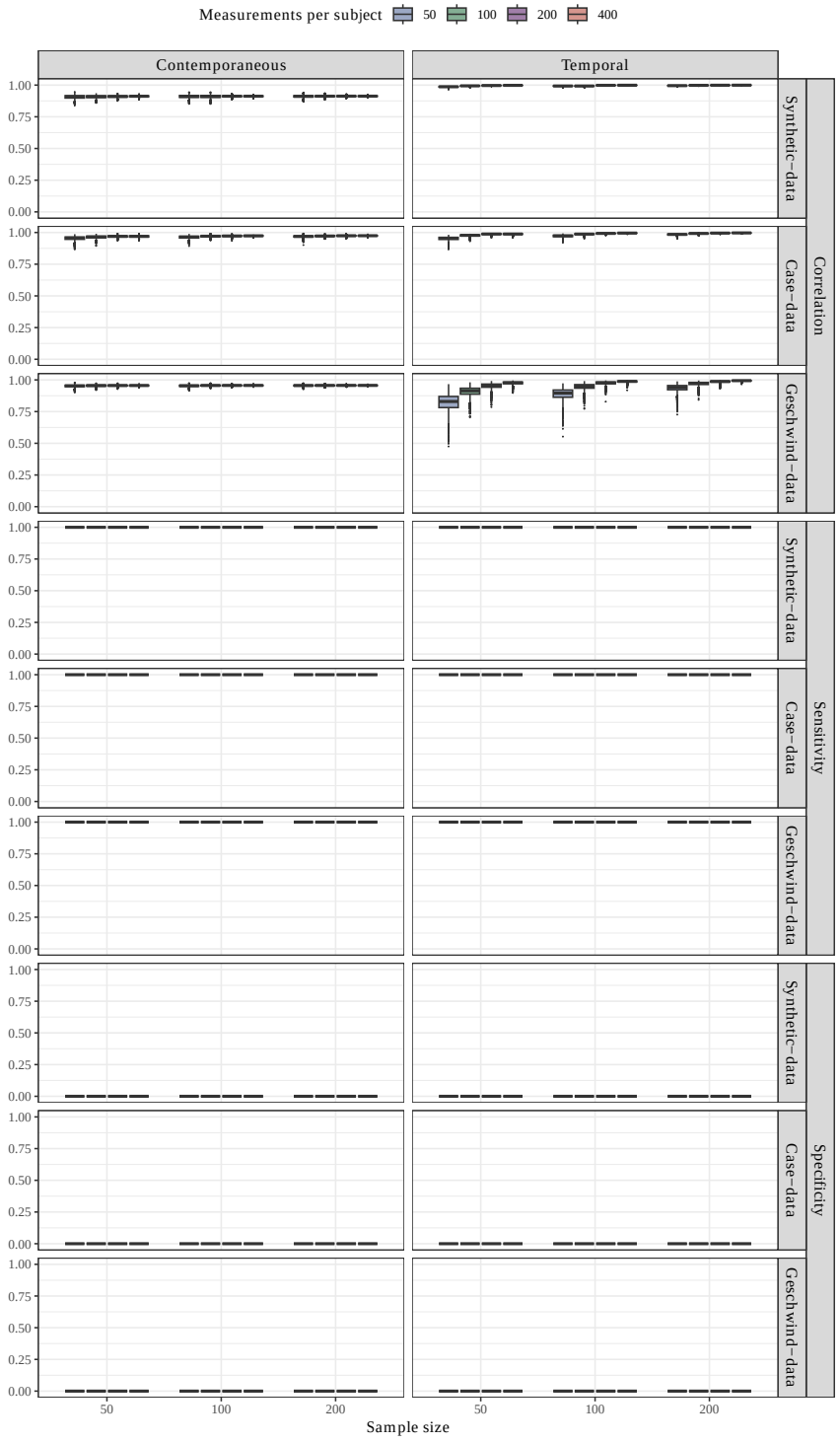


Figure 35: Sensitivity results for multilevel VAR. Correlation, sensitivity, and specificity are depicted for contemporaneous and temporal individual networks for synthetic-data, case-data, and Geschwind-data network structures. In line with previous simulation study results, sensitivity is 100% and specificity 0% as a result of all edges being estimated (see Epskamps et al., 2018. Therewith, the correlation between the data generating network model and the estimated network model is a better indication of the power of the estimation technique.

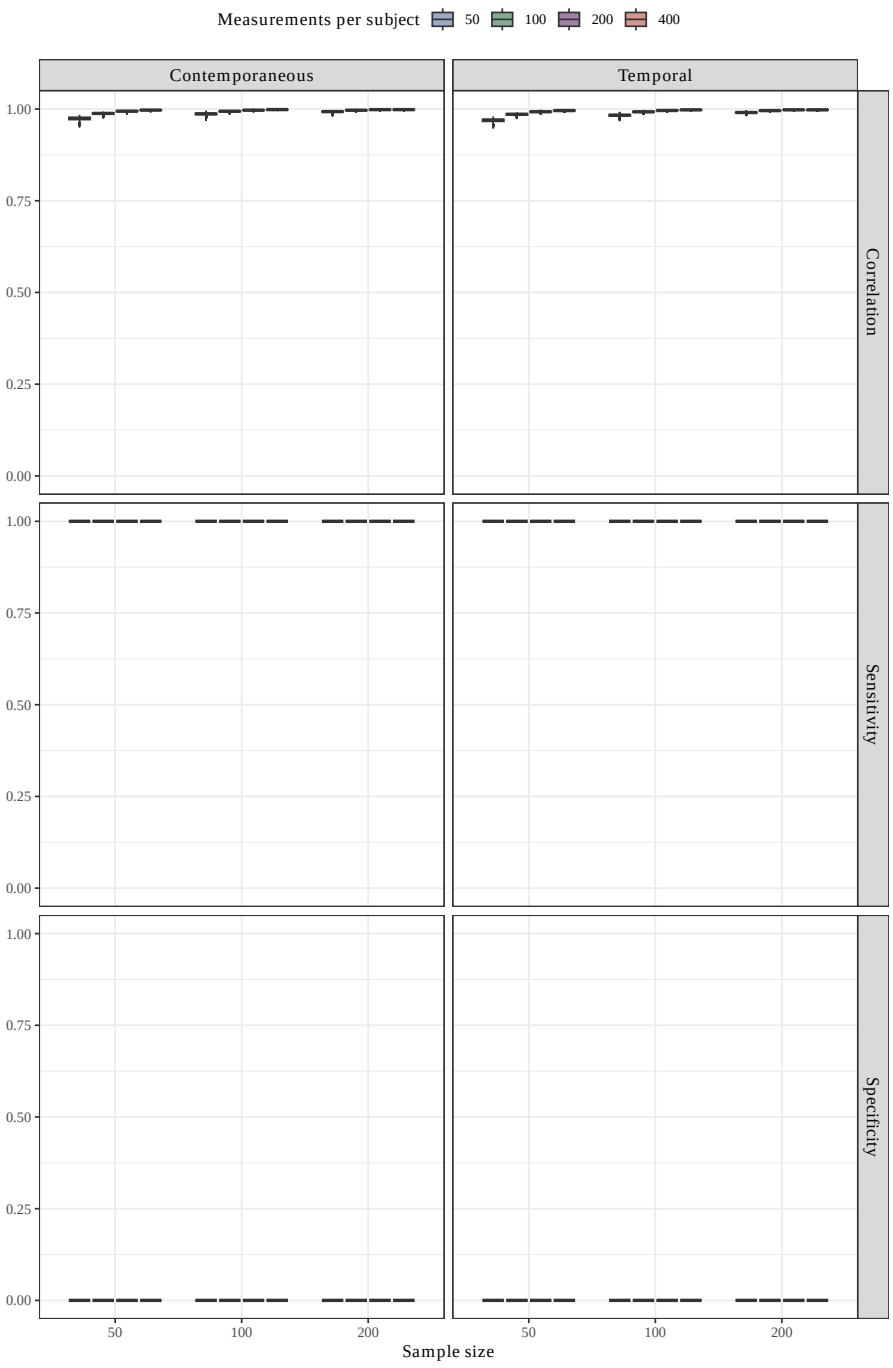


Figure 36: Sensitivity results for mlVAR when data was simulated under the large-synthetic-network condition. Portrayed are the correlation, sensitivity, and specificity for contemporaneous and temporal individual networks.

A.4.3 *GIMME*

Sensitivity result for GIMME.

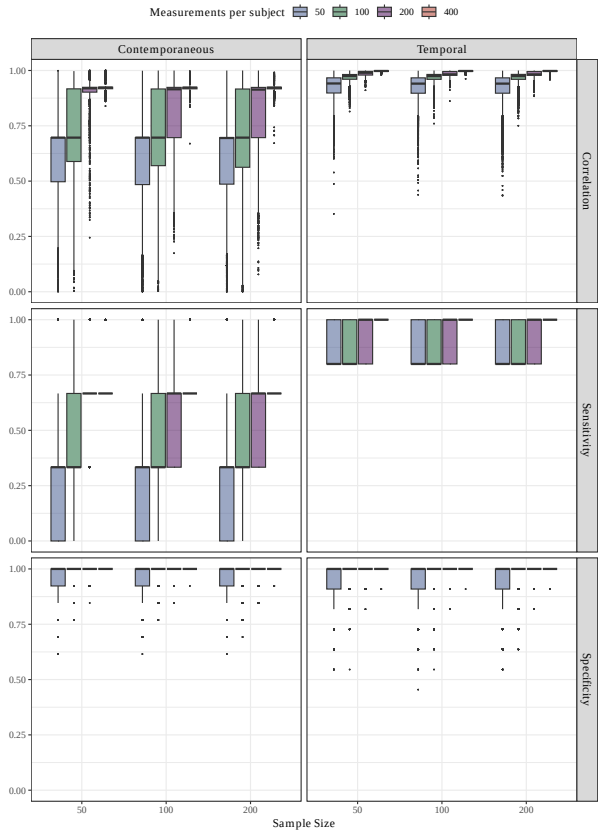


Figure 37: Sensitivity results for GIMME. Correlation, sensitivity, and specificity are depicted for contemporaneous and temporal individual networks as simulated under a SVAR network and estimated using GIMME. More details on the simulation procedure and parameter values of the data generating network structure can be found in supplement B. Important to note here is that the diagonal of the temporal matrix is taken into account when computing the sensitivity to make the comparison between sensitivity results for GIMME, graphicalVAR, and mlVAR more straightforward.

B

SUPPLEMENTS TO CHAPTER 3

In this appendix we provide a tutorial on the use of the *INIT* R package introduced in Chapter 3, details regarding the simulation study presented in Chapter 3, and present full results for all simulation conditions discussed in Chapter 3.

B.1 INIT TUTORIAL

In this tutorial we briefly show how the *INIT* R package can be used to test for (in)equalities in idiographic networks. Installing the current developmental version can be done by using the remotes-package:

```
remotes::install_github("RiaHoekstra/INIT")
```

After installing this developmental version, it is important to load INIT using the library function:

```
library("INIT")
```

INIT comes with an example dataset that can be loaded as follows:

```
# Load data:
data <- INIT::data
```

We will use this dataset to exemplify how INIT can be used to test for differences between idiographic networks.

INIT has several arguments that can be specified. However, for INIT to run with the default arguments, it simply needs (1) a dataset containing the data of the two individuals one wants to compare specified in the data argument, (2) the name of the column with the variable that identifies each individual, (3) and the column names indicating the variables to be modeled in the network:

```
INIT(data = data,
     idvar = colnames(data)[1],
     vars = colnames(data)[2:7])
```

Using the default settings of INIT equality constraints will be placed on both the temporal and contemporaneous saturated (i.e., fully connected) networks. INIT automatically provides an output summary of the results indicating which model fits the data best, displaying the AIC of the $\mathcal{M}_{\text{homogeneous}}$ and $\mathcal{M}_{\text{heterogeneous}}$, the difference between these two as

ΔAIC, a summary of the arguments used in INIT, and a summary of the estimated model.

One can change the default arguments by specifying them in the function. In addition to the `data`, `idvar`, and `vars` argument, a `dayvar` indicating the column name of the day variable and a `beepvar` indicating the beep variable can be specified. The `estimator` argument indicates the type of estimator to be use such as `"ML" "FIML"`. `network_type` indicates if a `"saturated"` or `"pruned"` network should be estimated. If a pruned network is estimated, one can choose the alpha level of the pruning (default sets to .05) by specifying `alpha`. `homogeneity_test` indicates the type of (in)equalities to be tested, where `"overall"`, refers to test both contemporaneous and temporal network structures for (in)equalities, `"contemporaneous"` to test contemporaneous network structures for (in) equalities, and `"temporal"` to test temporal network structures for (in) equalities. Specifying `save_models` as `TRUE`, means the estimated network models will be saved. One can easily change one or multiple of the default arguments, for example:

```
INIT(data = data,
     idvar = colnames(data)[1],
     vars = colnames(data)[2:7],
     estimator = "FIML",
     network_type = "pruned",
     homogeneity_test = "contemporaneous",
     save_models = TRUE)
```

To inspect output in more detail, results from INIT can be saved in an object:

```
# Save INIT results in an object:
res <- INIT(data = data,
            idvar = colnames(data)[1],
            vars = colnames(data)[2:7],
            save_models = TRUE)

# Inspect summary of results:
res

# Inspect estimated network structures:
res$network
```

This object can be saved, and when `save_models = TRUE`, the estimated temporal and contemporaneous networks can be inspected.

B.2 DETAILS SIMULATION STUDY

When simulating data under a GVAR model one needs to specify a B matrix (encoding the auto-regressive and cross-lagged effects that are used to create a temporal network) and a precision matrix K (which is used to construct the contemporaneous network). In this appendix, we present the B matrix and K matrix used as the true-data-generating mechanism in the simulation study.

B.2.1 *Synthetic simulation condition*

In the synthetic simulation condition, the data-generating mechanism was modeled as a sparse chain graph. In a chain graph, each node is connected to the subsequent node. Parameter values for the B matrix were sampled from a normal distribution with a mean of 0.30 and $sd = 0.10$ Parameter values for the k matrix were generated from a normal distribution with a mean of 0.20 and $sd = 0.26$. As parameter values were sampled, this means in each of the 1,000 generations, the model matrices were slightly different. Below is an example of the B and K model matrix used in the chain-graph condition:

$$B = \begin{bmatrix} .15 & 0 & 0 & 0 & 0 & -.58 \\ .18 & .29 & 0 & 0 & 0 & 0 \\ 0 & .38 & .23 & 0 & 0 & 0 \\ 0 & 0 & -.17 & .27 & 0 & 0 \\ 0 & 0 & 0 & -.48 & .19 & 0 \\ 0 & 0 & 0 & 0 & .26 & .24 \end{bmatrix},$$

$$K = \begin{bmatrix} 0 & -.35 & 0 & 0 & 0 & .34 \\ -.35 & 0 & .35 & 0 & 0 & 0 \\ 0 & .35 & 0 & .29 & 0 & 0 \\ 0 & 0 & .29 & 0 & .37 & 0 \\ 0 & 0 & 0 & .37 & 0 & .30 \\ .34 & 0 & 0 & 0 & .30 & 0 \end{bmatrix},$$

B.2.2 *Empirical simulation condition*

For the empirical simulation condition, we simulated data based on the two networks shown in the illustrative data example in the main part of

Chapter3, see Figure 10. Data on emotion dynamics and symptoms was collected by prof. Nierenberg & McNally and came from two outpatients diagnosed with bipolar disorder who were undergoing maintenance treatment at the Dauten Family Center for Bipolar Treatment Innovation at Massachusetts General Hospital (MGH).

On their time series data a GVAR model was estimated using the *psychonetrics* package in R [74]. After fitting a saturated network model, model selection was performed in the form of pruning with $\alpha = .05$. The resulting sparse network structures were used as the data-generating mechanism for the empirical simulation condition. Data was generated based on equation 10 using the `graphicalVARsim` function in R. Below the B matrix encoding the temporal network and the K matrix are shown for both participants x and y respectively.

$$B_x = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & .13 \\ 0 & .20 & 0 & 0 & 0 & .04 \\ 0 & 0 & .13 & 0 & 0 & 0 \\ 0 & 0 & 0 & .11 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$K_x = \begin{bmatrix} 6.16 & -1.90 & 1.47 & 0.87 & 0 & -2.21 \\ -1.90 & 3.56 & 0.96 & 0 & 0 & -0.37 \\ 1.47 & 0.96 & 2.88 & 0 & 0 & 0 \\ 0.87 & 0 & 0 & 1.95 & 0 & 0.55 \\ 0 & 0 & 0 & 0 & 1.230 & 0.62 \\ -2.21 & -0.37 & 0 & 0.55 & 0.62 & 3.76 \end{bmatrix},$$

$$B_y = \begin{bmatrix} 0 & 0 & -.08 & 0 & 0 & 0 \\ .14 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .23 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$K_y = \begin{bmatrix} 5.02 & -0.55 & 0 & 0.61 & 1.83 & -2.09 \\ -0.55 & 1.73 & 0.39 & 0 & 0.33 & 0 \\ 0 & 0.39 & 1.67 & -0.28 & -0.61 & 0 \\ 0.61 & 0 & -0.28 & 1.44 & 0 & 0 \\ 1.83 & 0.33 & -0.61 & 0 & 6.26 & 3.52 \\ -2.09 & 0 & 0 & 0 & 3.52 & 5.94 \end{bmatrix}.$$

The simulation code to generate the results from this appendix and Chapter 3 is available on the Open Science Framework (OSF) repository at: `https://osf.io/bqhgc/` [142].

B.3 FULL RESULTS FOR SATURATED NETWORK STRUCTURES

Figure 38, 39 and 40 contain the χ^2 test rejection rate, the AIC rejection rate and the BIC rejection rate when applying INIT to compare saturated network structures under all rewiring probabilities. The rewiring probability is taken as an indication of the amount of heterogeneity. The rewiring probability is taken as an indication of the amount of heterogeneity. Simulation was performed under several network conditions: two empirical networks of participant x and y, indicated by 'empirical x' and 'empirical y' and two synthetic networks in which the number of nodes was varied. 'synthetic 6' corresponds to the synthetic network containing 6 nodes, and 'synthetic 12' corresponds to the synthetic network containing 12 nodes. The color of the lines corresponds to the placement of equality constraints. Red indicates equality constraints have been placed on both the contemporaneous and temporal network structure. Blue indicates equality constraints have been placed on the contemporaneous network structure and green indicates equality constraints have been placed on the temporal network structure.

B.4 FULL RESULTS FOR PRUNED NETWORK STRUCTURES

Figure 41, 42 and 43 contain the χ^2 test rejection rate, the AIC rejection rate and the BIC rejection rate when applying INIT to compare pruned network structures under all rewiring probabilities. The rewiring probability is taken as an indication of the amount of heterogeneity. The rewiring probability is taken as an indication of the amount of heterogeneity. Simulation was performed under several network conditions: two empirical networks of participant x and y, indicated by 'empirical x' and 'empirical y' and two synthetic networks in which the number of nodes

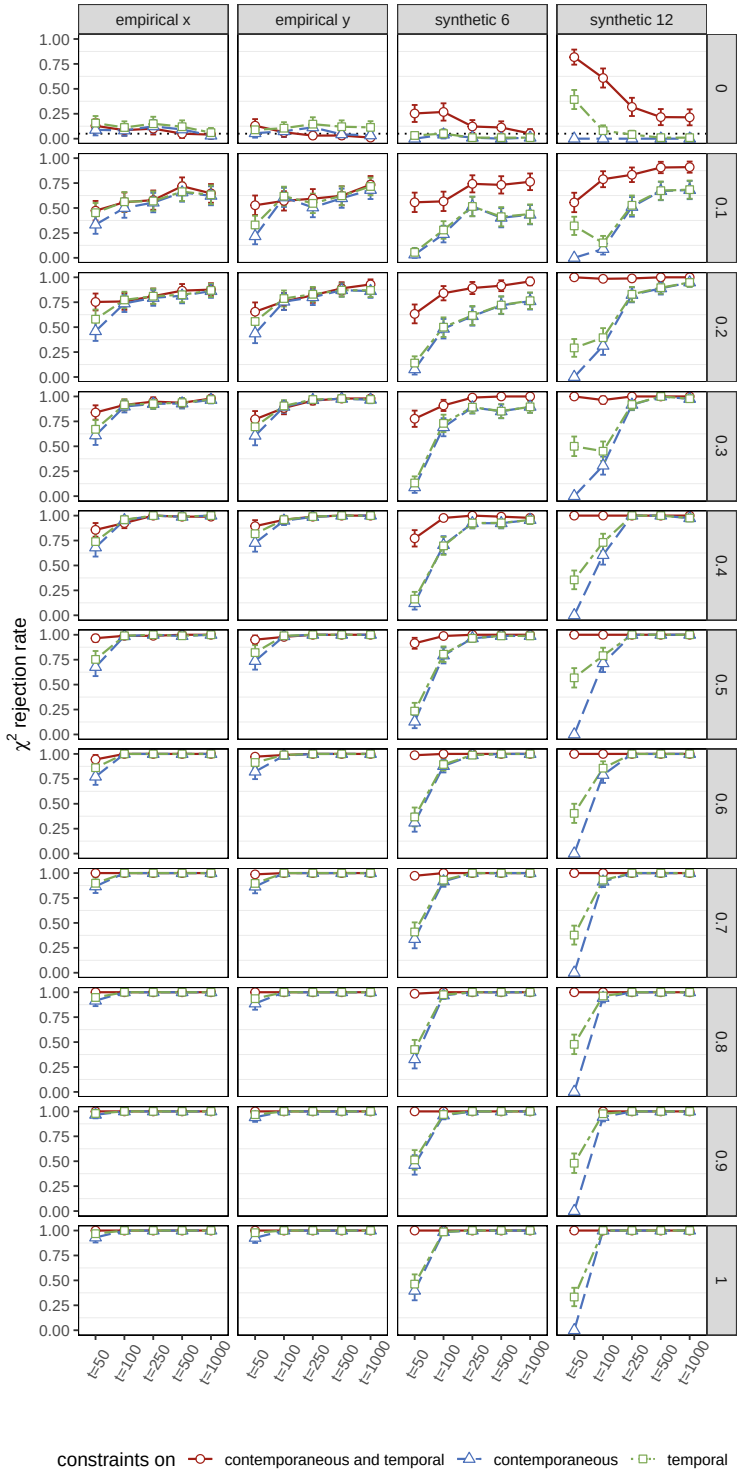


Figure 38: χ^2 test rejection rate results when applying INIT to compare saturated network structures under all rewiring probabilities.

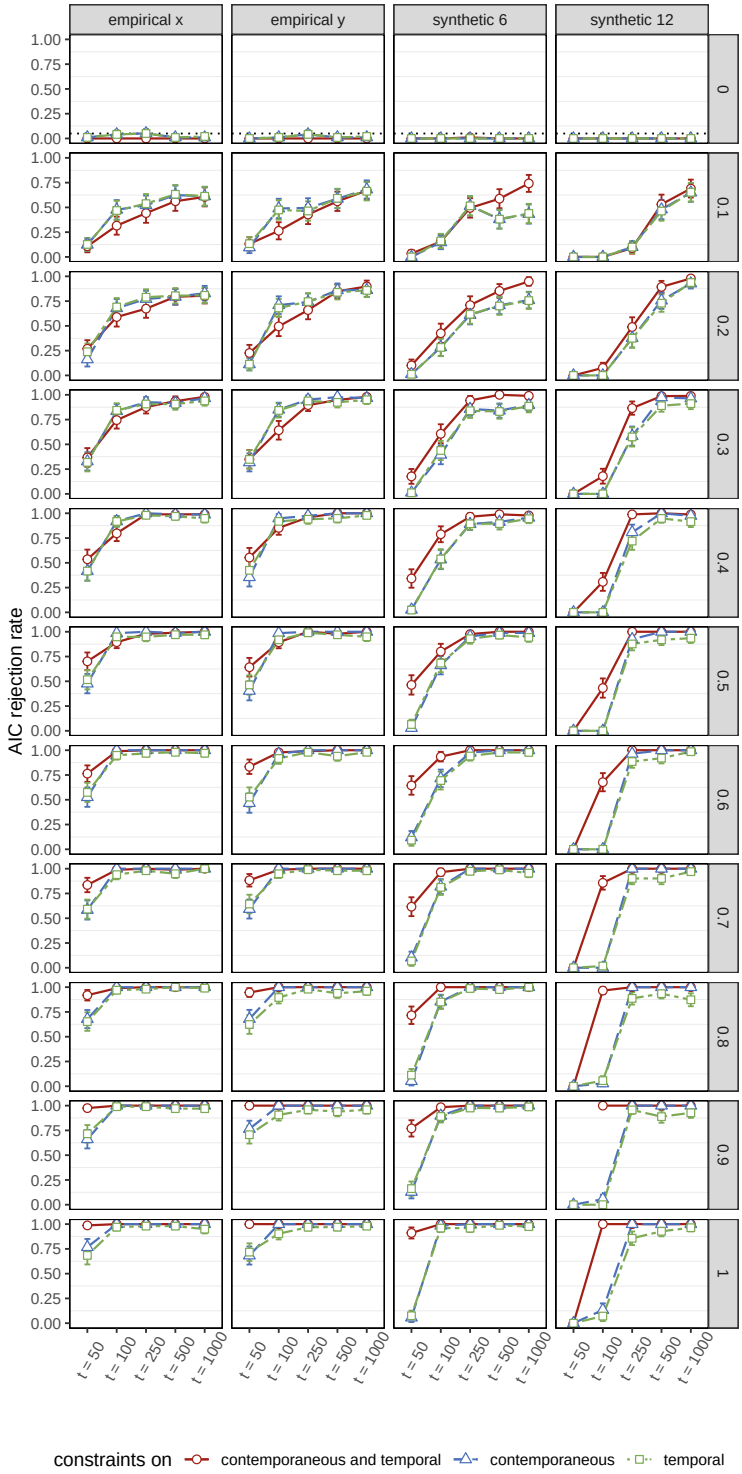


Figure 39: AIC rejection rate results when applying INIT to compare saturated network structures under all rewiring probabilities.

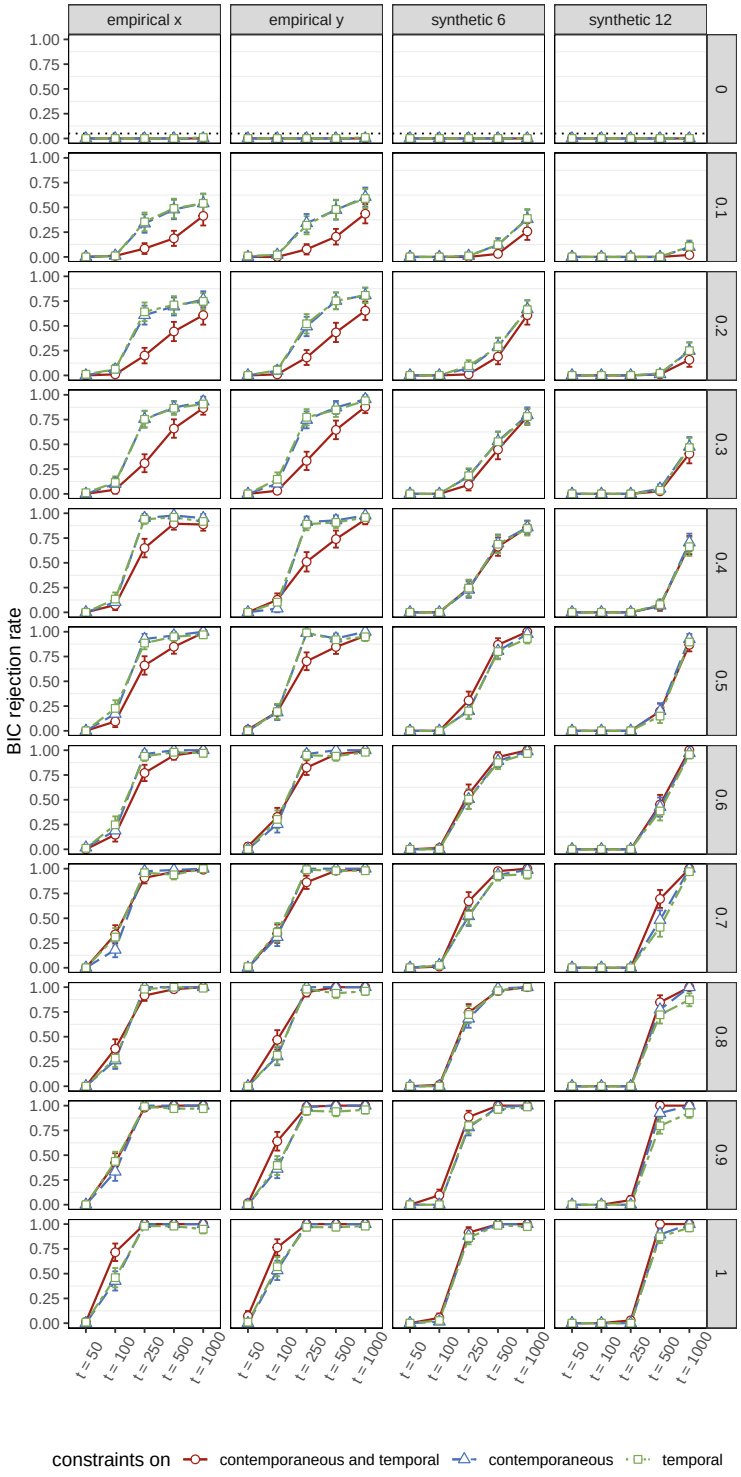


Figure 40: BIC rejection rate results when applying INIT to compare saturated network structures under all rewiring probabilities.

was varied. 'synthetic 6' corresponds to the synthetic network containing 6 nodes, and 'synthetic 12' corresponds to the synthetic network containing 12 nodes. The color of the lines corresponds to the placement of equality constraints. Red indicates equality constraints have been placed on both the contemporaneous and temporal network structure. Blue indicates equality constraints have been placed on the contemporaneous network structure and green indicates equality constraints have been placed on the temporal network structure.

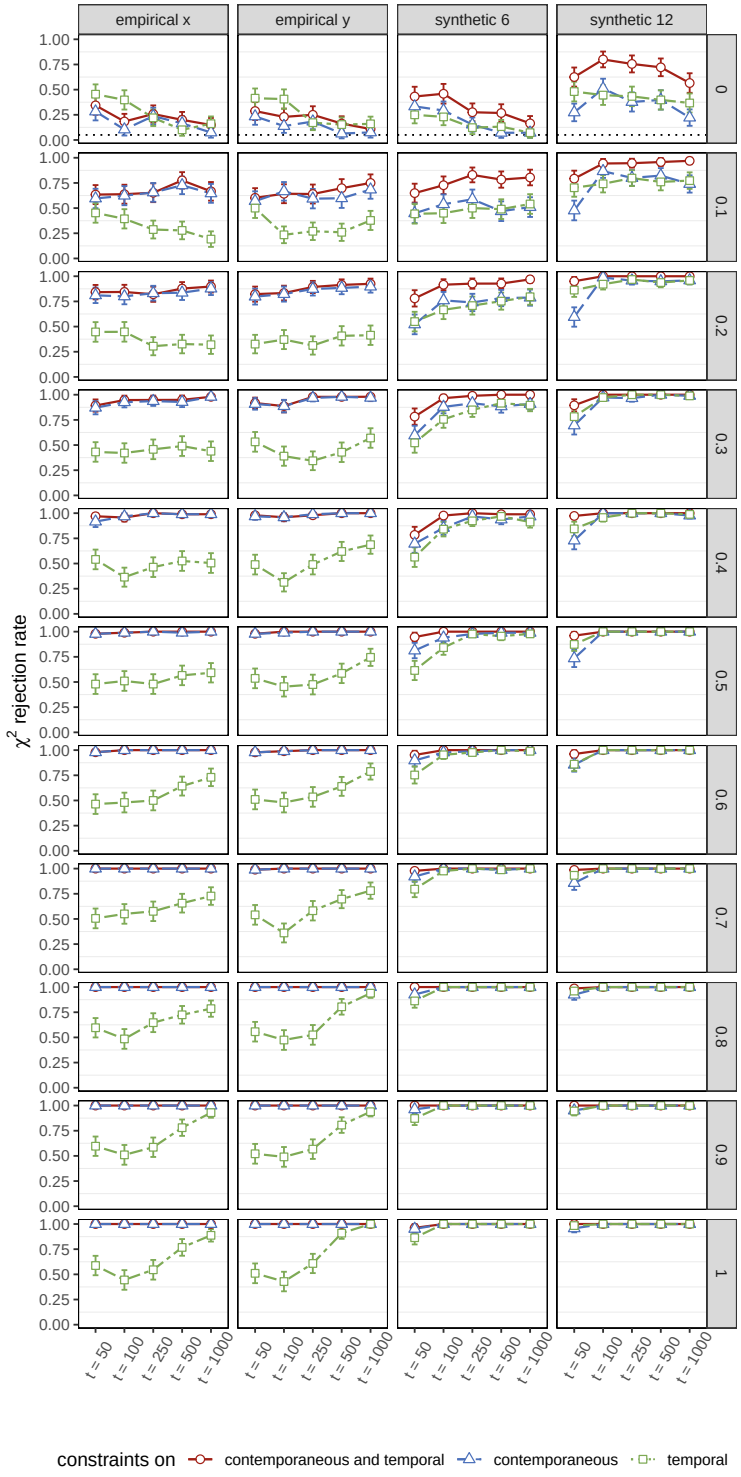


Figure 41: χ^2 test rejection rate results when applying INIT to compare *pruned* network structures under all rewiring probabilities.

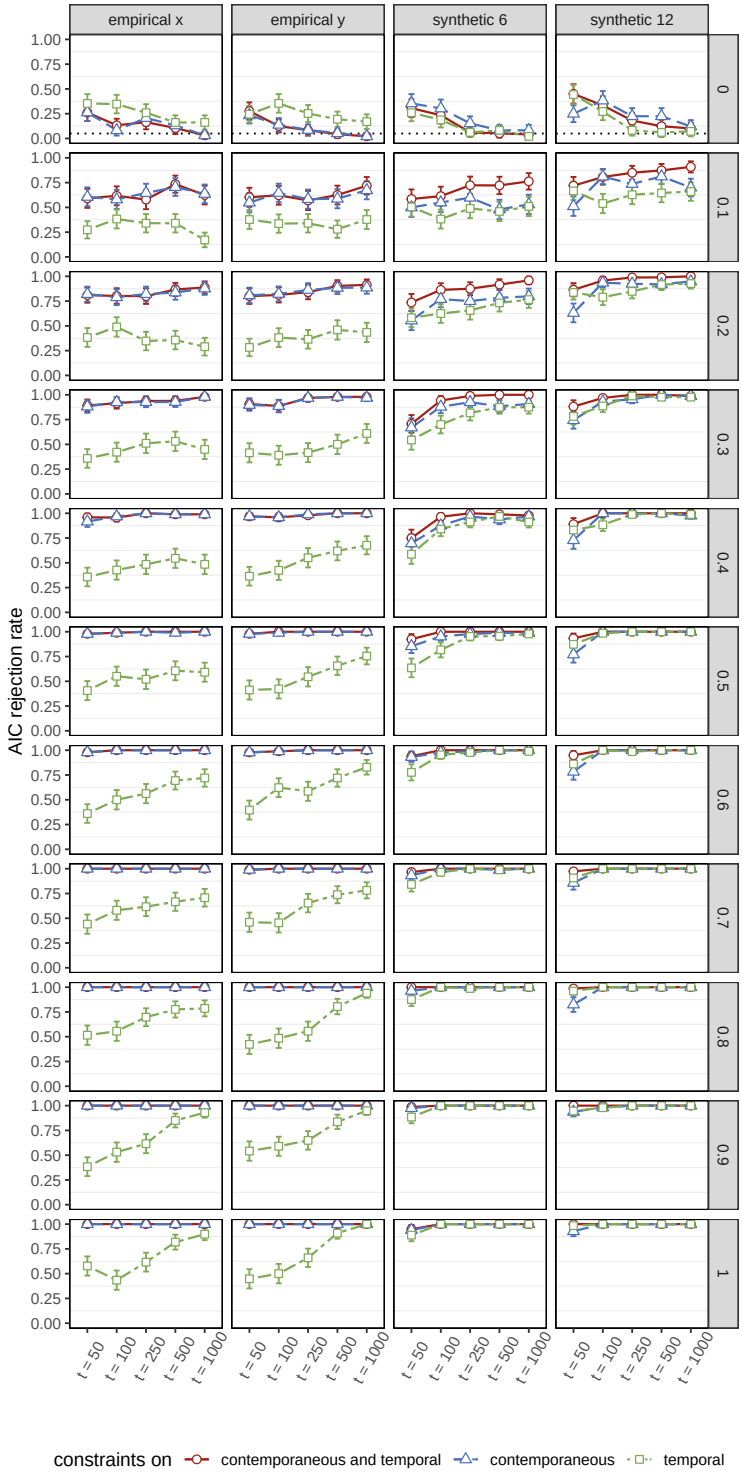


Figure 42: AIC rejection rate results when applying INIT to compare *pruned* network structures under all rewiring probabilities.

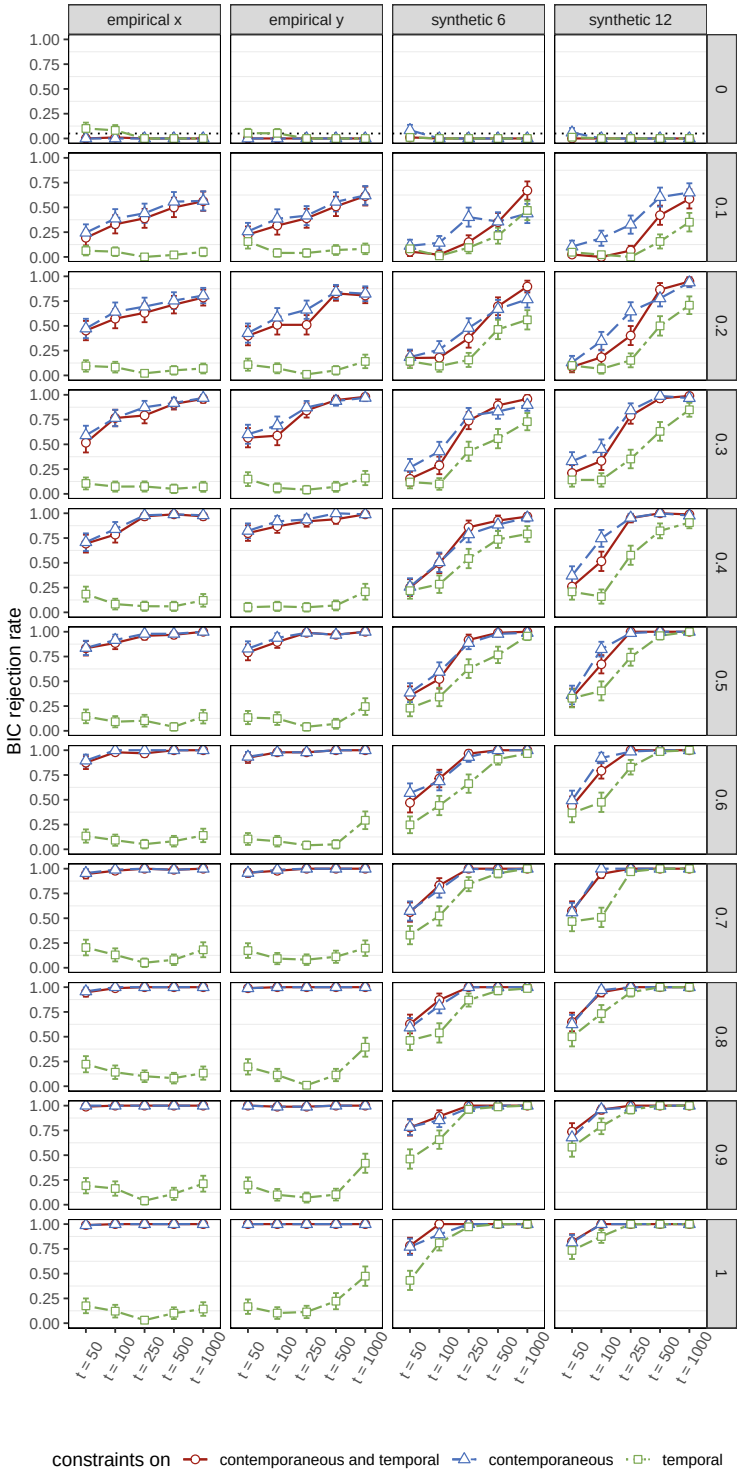


Figure 43: BIC rejection rate results when applying INIT to compare *pruned* network structures under all rewiring probabilities (0 to 1).

C

SUPPLEMENTS TO CHAPTER 4

In this appendix, we provide more details regarding the work discussed in Chapter 4. The appendix consists of four sections. First, we provide an overview of the most commonly used evaluation metrics in network psychometrics. Second, we present a (small) simulation study to determine optimal parameter values for the Ising model. Third, we present a simulation study within the Ising modeling framework. Fourth and final, we present a simulation study within the GVAR modeling framework

C.1 EVALUATION METRICS

Several metrics can be used to evaluate the performance of a network estimation technique. These metrics evaluate different aspects of the estimation procedure, such as the ability to retrieve edges, the accuracy of the edge weights, the accuracy of the centrality indices. Here we discuss the most commonly used metrics.

C.1.1 *Network accuracy*

SENSITIVITY Sensitivity is a metric that is used to determine the ability to identify true edges. That means sensitivity can be used to determine if the estimation technique is capable of picking up effects that we know are there. The sensitivity is expressed as the proportion of edges in the true network that is also included in the estimated network:

$$\text{sensitivity} = \frac{\text{\# of true present edges in the estimated network}}{\text{\# of total present edges in the true network}}$$

Here, 'true present edges' denotes edges that are shared between both the estimated network and the true network.

Various variants of sensitivity can be computed, including the 'signed sensitivity.' In signed sensitivity, the numerator considers only edges estimated with the same *sign* as in the true network. Put differently, if an edge in the true network is positive but is estimated as a negative edge in the estimated network, this edge no longer contributes to the count of true present edges in the estimated network.

Additional sensitivity variants encompass the *'top 50%'*, *'top 25%'*, and *'top 10%'* sensitivity. In these instances, both the numerator and denomi-

nator restrict their calculation to the top percentage of edges. For example, the 'top 50%' sensitivity, denotes the proportion of the 50% strongest edges in the true network that are also included in the estimated network.

High sensitivity signifies most existing edges were successfully detected, indicating the evaluated method excels in capturing effects that are truly present. In general, sensitivity increases when power increases [151]. Power can be increased by increasing the number of participants in cross-sectional analysis, increasing both the number of participants and time series in longitudinal multilevel analysis, or increasing the number of time series in idiographic longitudinal analysis. Conversely, a low sensitivity implies most existing edges were *not* detected, potentially leading to sub-optimal replication results and inflated visual heterogeneity across estimated networks [143, 195].

SPECIFICITY Specificity is a metric that can be used to determine the capacity to *avoid* the inclusion of false edges, i.e., the capability to correctly identify absent edges.

Specificity is quantified as the ratio of edges that are genuinely missing in the true network and accurately excluded from the estimated network:

of missing edges in the true network that were correctly not included in the estimated network:

$$\text{specificity} = \frac{\text{\# of true absent edges in the estimated network}}{\text{\# of total absent edges in the true network}}$$

Here 'true absent edges' refers to edges that are absent in both the true and the estimated network.

In principle, specificity should be high. When trends in specificity are encountered as a function of sample size, this can be an indication of problems with the estimation method [151].

PRECISION Precision is a metric to determine the accuracy of the estimation method. It measures the proportion of edges present in the estimated network that are also present in the true network:

$$\text{precision} = \frac{\text{\# of true present edges in the estimated network}}{\text{\# of edges in the estimated network}}$$

Here 'true present edges' entails edges that are shared between both the estimated network and the true network.

Similar to the sensitivity metric, precision allows for the calculation of various variants such as the *'top 50%'*, *'top 25%'*, and *'top 10%'* precision, which involve considering only the top % prominent edges when computing precision.

It is important to note specificity and precision are closely related: like specificity, precision tends to be low when numerous edges are included in the estimated network that are absent from the true network, resulting in a high number of false positive edges.

JACCARD INDEX The Jaccard index is the proportion of shared edges relative to the total number of edges:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Where A and B are matrices containing 0 when there is no edge present and 1 if there is an edge present. When we are mostly interested in the accuracy of the non-zero elements rather than the overall accuracy, the Jaccard index allows us to focus specifically on the overlapping edges between the two network structures.

C.1.2 *Edge accuracy*

Sensitivity, specificity, precision, and the Jaccard index are metrics to investigate the structure of networks by assessing whether edges are present or absent. These metrics, however, do not tell us something about the accuracy of the estimated edge weight. Nevertheless, we can think of many scenarios in which it is important to determine accuracy of the estimated edge weights. For one, edge weights play a pivotal role in determining the visual appearance of the network structure, and two, various other metrics that are often of interest such as network density and centrality metrics are contingent upon the accuracy of the estimated edge weights.

CORRELATION To investigate the accuracy of edge weights *correlations* are frequently employed as a metric. The Pearson correlation between true and estimated edge-weights evaluates their similarity, thereby providing an indication of the precision of the estimated edge weights. In addition the correlation between aboslute edge weights and non-zero edge weights can be computed as a means to compare edges weights between the estimated network structure and true network structure.

In general, a higher correlation coefficient implies a more accurate estimation of the edge weight. However, it is important to consider that the strength of the correlation coefficient is also determined by variance. In cases, where the true and estimated network structures are perfectly identical, exhibiting little to no variance, the correlation coefficient is close to zero. This means the strength of the correlation and the accuracy of the estimated edge-weights do not share a one-to-one correspondence. Con-

sequently, it is important to inspect additional metrics next to the Pearson correlation.

BIAS In addition to correlating edge weights to determine the accuracy of their estimation we can also investigate the bias in edge weights. For example we can compute the average absolute deviation (bias) between the true edge weight and the estimated edge weight. We can compute this bias for either all edges in both networks, or for only for the true edges that were included in the estimated network.

C.1.3 *Centrality accuracy*

To give an indication of important nodes in the network, centrality indices such as strength, closeness, betweenness, or expected influence are computed to provide insights into the prominence of nodes within a network [e.g., 27, 228, 39, 84].

CORRELATIONS To assess the accuracy of these centrality indices, Pearson correlations as well as Kendall correlations can be computed between the true centrality indices and the centrality indices as computed based on the estimated network structure [151].

TOP-RANKED Often, a primary focus often centers on nodes with the highest centrality indices [39]. If interested in the ability to correctly detect the nodes with the highest centrality indices, an examination of the top-ranked nodes is optional. This entails an investigation into the frequency with which the same nodes occupy the 'top 1', 'top 3', or 'top 5' position of most central nodes.

C.1.4 *Replicability*

The evaluation metrics discussed here can be extended to assess the replicability of network estimation procedures. In this context, the focus shifts from comparing the true network to an estimated network, to a comparison between two estimated networks. Thus, we are no longer interested in comparing the true data generating network to an estimated network, instead, we would want to compare two estimated networks with each other. Metrics such as the correlation between edge-weights, or the correlation between centrality indices are intuitive indicators to determine the replication of a network structures.

Furthermore, networks can be compared on network characteristics such as connectivity (i.e., percentage of actualized edges) and network density. The quality of the replication can be evaluated using correlation

metrics or metrics that gauge the dissimilarity between two matrices, such as the Jaccard index which measures the proportion of shared edges relative to the total number of edges.

Changes in the estimated edges can be quantified by calculating the percentage of change in edge weights, the proportion of replicated edges (i.e., edges that are included in both network structures), the proportion of non-replicated edges, and the proportion of edges unique to the replication set. Discrepancies in node centrality can be evaluated by correlating centrality indices, rank order correspondence, and assessing the consistency of the top 1, 3, and 5 most central nodes across the replicated networks.

C.2 DETERMINING PARAMETER VALUES FOR ISING MODEL

Generating a synthetic Ising model can be complicated, as many choices can be made such as determining the values for the parameters in the model. This is not an arbitrary choice, as we discussed in Chapter 4 that the encoding used (i.e., either $\{0, 1\}$ or $\{-1, 1\}$), and the parameter values (i.e., thresholds denoted as τ and edge weights denoted as ω), need to be specified in such a way that the constructed Ising model leads to generated data with enough *entropy*. To determine what combination of encoding, threshold parameters, and edge weights, leads to a structure that generates data with enough entropy, is the very reason to conduct a simulation study. In this appendix, we briefly explain why symmetric endocing is preferred when generating an Ising model. How this choice of encoding works best when specifying threshold parameters as 0, and determine what value of edge weights works well by means of a simulation study.

C.2.1 *Choosing the encoding and thresholds*

Typically, the dichotomous responses for the Ising model are either encoded using 0 and 1, which we term *binary encoding*, or -1 and 1, which we term *symmetric encoding*. Binary encoding is often the default in psychological research and psychometrics [135]. This also means that some of the software packages that estimate Ising models from psychological data assume binary encoded data (e.g., *IsingFit* requires binary data, and *mgm* prefers binary data for parameterization purposes [21, 132]). Symmetric encoding, on the other hand, is the default used in statistical physics when simulating data under the Ising model. Only a few R packages for psychological network estimation allow for symmetric encoded data (e.g., the *psychonetrics* package allows for both binary and symmetric encoded data [73]).

Important to note is that the two encodings are equivalent. This means that if we generate data using symmetric encoding, and subsequently transform the data to binary encoding (i.e., making all the -1 values 0) to fit an Ising model, we obtain an equivalent model. The obtained model parameters can easily be transformed into one-another. For example, we can transform Ising parameters based on the binary encoding $\{0, 1\}$ to parameters based on the symmetric encoding $\{-1, 1\}$ as follows:

$$\begin{aligned}\omega_{ij}^{(-1,1)} &= 0.25\omega_{ij}^{(0,1)} \\ \tau_i^{(-1,1)} &= 0.5\tau_i^{(0,1)} - 0.25\sum_j \omega_{ij}^{(0,1)}\end{aligned}$$

Here, the superscript denotes the encoding. Likewise, we can transform parameters based on the symmetric encoding to parameters based on the binary encoding as follows:

$$\begin{aligned}\omega_{ij}^{(0,1)} &= 4\omega^{(-1,1)} \\ \tau_i^{(0,1)} &= 2\tau_i^{(-1,1)} - 2\sum_j \omega_{ij}^{(-1,1)}\end{aligned}$$

These transformations to transform Ising parameters from one encoding to another are implemented in the `LinTransform()` function in the *IsingSampler* package [76]. These transformations can be very useful in data generation, as it, for example, allows one to generate symmetric encoded data (which typically is easier as explained below), transform the data to binary encoded data for a network estimation technique (e.g., *IsingFit*), fit a network in the binary domain, and then transform those results back to the symmetric domain to assess the performance of the estimated network parameters.

If the threshold parameters (τ) are chosen too low or too high, all responses will simply be identical (e.g., not endorsed or endorsed), and no information is present for statistical modeling. A very useful property of the symmetric encoding of the Ising model is the following:

$$\mathcal{E}\left(y_i|\tau_i^{(-1,1)} = 0\right) = 0.$$

That is, if the threshold parameter is 0, a symmetric encoded variable has an expected value of 0, meaning that it has a 50% probability on -1 and a 50% probability on 1. This property ensures we obtain some variation in the variables we simulate. Therefore, it is preferred to use symmetric encoding $\{-1, 1\}$ and set the threshold parameters of all variables to 0 for the data generating network structure.

Binary encoded data does not have this same property. Instead, setting all thresholds to 0 for binary encoded data could instead lead to almost all responses being 1 (depending on the strength of the network parameters in ω). We can derive from the transformations above that when we have binary encoded data, we require $\tau_i^{(0,1)}$ to be set to negative half the node strength (sum of the connected edge weights) to achieve the same result. For example, if node 1 has three connected edges each with a edge strength of 0.20 (e.g., $\omega_{12} = \omega_{13} = \omega_{14} = 0.20$, then the node strength of node 1 is $0.20 + 0.20 + 0.20 = 0.6$ and we would need to set $\tau_1 = -0.3$ to obtain a 50% probability of (not) endorsing the variable.

With the encoding and thresholds chosen, all that is left is to generate the network structure. To investigate what parameter values work well, we performed a small simulation study.

C.2.2 *Simulation set up*

We generated synthetics Ising models, varying the following conditions:

NUMBER OF NODES: 5 or 10

NETWORK STRUCTURE: random (25% or 50% of potential edges were included at random), chain (e.g., node 1 is connected to node 2, node 2 is connected to node 3, etc.), Curie-Weiss (a fully connected network).

EDGE WEIGHT: Different mean edge weights were simulated (the weights were chosen at random using a normal distribution with the set mean edge weight and a standard deviation of 0.1)

ENCODING: Binary or symmetric

THRESHOLDS: -1, 0, 1

As an evaluation metric we inspected the entropy of the specified Ising model using the `IsingEntrophy()` function from the *IsingSampler* package [76]. The simulation code used to conduct this study can be found on OSF: `https://osf.io/ygjn7/`.

C.2.3 *Results*

We have visualised the results in Figure 45. Results indicate that the entropy goes down as a function of the edge weight, meaning that the stronger connected the network the fewer response patters are generated. Entropy is at a maximum when all τ and ω are set to zero, but this is not an interesting structure to generate data under as this would simply

imply completely random (coin-flip) data. Entropy reduces to 0 in most settings, except in symmetric encoded data with all τ parameters set to 0 (our above described preferred simulation setting), in which case, due to symmetry, two response patterns eventually emerge (all variables -1 or all variables 1). In general, when generating symmetric encoded data with all threshold parameters set to 0 edge weights around 0.25 and 0.5 tend to work well.

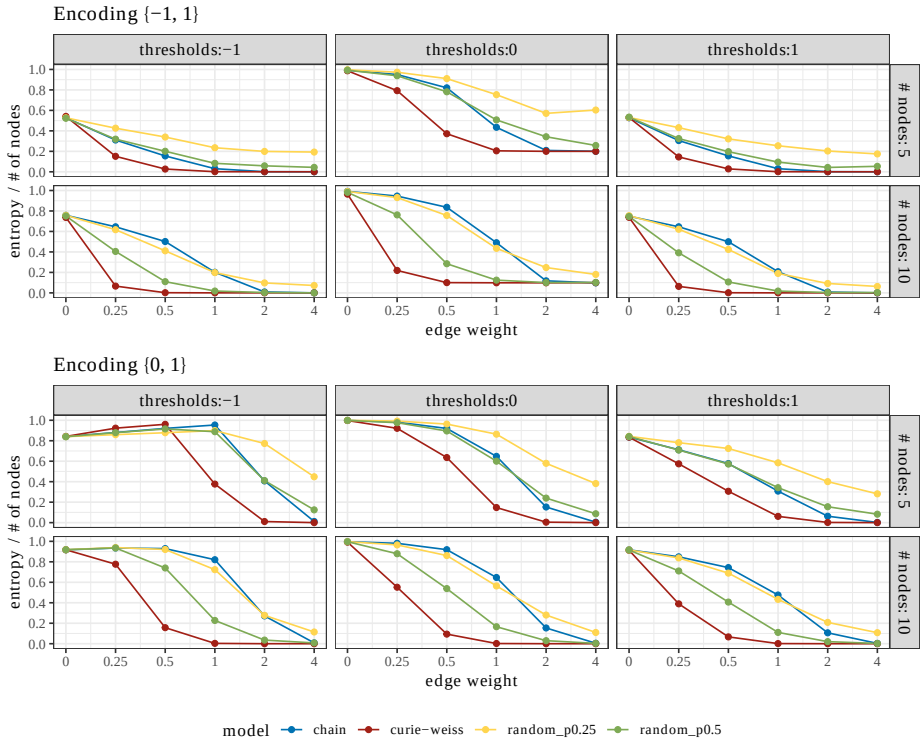


Figure 44: Results from Ising model simulation tutorial.

C.3 TUTORIAL: SIMULATION STUDY WITHIN THE ISING MODELING FRAMEWORK

In this tutorial, we exemplify how a simulation study can be set up within an Ising model framework to answer the question: *"To what extend can we accurately estimate an Ising model for reasonable sample sizes?"*. The annotated tutorial code can be found on OSF: `https://osf.io/ygjn7/`. In order to answer the research question, we varied the sample size condition (i.e., $n \in 100, 200, 300$), and obtained the sensitivity, specificity, and correlation as our evaluation metrics. See above for a description of the most com-

monly used evaluation metrics in network psychometrics, and Table 1 for all evaluation metrics implemented in the `evalNet()` function.

We use the `parSim()` function from the *parSim* package [78] to provide the skeleton for our simulation study. We specified our sample size condition (i.e., `sampleSize = c(100, 200, 300)`), and within the R expression we load the required R packages, specify the true data generating network, simulate data based upon this network, and estimate a network based on the simulated data to ultimately compare the estimated to the true network using our evaluation metrics. In the tutorial code we set the `reps` argument to 100 by means of a simplification, but ideally this number of repetitions would higher.

After we ran our simulation, we can analyze the results. We use the *tidyverse* packages *dplyr* and *tidyr* for data manipulation, and *ggplot2* for plotting the results [270, 268, 269, 267]. We display our results in a boxplot with our sample size condition on the x-axis the evaluation metrics (i.e., correlation, sensitivity, and specificity) are plotted on the y-axis. In order to plot our results with the `ggplot()` function, some data modifications need to be made. First we want to format our results in such a way that they are applicable to use within `ggplot()`. This means, we need to go from a wide data format to a long data format. We can use the `gather()` function from the *tidyr* package to do so. After cleaning up the data and creating the right format, we can use `ggplot()` to display the simulated results and start interpreting the results, see Figure 45 for the simulation results.

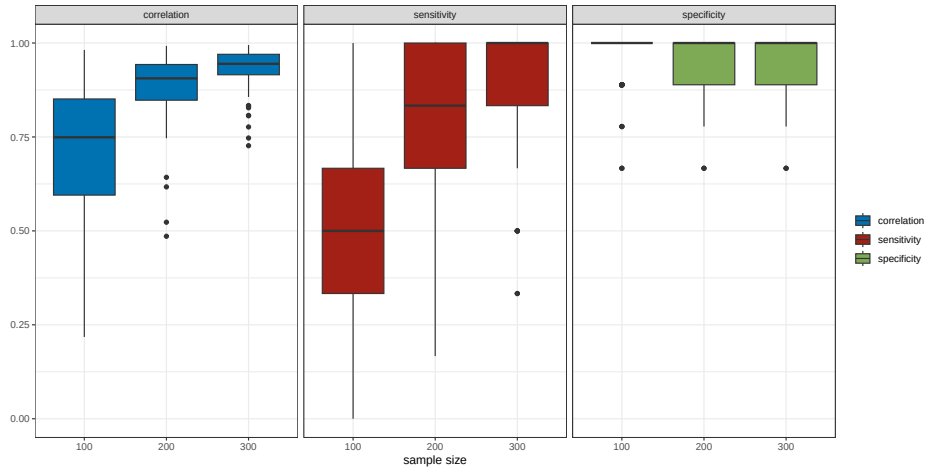


Figure 45: Results from Ising model simulation tutorial.

Our results indicate that the correlation between the estimated network model and the true data generating network model increases as a function of the sample size. In addition, the sensitivity and specificity increases.

Meaning more true effects are picked up and less false positives are included with increasing sample sizes, Figure 45.

C.4 TUTORIAL: SIMULATION STUDY WITHIN A GVAR MODELING FRAMEWORK

In this tutorial, we exemplify how a simulation study can be set up within a GVAR model framework to answer the question: *"To what extend can we accurately estimate a GVAR model for reasonable time series?"*. Annotated tutorial code can be found on OSF: `https://osf.io/ygjn7/`. In order to answer this research question, we varied the number of time series (i.e., $t \in 100, 200, 300$), and obtained the sensitivity, specificity, and correlation as our evaluation metrics for both the contemporaneous network, as well as the temporal network. See above for a description of the most commonly used evaluation metrics in network psychometrics, and Table 1 for all evaluation metrics implemented in the `evalNet()` function.

We use the `parSim()` function from the *parSim* package [78] to provide the skeleton for our simulation study. We specified our sample size condition (i.e., `nTime = c(100, 200, 300)`), and within the R expression we load the required R packages, specify the true data generating network, simulate data based upon this network, and estimate a network based on the simulated data to ultimately compare the estimated to the true network using our evaluation metrics. In the tutorial code we set the `reps` argument to 100 by means of a simplification, but ideally this number would be higher.

After we ran our simulation, we can analyze the results. We use the *tidyverse* packages *dplyr* and *tidyr* for data manipulation, and *ggplot2* for plotting the results [270, 268, 269, 267]. We display our results in a boxplot with our sample size condition on the x-axis the evaluation metrics (i.e., correlation, sensitivity, and specificity) are plotted on the y-axis. In order to plot our results with the `ggplot()` function, some data modifications need to be made. This means, we need to go from a wide data format to a long data format. We can use the `gather()` function from the *tidyr* package to do so. After cleaning up the data and creating the right format, we can use `ggplot()` to display the simulated results and start interpreting the results, see Figure 45 for the simulation results.

Results indicate that the correlation between the estimated network model and the true data generating network model goes up as the sample size increases. For the contemporaneous network model, both sensitivity and specificity increase. The estimated network structure is accurately estimated from $t = 250$ onward.

In addition, sensitivity is low when $t = 100$, and in general is the correlation, sensitivity lower for the temporal network as for the contempora-

neous network. This is a logical consequence of the number of parameters that need to be estimated for each of these network models. As the temporal network model is a directed network model, many more parameters need to be estimated, and that often requires more power, i.e., bigger sample sizes.

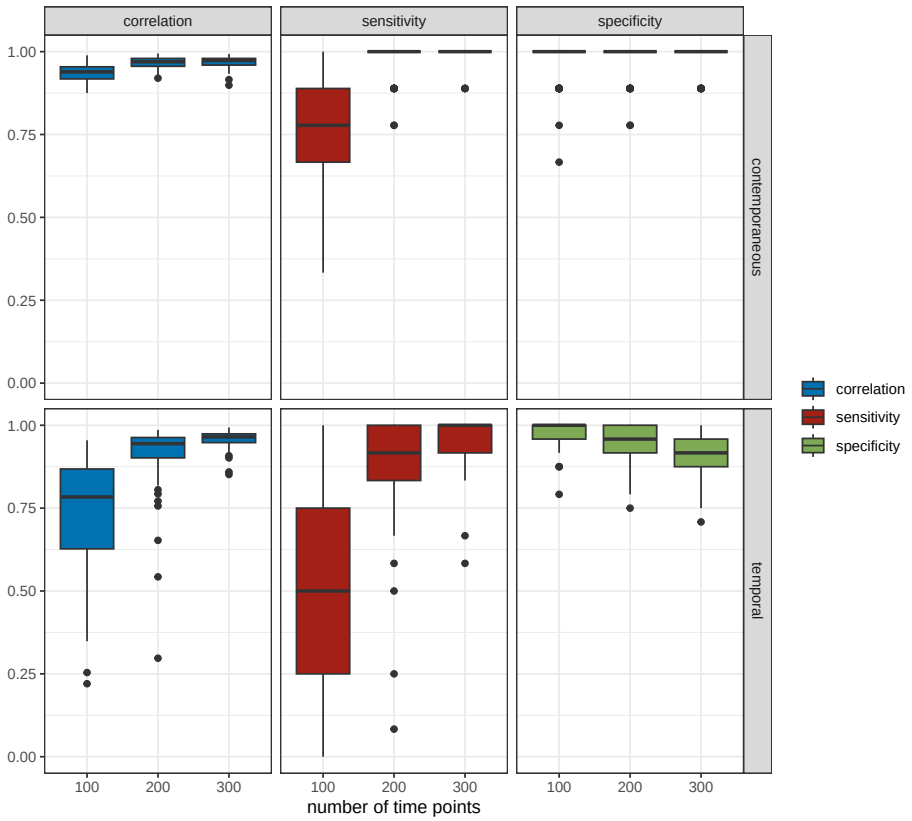


Figure 46: Results from GVAR simulation tutorial.

D

SUPPLEMENTS TO CHAPTER 5

D.1 PREPROCESSING STEPS

D.1.1 *Missing data*

Individuals had on average 9% missing data points for Diary 1 (baseline) and 6% missing data points for Diary 2 (1-year follow-up), ranging from 0% to 23%. Since there is no consensus in the literature on the best imputation strategy for imputing time series data, we tested six imputation strategies (MICE, Amelia, exponential moving average, linear moving average, Kalmal filter, and mean imputation) on our dataset (i.e., items used for analyses) in a previous study [253]. For this, we used data from six individuals with complete data. First, we randomly deleted 5%, 10% and 25% of their data points. Next, for 1000 iterations we performed the six imputation strategies on the newly created datasets. As the real data values were known for these six individuals, we compared the newly imputed data points to the real values and averaged and squared the difference to calculate mean squared errors per imputation strategy. Exponential moving average performed best on our dataset and was therefore chosen as imputation strategy.

D.1.2 *Imputation*

There is not only no consensus in the literature on which imputation strategy should be used for time series data, but there is also research lacking on whether data should be imputed at all in these types of data. As a sensitivity analysis, we created symptom networks with and without imputations and compared the covariance matrixes within individuals using correlations. This showed relatively high correlations for both Diary 1 (range: $r = .93$–1.00, mean: $r = .99$) and Diary 2 (range: $r = .61$–1.00, mean: $r = .99$) data, and thus imputation did not influence results greatly. To ensure maximum power, we therefore decided to impute missing data.

D.1.3 *Normality assumption*

One of the assumptions of the GVAR model, used to create idiographic the symptom networks, is that data is normally distributed. For some individuals and some items, data was heavily skewed. Therefore, data was

transformed with a nonparanormal copula transformation implemented in the *huge* package in R [187, 155]. Previous research has shown that this form of transformation works well for heavily skewed data in symptom networks [151, 195].

D.1.4 *Stationarity*

Another assumption of the GVAR model is that the data is trend-stationary. Plots showed a linear trend for most individuals and most items. Therefore, linear trends were removed from the data by detrending the data: domains were regressed against time and the residuals were entered in the GVAR model.

D.2 SENSITIVITY ANALYSIS

Using INIT, equality constraints can either be placed on unpruned networks, i.e., networks in which all edges are estimated to be non-zero, or on pruned networks, i.e., networks in which some edges have been shrunk to zero using a model selection procedure. This model selection procedure estimates sparse individual network structures with a high identification rate of the most substantial edges. In the analyses in Chapter 5, using INIT, equality constraints have been placed between the unpruned networks at T0 and T1 directly. As a sensitivity check, we have applied INIT on pruned network structures.

INIT makes use of the *psychonetrics* package in R to estimate individual networks using Maximum Likelihood estimation (ML). After estimating the fully connected network, the network can be pruned using several pruning techniques. For our sensitivity analysis we choose to apply recursive pruning at an alpha level of 0.05. This is a step-by-step procedure, i.e., in each iteration, non-significant edges are removed from the network, after which the remaining edges are re-estimated. Previous simulation studies have shown that GVAR estimation through ML combined with recursive pruning works well when estimating sparse individual network structures [151, 195].

Equality constraints are placed on the individual pruned networks at T0 and T1, using INIT. In order to place equality constraints between pruned network models, INIT makes use of a union model. The union model will add all edge-weights to both networks at T0 and T1 that are estimated to be non-zero in at least one of the networks of T0 and T1. A union model is a necessary step in the procedure because equality constraints cannot be placed directly between an edge that is estimated to be non-zero in one network and estimated to be zero in another. INIT follows the same logic as for the unpruned network case: an optimal model

fit needs to be ensured such that the estimated network parameters of an individual at T0 are equal to their network structure at T1. When placing equality constraints on pruned networks, model fit is best evaluated with the BIC, see simulation results presented in the next section and simulation results as discussed in Hoekstra et al. [144].[1]

Taking the Bayesian Information Criterion (BIC) as an indicator, the model fit of the network structures with equality constraints between T0 and T1 is compared to the model fit without imposing constraints on the network structures at T0 and T1. This results in either a better model fit for an equal network model over time, thus with constraints, or for a different network model over time, thus without constraints.

D.2.1 *Results*

Applying INIT on pruned network models showed that for 58 individuals out of 66, the model with equality constraints showed the best model fit according to the BIC. The other 8 individuals, the model without equality constraints fit better to the data. Applying INIT to pruned network models was more conservative in the number of differences estimated as 7 individuals who were placed in the "unstable" group would be labeled "stable" based on these results. Minor differences are to be expected as the BIC is more conservative. Further simulation studies need to be done in order to assess the exact number of differences between the two application procedures of INIT are to be expected.

As an indication of the degree to which network parameter estimations of T0 and T1 resemble each other, we inspected Pearson correlation between the unconstrained unpruned network models at T0 and T1. In addition, we inspected this correlation for the stable and unstable group. Important to note here is that a high Pearson correlation reflects a strong linear relation between parameter estimates (i.e., differences in estimated edge weights at time point T1 can be predicted from differences at T0), but that a low Pearson correlation merely indicates the lack of such a linear relation; a low correlation may either indicate large differences between the networks at T0 and T1 or result from the fact that at one or both time points, all edges have approximately the same weight. As an

1 When placing equality constraints on pruned networks, the AIC turned out to be too liberal, meaning the AIC was in favor of the model without equality constraints even if the network structures were highly stable. The BIC adds a penalty for model complexity and therefore proved to be the best information criteria to guide decision regarding model equality in light of structural network differences when placing equality constraints on pruned network models. It is important to note here, due to the additional penalty to model complexity the BIC entails, using the INIT on pruned network models is likely to result in a more cautious estimation of the differences present.

additional sensitivity check, we compared the results from INIT to the correlation analysis.

The correlation between the edge weights of unpruned contemporaneous networks at T0 and T1 for the whole sample was, on average, equal to $r = .66$ (sd = .25, range: $r = .09 - .93$). When split, the mean correlation between unpruned contemporaneous networks at T0 and T1 was estimated at $r = .69$ (sd = .16, range: $r = .09 - .93$) in the stable group and $r = .55$ (sd = .17, range: $r = .21 - .72$) for the unstable group. While results from the INIT and correlation analysis generally align, this is not necessarily the case for each individual specifically, as shown by the large range of correlations observed across individuals. Further inspection of the correlations showed that for individuals from the stable group who had a low correlation, this was often due to a sparse network model. When many edges are set to zero, as is the case in a sparse network model, the covariance matrix has a small standard deviation (i.e., there are few differences between the parameter estimates), leading to a low correlation via a restriction of range effect; this can happen even though two networks are almost identical.

D.3 SMALL SIMULATION STUDY TO TEST PERFORMANCE INIT

To test the power to detect small differences in individual network structures using INIT we performed a simulation study. INIT places equality constrains on the network model such that an optimal fit needs to be ensured in such a way that the network of an individual at a previous time point is equal to the network model at another time point. The fit of this model is then compared to the model fit of a network model in which the parameters of both network models for the individual are freely estimated.

For this simulation study we simulated data from a chain graph as this structure is commonly used in simulation studies [82]. We varied the number of time-points, the number of variables, and the amount of rewiring. Rewiring is taken as an indication of the number of differences between network models; where a rewiring of 0 means equal network models and a rewiring of 1 means all original edges are rewired, thus resulting in completely different network models. We varied the number of time points from $t = 50$ to $t = 500$ ($t \in \{50, 100, 250, 500\}$), creating a range of plausible amount of time points to desirable amount of time points. Number of variables were 6 or 12, corresponding to the size of the network structure. We increased the amount of rewiring with steps of .1 from 0 to 1 ($rewire \in \{0, 0.1, 0.2, 0.3, 0.5, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$). In addition, we preformed the test on unpruned network models as well as on pruned network models. We inspected the AIC rejection rate and

the BIC rejection rate. The simulation study was performed in R (version 4.1.0) [220].

D.3.1 *Results*

As can be seen in Figure 47, the AIC rejection rate is better to use when imposing equality constraints on saturated (i.e., fully connected) network models, whereas the BIC rejection rate is better to use when imposing equality constraints when comparing pruned network models, see Figure 48. Furthermore, the test is sensitive to pick up even minor differences in network structure ($rewire = 0.1$) when $t = 100$. Bigger differences ($rewire = 0.2$) are picked up when $t = 50$. The more differences and the higher the number of time points, the more power INIT has to determine the presence of differences in the network structure.

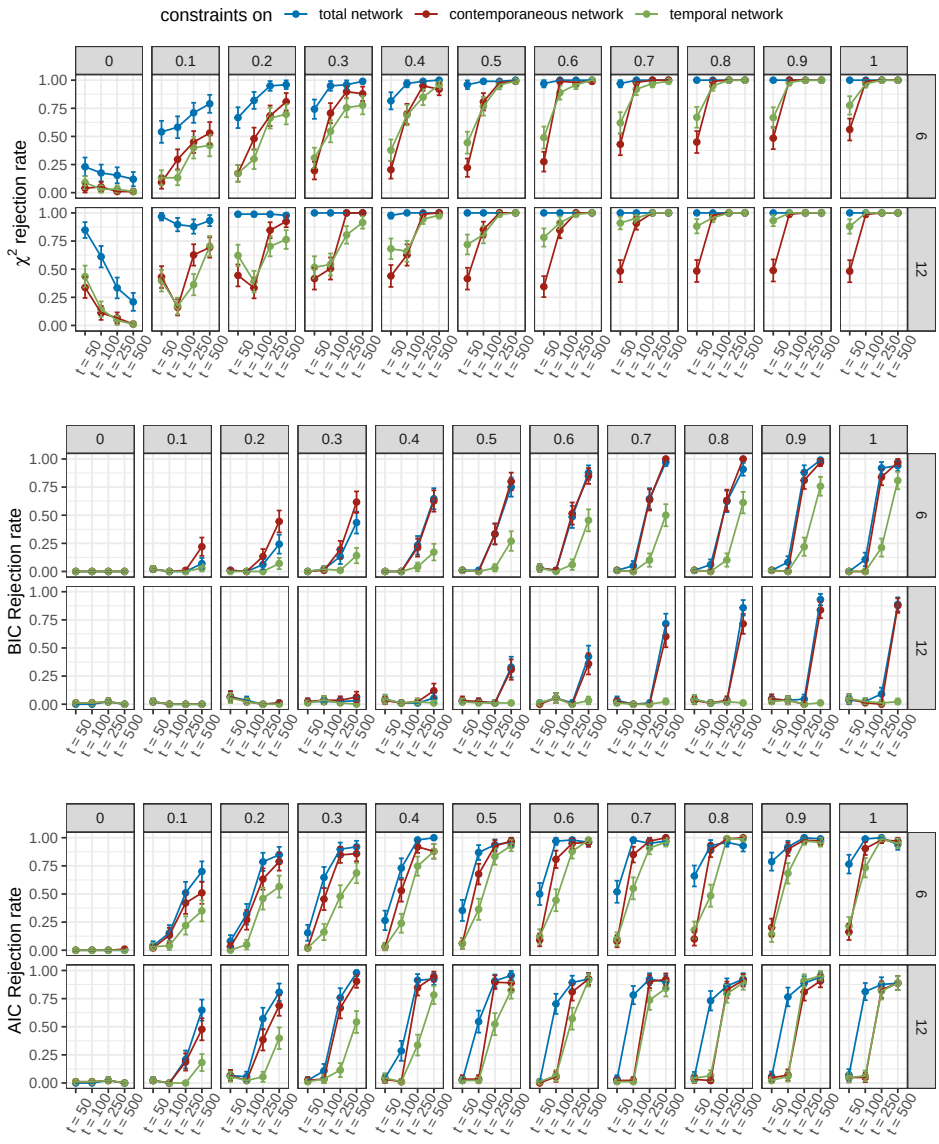


Figure 47: Simulation results for the INIT when equality constraints are placed on unpruned networks. We inspected the Chi-square rejection rate, the BIC and the AIC. The upper panels indicate the amount or rewiring between two network structures, from 0 to 1. The right panels indicate the number of variables in the network structure, either 6 or 12. Results show the AIC performed best when placing equality constraints on saturated networks.

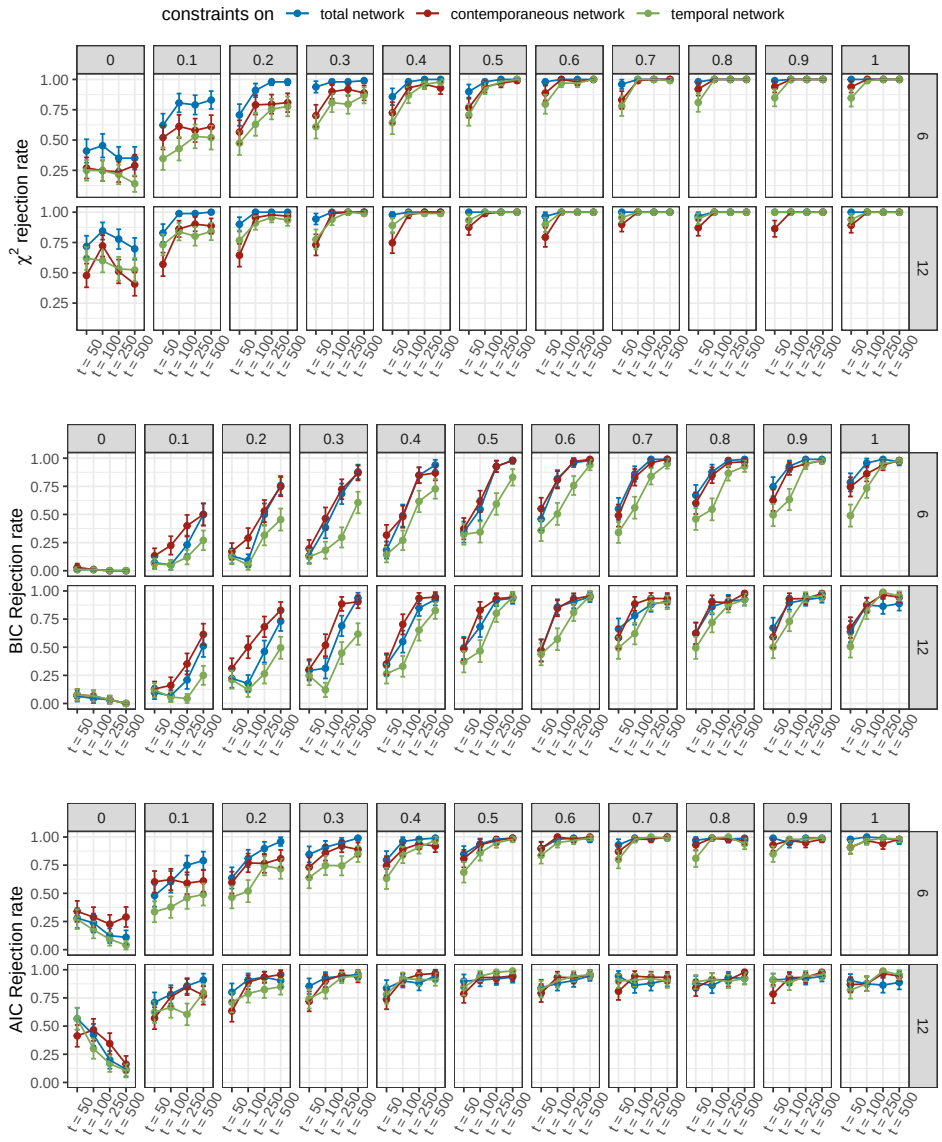


Figure 48: Simulation results for the INIT when equality constraints are placed on pruned network using a union model. We inspected the Chi-square rejection rate, the BIC and the AIC. The upper panels indicate the amount or rewiring between two network structures, from 0 to 1. The right panels indicate the number of variables in the network structure, either 6 or 12. Results show the BIC performed best when placing equality constraints on pruned networks using union models.

E

SUPPLEMENTS TO CHAPTER 6

This appendix contains supplemental material to Chapter 6. In particular, here we provide sample characteristics of the full sample, details regarding prepossessing the data, and discuss the performed sensitivity and stability checks.

E.1 SAMPLE CHARACTERISTICS FOR ORIGINAL FULL SAMPLE

As our inclusion criteria lead to a significant reduction ($n = 228$) of the original full sample ($n = 1355$), we inspect deviations from the selected sample and the original full sample on sample characteristics and mean and standard deviation for the PHQ-9 and for each of the PANAS variables, see Table 11 and 12 respectively.

E.2 INSPECTING MEASUREMENT OCCASIONS

One of the assumptions of the multilevel VAR model is equal time intervals between measurement occasions. The full version of the daily surveys, which contained assessments of mood and depressive complaints, were sent out on the first three days and afterward on two randomly selected days of the week, therewith violating the assumption of equal time intervals between measurement occasions. Inspecting the response rate of the full survey over the study period showed a clear three-day interval pattern; see panel (a) of Figure 49. In addition, we see an apparent occurrence of peaks within these three-day intervals. Therefore, we decided to group days into measurement occasions defined by a three-day window. This insured data at every measurement occasion; see panel (b) of Figure 49.

Inspecting the number of participants who completed assessments compared to the number of participants who completed measurement occasions, we found no significant loss of participants, see Figure 50. In line with recommendations for the multilevel VAR model, we selected participants with at least twenty measurement occasions [159].

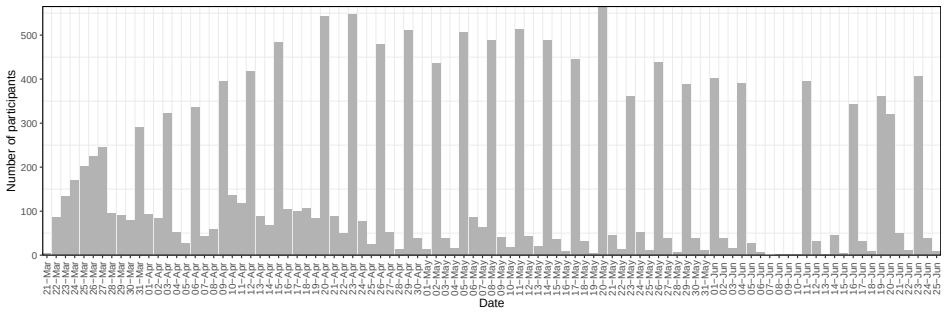
The above-described procedure resulted in a final sample of $n = 228$. Each participant has completed data for a minimum of 20 and a maximum of 33 measurement occasions. On average, participants had 9.4 ($SD = 2.7$) missed measurement occasions. Thus, on average, 28% of a participant's data is missing. It is important to note that within each of

Characteristics	Mean	SD
Age	44.8	19.2
	n	**Percentage**
Gender		
Female	1112	82.1%
Male	268	19.8%
Race/ethnicity		
African-American	29	2.1%
Asian	164	11.9%
White	1128	81.6%
Hispanic/Latinx	70	5.1%
Native Hawaiian or other Pacific Islander	3	0.2%
American Indian/Alaska Native	5	0.3%
More than one race	32	2.3%
Unknown	6	0.4%
Prefer not to say	11	0.8%
Cultural Background		
North-America	1130	87.1%
South-America	44	3.2%
Africa	3	0.2%
Asia	59	4.4%
Europe	109	8.0%
Oceania	33	2.4%
Annual household income		
$0-$25,000	144	10.6%
$25,001-$50,000	203	15.0%
$50,001-$75,000	251	18.5%
$75,001-$100,000	224	16.5%
$100,001-$150,000	256	18.9%
$150,001-$250,000	184	13.6%
$250,000+	118	8.7%
Education		
High School Diploma	12	0.9%
Some college	41	3%
College degree	214	15.8%
Some post-bacc education	141	10.4%
Graduate, medical or professional degree	587	43.3%

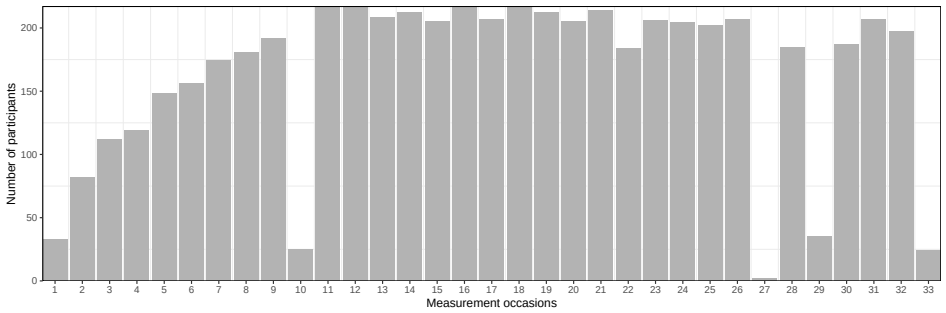
Table 11: Sample characteristics of the original full sample ($n = 1355$).

Variable	Mean	SD
PHQ9	7.81	5.09
PANAS		
Inspired	2.01	1.07
Alert	2.8	1.11
Excited	1.70	0.92
Enthusiastic	1.98	1.07
Determined	2.52	1.19
Afraid	1.98	1.1
Upset	1.89	1.01
Nervous	2.36	1.16
Scared	2.12	1.08
Distressed	2.19	1.06

Table 12: Mean and standard deviation per variable of total sample ($n = 1355$).



(a) Response rate for the full study period

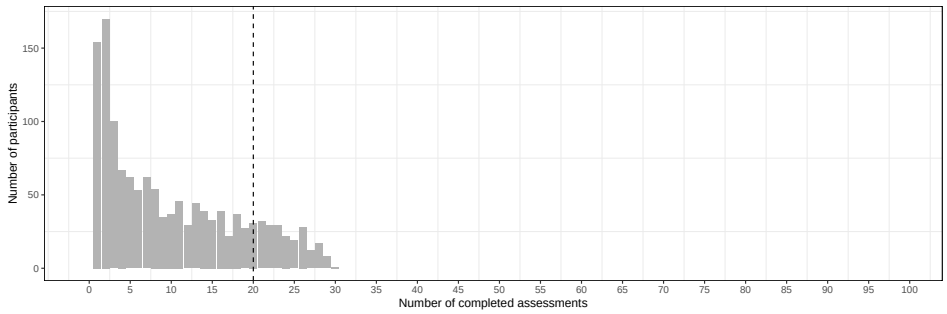


(b) Response rate per measurement occasion

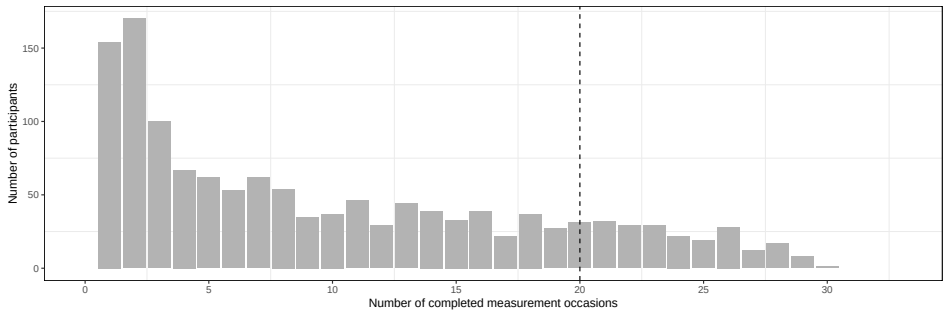
Figure 49: The response rate for the full survey over the study period from March 20th, 2020 until June 26th, 2020.

the measurement occasions, there is no missing data. That is, for each registered response, data on both the PHQ-9 and the PANAS are completed as these were implemented in the same questionnaire with forced responses on all questions. Thus, the average of 28% missing data, relates

to complete missed assessments between participants' first questionnaire and their last completed questionnaire.



(a) Completed assessments



(b) Completed measurement occasions

Figure 50: The upper panel shows the number of participants completed assessments. The lower panel shows the number of participants and the number of completed measurement occasions. The dotted line represents the cut-off value of 20 measurement occasions recommended to estimate a MLVAR network model.

E.3 SENSITIVITY CHECKS

We observed a strong correlation between the maximum absolute change in PHQ-9 score and the person-specific temporal network density ($r = 0.77$). Interestingly, when splitting the change in PHQ-9 score into alleviation or aggravation, we saw that this pattern is present in both directions: a larger change in PHQ-9 score is associated with a more strongly connected temporal network. In order to inspect the robustness of our findings, we conducted two sensitivity checks.

E.3.1 *Detrending*

As a first sensitivity check, we inspected whether trends in the data influence the observed correlation. To investigate this, we first tested for trends

in the data using the kpss.test() function in R. The kpss.test revealed significant trends in the data. Next, we detrended individual data by fitting linear regression models on each variable, regressing out a linear trend on measurement occasion, at an alpha of 0.05. Afterward, we estimated the multilevel VAR model on the detrended data. We computed density measures for each person-specific network and correlated this with their absolute maximum change in PHQ-9 scores. Results indicated a slightly stronger correlation ($r = 0.79$) between person-specific temporal network density and absolute maximum change in PHQ-9 score, see Figure 51. We, therefore, conclude trends in the data do not influence the main conclusions made in our manuscript.

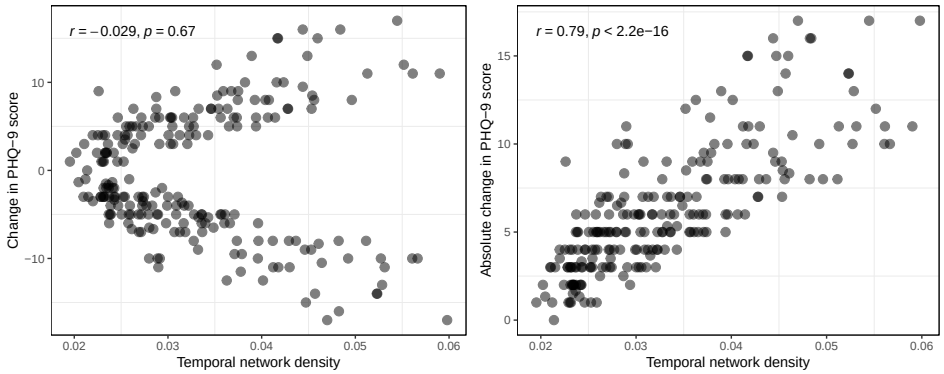


Figure 51: Correlation between person-specific temporal network density and the (absolute) maximum change in PHQ-9 score after detrending the data.

E.3.2 *Network density for affect items and PHQ-9 change score*

As a second sensitivity check, we investigated the possibility that the correlation was driven by the fact that the network includes the PHQ-9 items such that person-specific temporal network density is, at least in part, based on the same information (i.e., PHQ-9 items) that is also used to relate the density to (i.e., the absolute maximum change in PHQ-9 total score). Therefore, we investigated the relation between network density and maximum change in PHQ-9 score when including only the affect states into the networks.

We re-estimated the mlVAR model including only positive and negative affect items. Based on the person-specific temporal network models, we re-calculated network density for each person (i.e., the average absolute strength of the temporal associations). We correlated this person-specific network density with their absolute maximum change in PHQ-9 total score. As can be expected, removing the PHQ-9 items from the network decreased the strength of the observed correlation. However, we

still clearly observed a positive association between person-specific temporal network density and absolute change in PHQ-9 score ($r = 0.4$), see Figure 52.

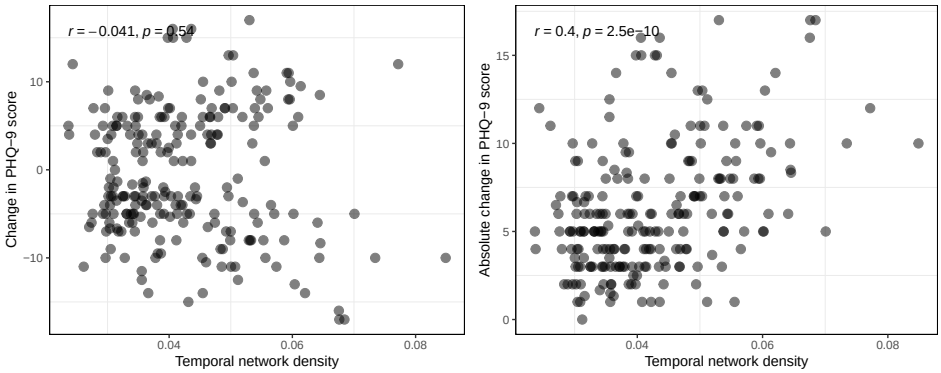


Figure 52: Correlation between person-specific temporal network density including only positive and negative affect and the (absolute) maximum change in PHQ-9 score.

E.3.3 *Stability checks*

In addition to conducting two sensitivity checks, we inspected the stability of the correlation between the maximum absolute change in PHQ-9 score and the person-specific temporal network density. To inspect the stability of our results, we re-estimated 100 mlVAR networks that included a random selection of 80% of the original data, that is using the data of 182 participants. For each of these re-estimated networks we calculate person-specific network density and correlate the network density with a change in PHQ-9 score. Results indicated, the correlation between the maximum (absolute) change in PHQ-9 score and the person-specific temporal network density was stable. The median correlation between person-specific temporal network density and maximum change in PHQ-9 score (median $r = 0.77$) was similar to the original value ($r = 0.77$). The median correlation between person-specific network density and their maximum change score re-estimating network structures (median $r = -0.04$), was identical to the original value ($r = -0.04$). In addition, we found narrow intervals around these median values, see Figure 53.

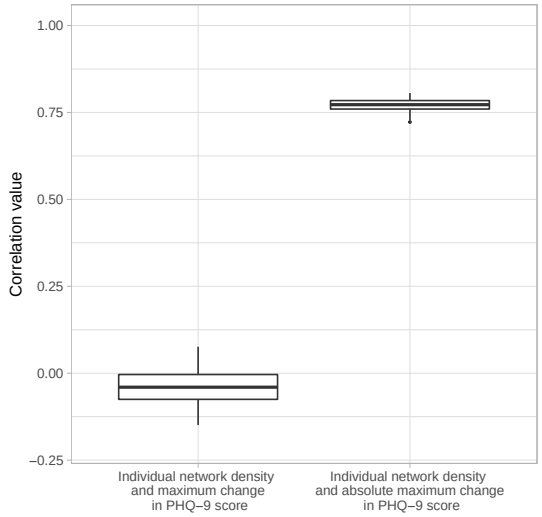


Figure 53: Boxplot showing the correlation between individual temporal network density and the (absolute) maximum change in PHQ-9 score when re-estimating mlVAR networks using 80% of the data.

F

SUPPLEMENTS TO CHAPTER 7

F.1 CONSTRUCTING STABILITY LANDSCAPES FOR THE GVAR MODEL

Let $\mathbf{y}_t$ represent an n_v length vector with responses. We can model these responses using a lag-1 Graphical Vector Autoregressive (GVAR) model:

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{B}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\zeta}_t.$$

Here, the *transpose* of $\mathbf{B}$ is often used to draw a temporal network, and $\boldsymbol{\mu}$ represents the mean or expected value of $\mathbf{y}$. We assume that $\boldsymbol{\zeta}_t$ is normally distributed:

$$\boldsymbol{\zeta}_T \sim \mathcal{N}(\mathbf{o}, \boldsymbol{\Sigma}).$$

Where the matrix $\boldsymbol{\Sigma}$ represents the variance–covariance matrix of $\boldsymbol{\zeta}$. This matrix can further be modeled as a Gaussian Graphical Model [GGM 82]:

$$\boldsymbol{\Sigma} = \boldsymbol{\Delta}(\mathbf{I} - \boldsymbol{\Omega})^{-1}\boldsymbol{\Delta},$$

in which $\boldsymbol{\Omega}$ represents the partial correlation matrix with zeroes on the diagonal, which is typically used to draw a *contemporaneous* network. The $\boldsymbol{\Delta}$ matrix is a diagonal scaling matrix that controls the variance.

Of note, $\boldsymbol{\zeta}$ represents the *innovation* in the model and not as commonly stated a residual. A residual would indicate measurement error or something orthogonal to the process of interest. The innovation term, however, is only orthogonal to previous time points. Let k be an integer, then it follows that:

$$\operatorname{cov}(\boldsymbol{\zeta}_T, \mathbf{y}_{T+k}) = \begin{cases} = \mathbf{O} & \text{if } k < 0 \\ \neq \mathbf{O} & \text{if } k \geqslant 0 \end{cases}$$

Because $\boldsymbol{\zeta}$ is the only exogenous random variable in this model, the model is linear, and $\boldsymbol{\zeta}$ is Gaussian, it follows that $\mathbf{y}$ is itself also Gaussian. We will assume the VAR model is *stationary*; which means that we assume the expected values and variance–covariance structure of $\mathbf{y}$ is the same for any time point t:

$$\begin{aligned} \mathbb{E}(\mathbf{y}_T) &= \mathbb{E}(\mathbf{y}_{T+k}) = \boldsymbol{\mu} \quad \forall_k \\ \operatorname{var}(\mathbf{y}_T) &= \operatorname{var}(\mathbf{y}_{T+k}) = \boldsymbol{\Sigma} \quad \forall_k, \end{aligned}$$

in which $\mathbb{E}(\ldots)$ represents the expected value and $\text{var}(\ldots)$ the variance–covariance matrix. We can then see that:

$$\begin{aligned}
\mathbb{E}(\mathbf{y}_T) &= \mathbb{E}\left[\boldsymbol{\mu} + \mathbf{B}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\zeta}_t\right] \\
&= \boldsymbol{\mu} + \mathbf{B}\mathbb{E}(\mathbf{y}_{t-1}) - \mathbf{B}\mathbb{E}(\boldsymbol{\mu}) + \mathbb{E}(\boldsymbol{\zeta}_t) \\
&= \boldsymbol{\mu} + \mathbf{B}\mathbb{E}(\mathbf{y}_{t-1}) - \mathbf{B}\mathbb{E}(\boldsymbol{\mu}) + \mathbf{o} \\
&= \boldsymbol{\mu} + \mathbf{B}\boldsymbol{\mu} - \mathbf{B}\boldsymbol{\mu} + \mathbf{o} \\
&= \boldsymbol{\mu}.
\end{aligned}$$

$$\begin{aligned}
\text{vec}(\boldsymbol{\Sigma} - \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top) &= \text{vec}(\boldsymbol{\Sigma}^{(\zeta)}) \\
\text{vec}(\boldsymbol{\Sigma}) - (\mathbf{B} \otimes \mathbf{B})\text{vec}(\boldsymbol{\Sigma}) &= \text{vec}(\boldsymbol{\Sigma}^{(\zeta)}) \\
(\mathbf{I} - \mathbf{B} \otimes \mathbf{B})\,\text{vec}(\boldsymbol{\Sigma}) &= \text{vec}(\boldsymbol{\Sigma}^{(\zeta)}) \\
\text{vec}(\boldsymbol{\Sigma}) &= (\mathbf{I} - \mathbf{B} \otimes \mathbf{B})^{-1}\,\text{vec}(\boldsymbol{\Sigma}^{(\zeta)}) \\
\boldsymbol{\Sigma} &= \text{mat}\left[(\mathbf{I} - \mathbf{B} \otimes \mathbf{B})^{-1}\,\text{vec}(\boldsymbol{\Sigma}^{(\zeta)})\right].
\end{aligned}$$

In this derivation, vec represents the vectorization operator (stacking the columns), mat its inverse that takes columns and makes a matrix (basically the `matrix` function in R), and $\otimes$ represents the Kronecker product.

With these expressions at hand, we can readily form the stationary density function of $\mathbf{y}$ and use the negative log density to draw a stability landscape based on the item score.

We can use this to obtain the stability landscape for the sum score of the variables in the GVAR model. Let s_y be the sum-score at time-point t:

$$s_t = \mathbf{1}^\top \mathbf{y}_t = \sum_{v=1}^{n_v} y_{tv},$$

in which $\mathbf{1}$ is a vector of ones. We can then obtain the mean and variance of the sum-score:

$$\begin{aligned}
\mu^{(s)} &= \mathbf{1}^\top \boldsymbol{\mu} = \sum_{v=1}^{n_v} \mu_v \\
\boldsymbol{\Sigma}^{(s)} &= \mathbf{1}^\top \boldsymbol{\Sigma} \mathbf{1} = \sum_{i=1}^{n_v} \sum_{j=1}^{n_v} \sigma_{ij}
\end{aligned}$$

F.2 SIMULATION STUDY OF THE ISING MODEL WITHIN DIFFERENT PARAMETER SPACE

As the dynamics of the Ising model are highly dependent upon the parameterization of the model, here we present an alternative simulation study. In contrast to the fully connected network model in Chapter 7, the Ising network in this simulation is following small-world properties and is therefore more sparse (i.e., less edges present).

F.2.1 *Constructing the Ising network*

To determine the network structure (i.e., whether an edge is present or absent) of the baseline Ising network, ($\boldsymbol{\omega}$), we use the `sample_smallworld()` function from the *igraph* package [58]. Next, we sample the exact edge weights from a Gaussian distribution with $\mu = 1$ and $sd = .3$. We multiply the baseline network with a constant, so-called *multiplier*. The multiplier determines the connectivity of the network, where we have three connectivity conditions: we multiply the network either by .5, 1 or 2. This results in a less strongly connected network, the baseline network, and a more strongly connected network. In addition, we study three different threshold conditions, namely $\tau = -1$ (tendency of all symptoms to be off), $\tau = 0$ (no tendency), and $\tau = 1$ (tendency of all symptoms to be on). As the dynamics of the Ising model are dependent upon the encoding of the nodes, we construct stability landscapes for two different type of network models: Ising model in the $\{-1, 1\}$ domain, the Ising model in the $\{0, 1\}$ domain.

F.2.2 *Results of the Ising model dynamics within the {-1,1} domain*

Figure 54 shows nine stability landscapes with varying parameter values of the Ising model within the {-1,1} domain. The columns indicate the difference in thresholds (i.e., $\tau \in \{-1, 0, 1\}$), while the rows indicate the difference in the multiplier that was used to either make the baseline network sparser (i.e., $multiplier = 0.5$), or denser (i.e., $multiplier = 2$).

The Ising model within the $\{-1, 1\}$ domain with sufficiently strong connectivity has two stable states. The threshold parameter, (τ), determines which of the two stable states is the most likely to occur. When all threshold parameters are negative, the most likely stable state is the state in which all symptoms are "deactiviated" indicated by the stable state at a sum score of 0, see left column in Figure 54. In contrast, when all threshold parameters are positive, the most likely stable state is the state in which all symptoms are "activated" as indicated by the stable state at a sum score of 6, see right column in Figure 54. When all thresholds are set to zero, the Ising model shows two equally likely stable states: one basin appears around the sum score of 0 and one basin appears around a sum score of 6, see middle column in Figure 54.

The connectivity of the Ising model determines (1) the number of stable states (one or two), and (2) the stability of the stable states. Given thresholds of zero, with increasing connectivity *both* stable states become more resilient to perturbations. In addition, when the system is not near a tipping point, higher connectivity implies the the system will return to its original stable state more quickly after a perturbation (see the bot-

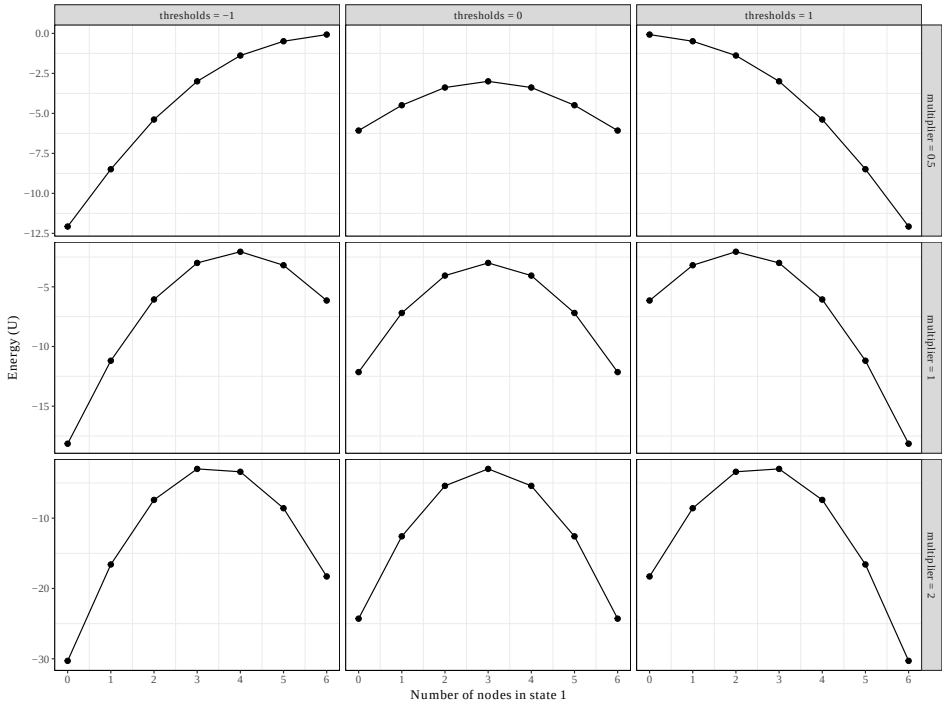


Figure 54: The potential stability landscapes for Ising model, where absent symptoms are coded as −1 and present symptoms as 1. We varied the density of the network and the thresholds. The x-axis represents the state of the system, namely the sum score of the variables, and the y-axis the potential (U) of that state. The lower the potential, the more stable the state is.

tom row in Figure 54). With decreasing connectivity the system responds more strongly to perturbations, as indicated by a more shallow basin, and the system would more slowly return to its current stable state after a perturbation (see the top row in Figure 54). Furthermore, as the connectivity increases, a pattern of increasing discontinuity arises. That is, at lower levels of connectivity, the Ising model may have only one stable state and exhibit a relatively smooth transition between different levels of symptom activation. However, as the connectivity increases a tipping point (as indicated in the landscape by a hill) emerges. This is the classic dynamical behaviour of the cusp catastrophe [92], where connectivity plays the role of splitting variable and external perturbations act as the normal variable.

Thus, together the threshold and connectivity parameters determine the shape of the landscape. When connectivity is high ($multiplier = 2$), the system is more "stuck" in a stable state and subsequently shows less variation around the stable state as indicated by the steepness of the basin. However, which of the two states is more likely is completely dependent upon the threshold parameter (τ), see lower panel in Figure 54.

If connectivity is low, the system shows more variation around the stable states, see upper panel in Figure 54.

F.2.3 *Results of the Ising model dynamics within the {0,1} domain*

Figure 55 shows nine stability landscapes with varying parameter values of the Ising model within the {0,1} domain. The columns indicate the difference in threshold parameters (i.e., $\tau \in \{-1, 0, 1\}$), while the rows indicate the difference in the multiplier that was used to either make the baseline network sparser (i.e., $\mathtt{multiplier} = 0.5$), or denser (i.e., $\mathtt{multiplier} = 2$).

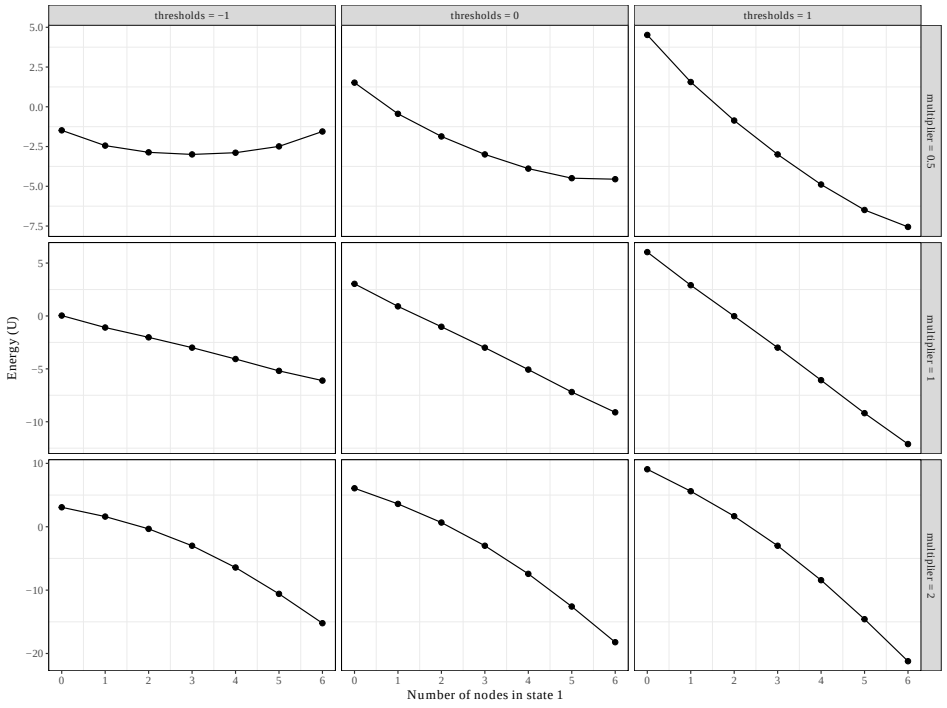


Figure 55: The potential stability landscapes for Ising model, where absent symptoms are coded as 0 and present symptoms as 1. We varied the density of the network and the thresholds. The x-axis represents the state of the system, namely the sum score of the variables, and the y-axis the potential (U) of that state. The lower the potential, the more stable the state is.

In the selected network model, for the $\{0, 1\}$ domain the Ising model dynamics have only one stable state, in which all symptoms are activated, i.e., sum score of 6. Either increasing thresholds or network connectivity will cause the stable state to become more stable. Only the upper left panel in Figure 54 with negative thresholds and low connectivity shows a landscape where the stable state is not located at the maximum symptom activation.

The results show that within the $\{0, 1\}$ domain, both the threshold parameter (τ) and the network connectivity determine the depth and steepness of the stable state, see Figure 55. When threshold parameters are positive, the stable state is more deep and steep, while when threshold parameters are negative the stable state is less deep and steep. The same effect, but stronger, occurs for increasing network connectivity, i.e., increasing the network connectivity results in a more deep and steep unhealthy stable state.

G

REFERENCES

[1] F. Abegaz and E. Wit. "Sparse time series chain graphical models for reconstructing genetic networks." In: *Biostatistics* 14.3 (2013), pp. 586–599.

[2] J. Adolf, N. K. Schuurman, P. Borkenau, D. Borsboom, and C. V. Dolan. "Measurement invariance within and between individuals: A distinct problem in testing the equivalence of intra-and inter-individual model structures." In: *Frontiers in Psychology* 5.1 (2014), pp. 1–14.

[3] K. E. Adolph, S. R. Robinson, J. W. Young, and F. Gill-Alvarez. "What is the shape of developmental change?" In: *Psychological review* 115.3 (2008), pp. 527–543.

[4] S. H. Aggen, M. C. Neale, and K. S. Kendler. "DSM criteria for major depression: evaluating symptom patterns using latent-trait item response models." In: *Psychological Medicine* 35 (2005), pp. 475–487.

[5] C. J. Albers and L. F. Bringmann. "Inspecting gradual and abrupt changes in emotion dynamics with the time-varying change point autoregressive model." In: *European Journal of Psychological Assessment* 36.3 (2020), pp. 492–499.

[6] W. A. Arrindell and J. H. M. Ettema. *SCL-90. Handleiding bij een multidimensionele psychopathologie-indicator*. Swets & Zeitlinger, 2003.

[7] K. A. Barford, P. Koval, P. Kuppens, and L. D. Smillie. "When good feelings turn mixed: Affective dynamics and big five trait predictors of mixed emotions in daily life." In: *European Journal of Personality* 34.3 (2020), pp. 393–411.

[8] C. Beard, A. J. Millner, M. J. Forgeard, E. I. Fried, K. J. Hsu, M. T. Treadway, C. V. Leonard, S. Kertz, and T. Björgvinsson. "Network analysis of depression and anxiety symptom relationships in a psychiatric sample." In: *Psychological Medicine* 46.16 (2016), pp. 3359–3369.

[9] E. D. Beck and J. J. Jackson. "Consistency and change in idiographic personality: A longitudinal ESM network study." In: *Journal of Personality and Social Psychology* 118.5 (2020), pp. 1080–1100.

[10] E. D. Beck and J. J. Jackson. "Consistency and change in idiographic personality: A longitudinal ESM network study." In: *Journal of Personality and Social Psychology* 118.5 (2020), pp. 1080–1100.

[11] A. M. Beltz and K. M. Gates. "Network mapping with GIMME." In: *Multivariate Behavioral Research* 52.6 (2017), pp. 789–804.

[12] A. M. Beltz, A. G. C. Wright, B. N. Sprague, and P. C. M. Molenaar. "Bridging the nomothetic and idiographic approaches to the analysis of clinical data." In: *Assessment* 23.4 (2016), pp. 447–458.

[13] A. Berta, C. M. Ángel, G.-S. Clara, and H. Rubén. "A bibliometric analysis of 10 years of research on symptom networks in psychopathology and mental health." In: *Psychiatry Research* 308 (2022), pp. 1–9.

[14] T. F. Blanken, A. M. Isvoranu, and S. Epskamp. "Estimating network structures using model selection." In: *Network Psychometrics with R: A Guide for Behavioral and Social Scientists*. Ed. by A. M. Isvoranu, S. Epskamp, L. J. Waldorp, and D. Borsboom. Routledge, 2022, pp. 111–132.

[15] N. Bolger and K. S. Zee. "Heterogeneity in temporal processes: Implications for theories in health psychology." In: *Health and Well-Being* 11.2 (2019), pp. 198–201.

[16] S. H. Booij, M. Wichers, P. De Jonge, S. Sytema, J. Van Os, L. Wunderink, and J. T. Wigman. "Study protocol for a prospective cohort study examining the predictive potential of dynamic symptom networks for the onset and progression of psychosis: the Mapping Individual Routes of Risk and Resilience (Mirorr) study." In: *BMJ Open* 8.1 (2018), pp. 1–12.

[17] C. D. van Borkulo, D. Borsboom, S. Epskamp, T. F. Blanken, L. Boschloo, R. A. Schoevers, and L. J. Waldorp. "A new method for constructing networks from binary data." In: *Scientific reports* 4.1 (2014), p. 5918.

[18] C. D. van Borkulo, D. Borsboom, and R. A. Schoevers. "Group-level symptom networks in depression—reply." In: *JAMA Psychiatry* 73.4 (2016), pp. 411–412.

[19] C. D. van Borkulo, L. Boschloo, D. Borsboom, B. W. Penninx, L. J. Waldorp, and R. A. Schoevers. "Association of symptom network structure with the course of depression." In: *JAMA Psychiatry* 72.12 (2015), pp. 1219–1226.

[20] C. D. van Borkulo, L. Boschloo, J Kossakowski, P. Tio, R. A. Schoevers, D. Borsboom, and L. J. Waldorp. "Comparing network structures on three aspects: A permutation test." In: *Psychological Methods* 28 (6 2024), pp. 1273–1285.

[21] C. D. van Borkulo and S. Epskamp. *IsingFit*. R package version 0.4. 2023. URL: https://cran.r-project.org/package=IsingFit.

[22] D. Borsboom. "Psychometric perspectives on diagnostic systems." In: *Journal of Clinical Psychology* 64.9 (2008), pp. 1089–1108.

[23] D. Borsboom. "A network theory of mental disorders." In: *World Psychiatry* 16 (1 2017), pp. 5–13.

[24] D. Borsboom. "Mental disorders, network models, and dynamical systems." In: *Philosophical Issues in Psychiatry IV: Psychiatric Nosology DSM-5 (International Perspectives in Philosophy and Psychiatry)* (2017), pp. 80–97.

[25] D. Borsboom, A. O. Cramer, E. I. Fried, A. M. Isvoranu, D. J. Robinaugh, J. Dalege, and H. L. van der Maas. "Network perspectives." In: *Network Psychometrics with R*. Routledge, 2022, pp. 9–27.

[26] D. Borsboom, A. O. Cramer, and A. Kalis. "Brain disorders? Not really: Why network structures block reductionism in psychopathology research." In: *Behavioral and Brain Sciences* 42 (2019), e2.

[27] D. Borsboom, A. O. Cramer, V. D. Schmittmann, S. Epskamp, and L. J. Waldorp. "The small world of psychopathology." In: *PloS ONE* 6.11 (2011), e27407.

[28] D. Borsboom, A. O. Cramer, et al. "Network analysis: An integrative approach to the structure of psychopathology." In: *Annual Review of Clinical Psychology* 9.1 (2013), pp. 91–121.

[29] D. Borsboom, M. K. Deserno, M. Rhemtulla, S. Epskamp, E. I. Fried, R. J. McNally, D. J. Robinaugh, M. Perugini, J. Dalege, G. Costantini, et al. "Network analysis of multivariate data in psychological science." In: *Nature Reviews Methods Primers* 1.1 (2021), pp. 1–18.

[30] D. Borsboom, E. I. Fried, S. Epskamp, L. J. Waldorp, C. D. van Borkulo, H. L. van der Maas, and A. O. Cramer. "False alarm? A comprehensive reanalysis of "Evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger (2017)." In: *Journal of Abnormal Psychology* 126.7 (2017), pp. 989–999.

[31] D. Borsboom, H. L. van der Maas, J. Dalege, R. A. Kievit, and B. D. Haig. "Theory construction methodology: A practical framework for building theories in psychology." In: *Perspectives on Psychological Science* 16.4 (2021), pp. 756–766.

[32] D. Borsboom, G. J. Mellenbergh, and J. Van Heerden. "The theoretical status of latent variables." In: *Psychological Review* 110.2 (2003), pp. 203–219.

[33] E. H. Bos and R. B. Wanders. "Group-level symptom networks in depression." In: *JAMA Psychiatry* 73.4 (2016), p. 411.

[34] F. M. Bos, E. I. Fried, S. D. Hollon, L. F. Bringmann, S. Dimidjian, R. J. DeRubeis, and C. L. Bockting. "Cross-sectional networks of depressive symptoms before and after antidepressant medication treatment." In: *Social Psychiatry and Psychiatric Epidemiology* 53 (2018), pp. 617–627.

[35] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: Forecasting and control*. San Francisco, CA: Holden-Day, 1970.

[36] L. F. Bringmann, M. Helmich, M. Eronen, and M. Voelkle. "Complex Systems Approaches to Psychopathology." In: *Oxford Textbook of Psychopathology* (2023), pp. 103–122.

[37] L. F. Bringmann. "Person-specific networks in psychopathology: Past, present, and future." In: *Current Opinion in Psychology* 41 (2021), pp. 59–64.

[38] L. F. Bringmann, C. Albers, C. Bockting, D. Borsboom, E. Ceulemans, A. Cramer, S. Epskamp, M. I. Eronen, E. Hamaker, P. Kuppens, et al. "Psychopathological networks: Theory, methods and practice." In: *Behaviour Research and Therapy* 149 (2022).

[39] L. F. Bringmann, T. Elmer, S. Epskamp, R. W. Krause, D. Schoch, M. Wichers, J. T. Wigman, and E. Snippe. "What do centrality measures measure in psychological networks?" In: *Journal of Abnormal Psychology* 128.8 (2019), pp. 892–903.

[40] L. F. Bringmann, N. Vissers, M. Wichers, N. Geschwind, P. Kuppens, F. Peeters, D. Borsboom, and F. Tuerlinckx. "A network approach to psychopathology: New insights into clinical longitudinal data." In: *PloS ONE* 8.4 (2013), e60188.

[41] J. Burger, R. H. A. Hoekstra, A. C. Mansueto, and S. Epskamp. "Network estimation from time series and panel data." In: *Network Psychometrics with R: A Guide for Behavioral and Social Scientists*. Ed. by A. M. Isvoranu, S. Epskamp, L. J. Waldorp, and D. Borsboom. Routledge, 2022, pp. 169–192.

[42] J. Burger, D. C. V. D. Veen, D. J. Robinaugh, R. Quax, H. Riese, R. A. Schoevers, and S. Epskamp. "Bridging the gap between complexity science and clinical practice by formalizing idiographic theories: A computational model of functional analysis." In: *BMC Medicine* 18.99 (2020), pp. 1–18.

[43] J. Burger, D. C. van der Veen, D. J. Robinaugh, R. Quax, H. Riese, R. A. Schoevers, and S. Epskamp. "Bridging the gap between complexity science and clinical practice by formalizing idiographic theories: a computational model of functional analysis." In: *BMC Medicine* 18 (2020), pp. 1–18.

[44] K. P. Burnham and D. R. Anderson. "Multimodel inference: understanding AIC and BIC in model selection." In: *Sociological Methods & Research* 33.2 (2004), pp. 261–304.

[45] S. Calugi, L. Dametti, M. Chimini, A. Dalle Grave, and R. Dalle Grave. "Change in eating-disorder psychopathology network structure in patients with anorexia nervosa treated with intensive cognitive behavior therapy." In: *International Journal of Eating Disorders* 54.10 (2021), pp. 1800–1809.

[46] G. Chen, D. R. Glen, Z. S. Saad, J. P. Hamilton, M. E. Thomason, I. H. Gotlib, and R. W. Cox. "Vector autoregression, structural equation modeling, and their synthesis in neuroimaging data analysis." In: *Computers in Biology and Medicine* 41.12 (2011), pp. 1142–1155.

[47] J. Chen and Z. Chen. "Extended Bayesian information criteria for model selection with large model spaces." In: *Biometrika* 95.3 (2008), pp. 759–771.

[48] S. Chib and E. Greenberg. "Understanding the metropolis-hastings algorithm." In: *The American Statistician* 49.4 (1995), pp. 327–335.

[49] W. S. Cleveland. "Robust locally weighted regression and smoothing scatterplots." In: *Journal of the American Statistical Association* (1979), pp. 829–836.

[50] M. Constantin, N. K. Schuurman, and J. Vermunt. "A general Monte Carlo method for sample size analysis in the context of network models." In: *Psychological Methods* (2023), Advance online publication.

[51] G. Costantini, S. Epskamp, D. Borsboom, M. Perugini, R. Mõttus, L. J. Waldorp, and A. O. Cramer. "State of the aRt personality research: A tutorial on network analysis of personality data in R." In: *Journal of Research in Personality* 54 (2015), pp. 13–29.

[52] G. Costantini, J. Richetin, E. Preti, E. Casini, S. Epskamp, and M. Perugini. "Stability and variability of personality networks. A tutorial on recent developments in network psychometrics." In: *Personality and Individual Differences* 136 (2019), pp. 68–78.

[53] A. O. Cramer, C. D. van Borkulo, E. J. Giltay, H. L. Van Der Maas, K. S. Kendler, M. Scheffer, and D. Borsboom. "Major depression as a complex dynamic system." In: *PloS ONE* 11.12 (2016), e0167490.

[54] A. O. Cramer, L. J. Waldorp, H. L. Van Der Maas, and D. Borsboom. "Comorbidity: A network perspective." In: *Behavioral and Brain Sciences* 33.2-3 (2010), pp. 137–150.

[55] A. O. Cramer, L. J. Waldorp, H. L. Van Der Maas, and D. Borsboom. "Complex realities require complex theories: Refining and extending the network approach to mental disorders." In: *Behavioral and Brain Sciences* 33.2-3 (2010), pp. 178–193.

[56] I. Cribben, R. Haraldsdottir, L. Y. Atlas, T. D. Wager, and M. A. Lindquist. "Dynamic connectivity regression: determining state-related changes in brain connectivity." In: *Neuroimage* 61.4 (2012), pp. 907–920.

[57] L. J. Cronbach. "Coefficient alpha and the internal structure of tests." In: *Psychometrika* 16.3 (1951), pp. 297–334.

[58] M. G. Csardi. "Package 'igraph'." In: *Last Accessed* 3.09 (2013), p. 2013.

[59] J. Cui, F. Hasselman, and A. Lichtwarck-Aschoff. "Unlocking nonlinear dynamics and multistability from intensive longitudinal data: A novel method." In: *Psychological Methods* (2023), Advance online publication.

[60] J. Cui, A. Lichtwarck-Aschoff, M. Olthof, T. Li, and F. Hasselman. "From metaphor to computation: Constructing the potential landscape for multivariate psychological formal models." In: *Multivariate Behavioral Research* (2022), pp. 1–19.

[61] J. Cui, G. Lunansky, A. Lichtwarck-Aschoff, N. Mendoza, and F. Hasselman. "Quantifying the Stability Landscapes of Psychological Networks." In: *PsyArXiv* (2023).

[62] T. J. Cunningham, E. C. Fields, and E. A. Kensinger. "Boston College daily sleep and well-being survey data during early phase of the COVID-19 pandemic." In: *Scientific Data* 8.1 (2021), pp. 1–6.

[63] J. E. Curtiss, M. Pinaire, D. Fulford, R. J. McNally, and S. G. Hofmann. "Temporal and contemporaneous network structures of affect and physical activity in emotional disorders." In: *Journal of Affective Disorders* 315 (2022), pp. 139–147.

[64] F. Dablander and M. Hinne. "Node centrality measures are a poor substitute for causal inference." In: *Scientific Reports* 9.1 (2019), p. 6846.

[65] J. Dalege, D. Borsboom, F. van Harreveld, and H. L. van der Maas. "Network analysis on attitudes: A brief tutorial." In: *Social Psychological and Personality Science* 8.5 (2017), pp. 528–537.

[66] J. Dalege, D. Borsboom, F. Van Harreveld, H. Van den Berg, M. Conner, and H. L. Van der Maas. "Toward a formalized account of attitudes: The Causal Attitude Network (CAN) model." In: *Psychological Review* 123.1 (2016), pp. 2–22.

[67] S. De Vos, K. J. Wardenaar, E. H. Bos, E. C. Wit, M. E. Bouwmans, and P. De Jonge. "An investigation of emotion dynamics in major depressive disorder patients and healthy persons using sparse longitudinal networks." In: *PloS ONE* 12.6 (2017), e0178586.

[68] N. van Dongen, R. van Bork, A. Finnemann, H. van der Maas, D. Robinaugh, J. M. B. Haslbeck, J. de Ron, J. Sprenger, and D. Borsboom. "Productive Explanation: A Framework for Evaluating Explanations in Psychological Science." In: *PsyArXiv* (2022).

[69] C. Dormann and M. A. Griffin. "Optimal time lags in panel studies." In: *Psychological Methods* 20.4 (2015), pp. 489–505.

[70] O. V. Ebrahimi, D. Borsboom, R. H. A. Hoekstra, S. Epskamp, E. G. Ostinelli, J. A. Bastiaansen, and A. Cipriani. "Towards precision in the diagnostic profiling of patients: Leveraging symptom dynamics in the assessment of major depressive disorder." In: *The British Journal of Psychiatry* 224.5 (2023), pp. 157–163.

[71] O. V. Ebrahimi, J. Burger, A. Hoffart, and S. U. Johnson. "Within- and across-day patterns of interplay between depressive symptoms and related psychopathological processes: a dynamic network approach during the COVID-19 pandemic." In: *BMC Medicine* 19.1 (2021), pp. 1–17.

[72] S. Epskamp, A. M. Isvoranu, and M. Cheung. "Meta-analytic Gaussian network aggregation." In: *Psychometrika* 87.1 (2022), pp. 12–46.

[73] S. Epskamp. "Psychometric network models from time-series and panel data." In: *Psychometrika* 85.1 (2020), pp. 206–231.

[74] S. Epskamp. *psychonetrics: Structural Equation Modeling and Confirmatory Network Analysis*. R package version 0.10. 2021. URL: `http://psychonetrics.org/`.

[75] S. Epskamp. *qgraph: Graph Plotting Methods, Psychometric Data Visualization and Graphical Model Estimation*. R package version 1.9.3. 2022. URL: `https://cran.r-project.org/package=qgraph`.

[76] S. Epskamp. *IsingSampler*. R package version 10.2.3. 2023. URL: `https://cran.r-project.org/package=IsingSampler`.

[77] S. Epskamp. *graphicalVAR: Graphical VAR for experience sampling data*. R package version 0.3.3. 2023. URL: `https://cran.r-project.org/package=graphicalVAR`.

[78] S. Epskamp. *parSim: Parallel Simulation Studies*. R package version 0.1.5. 2023. URL: https://cran.r-project.org/package=parSim.

[79] S. Epskamp. *psychonetrics*. R package version 0.11.5. 2023. URL: https://cran.r-project.org/package=psychonetrics.

[80] S. Epskamp, C. D. van Borkulo, D. C. van der Veen, M. N. Servaas, A. M. Isvoranu, H. Riese, and A. O. Cramer. "Personalized network modeling in psychopathology: The importance of contemporaneous and temporal Connections." In: *Clinical Psychological Science* 6.3 (2018), pp. 416–427.

[81] S. Epskamp, C. D. van Borkulo, D. C. van der Veen, M. N. Servaas, A. M. Isvoranu, H. Riese, and A. O. Cramer. "Personalized network modeling in psychopathology: The importance of contemporaneous and temporal connections." In: *Clinical Psychological Science* 6.3 (2018), pp. 416–427.

[82] S. Epskamp, D. Borsboom, and E. I. Fried. "Estimating psychological networks and their accuracy: A tutorial paper." In: *Behavior Research Methods* 50.1 (2018), pp. 195–212.

[83] S. Epskamp, M. K. Deserno, and L. F. Bringmann. *mlVAR: Multi-Level Vector Autoregression*. R package version 0.5.1. 2023. URL: https://cran.r-project.org/package=mlVAR.

[84] S. Epskamp and E. I. Fried. "A tutorial on regularized partial correlation networks." In: *Psychological Methods* 23.4 (2018), pp. 617–212.

[85] S. Epskamp and E. I. Fried. *bootnet: Bootstrap Methods for Various Network Estimation Routines*. R package version 1.5.6. 2023. URL: https://cran.r-project.org/package=bootnet.

[86] S. Epskamp, R. H. A. Hoekstra, J. Burger, and L. J. Waldorp. "Longitudinal Design choices: Relating data to analysis." In: *Network Psychometrics with R: A Guide for Behavioral and Social Scientists*. Ed. by A. M. Isvoranu, S. Epskamp, L. J. Waldorp, and D. Borsboom. Routledge, 2022, pp. 157–168.

[87] S. Epskamp, G. K. J. Maris, L. J. Waldorp, and D. Borsboom. "Network Psychometrics." In: *Handbook of psychometrics*. Ed. by P. Irwing, D. Hughes, and T. Booth. New York, NY: Wiley, 2016.

[88] S. Epskamp, M. Rhemtulla, and D. Borsboom. "Generalized network psychometrics: Combining network and latent variable models." In: *Psychometrika* 82 (2017), pp. 904–927.

[89] S. Epskamp, L. J. Waldorp, R. Mõttus, and D. Borsboom. "The Gaussian graphical model in cross-sectional and time-series data." In: *Multivariate Behavioral Research* 53.4 (2018), pp. 453–480.

[90] M. I. Eronen. "The levels problem in psychopathology." In: *Psychological Medicine* 51.6 (2021), pp. 927–933.

[91] J. Fan, H. Liu, Y. Ning, and H. Zou. "High dimensional semiparametric latent graphical model for mixed data." In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79.2 (2017), pp. 405–421.

[92] A. Finnemann, D. Borsboom, S. Epskamp, and H. L. J. van der Maas. "The Theoretical and Statistical Ising Model: A Practical Guide in R." In: *Psych* 3.4 (2021), pp. 594–618.

[93] A. Finnemann, D. Borsboom, S. Epskamp, and H. L. van der Maas. "The theoretical and statistical Ising model: A practical guide in R." In: *Psych* 3.04 (2021), pp. 593–617.

[94] A. J. Fisher, J. D. Medaglia, and B. F. Jeronimus. "Lack of group-to-individual generalizability is a threat to human subjects research." In: *Proceedings of the National Academy of Sciences* 115.27 (2018), E6106–E6115.

[95] A. J. Fisher, J. W. Reeves, G. Lawyer, J. D. Medaglia, and J. A. Rubel. "Exploring the idiographic dynamics of mood and anxiety via network analysis." In: *Journal of Abnormal Psychology* 126.8 (2017), pp. 1044–1056.

[96] Z. F. Fisher, K. A. Bollen, and K. M. Gates. "A limited information estimator for dynamic factor models." In: *Multivariate Behavioral Research* 54.2 (2019), pp. 246–263.

[97] M. K. Forbes, A. G. Wright, K. E. Markon, and R. F. Krueger. "Evidence that psychopathology symptom networks have limited replicability." In: *Journal of Abnormal Psychology* 126.7 (2017), pp. 969–988.

[98] M. K. Forbes, A. G. Wright, K. E. Markon, and R. F. Krueger. "Further evidence that psychopathology networks have limited replicability and utility: Response to Borsboom et al.(2017) and Steinley et al.(2017)." In: *Journal of Abnormal Psychology* 126.7 (2017), 1011––1016.

[99] M. K. Forbes, A. G. Wright, K. E. Markon, and R. F. Krueger. "The network approach to psychopathology: Promise versus reality." In: *World Psychiatry* 18.3 (2019), pp. 272–273.

[100] M. K. Forbes, A. G. Wright, K. E. Markon, and R. F. Krueger. "Quantifying the reliability and replicability of psychopathology network characteristics." In: *Multivariate Behavioral Research* 56.2 (2021), pp. 224–242.

[101] E. I. Fried. "Moving forward: How depression heterogeneity hinders progress in treatment and research." In: *Expert Review of Neurotherapeutics* 17.5 (2017), pp. 423–425.

[102] E. I. Fried, C. D. van Borkulo, A. O. Cramer, L. Boschloo, R. A. Schoevers, and D. Borsboom. "Mental disorders as networks of problems: A review of recent insights." In: *Social Psychiatry and Psychiatric Epidemiology* 52.1 (2017), pp. 1–10.

[103] E. I. Fried, C. D. van Borkulo, and S. Epskamp. "On the importance of estimating parameter uncertainty in network psychometrics: A Response to Forbes et al.(2019)." In: *Multivariate Behavioral Research* 56.2 (2021), pp. 243–248.

[104] E. I. Fried, M. B. Eidhof, S. Palic, G. Costantini, H. M. Huisman-van Dijk, C. L. Bockting, I. Engelhard, C. Armour, A. B. Nielsen, and K.-I. Karstoft. "Replicability and generalizability of posttraumatic stress disorder (PTSD) networks: A cross-cultural multisite study of PTSD symptoms in four trauma patient samples." In: *Clinical Psychological Science* 6.3 (2018), pp. 335–351.

[105] E. I. Fried, F. Papanikolaou, and S. Epskamp. "Mental health and social contact during the COVID-19 pandemic: an ecological momentary assessment study." In: *Clinical Psychological Science* 10.2 (2022), pp. 340–354.

[106] J. Friedman, T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation with the graphical lasso." In: *Biostatistics* 9.3 (2008), pp. 432–441.

[107] M. R. Frumkin, M. L. Piccirillo, E. D. Beck, J. T. Grossman, and T. L. Rodebaugh. "Feasibility and utility of idiographic models in the clinic: a pilot study." In: *Psychotherapy Research* 31.4 (2021), pp. 520–534.

[108] K. M. Gates and P. C. M. Molenaar. "Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples." In: *NeuroImage* 63.1 (2012), pp. 310–319.

[109] K. M. Gates, P. C. Molenaar, F. G. Hillary, N. Ram, and M. J. Rovine. "Automatic search for fMRI connectivity mapping: An alternative to Granger causality testing using formal equivalences among SEM path modeling, VAR, and unified SEM." In: *NeuroImage* 50.3 (2010), pp. 1118–1125.

[110] G. J. Geldhof, K. J. Preacher, and M. J. Zyphur. "Reliability estimation in a multilevel confirmatory factor analysis framework." In: *Psychological Methods* 19.1 (2014), p. 72.

[111] N. Geschwind, F. Peeters, M. Drukker, J. van Os, and M. Wichers. "Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: A randomized controlled trial." In: *Journal of Consulting and Clinical Psychology* 79.5 (2011), pp. 618–628.

[112] P. Gilbert, K. McEwan, R. Mitra, L. Franks, A. Richter, and H. Rockliff. "Feeling safe and content: A specific affect regulation system? Relationship to depression, anxiety, stress, and self-criticism." In: *The Journal of Positive Psychology* 3.3 (2008), pp. 182–191.

[113] C. J. W. Granger. "Investigating causal relations by econometric models and cross-spectral methods." In: *Econometrica Journal of the Econometric Society* 37.3 (1969), pp. 424–438.

[114] P. E. Greenberg, A.-A. Fournier, T. Sisitsky, C. T. Pike, and R. C. Kessler. "The economic burden of adults with major depressive disorder in the United States (2005 and 2010)." In: *The Journal of Clinical Psychiatry* 76.2 (2015), 155—162.

[115] R. N. Groen, E. Snippe, L. F. Bringmann, C. J. Simons, J. A. Hartmann, E. H. Bos, and M. Wichers. "Capturing the risk of persisting depressive symptoms: A dynamic network investigation of patients' daily symptom experiences." In: *Psychiatry Research* 271 (2019), pp. 640–648.

[116] R. E. Grolli, M. E. D. Mingoti, A. G. Bertollo, A. R. Luzardo, J. Quevedo, G. Z. Reus, and Z. M. Ignacio. "Impact of COVID-19 in the mental health in elderly: psychological and biological updates." In: *Molecular Neurobiology* 58 (2021), pp. 1905–1916.

[117] J. J. Gross. "Emotion regulation: Past, present, future." In: *Cognition & Emotion* 13.5 (1999), pp. 551–573.

[118] G. Grynkiv and L. Stentoft. "Stationary threshold vector autoregressive models." In: *Journal of Risk and Financial Management* 11.45 (2018), pp. 1–23.

[119] J. Guo, G. James, E. Levina, G. Michailidis, and J. Zhu. "Principal component analysis with sparse fused loadings." In: *Journal of Computational and Graphical Statistics* 19.4 (2010), pp. 930–946.

[120] S. de Haan-Rietdijk, M. C. Voelkle, L. Keijsers, and E. L. Hamaker. "Discrete-vs. continuous-time modeling of unequally spaced experience sampling method data." In: *Frontiers in Psychology* 8 (2017), pp. 1–19.

[121] K. A. Hallgren. "Conducting simulation studies in the R programming environment." In: *Tutorials in Quantitative Methods for Psychology* 9.2 (2013), pp. 43–60.

[122] M. N. Hallquist, A. G. C. Wright, and P. C. M. Molenaar. "Problems with centrality measures in psychopathology symptom networks: Why network psychometrics cannot escape psychometric theory." In: *Multivariate Behavioral Research* 56.2 (2021), pp. 199–223.

[123] E. L. Hamaker. "Why researchers should think "within-person": A pragmatic rationale." In: *Handbook of research methods for studying daily life*. New York, NY: The Guilford Press, 2012, pp. 21–45.

[124] E. L. Hamaker and C. V. Dolan. "Idiographic data analysis: Quantitative methods—from simple to advanced." In: *Dynamic Process Methodology in the Social and Developmental Sciences*. Ed. by J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, and N. Chaudhary. New York, NY: Springer–Verlag, 2009, pp. 191–216.

[125] E. L. Hamaker, C. V. Dolan, and P. C. Molenaar. "On the nature of SEM estimates of ARMA parameters." In: *Structural Equation Modeling* 9.3 (2002), pp. 347–368.

[126] E. L. Hamaker, C. V. Dolan, and P. C. Molenaar. "Statistical modeling of the individual: Rationale and application of multivariate stationary time series analysis." In: *Multivariate Behavioral Research* 40.2 (2005), pp. 207–233.

[127] E. L. Hamaker and R. P. P. P. Grasman. "To center or not to center? Investigating inertia with a multilevel autoregressive model." In: *Frontiers in Psychology* 5 (2015), pp. 1–5.

[128] E. L. Hamaker, R. P. Grasman, and J. H. Kamphuis. "Modeling BAS dysregulation in bipolar disorder: Illustrating the potential of time series analysis." In: *Assessment* 23.4 (2016), pp. 436–446.

[129] E. L. Hamaker and M. Wichers. "No time like the present: Discovering the hidden dynamics in intensive longitudinal data." In: *Current Directions in Psychological Science* 26.1 (2017), pp. 10–15.

[130] J. Hamilton. *Time series analysis*. Princeton University Press, 1994.

[131] A. Hardy, R. Emsley, D. Freeman, P. Bebbington, P. A. Garety, E. E. Kuipers, G. Dunn, and D. Fowler. "Psychological mechanisms mediating effects between trauma and psychotic symptoms: the role of affect regulation, intrusive trauma memory, beliefs, and depression." In: *Schizophrenia Bulletin* 42.suppl. no. 1 (2016), pp. S34–S43.

[132] J. M. B. Haslbeck. *mgm*. R package version 1.2-14. 2023. URL: `https://cran.r-project.org/package=IsingFit`.

[133] J. M. B. Haslbeck. "Estimating group differences in network models using moderation analysis." In: *Behavior Research Methods* (2022), pp. 1–19.

[134] J. M. B. Haslbeck, L. F. Bringmann, and L. J. Waldorp. "A tutorial on estimating time-varying vector autoregressive models." In: *Multivariate Behavioral Research* 56.1 (2021), pp. 120–149.

[135] J. M. B. Haslbeck, S. Epskamp, M. Marsman, and L. J. Waldorp. "Interpreting the Ising model: The input matters." In: *Multivariate Behavioral Research* 56.2 (2021), pp. 303–313.

[136] J. M. B. Haslbeck, O. Ryan, and F. Dablander. "Multimodality and skewness in emotion time series." In: *Emotion* 23.8 (2023), pp. 2117–2141.

[137] J. M. B. Haslbeck, O. Ryan, H. L. van der Maas, and L. J. Waldorp. "Modeling Change in Networks." In: *Network Psychometrics with R: A Guide for Behavioral and Social Scientists*. Ed. by A. M. Isvoranu, S. Epskamp, L. J. Waldorp, and D. Borsboom. Routledge, 2022, pp. 193–209.

[138] A. M. Hayes, C. Yasinski, J. B. Barnes, and C. L. Bockting. "Network destabilization and transition in depression: New methods for studying the dynamics of therapeutic change." In: *Clinical Psychology Review* 41 (2015), pp. 27–39.

[139] M. T. Heino, D. Proverbio, G. Marchand, K. Resnicow, and N. Hankonen. "Attractor landscapes: A unifying conceptual model for understanding behaviour change across scales of observation." In: *Health Psychology Review* 17.4 (2023), pp. 655–672.

[140] T. R. Henry, D. J. Robinaugh, and E. I. Fried. "On the control of psychological networks." In: *Psychometrika* 87.1 (2022), pp. 188–213.

[141] I. B. Hickie, E. M. Scott, D. F. Hermens, S. L. Naismith, A. J. Guastella, M. Kaur, A. Sidis, B. Whitwell, N. Glozier, T. Davenport, et al. "Applying clinical staging to young people who present for mental health care." In: *Early Intervention in Psychiatry* 7.1 (2013), pp. 31–43.

[142] R. H. A. Hoekstra, S. Epskamp, D. Borsboom, and R. J. McNally. *Testing similarity in longitudinal networks: The Individual Network Invariance Test (INIT)*. Open Science Framework. 2023, October 13.

[143] R. H. A. Hoekstra, S. Epskamp, and D. Borsboom. "Heterogeneity in Individual Network Analysis: Reality or Illusion?" In: *Multivariate Behavioural Research* 58.4 (2023), 762—786.

[144] R. H. A. Hoekstra, S. Epskamp, A. A. Nierenberg, D. Borsboom, and R. J. McNally. "Testing similarity in longitudinal networks: The Individual Network Invariance Test (INIT)." In: *Psychological Methods* (2024), Advance online publication.

[145] S. G. Hofmann, J. Curtiss, and R. J. McNally. "A complex network perspective on clinical science." In: *Perspectives on Psychological Science* 11.5 (2016), pp. 597–605.

[146] K. Hoorelbeke, N. Van den Bergh, M. Wichers, and E. H. Koster. "Between vulnerability and resilience: A network analysis of fluctuations in cognitive risk and protective factors following remission from depression." In: *Behaviour Research and Therapy* 116 (2019), pp. 1–9.

[147] K. B. Huth, J. de Ron, A. E. Goudriaan, J. Luigjes, R. Mohammadi, R. J. van Holst, E.-J. Wagenmakers, and M. Marsman. "Bayesian analysis of cross-sectional networks: A tutorial in R and JASP." In: *Advances in Methods and Practices in Psychological Science* 6.4 (2023), pp. 1–18.

[148] K. Huth and S. Keetelaar. *easybgm*. R package version 0.1.1. 2023. URL: `https://cran.r-project.org/package=easybgm`.

[149] A. M. Isvoranu and S. Epskamp. "Which Estimation Method to Choose in Network Psychometrics? Deriving Guidelines for Applied Researchers." In: *Psychological Methods* 28.4 (2023), pp. 925–946.

[150] A. M. Isvoranu, C. D. van Borkulo, L.-L. Boyette, J. T. Wigman, C. H. Vinkers, D. Borsboom, and G. Investigators. "A network approach to psychosis: pathways between childhood trauma and psychotic symptoms." In: *Schizophrenia Bulletin* 43.1 (2017), pp. 187–196.

[151] A. M. Isvoranu and S. Epskamp. "Which estimation method to choose in network psychometrics? Deriving guidelines for applied researchers." In: *Psychological Methods* 28.4 (2023), pp. 925–946.

[152] A. M. Isvoranu, S. Epskamp, and M. W.-L. Cheung. "Network models of posttraumatic stress disorder: A meta-analysis." In: *Journal of Abnormal Psychology* 130.8 (2021), 841–861.

[153] A. M. Isvoranu, S. Epskamp, L. Waldorp, and D. Borsboom. *Network psychometrics with R: A guide for behavioral and social scientists*. Routledge, 2022.

[154] D. V. Jeste, B. W. Palmer, D. C. Rettew, and S. Boardman. "Positive psychiatry: its time has come." In: *The Journal of Clinical Psychiatry* 76.6 (2015), pp. 675–683.

[155] H Jiang, X Fei, H Lui, K Roeder, J Lafferty, L Wasserman, L Xingguo, and T Zhao. *huge: High-Dimensional Undirected Graph Estimation*. R package version 1.3.4. 2021. URL: `https://cran.r-project.org/package=huge`.

[156] A. Jongeneel, G. Aalbers, I. Bell, E. I. Fried, P. Delespaul, H. Riper, M. Van Der Gaag, and D. Van Den Berg. "A time-series network approach to auditory verbal hallucinations: Examining dynamic interactions using experience sampling methodology." In: *Schizophrenia Research* 215 (2020), pp. 148–156.

[157] J. Joormann and I. H. Gotlib. "Emotion regulation in depression: Relation to cognitive inhibition." In: *Cognition & Emotion* 24.2 (2010), pp. 281–298.

[158] J. Joormann and C. H. Stanton. "Examining emotion regulation in depression: A review and future directions." In: *Behaviour Research and Therapy* 86 (2016), pp. 35–49.

[159] D. G. Jordan, E. S. Winer, and T. Salem. "The current status of temporal network analysis for clinical science: Considerations as the paradigm shifts?" In: *Journal of Clinical Psychology* 76 (2020), pp. 1591–1612.

[160] R. Kalisch, A. O. Cramer, H. Binder, J. Fritz, I. Leertouwer, G. Lunansky, B. Meyer, J. Timmer, I. M. Veer, and A.-L. Van Harmelen. "Deconstructing and reconstructing resilience: a dynamic network approach." In: *Perspectives on Psychological Science* 14.5 (2019), pp. 765–777.

[161] A. Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.5.0. 2022. URL: `https://cran.r-project.org/package=ggpubr`.

[162] K. R. Kaufman, E. Petkova, K. S. Bhui, and T. G. Schulze. "A global needs assessment in times of a global crisis: world psychiatry response to the COVID-19 pandemic." In: *BJPsych Open* 6.3 (2020), e48.

[163] S. W. Kelley, A. J. Fisher, C. T. Lee, E. Gallagher, A. K. Hanlon, I. H. Robertson, and C. M. Gillan. "Elevated emotion network connectivity is associated with fluctuations in depression." In: *Proceedings of the National Academy of Sciences* 120.45 (2023), pp. 1–8.

[164] K. S. Kendler. "Explanatory models for psychiatric illness." In: *American Journal of Psychiatry* 165.6 (2008), pp. 695–702.

[165] K. S. Kendler. "The nature of psychiatric disorders." In: *World Psychiatry* 15.1 (2016), pp. 5–12.

[166] K. S. Kendler and C. A. Prescott. "A population-based twin study of lifetime major depression in men and women." In: *Archives of General Psychiatry* 56.1 (1999), pp. 39–44.

[167] K. S. Kendler, P. Zachar, and C. Craver. "What kinds of things are psychiatric disorders?" In: *Psychological Medicine* 41.6 (2011), pp. 1143–1150.

[168] G. K. Khazanov and A. M. Ruscio. "Is low positive emotionality a specific risk factor for depression? A meta-analysis of longitudinal studies." In: *Psychological Bulletin* 142.9 (2016), 991—1015.

[169] R. A. Kievit, W. E. Frankenhuis, L. J. Waldorp, and D. Borsboom. "Simpson's paradox in psychological science: A practical guide." In: *Frontiers in Psychology* 4 (2013), pp. 1–14.

[170] J. Kim, W. Zhu, L. Chang, P. M. Bentler, and T. Ernst. "Unified structural equation modeling approach for the analysis of multisubject, multivariate functional MRI data." In: *Human Brain Mapping* 28.2 (2007), pp. 85–93.

[171] R. B. Kline. *Principles and Practice of Structural Equation Modeling*. Guilford publications, 2015.

[172] A. Klippel, W. Viechtbauer, U. Reininghaus, J. Wigman, C. van Borkulo, MERGE, I. Myin-Germeys, and M. Wichers. "The cascade of stress: a network approach to explore differential dynamics in populations varying in risk for psychosis." In: *Schizophrenia Bulletin* 44.2 (2018), pp. 328–337.

[173] L. von Klipstein, H. Riese, D. C. van der Veen, M. N. Servaas, and R. A. Schoevers. "Using person-specific networks in psychotherapy: challenges, limitations, and how we could use them anyway." In: *BMC Medicine* 18.1 (2020), pp. 1–8.

[174] D. Kocevska, T. F. Blanken, E. J. Van Someren, and L. Rösler. "Sleep quality during the COVID-19 pandemic: not one size fits all." In: *Sleep Medicine* 76 (2020), pp. 86–88.

[175] M. Konings, M. Bak, M. Hanssen, J. Van Os, and L. Krabbendam. "Validity and reliability of the CAPE: a self-report instrument for the measurement of psychotic experiences in the general population." In: *Acta Psychiatrica Scandinavica* 114.1 (2006), pp. 55–61.

[176] S. L. Koole and N. B. Jostmann. "Getting a grip on your feelings: effects of action orientation and external demands on intuitive affect regulation." In: *Journal of Personality and Social psychology* 87.6 (2004), p. 974.

[177] J. Kossakowski, P. Groot, J. M. B. Haslbeck, D. Borsboom, and M. Wichers. "Data from 'critical slowing down as a personalized early warning signal for depression'." In: *Journal of Open Psychology Data* 5 (2017), pp. 1–3.

[178] K. Kroenke, R. L. Spitzer, and J. B. Williams. "The PHQ-9: validity of a brief depression severity measure." In: *Journal of General Internal Medicine* 16.9 (2001), pp. 606–613.

[179] R. Kroeze, M. N. Servaas, J. A. Bastiaansen, R. C. O. Voshaar, D. Borsboom, H. G. Ruhe, R. A. Schoevers, H. Riese, et al. "Personalized feedback on symptom dynamics of psychopathology: A proof-of-principle study." In: *Journal for Person-Oriented Research* 3.1 (2017), pp. 1–10.

[180] W. van Lancker and Z. Parolin. "COVID-19, school closures, and child poverty: a social crisis in the making." In: *The Lancet Public Health* 5.5 (2020), pp. e243–e244.

[181] S. T. Lane and K. M. Gates. "Automated selection of robust individual-level structural equation models for time series data." In: *Structural Equation Modeling* 24.5 (2017), pp. 768–782.

[182] S. T. Lane et al. *gimme: Group iterative multiple model estimation.* R package version 0.7-16. 2024. URL: `https://cran.r-project.org/package=gimme`.

[183] S. L. Lauritzen. *Graphical models*. Oxford: Clarendon Press, 1996.

[184] I. A. van de Leemput, M. Wichers, A. O. Cramer, D. Borsboom, F. Tuerlinckx, P. Kuppens, E. H. Van Nes, W. Viechtbauer, E. J. Giltay, S. H. Aggen, et al. "Critical slowing down as early warning for the onset and termination of depression." In: *Proceedings of the National Academy of Sciences* 111.1 (2014), pp. 87–92.

[185] C. A. Levinson, R. A. Hunt, C. Christian, B. M. Williams, A. C. Keshishian, I. A. Vanzhula, and C. Ralph-Nearman. "Longitudinal group and individual networks of eating disorder symptoms in individuals diagnosed with an eating disorder." In: *Journal of Abnormal Psychology* 131.1 (2021), pp. 58–72.

[186] C. A. Levinson, R. A. Hunt, A. C. Keshishian, M. L. Brown, I. Vanzhula, C. Christian, L. C. Brosof, and B. M. Williams. "Using individual networks to identify treatment targets for eating disorder treatment: a proof-of-concept study and initial data." In: *Journal of Eating Disorders* 9.1 (2021), pp. 1–18.

[187] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. "High-dimensional semiparametric Gaussian copula graphical models." In: *The Annals of Statistics* 40.4 (2012), pp. 2293–2326.

[188] S. Liu, K. M. Gates, and E. Ferrer. "Homogeneity Assumptions in the Analysis of Dynamic Processes." In: *Multivariate Behavioral Research* (2023), pp. 1–11.

[189] B. Löwe, J. Unützer, C. M. Callahan, A. J. Perkins, and K. Kroenke. "Monitoring depression treatment outcomes with the patient health questionnaire-9." In: *Medical Care* (2004), pp. 1194–1201.

[190] G. Lunansky, C. D. van Borkulo, T. Blanken, A. Cramer, and D. Borsboom. "Bouncing back from life's perturbations: Formalizing psychological resilience from a complex systems perspective." In: *PsyArXiv* (2022).

[191] G. Lunansky, R. H. A. Hoekstra, and T. F. Blanken. "Disentangling the Role of Affect in the Evolution of Depressive Complaints Using Complex Dynamical Networks." In: *Collabra: Psychology* 9.1 (2023), pp. 1–16.

[192] D. M. Lydon-Staley, M. Xia, H. W. Mak, and G. Fosco. "Adolescent emotion network dynamics in daily life and implications for depression." In: *Journal of Abnormal Child Psychology* 47 (2019), pp. 717–729.

[193] A. Mackinnon, A. F. Jorm, H. Christensen, A. E. Korten, P. A. Jacomb, and B. Rodgers. "A short form of the Positive and Negative Affect Schedule: Evaluation of factorial validity and invariance across demographic variables in a community sample." In: *Personality and Individual Differences* 27.3 (1999), pp. 405–416.

[194] M. Maj. "The need for a conceptual framework in psychiatry acknowledging complexity while avoiding defeatism." In: *World Psychiatry* 15.1 (2016), pp. 1–2.

[195] A. C. Mansueto, R. W. Wiers, J. van Weert, B. C. Schouten, and S. Epskamp. "Investigating the feasibility of idiographic network models." In: *Psychological Methods* 28.5 (2023), 1052–1068.

[196] M Marsman, D Borsboom, J Kruis, S Epskamp, R. v. van Bork, L. Waldorp, H. v. d. Maas, and G Maris. "An introduction to network psychometrics: Relating Ising network models to item response theory models." In: *Multivariate Behavioral Research* 53.1 (2018), pp. 15–35.

[197] E. McElroy, E. Napoleone, M. Wolpert, and P. Patalay. "Structure and connectivity of depressive symptom networks corresponding to early treatment response." In: *eClinicalMedicine* 8 (2019), pp. 29–36.

[198] R. J. McNally. "Can network analysis transform psychopathology?" In: *Behaviour Research and Therapy* 86 (2016), pp. 95–104.

[199] R. J. McNally. "Network analysis of psychopathology: controversies and challenges." In: *Annual Review of Clinical Psychology* 17 (2021), pp. 31–53.

[200] P. C. M. Molenaar. "State space techniques in structural equation modeling: Transformation of latent variables in and out of latent variable models." 2003. URL: `http://www.hhdev.psu.edu/hdfs/faculty/docs/StateSpaceTechniques.pdf`.

[201] P. C. M. Molenaar. "A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever." In: *Measurement* 2.4 (2004), pp. 201–218.

[202] P. C. Molenaar. "On the necessity to use person-specific data analysis approaches in psychology." In: *European Journal of Developmental Psychology* 10.1 (2013), pp. 29–39.

[203] P. C. Molenaar and J. R. Nesselroade. "A comparison of pseudo-maximum likelihood and asymptotically distribution-free dynamic factor analysis parameter estimation in fitting covariance-structure models to block-Toeplitz matrices representing single-subject multivariate time-series." In: *Multivariate Behavioral Research* 33.3 (1998), pp. 313–342.

[204] T. P. Morris, I. R. White, and M. J. Crowther. "Using simulation studies to evaluate statistical methods." In: *Statistics in Medicine* 38.11 (2019), pp. 2074–2102.

[205] I. Murray. *Advances in Markov chain Monte Carlo methods*. University of London, University College London (United Kingdom), 2007.

[206] S. Nestler and S. Humberg. "GIMME's ability to recover group-level path coefficients and individual-level path coefficients." In: *Methodology* 17.1 (2021), pp. 58–91.

[207] D. Nguyen, E. E. Naffziger, and K. C. Berridge. "Positive affect: nature and brain bases of liking and wanting." In: *Current Opinion in Behavioral Sciences* 39 (2021), pp. 72–78.

[208] M. Olthof, F. Hasselman, F. Oude Maatman, A. M. Bosman, and A. Lichtwarck-Aschoff. "Complexity theory of psychopathology." In: *Journal of Psychopathology and Clinical Science* 132.3 (2023), pp. 314–323.

[209] M. Olthof, F. Hasselman, G. Strunk, M. van Rooij, B. Aas, M. A. Helmich, G. Schiepek, and A. Lichtwarck-Aschoff. "Critical fluctuations as an early-warning signal for sudden gains and losses in patients receiving psychotherapy for mood disorders." In: *Clinical Psychological Science* 8.1 (2020), pp. 25–35.

[210] T. Opsahl, F. Agneessens, and J. Skvoretz. "Node centrality in weighted networks: Generalizing degree and shortest paths." In: *Social Networks* 32.3 (2010), pp. 245–251.

[211] T. H. Oreel, D. Borsboom, S. Epskamp, I. D. Hartog, J. E. Netjes, P. T. Nieuwkerk, J. P. Henriques, M. Scherer-Rath, H. W. van Laarhoven, and M. A. Sprangers. "The dynamics in health-related quality of life of patients with stable coronary artery disease were revealed: A network analysis." In: *Journal of Clinical Epidemiology* 107 (2019), pp. 116–123.

[212] W. H. Organization. *Depression*. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/depression.

[213] J. J. Park, S.-M. Chow, Z. F. Fisher, and P. Molenaar. "Affect and personality: Ramifications of modeling (non-) directionality in dynamic network models." In: *European Journal of Psychological Assessment* 36.6 (2020), 1009—1023.

[214] M. L. Pe, K. Kircanski, R. J. Thompson, L. F. Bringmann, F. Tuerlinckx, M. Mestdagh, J. Mata, S. M. Jaeggi, M. Buschkuehl, J. Jonides, et al. "Emotion-network density in major depressive disorder." In: *Clinical Psychological Science* 3.2 (2015), pp. 292–300.

[215] B. Pfefferbaum and C. S. North. "Mental health and the Covid-19 pandemic." In: *New England Journal of Medicine* 383.6 (2020), pp. 510–512.

[216] M. L. Piccirillo, E. D. Beck, and T. L. Rodebaugh. "A clinician's primer for idiographic research: considerations and recommendations." In: *Behavior Therapy* 50.5 (2019), pp. 938–951.

[217] M. L. Piccirillo and T. L. Rodebaugh. "Personalized networks of social anxiety disorder and depression and implications for treatment." In: *Journal of Affective Disorders* 298 (2022), pp. 262–276.

[218] J Pinheiro, D Bates, S DebRoy, and S. D. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1. 2020. URL: https://CRAN.R-project.org/package=multilevelTools.

[219] D. Proudman, P. Greenberg, and D. Nellesen. "The growing burden of major depressive disorders (MDD): implications for researchers and policy makers." In: *Pharmacoeconomics* 39 (2021), pp. 619–625.

[220] R Core Team. "R: A Language and Environment for Statistical Computing 55." In: *R Development Core Team* 55 (2015), 275—286.

[221] L. R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition." In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[222] J. W. Reeves and A. J. Fisher. "An examination of idiographic networks of posttraumatic stress disorder symptoms." In: *Journal of Traumatic Stress* 33.1 (2020), pp. 84–95.

[223] J. Revord, K. Sweeny, and S. Lyubomirsky. "Categorizing the function of positive emotions." In: *Current Opinion in Behavioral Sciences* 39 (2021), pp. 93–97.

[224] H Riese, L Von Klipstein, R. Schoevers, D. Van Der Veen, and M. Servaas. "Personalized ESM monitoring and feedback to support psychological treatment for depression: a pragmatic randomized controlled trial (Therap-i)." In: *BMC Psychiatry* 21 (2021), pp. 1–11.

[225] B. Ripley and W. Venables. *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. R package version 7.3.18. 2023. URL: https://cran.r-project.org/package=nnet.

[226] D. J. Robinaugh, J. M. B. Haslbeck, O. Ryan, E. I. Fried, and L. J. Waldorp. "Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction." In: *Perspectives on Psychological Science* 16.4 (2021), pp. 725–743.

[227] D. J. Robinaugh, R. H. A. Hoekstra, E. R. Toner, and D. Borsboom. "The network approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research." In: *Psychological Medicine* 50.3 (2020), pp. 353–366.

[228] D. J. Robinaugh, A. J. Millner, and R. J. McNally. "Identifying highly influential nodes in the complicated grief network." In: *Journal of Abnormal Psychology* 125.6 (2016), pp. 747–757.

[229] D. J. Robinaugh, A. J. Millner, and R. J. McNally. "Supplemental Material for Identifying Highly Influential Nodes in the Complicated Grief Network." In: *Journal of Abnormal Psychology* 125.6 (2016), pp. 747–757.

[230] D. Robinaugh, J. M. B. Haslbeck, L. Waldorp, J. Kossakowski, E. I. Fried, A. Millner, R. J. McNally, O. Ryan, J. de Ron, H. van der Maas, et al. "Advancing the network theory of mental disorders: A computational model of panic disorder." In: (2019).

[231] M. Rodriguez, G. Aalbers, and R. J. McNally. "Idiographic network models of social media use and depression symptoms." In: *Cognitive Therapy and Research* 46.1 (2022), pp. 124–132.

[232] A. Roefs, E. I. Fried, M. Kindt, C. Martijn, B. Elzinga, A. W. Evers, R. W. Wiers, D. Borsboom, and A. Jansen. "A new science of mental disorders: Using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology." In: *Behaviour Research and Therapy* 153 (2022), p. 104096.

[233] Y. Rosseel. "lavaan: An R Package for Structural Equation Modeling." In: *Journal of Statistical Software* 48.2 (2012), pp. 1–36.

[234] J. M. Round, C. Lee, J. G. Hanlon, E. Hyshka, J. R. Dyck, and D. T. Eurich. "Changes in patient health questionnaire (PHQ-9) scores in adults with medical authorization for cannabis." In: *BMC Public Health* 20.1 (2020), pp. 1–10.

[235] O. Ryan, L. F. Bringmann, and N. K. Schuurman. "The challenge of generating causal hypotheses using network models." In: *Structural Equation Modeling: A Multidisciplinary Journal* (2022), pp. 1–18.

[236] O. Ryan and E. L. Hamaker. "Time to intervene: A continuous-time approach to network analysis and centrality." In: *Psychometrika* 87.1 (2022), pp. 214–252.

[237] O. Ryan, J. M. B. Haslbeck, and L. Waldorp. "Non-stationarity in time-series analysis: Modeling stochastic and deterministic trends." In: *PsyArXiv* (2023).

[238] H. M. Schenk, B. F. Jeronimus, L. van der Krieke, E. H. Bos, P. de Jonge, and J. G. Rosmalen. "Associations of positive affect and negative affect with allostatic load: a lifelines cohort study." In: *Psychosomatic Medicine* 80.2 (2018), pp. 160–166.

[239] R. Schoevers, C. van Borkulo, F Lamers, M. Servaas, J. Bastiaansen, A. Beekman, A. Van Hemert, J. Smit, B. Penninx, and H Riese. "Affect fluctuations examined with ecological momentary assessment in patients with current or remitted depression and anxiety disorders." In: *Psychological Medicine* 51.11 (2021), pp. 1906–1915.

[240] U. Scholz. "It's time to think about time in health psychology." In: *Applied Psychology: Health and Well-Being* 11.2 (2019), pp. 173–186.

[241] M. J. Schreuder, R. N. Groen, J. T. Wigman, C. A. Hartman, and M. Wichers. "Measuring psychopathology as it unfolds in daily life: Addressing key assumptions of intensive longitudinal methods in the TRAILS TRANS-ID study." In: *BMC Psychiatry* 20 (2020), pp. 1–14.

[242] M. Schultzberg and B. Muthén. "Number of subjects and time points needed for multilevel time-series analysis: A simulation study of dynamic structural equation modeling." In: *Structural Equation Modeling: A Multidisciplinary Journal* 25.4 (2018), pp. 495–515.

[243] N. K. Schuurman, J. H. Houtveen, and E. L. Hamaker. "Incorporating measurement error in n = 1 psychological autoregressive modeling." In: *Frontiers in Psychology* 6 (2015), pp. 1–15.

[244] L. Schweren, C. D. van Borkulo, E. Fried, and I. M. Goodyer. "Assessment of symptom network density as a prognostic marker of treatment response in adolescent depression." In: *JAMA Psychiatry* 75.1 (2018), pp. 98–100.

[245] K. E. Shin, M. G. Newman, and N. C. Jacobson. "Emotion network density is a potential clinical marker for anxiety and depression: Comparison of ecological momentary assessment and daily diary." In: *British Journal of Clinical Psychology* 61 (2022), pp. 31–50.

[246] E. H. Simpson. "The interpretation of interaction in contingency tables." In: *Journal of the Royal Statistical Society* 13.2 (1951), pp. 238–241.

[247] I. A. Smits, M. E. Timmerman, D. P. Barelds, and R. R. Meijer. "The Dutch symptom checklist-90-revised." In: *European Journal of Psychological Assessment* 31 (4 2014), pp. 263–271.

[248] E. Snippe, W. Viechtbauer, N. Geschwind, A. Klippel, P. de Jonge, and M. Wichers. "The impact of treatments for depression on the dynamic network structure of mental states: Two randomized controlled trials." In: *Scientific Reports* 7.1 (2017), p. 46523.

[249] H. van Steenbergen, E. R. de Bruijn, A. C. van Duijvenvoorde, and A.-L. van Harmelen. "How positive affect buffers stress responses." In: *Current Opinion in Behavioral Sciences* 39 (2021), pp. 153–160.

[250] A. Steptoe. "Happiness and health." In: *Annual Review of Public Health* 40 (2019), pp. 339–359.

[251] E. Sturt. "Hierarchical patterns in the distribution of psychiatric symptoms." In: *Psychological Medicine* 11.4 (1981), pp. 783–794.

[252] R. Tibshirani. "Regression shrinkage and selection via the lasso." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[253] S. van der Tuin, S. E. Balafas, A. J. Oldehinkel, E. C. Wit, S. H. Booij, and J. T. Wigman. "Dynamic symptom networks across different at-risk stages for psychosis: An individual and transdiagnostic perspective." In: *Schizophrenia Research* 239 (2022), pp. 95–102.

[254] S. van der Tuin, R. N. Groen, S. Castro-Alvarez, A. J. Oldehinkel, S. H. Booij, and J. T. Wigman. "Group, Subgroup, and Person-Specific Symptom Associations in Individuals at Different Levels of Risk for Psychosis: A Combination of Theory-based and Data-driven Approaches." In: *Schizophrenia Bulletin Open* 2.1 (2021), sgab047.

[255] M. Vervaet, L. Puttevils, R. H. A. Hoekstra, E. Fried, and M.-A. Vanderhasselt. "Transdiagnostic vulnerability factors in eating disorders: A network analysis." In: *European Eating Disorders Review* 29.1 (2021), pp. 86–100.

[256] M. C. Voelkle and J. H. Oud. "Continuous time modelling with individually varying time intervals for oscillating and non-oscillating processes." In: *British Journal of Mathematical and Statistical Psychology* 66.1 (2013), pp. 103–126.

[257] J. Wang, L. Xu, and E. Wang. "Potential landscape and flux framework of nonequilibrium networks: Robustness, dissipation, and coherence of biochemical oscillations." In: *Proceedings of the National Academy of Sciences* 105.34 (2008), pp. 12271–12276.

[258] L. P. Wang and S. E. Maxwell. "On disaggregating between-person and within-person effects with longitudinal data using multilevel models." In: *Psychological Methods* 20.1 (2015), pp. 63–83.

[259] D. Watson, L. A. Clark, and A. Tellegen. "Development and validation of brief measures of positive and negative affect: the PANAS scales." In: *Journal of Personality and Social Psychology* 54.6 (1988), 1063—1070.

[260] D. J. Watts and S. H. Strogatz. "Collective dynamics of 'small-world'networks." In: *Nature* 393.6684 (1998), pp. 440–442.

[261] M. Wichers, F. Peeters, N. Geschwind, N. Jacobs, C. Simons, C. Derom, E. Thiery, P. H. Delespaul, and J. van Os. "Unveiling patterns of affective responses in daily life may improve outcome prediction in depression: A momentary assessment study." In: *Journal of Affective Disorders* 124.1-2 (2010), pp. 191–195.

[262] M. Wichers. "The dynamic nature of depression: A new micro-level perspective of mental disorder that meets current challenges." In: *Psychological Medicine* 44.7 (2014), pp. 1349–1360.

[263] M. Wichers, P. C. Groot, E. Psychosystems group, and E. group. "Critical slowing down as a personalized early warning signal for depression." In: *Psychotherapy and Psychosomatics* 85.2 (2016), pp. 114–116.

[264] M. Wichers, I. Myin-Germeys, N. Jacobs, F. Peeters, G. Kenis, C. Derom, R. Vlietinck, P. Delespaul, and J. Van Os. "Genetic risk of depression and stress-induced negative affect in daily life." In: *The British Journal of Psychiatry* 191.3 (2007), pp. 218–223.

[265] M. Wichers, M. J. Schreuder, R. Goekoop, and R. N. Groen. "Can we predict the direction of sudden shifts in symptoms? Transdiagnostic implications from a complex systems perspective on psychopathology." In: *Psychological Medicine* 49.3 (2019), pp. 380–387.

[266] M. Wichers, J. Wigman, and I. Myin-Germeys. "Micro-level affect dynamics in psychopathology viewed from complex dynamical system theory." In: *Emotion Review* 7.4 (2015), pp. 362–367.

[267] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: `https://ggplot2.tidyverse.org`.

[268] H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A grammar of data manipulation*. R package version 1.1.4. 2023. URL: `https://CRAN.R-project.org/package=dplyr`.

[269] H. Wickham and G. Grolemund. *R for data science: Import, tidy, transform, visualize, and model data*. R package version 1.3.0. 2023. URL: `https://CRAN.R-project.org/package=tidyr`.

[270] H. Wickham et al. "Welcome to the tidyverse." In: *Journal of Open Source Software* 4.43 (2019), p. 1686.

[271] J. Wigman, J Van Os, D Borsboom, K. Wardenaar, S Epskamp, A. Klippel, W. Viechtbauer, I. Myin-Germeys, M Wichers, et al. "Exploring the underlying structure of mental disorders: cross-diagnostic differences and similarities from a network perspective using both a top-down and a bottom-up approach." In: *Psychological Medicine* 45.11 (2015), pp. 2375–2387.

[272] J. T. Wigman, S. van der Tuin, D. van den Berg, M. K. Muller, and S. H. Booij. "Mental health, risk and protective factors at micro-and macro-levels across early at-risk stages for psychosis: The Mirorr study." In: *Early Intervention in Psychiatry* 17.5 (2023), pp. 478–494.

[273] J. Wiley. *multilevelTools: Multilevel and Mixed Effects Model Diagnostics and Effect Sizes*. R package version 0.1.1. 2020. URL: `https://CRAN.R-project.org/package=multilevelTools`.

[274] D. R. Williams. "Learning to live with sampling variability: Expected replicability in partial correlation networks." In: *Psychological Methods* 27.4 (2022), pp. 606–621.

[275] D. R. Williams, P. Rast, L. R. Pericchi, and J. Mulder. "Comparing Gaussian graphical models with the posterior predictive distribution and Bayesian model selection." In: *Psychological Methods* 25.5 (2020), 653—672.

[276] D. R. Williams, M. Rhemtulla, A. C. Wysocki, and P. Rast. "On nonregularized estimation of psychological networks." In: *Multivariate Behavioral Research* 54.5 (2019), pp. 719–750.

[277] A. G. Wright and W. C. Woods. "Personalized models of psychopathology." In: *Annual Review of Clinical Psychology* 16 (2020), pp. 49–74.

[278] A. Wysocki, M. Rhemtulla, R. van Bork, and A. Cramer. "Cross-lagged network models." In: (2022).

[279] J. Yin and H. Li. "A sparse conditional Gaussian graphical model for analysis of genetical genomics data." In: *The Annals of Applied Statistics* 5.4 (2011), pp. 2630–2650.

[280] T. Zander-Schellenberg, I. M. Collins, M. Miché, C. Guttmann, R. Lieb, and K. Wahl. "Does laughing have a stress-buffering effect in daily life? An intensive longitudinal study." In: *PloS ONE* 15.7 (2020), e0235851.

H

SUMMARY

As the aim of psychology research is to understand, explain, and predict human behavior and mental processes, psychological science is inherently concerned with dynamic processes that vary among individuals. Traditionally, psychological research has predominantly utilized aggregate models to do so by generalizing findings from aggregate models to individuals. However, these aggregate models are not directly translatable to individuals. That a certain model holds true at the aggregate level does not mean we should expect this model to hold true for each individual. An aggregate model is just that, it demonstrates what holds for the group on average, but these aggregate results do not necessarily hold for each individual within the group. Acknowledging this limitation has led to a significant shift in psychological research from predominantly focusing on aggregate—nomothetic—models to estimating idiographic models. These idiographic models allow for a more detailed exploration of the complex psychological dynamics specific to an individual's experience.

This shift is particularly evident in the network approach to psychopathology, where mental disorders are conceptualized as networks of complex causal interactions between symptoms. As each individual is conceptualized by their own symptom network, this approach naturally lends itself to study how these complex causal interactions between symptoms unfold for a given individual and how they might differ from person to person. However, the major challenge that remains in this approach is to effectively quantify the individual differences between idiographic network models. Addressing this challenge is crucial for advancing our theoretical understanding of psychopathology and developing (personalized), effective interventions. In this thesis, we aimed to assess individual differences in idiographic network models and investigate how we can use idiographic network modeling techniques to help us unravel and understand the individual differences observed between these idiographic network models.

With the rise in popularity of idiographic network modeling, the comparison of these idiographic network structures has led many researchers to claim there is a tremendous amount of heterogeneity present between individual networks. That is, individuals seemed to vary greatly with respect to their idiographic network structure. However, in this thesis, we identified an important pitfall when comparing such estimated idiographic network models: a tendency to interpret *all* variability between in-

dividual network structures as heterogeneity, while some of this variability is the result of fluctuations in the data created by noise. In a thought experiment, we showed that even if the generating network structure is the same for two individuals, network structures estimated from their data might differ substantially. Combining sampling variation with overall conservative estimation procedures as we do in network estimation (high specificity and lower sensitivity at low sample sizes) will inevitably lead to a mismatch in the presence and absence of estimated edges and varying edge weights across individuals. Therefore, we should not forget that statistical models are subject to the reliability of their estimation, and in order to separate illusionary from true heterogeneity when comparing idiographic network models, we should ensure high sensitivity and specificity.

Simulation studies offer a general framework to determine the sensitivity and specificity of various (idiographic) network estimation techniques. We exemplified the setup and execution of simulation studies within the network modeling framework in a tutorial in this thesis. We aimed to enable a broader range of researchers to examine the accuracy of network estimation techniques to determine to what extent model differences can be attributed to estimation accuracy rather than genuine individual differences. We used such simulation study design to determine the extent to which popular idiographic network modeling techniques can separate illusionary (i.e., sampling variation) from real heterogeneity (i.e., individual differences). Our results showed that low statistical power always limits the validity of conclusions regarding heterogeneity for all tools investigated. At low sample sizes, an applied researcher is likely to erroneously conclude that a sample is heterogeneous even if the sample is homogeneous.

In response to these results, we proposed a rigorous way of testing if heterogeneity between idiographic network structures is plausible given the data at hand. To that end, we employed ideas of invariance testing from structural equation modeling (SEM) to propose a systematic model comparison approach, where a model in which equality between idiographic network structures is assumed is compared to a model in which inequality between idiographic network structures is assumed. We deemed this practice the Individual Network Invariance Test (INIT). INIT can be interpreted as a null hypothesis test to compare (strict) homogeneity (i.e., edges between nodes and their parameter values are constant across individuals) to heterogeneity (i.e., edges between nodes and their parameter values are allowed to vary across individuals). INIT takes homogeneity rather than heterogeneity as the null model. As a result, INIT is a more conservative practice that errs on the side of caution, only deem-

ing the variance in the data as individual differences when the data provide enough evidence to do so.

Using this new proposed technique, we investigated heterogeneity within idiographic network structures over time and investigated the relationship between a change in network structure and changes in psychopathology severity. To examine this relationship, we determined the stability of individual symptom networks over a one-year period and identified whether network stability was associated with psychopathology severity. Using the newly developed INIT methodology, we identified individuals with a stable network over time and those with an unstable network over time. We tested whether these groups differed in their absolute change in psychopathology severity during that same period. Most of our sample showed a stable symptom network over time, while most individuals showed decreased psychopathological severity. Thus, our results show no relationship between a change in psychopathology severity and a change in individual network structure over time.

To shed further light on heterogeneity in symptom dynamics, we used several techniques in a different study, including network analysis. Specifically, we aimed to disentangle heterogeneity in depressive complaints and affect by investigating both psychopathology severity and stability. We used descriptive tools such as visual inspection of the time-series data to investigate how the evaluation of depressive complaints relates to affect fluctuations. We revealed that affect fluctuations show different patterns for individuals with an aggravation or alleviation in depressive complaints, as well as for individuals with consistent no depressive complaints, occasional depressive complaints, and consistent depressive complaints. Furthermore, examining person-specific network models revealed that idiographic networks' connectivity (i.e., the strength and number of relations in the network model) was strongly related to a change (aggravation or alleviation) in depressive complaints. Thus, combining the above-discussed results, heterogeneity in idiographic network models seems to relate to psychopathology stability rather than severity.

A clear understanding of model-implied behavior is essential to unravel the underlying mechanisms of mental disorders and the potential role individual differences play with respect to these dynamics. To elucidate heterogeneity in idiographic network models, we used stability landscapes as a novel way to systematically investigate the model-implied behavior of several idiographic network estimation techniques. By mapping the two most essential postulates of the network approach to estimated idiographic network models, we revealed that both the estimation technique employed and the model parameters matter a great deal for the dynamics the idiographic network model can display. We confirmed our previously found results: for the most commonly used idiographic network model-

ing technique, i.e., the VAR model, idiographic network structures relate to psychopathology stability rather than severity.

This thesis represents a step forward in assessing and understanding heterogeneity in idiographic network models in particular and mental disorders more broadly. It highlights the need for more sophisticated models to capture the complex dynamics of individual experiences in psychopathology and underscores the need for a deeper insight into the dynamical systems that underlie mental health.

I

NEDERLANDSE SAMENVATTING

Het doel van psychologisch onderzoek is het begrijpen, verklaren en voorspellen van menselijk gedrag en mentale processen. Daarmee houdt de psychologische wetenschap zich inherent bezig met dynamische processen die van persoon tot persoon verschillen. Hoewel dynamische processen zich afspelen binnen een persoon, heeft psychologisch onderzoek zich voornamelijk gericht op het gebruik van geaggregeerde modellen. Deze geaggregeerde modellen worden vervolgens geacht zich te vertalen naar een individu. Echter, dat een bepaald model op geaggregeerd niveau geldt, betekent niet dat we mogen verwachten dat ditzelfde model voor elk individu geldt. Verschillende onderzoekers erkennen deze beperking en hebben hun focus verlegd van deze geaggregeerde modellen naar idiografische—individuele—modellen. Idiografische modellen stellen ons in staat om de complexe psychologische dynamiek en ervaring van een individu te modelleren.

De verschuiving van een focus op geaggregeerde—nomothetische—modellen naar idiografische modellen is vooral merkbaar binnen de netwerkbenadering van psychopathologie. In de netwerkbenadering van psychopathologie, worden psychische stoornissen geconceptualiseerd als een netwerk van complexe causale interacties tussen symptomen. Omdat elk individu zijn eigen symptoom netwerk kent, leent deze benadering zich uitermate om te bestuderen hoe deze complexe interacties tussen symptomen zich voor een bepaald individu ontvouwen en hoe deze van persoon tot persoon kunnen verschillen. De uitdaging blijft echter hoe de individuele verschillen tussen idiografische netwerkmodellen effectief kunnen worden gekwantificeerd. Het aangaan van deze uitdaging is essentieel om ons theoretische begrip van psychopathologie te vergroten en om te bepalen of deze idiografische netwerkmodellen de weg kunnen vrijmaken voor meer gepersonaliseerde en effectieve interventies. In deze these richten we ons op hoe we individuele verschillen in idiografische netwerkmodellen kunnen toetsen, en hoe we idiografische netwerkmodelleringstechnieken kunnen gebruiken om ons te helpen de individuele verschillen die we tussen deze idiografische netwerkmodellen vinden, te begrijpen.

De toenemende populariteit van idiografische netwerkmodellen heeft ertoe geleid dat veel onderzoekers beweren dat er een enorme hoeveelheid heterogeniteit bestaat tussen individuele netwerken. Dat wil zeggen dat individuen sterk lijken te variëren met betrekking tot hun idiografis-

che netwerkstructuur. In dit proefschrift hebben we echter een belangrijke valkuil geïdentificeerd bij het vergelijken van dergelijke geschatte idiografische netwerkmodellen: de neiging om alle variabiliteit tussen individuele netwerkstructuren als heterogeniteit te interpreteren, terwijl een deel van deze variabiliteit het resultaat is van fluctuaties in de data gecreëerd door ruis. In een gedachte-experiment hebben we aangetoond dat zelfs als de genererende netwerkstructuur voor twee individuen hetzelfde is, de netwerkstructuren die op basis van hun data worden geschat, aanzienlijk kunnen verschillen. Het combineren van steekproefvariatie met algemene conservatieve schattingsprocedures, zoals gebruikelijk is bij netwerkschatting (hoge specificiteit en lage sensitiviteit) zal onvermijdelijk leiden tot een mismatch in de aan- en afwezigheid van geschatte relaties tussen variabelen en variërende gewichten in de relaties tussen variabelen tussen individuen. Het is daarom van belang niet te vergeten dat statistische modellen onderworpen zijn aan de betrouwbaarheid van hun schatting, en om illusoire van echte heterogeniteit te scheiden bij het vergelijken van idiografische netwerkmodellen. Om dat te bewerkstelligen, moeten we zorgen dat we een een hoge sensitivieit en specificiteit waarborgen in onze procedures.

Simulatiestudies bieden een algemeen raamwerk om de sensitiviteit en specificiteit van verschillende (idiografische) netwerkschattingstechnieken te bepalen. We illustreerden de opzet en uitvoering van simulatiestudies binnen de netwerkbenadering in een tutorial. Ons doel was om een breder scala aan onderzoekers in staat te stellen de nauwkeurigheid van netwerkschattingstechnieken te onderzoeken om te bepalen in welke mate modelverschillen kunnen worden toegeschreven aan de nauwkeurigheid van de schattingen in plaats van aan individuele verschillen. We hebben een dergelijk simulatiestudieontwerp zelf gebruikt om te bepalen in welke mate populaire idiografische netwerkmodelleringstechnieken illusoire heterogeniteit (i.e., steekproefvariatie) kunnen scheiden van echte heterogeniteit (i.e., individuele verschillen). Onze resultaten toonden aan dat een lage statistische power de geldigheid van conclusies over heterogeniteit voor alle onderzochte tools dusdanig beperkt. Bij het gebruik van de onderzochte technieken in combinatie met een kleine steekproef zal een onderzoeker waarschijnlijk ten onrechte concluderen dat een steekproef heterogeen is, zelfs als de steekproef homogeen is.

Als reactie op deze resultaten hebben we in deze these een rigoureuze manier voorgesteld om te testen of heterogeniteit tussen idiografische netwerkstructuren plausibel is gegeven de data. Daartoe hebben we ideeën van invariantietesten gebruikt als een systematische benadering van modelvergelijking, waarbij een model waarin gelijkheid tussen idiografische netwerkstructuren wordt verondersteld, wordt vergeleken met een model waarin ongelijkheid tussen idiografische netwerkstructuren wordt veron-

dersteld. uitgegaan van. We noemden deze test de Individual Network Invariantie Test (INIT). INIT kan worden geïnterpreteerd als een nulhypothesetest om (strikte) homogeniteit (i.e., relaties tussen variabelen en hun parameterwaarden zijn hetzelfde voor ieder individuen) te vergelijken met heterogeniteit (i.e., relaties tussen variabelen en hun parameterwaarden mogen verschillen voor ieder individie). INIT neemt homogeniteit in plaats van heterogeniteit als nulmodel. Als gevolg hiervan is INIT een meer conservatieve toets die het zekere voor het onzekere neemt en de variantie in de gegevens alleen als individuele verschillen beschouwt wanneer de data daartoe voldoende bewijs leveren.

Met behulp van deze nieuw voorgestelde techniek onderzochten we heterogeniteit binnen idiografische netwerkstructuren over tijd en of we een verandering in deze structuren konden relateren aan veranderingen in de ernst van psychopathologie. Om deze relatie te onderzoeken, bepaalden we de stabiliteit van individuele symptoom netwerken over een periode van een jaar gebruik makend van INIT, en identificeerden we of netwerkstabiliteit geassocieerd was met een verandering in de ernst van psychopathologie. Het grootste deel van onze steekproef vertoonde in de loop van de tijd een stabiel symptoomnetwerk, terwijl de meeste individuen een vermindering in psychopathologische ernst vertoonden. Onze resultaten laten dus geen verband zien tussen een verandering in de ernst van de psychopathologie en een verandering in de individuele netwerkstructuur in de loop van de tijd.

Om een breeder licht te werpen op heterogeniteit in symptoomdynamiek, hebben we in een volgend onderzoek verschillende technieken gebruikt. Concreet wilden we de heterogeniteit in depressieve klachten en affect ontrafelen door zowel naar de ernst als de stabiliteit van psychopathologie te kijken. We gebruikten beschrijvende technieken zoals visuele inspectie van de tijdseriedata om te onderzoeken hoe de evaluatie van depressieve klachten zich verhouden tot affectfluctuaties. We hebben laten zien dat affectfluctuaties verschillende patronen laten zien bij individuen met een verergering of verlichting van depressieve klachten, maar ook bij individuen met aanhoudend afwezigheid van depressieve klachten, incidenteel depressieve klachten en aanhoudende depressieve klachten. Bovendien bleek uit het inspecteren van de individuele netwerkmodellen dat de connectiviteit van individuele netwerken (d.w.z. de sterkte en het aantal relaties in het netwerkmodel) sterk gerelateerd was aan een verandering (verergering of verlichting) van depressieve klachten. Als we het hierboven besproken resultaat en het hier besproken resultaat samen nemen, lijkt heterogeniteit in idiografische netwerkmodellen dus eerder verband te houden met de stabiliteit van de psychopathologie dan met de ernst van de psychopathologie.

Een duidelijk begrip van modelgeïmpliceerd gedrag is essentieel om de onderliggende mechanismen van psychische stoornissen te ontrafelen en de potentiële rol die individuele verschillen spelen met betrekking tot deze dynamiek. Om de heterogeniteit in idiografische netwerkmodellen te verhelderen, hebben we stabiliteitslandschappen gebruikt als een nieuwe manier om systematisch het model-geïmpliceerde gedrag van verschillende idiografische netwerkschattingstechnieken te bestuderen. Door de twee meest essentiële postulaten van de netwerkbenadering in kaart te brengen en te relateren aan geschatte idiografische netwerkmodellen, hebben we onthuld dat zowel de gebruikte schattingstechniek als de modelparameters van groot belang zijn voor de dynamiek die het idiografische netwerkmodel weergeeft. We bevestigden onze eerder gevonden resultaten: voor de meest gebruikte techniek voor het modelleren van idiografische netwerken, namelijk het VAR-model, hebben idiografische netwerkstructuren eerder betrekking op de stabiliteit van psychopathologie dan op de ernst.

Dit proefschrift behelst een stap voorwaarts in het beoordelen en begrijpen van heterogeniteit in idiografische netwerkmodellen in het bijzonder, en psychische stoornissen in bredere zin. Het benadrukt de behoefte aan meer geavanceerde modellen om de complexe dynamiek van individuele ervaringen in de psychopathologie te modelleren en de noodzaak tot een dieper inzicht in de dynamische systemen die ten grondslag liggen aan psychpatologie.

DANKWOORD — ACKNOWLEDGMENTS

Dit proefschrift was niet tot stand gekomen zonder de steun, aanmoediging en liefde die ik van de mensen om me heen heb mogen ontvangen. In het bijzonder wil ik een aantal van jullie bedanken.

Allereerst wil ik mijn promotoren *Denny & Sacha* bedanken. Soms heb je in je leven het geluk dat iemand meer in je ziet dan je zelf doet. *Denny*, het staat buiten kijf dat ik inhoudelijk ontzettend veel van je heb mogen en kunnen leren, maar ik waardeer bovenal je vermogen om een omgeving the creëren waarin ik kon groeien. Je luisterde naar wat ik nodig had tijdens dit traject en gaf me de ruimte om mijn eigen pad te bewandelen. Jouw aanmoediging heeft me geholpen om te gaan daar waar ik wilde gaan, ook als ik zelf nog niet helemaal durfde. Jouw vertrouwen was genoeg voor twee. *Sacha*, alles waar ik niet uit kom, zie jij helder, en alles wat ik (nog) niet weet, weet jij. Je statistische kennis is onuitputtelijk, en ik hoop ooit een fractie van jouw expertise te hebben. Het was een plezier om samen met jou en *Adela* de network analysis winter- en zomerscholen te organiseren, en nog meer om die steevast af te sluiten ergens op het Rembrandtplein. Ik wil je bedanken voor je hulp, ook vanuit Singapore. *Denny & Sacha*, het was een voorrecht om mijn promotietraject met jullie te mogen doorlopen, en ik hoop dat ik ook na dit proefschrift nog lang van jullie mag blijven leren.

To my promotion committee, *Ellen Hamaker*, *Han van der Maas*, *Richard McNally*, *Tessa Blanken*, *Maarten Marsman* & my guest opponent *Laura Bringmann*, thank you for taking the time to review my thesis. I am thrilled and honored to defend my thesis in front of people who's work has been and continues to be such an inspiration to me.

Mijn paranifmen, lieve *Femke* & *Zenab*, wat een geluk dat jullie ook nu aan mijn zijde staan. Dank jullie voor het delen van jullie leven met me. Dat een half woord genoeg is, maar er ruimte is voor urenlange dialogen. Voor de gesprekken die soms als intense sessies voelen, voor de borrels, het uithuilen, het schatterlachen, de nieuwe perspectieven en de liefde. Alles is beter, leuker en lichter met jullie naast me.

I would like to thank all my (former) colleagues from the *Psychological Methods programme group*. Throughout the years and different roles, it has been a pleasure working with you. From being a student at this department to writing papers together, organizing events, or just bumping

into you at the pantry or at one of the 'borrels'. Because of the atmosphere you created, I was happy to go to work. *Han*, bedankt voor je steun en betrokkenheid tijdens dit traject. Je wijze (en pragmatische) adviezen hebben me enorm gehlopen, en ik heb genoten van onze gezellige gesprekken—die maakten de dagen leuker. Bedankt dat je altijd even tijd maakte. *Dora*, hoe lang het to-do lijstje ook is, door jou kalmte is het organiseren van onderwijs met jou en *Lisa* niet alleen effectief, maar vooral ook ontzettend gezellig. *Abe*, *Lisa*, *Riet*, *Johnny* & *Marie* ik heb met plezier onderwijs met jullie gedraaid en daarin zoveel van jullie mogen leren. Dank voor het vertrouwen en jullie steun tijdens het lesgeven dat ik vaak vooral toch erg spannend vond. Door jullie hulp sta ik nu met veel plezier voor een collegezaal. *Adela*, *Julian*, both of you have moved on to bigger things in life on opposite sides of the world, but I cherishes the memories of sharing an office with you, organizing winter and summer schools, and having you both as a colleague.

In addition to working with wonderful colleagues from my own department, I have been fortunate to collaborate with excellent researchers from other universities. To each of you, I am grateful. In particular, dear *Richard*, I would like to thank you for taking me under your wings. My research visit to your lab was a truly life-changing experience. No matter how busy you were, you always made me feel like you had all the time in the world for me. I cherish our discussions, your book recommendations, and especially your lively storytelling—reenactments including. Your students are lucky you didn't end up in Hollywood. *Shean* and *Ben*, thank you for making my stay in Boston unforgettable. From showing me around the city, the office chit-chats, and of course, the poetry lunches. You made my stay truly special. I hope our paths will cross again. Lieve *Sara*, soms zijn er van die samenwerkingen waarbij het vanaf het eerste moment klikt. Dat je de meest fantastische 'nuchtere boer' bent, wist ik al uit onze (online) samenwerking, maar samen borrelen bleek nog zoveel leuker. Ik heb enorm genoten van onze samenwerking en ben trots op het resultaat dat nu in dit proefschrift staat.

Door de jaren heen vervaagden de grenzen tussen werk en vriendschap. Sommige vrienden werden waardevolle collega's, en sommige collega's groeiden uit tot dierbare vrienden. Ik voel me gelukkig dat ik ten allen tijde omringd was door zoveel persoonlijke én professionele steun. Dear *Jill*, *Karoline*, *Maarten*, & *Michelle*, I am so grateful for the friendship we built during the final years of my PhD. You made everything more enjoyable and fun. Dear *Jill*, *Karoline*, & *Maarten*, I still refer to our trip to Italy as a holiday, even though its purpose was to teach a workshop and attend a summer school on complexity. Nothing feels complex with the four of you around me. Sharing an office with you has has been one

of the highlights of this journey. Seeing the lights on, knowing one of you would be there, always made me excited to step inside. I have loved our work-related discussions and collaborations almost as much as our non-work related conversations. I value our friendship deeply and I can't wait to see each of you shine when it is your turn to defend your thesis! Lieve *Lisa*, je coachte me al door mijn bachelor these, en nog steeds coach jij mij overal doorheen. Ik ben onzettend blij met jou als collega en dankbaar voor onze vriendschap. Lieve *Tessa*, enorm bedankt voor je adviezen met een koffie of een biertje. Ik bewonder je & ik leer zo veel van jou. Lieve *Gaby*, onze vriendschap ontstond al snel tijdens de studie, toen ik doorhad hoe jij altijd zegt waar het op staat en je je niet zomaar door autoriteit laat imponeren. Van mijn beste studiemaatje in de collegebanken groeiden we uit tot collega's die samen op kantoor promoveerden. Ik ben dankbaar dat we al die ervaringen met elkaar mochten delen en dat we dat nog steeds doen. Lieve *Simone*, als huisgenoot en vriendin heb jij dit promotie traject met al zijn ups en downs van dichtbij meegemaakt. Wat een geluk om die momenten met je te hebben kunnen delen. Dank je voor je eindeloze bemoedigende woorden, je luisterende oor en vooral voor het niet alleen tolereren, maar zelfs aanmoedigen van uitbundige Celine Dion singalongs. Lieve *Zoë*, wat een plezier om jou na jaren weer in mijn leven te hebben, en dat het voelt alsof je nooit echt bent weg geweest. Bij jou vind ik altijd herkenning en begrip. Lieve *Meike*, als mijn allereerste psychologie studiemaatje was je er vanaf het begin bij, en wat een geluk dat ik jou naast me had. Jouw, voor mij verbazingwekkende, talent om met iedereen een connectie te maken blijft me fascineren. Dankzij jou eindigde elk avondje borrelen in de meest onverwachte avonturen. Mijn lieve PML guapas: *Femke, Frank, Gaby, Jessica, Koen, Leonie, Rens, Sofieke, Zenab*, ik had tijdens deze periode vaker samen willen proosten dan is gelukt. Het maakt me nostalgisch naar de vodka-woensdagen, de Looiersgracht en die goede oude Roeter-tijd. Fijn dat we op de belangrijke momenten (groot & klein) elkaar nog altijd opzoeken.

Meen lieve familie, julie spreek ik ut liefste an in de taol die ut dichtste bij meen arte legt: ut Urrekers. Lieve *Martina*, wat een gelok as je zus ok je moat is. Ik wil joe bedanken vor de arrege bult lol die we eawen samen, je weze adviezen in je leusterend oor. Ik bin verekte bleede dat jie meen zus binnen. Lieve *Evert*, dank je wel dat je alteed een stark veraol vor m'n kloar eawen in mij alteed an ut lachen wieten te maken. Lieve *Mirjam*, lieve *Amy*, wat een gelok om julie z'n tante te mugen wezen. Ik kiek er noar ut om julie groot te zien worren. Lieve *Josephina*, ik zou niet wieten wat ik zonger joe zou moeten. Alles kuun ik mit joe dielen, in samen mit joe kuun ik alles dragen. Lieve *va*, lieve *moe*, ur is zoo verrekte vuul woar ik julie vor wil bedanken, in ur is zoo vuul woaromme ik groos op julie bin. Boven alles eaw ik inmins vuul bewondering vor julie

z'n duurzettingsvermugen. Dat et mij ut vertrouwen egieven dat je mit volouwen in arde warreken vuul kunen berikken. Julie zeggen zoo vaak: "Wat eawen we ut ok goed è", ok as niet alles vor de wiend goat. Dat et mij elaard dat ur alteed wel wat is om bleede in dankboar vor te weezen in ut leven. Dank jelui wel vor alles wat julie mij eawen mie-egieven, ik ouw groot van julie.

K

PUBLICATIONS

Chapter 2 is published as:
Hoekstra, R. H. A., Epskamp, S., & Borsboom, D. (2023). Heterogeneity in Individual Network Analysis: Reality or Illusion?. *Multivariate Behavioral Research*, *58*(4), 762-786.

RHAH, SE, and DB developed the conceptual design of the simulation study. RHAH conducted the simulation study and wrote the article. SE and DB supervised the analyses and revised the article.

Chapter 3 is published as:
Hoekstra, R. H. A., Epskamp, S., Borsboom, D. Nierenberg, A. A., & McNally, R. J. (*2024*). Testing similarity in longitudinal networks: The Individual Network Invariance Test (INIT). *Psychological Methods*. Advance Online Publication.

RHAH, SE, and DB developed the statistical technique and the conceptual design of the simulation study. RHAH conducted the simulation study, developed the R-package, analyzed the data, and wrote the article. AAN and RJM provided the empirical data. SE, DB, and RJM supervised the analyses and revised the article.

Chapter 4 is submitted as:
Hoekstra, R. H. A., Epskamp., S., Finnemann, A. T. K., & Borsboom, D. (*under review*). Unlocking the potential of simulation studies in network psychometrics: A tutorial.

RHAH drafted the conceptual design, developed additional R functions, and wrote the article. ATKF and SE, wrote the Ising model section. SE conducted the simulation study. SE and DB supervised and revised the article.

Chapter 5 is published as:
van der Tuin, S.*, **Hoekstra, R. H. A.***, Booij, S. H., Oldehinkel, A. J., Wardenaar, K.J., van den Berg, D., Borsboom, D. & Wigman, J. T. W. (2023). Relating stability of individual dynamical networks to change in psychopathology. *PloS ONE*, *18*(11), e0293200.

SvdT, RHAH, DB, and JTWW developed the conceptual design of the study and deliberated on the statistical analyses. SvdT and RHAH conducted the statistical

analyses and wrote the article. SHB, DvdB, and JTWW collected the data. AJO, DB, and JTWW supervised the statistical analyses. SHB, AJO, KJW, DvdB, DB, and JTWW revised the article.

Chapter 6 is published as:
Lunansky, G.*, **Hoekstra, R. H. A.***, & Blanken, T. F. (2023). Disentangling the Role of Affect in the Evolution of Depressive Complaints Using Complex Dynamical Networks. *Collabra: Psychology*, *9*(1), 74841.

GL, RHAH, and TFB designed the study, conducted the statistical analyses, and wrote and revised the manuscript.

Chapter 7 is submitted as:
Hoekstra, R. H. A.*, de Ron, J.*, Epskamp, S., Robinaugh D. J., & Borsboom, D. (*under review*). Mapping the dynamics of idiographic network models to psychopathology using stability landscapes.

RHAH and JdR designed the study, conducted the analyses, and wrote the manuscript. SE and DB supervised the analyses. SE, DJR, and DB revised the manuscript.

OTHER PUBLICATIONS

Briganti, G., Scutari, M., Epskamp, S., Borsboom, D., **Hoekstra, R. H. A.**, Golino, H. F., Christensen, A. P., Morvan, Y., Ebrahimi, O. V., Costantini, G., Heeren, A., de Ron, J., Bringmann, L. F., Huth, K., Haslbeck, J. M. B. Isvoranu, A. M., Marsman, M., Blanken, T., Gilbert, A., Henry, T.R., Fried, (2024). Network analysis: An overview for mental health research. *International Journal of Methods in Psychiatric Research*, 33(4), e2034.

Ebrahimi. O. V., Borsboom, D., **Hoekstra, R. H. A.**, Tomlinson, A., Ostinelli, E., O., Epskamp, S., Bastiaansen, J. A., Cipriani, A. (2024). Towards precision in the diagnostic profiling of patients: Leveraging symptom dynamics in the assessment and treatment of mental health disorders. *The British Journal of Psychiatry*, 224, 157–163.

Burger, J.*, Isvaranu, A. M.*, ..., **Hoekstra, R. H. A.**, ... & Blanken, T. F. (2023). Reporting standards for psychological network analysis in cross-sectional data. *Psychological Methods.*, *28*(4), 806–824.

Burger, J., **Hoekstra, R. H. A.**, Mansueto, A. C., & Epskamp, S. (2022). Chapter 10. Network estimation from time-series and panel data. In Isvoranu, A.M, Epskamp, S., Waldorp, L.J. & Borsboom, D. (Eds.). *Network Psychometrics with R: A Guide for Behavioral and Social Researchers*. Rout-

ledge, Taylor & Francis Group.

Epskamp, S., **Hoekstra, R. H. A.**, Burger, J., & Waldorp, L. (2022). Chapter 9. Longitudinal data analysis and vector auto-regressive models. In Isvoranu, A.M, Epskamp, S., Waldorp, L.J. & Borsboom, D. (Eds.). *Network Psychometrics with R: A Guide for Behavioral and Social Researchers*. Routledge, Taylor & Francis Group.

Sanders, J. G., Zomer, C., **Hoekstra, R. H. A.**, de Ron, J., Blanken, T. F., Epskamp, S., van Dijken, K.S., Gerkema, M. H., Hart, L., Visser, O., Borsboom, D., & de Bruin, M. (2022). Ondersteuning van thuisisolatie en quarantaine tijdens de pilot grootschalig testen in de gemeente Bunschoten en gemeente Dronten *Report for the Dutch Ministry of Health, Welfare and Sport.*

Sanders, J. G., Zomer, C., **Hoekstra, R. H. A.**, Blanken, T. F., Gerkema, M. H., Hart, L., Visser, O., Borsboom, D., & de Bruin, M. (2021). Verhogen van deelname aan de pilot grootschalig testen in Gemeente Bunschoten. *Report for the Dutch Ministry of Health, Welfare and Sport.*

Sanders, J. G., Zomer, C., **Hoekstra, R. H. A.**, de Ron, J., Blanken, T. F., Epskamp, S., van Dijken, K.S., Gerkema, M. H., Hart, L., Visser, O., Borsboom, D., & de Bruin, M. (2021). Verhogen van testdeelname tijdens de pilot grootschalig testen in de gemeente Dronten en gemeente Bunschoten. *Rijksinstituut voor Volksgezongheid en Milieu RIVM.*

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

Vervaet, M., Puttevils, L., Vanderhasselt, M., **Hoekstra, R. H. A.**, & Fried, E. I., (2020). Transdiagnostic vulnerability factors in eating disorders: A network analysis. *European Eating Disorder Review, 29*(1), 86–100.

Derks, K., Burger, J., van Doorn, J., ..., **Hoekstra, R. H. A.**, ... & Wagenmakers, E. J. (2018). Network models to organize a dispersed literature: The case of misunderstanding analysis of covariance. *Journal of European Psychology Students*, *9*(1), 48–57.

Hoekstra, R. H. A., Kossakowski, J. J., & van der Maas, H. L. (2018). Psychological perturbation data on attitudes towards the consumption of meat. *Journal of Open Psychology Data*, *6*(1), 3.

*authors contributed equally