Optimizing Walktrap's Community Detection in Networks Using the Total Entropy Fit Index

Laura Jamison[1], Hudson F. Golino[1], & Alexander P. Christensen[2]

[1] University of Virginia

[2] University of Pennsylvania

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

Correspondence concerning this article should be addressed to Laura Jamison, Postal address. E-mail: lj5yn@virginia.edu

Abstract

Exploratory graph analysis (EGA) is used to estimate the structural organization of variables, uncovering latent dimensions as clusters of nodes. EGA first estimates a weighted network then uses the Walktrap algorithm to detect clusters of nodes. The Walktrap algorithm uses random walks to estimate the topography of a graph. The number of random walks taken (t) is typically set statically. However, the impact of t and the properties determining its optimization have yet to be fully researched. The present study proposes and tests a new approach optimizing t by iteratively varying t and employ *total entropy fit index* as a fit index to identify the number of steps that best fit the data using a Monte-Carlo simulation varying data structure characteristics. Results indicate that the proposed method is most effective for a higher number of variables per factor and when variables are polytomous. Varying t is important as spurious connections are introduced between communities. An empirical example using the Developmental Coordination Disorder Questionnaire is shown demonstrating improved measure interpretation by optimizing the Walktrap algorithm. The paper finishes with a discussion about the relevance of the findings and future directions for research.

Keywords: keywords

Word count: X

Optimizing Walktrap's Community Detection in Networks Using the Total Entropy Fit Index

Introduction

Within psychological research, network modeling approaches have been steadily gaining popularity across clinical psychology (Borsboom, 2017; McNally, 2016), developmental psychology (Dijkstra, Cillessen, & Borch, 2013), psychopathology (Bringmann et al., 2013), and in particular, psychometrics (Costantini et al., 2019; Golino, Shi, et al., 2020; Marsman et al., 2018). Within network psychometrics, a common goal of research is to estimate the structural organization of variables (e.g., items in a survey or test) by uncovering latent dimensions as clusters of nodes in weighted networks, a general approach termed *exploratory graph analysis* (**EGA**; Golino & Epskamp, 2017; Christensen et al., 2019b, 2019c; Golino, Shi, et al., 2020). EGA is an innovative approach for dimensionality assessment and reduction that starts by estimating a network (Golino, Shi, et al., 2020) and then uses the Walktrap algorithm (Pons & Latapy, 2006) to detect clusters of nodes.

The Walktrap algorithm is a modularity-based approach (similar to cluster analysis), shown to outperform other algorithms (e.g., Fast Greedy, Newman's Spectral Approach) when using correlation matrices and sparse count networks (Christensen, Garrido, & Golino, 2021; Gates, Henry, Steinley, & Fair, 2016; Orman & Labatut, 2009) and has been repeatedly found to successfully uncover community structure in both small and large networks (Golino, Shi, et al., 2020; Pons & Latapy, 2006; Yang, Algesheimer, & Tessone, 2016). In the area of dimensionality analysis and reduction, when EGA is used with the Walktrap algorithm, it has shown to perform above and beyond other methods used in factor analysis when the data generating mechanism is a factor model (Golino & Epskamp, 2017). These findings make the Walktrap algorithm an attractive choice for substantive psychological research, from neuroscience (Gates et al., 2016) to the study of individual differences (Golino, Shi, et al., 2020).

The Walktrap algorithm has been used in many applications in psychology. For example, the group iterative multilevel model estimation (*GIMME*) uses the Walktrap algorithm as a part of a process designed to recover connections and directionality within regions of interest from fMRI data (Gates & Molenaar, 2012). Another method focuses on detecting communities within networks using Cohen's $\varkappa$ for clustering social network data (Hoffman, Steinley, Gates, Prinstein, & Brusco, 2018), while EGA aims to estimate the number of dimensions in multivariate data (Golino & Epskamp, 2017; Golino, Shi, et al., 2020). In each application, the Walktrap algorithm uses a series of random walks to define one important characteristic of the topography of a graph: the number and composition of communities (i.e., clusters of nodes or variables). The algorithm begins with a square matrix, the values of which indicate the relationship between units of analysis. In psychology, this matrix is typically made up of (partial) correlations between variables which form weighted, undirected graphs when modeled using network techniques.

A network is considered to have a good community structure when the average edge weight within a community is higher than the edge weights between that community's nodes and nodes in other communities (Newman, 2006; Pons & Latapy, 2006). The Walktrap algorithm capitalizes directly on this definition of good community structure by using a series of random walks. Starting in a given node, the algorithm repeatedly moves along the edges connecting that node to its neighbors. A probability function determines where it is more likely to "walk" to a node with a higher degree than a node with a lesser degree. In this way, the process will get "trapped" within a community because it is less probable for it to move to a node that does not belong in that community.

The number of random walks (t) taken is generally set statically as an empirical compromise to computational efficiency to make sure algorithm run time is reasonable (Pons & Latapy, 2006). Pons and Latapy (2006) recommend taking steps $t = 4$ or $t = 5$ as the most computationally efficient approach with the least empirical compromise. Typically, a random walk of $t = 4$ is used in many applications (Gates et al., 2019; Golino, Shi, et al., 2020). Pons and Latapy (2006) state that t must be large enough to adequately capture the topography of the graph, but if t is too large, then the probability of transitioning from one node to another depends solely on the degree of the second node. As sparsity increases, t can also increase as the convergence speed of the algorithm increases, and conversely t should decrease as density increases (Pons & Latapy, 2006). However, the impact of t and the properties determining its optimization have yet to be fully researched. This is especially pressing in the network psychometric literature which uses a range of data structures from many subfields of psychology, thus making a “one solution fits all” approach unlikely.

The goal of the current paper is to propose and test a new approach to optimize the number of steps of the Walktrap algorithm, which could potentially improve its accuracy to identify groups of variables in weighted networks. Instead of using a predetermined number of steps, we iteratively vary the number of steps (from 3 to 10) and employ a novel fit index termed *total entropy fit index* (**TEFI**; Golino, Moulder, et al., 2020) to identify the number of steps that best fit the data. A Monte-Carlo simulation is implemented to verify if our optimization approach improves the capacity of the Walktrap algorithm to estimate the number of factors (clusters of nodes) in weighted networks. We controlled several important characteristics: sample size, number of variables per factor, factor loadings, interfactor correlation, type of variable, link probability, type of correlation and network estimation method. The paper is organized as follows: first we will present a general overview of network model estimation used in this study followed by an in depth review of the Walktrap algorithm. Then, we will discuss the proposed method as well as the methods and metrics used to test it. Finally, an empirical example is shown to demonstrate how our optimization approach improves the interpretation of the final partition of the network into distinct communities or factors using data from the Developmental Coordination Disorder Questionnaire (DCDQ: Schoemaker et al., 2006).

Estimating Factors in Network Psychometrics

Network Model Estimation. To estimate the number and composition of factors in the network psychometrics literature, EGA is used. EGA uses two main network estimation methods: the graphical least absolute shrinkage and selection operator (glasso; Friedman, Hastie, & Tibshirani, 2008) and triangulated maximally filtered graph (TMFG; Massara, Di Matteo, & Aste, 2016).

The *glasso* is a commonly used method for estimating networks that are known as Gaussian Graphical Models (GGM) (Lauritzen, 1996). The original input matrix and the edges of the network are made up of partial correlations between variables, in other words the correlation between variables after conditioning on all other variables in the network. The lasso operator shrinks coefficients to zero (to account for spurious relationships and control for overfitting to the data). This creates a sparse network that can be formed at different levels between a completely connected network to an entirely unconnected network. As each network in this range is estimated, the extended Bayesion information criterion (EBIC) (Chen & Chen, 2012) is computed. The network with the lowest EBIC is selected (Epskamp et al., 2018, 2018; Foygel & Drton, 2010). The EBIC has a hyperparameter that provides a penalization for more complicated models to help control for overfitting to the data (Epskamp & Fried, 2018). Typically, this hyperparameter (γ) is set to 0.5 (Foygel & Drton, 2010). Lower values of γ provide greater sensitivity but may reduce specificity (Williams, Rhemtulla, Wysocki, & Rast, 2019). As such, EGA starts off using $\gamma = 0.5$. If the network has disconnected nodes, EGA will continue to lower the value of γ until this is no longer the case.

The TMFG algorithm is another commonly used method for network estimation that works by constraining the number of zero-order correlations included in the network to be $3n - 6$, where n is the number of variables (Christensen et al., 2019a; Golino, Shi, et al., 2020; Massara et al., 2016). The algorithm begins by connecting together the four variables that have the highest correlation sum to all other variables. Iteratively, variables are added to this network based on the highest correlation sum of three variables with nodes already contained in the network.

Walktrap Algorithm. After estimating a weighted network, the EGA technique uses the Walktrap algorithm to uncover the number and composition of latent factors, represented in networks as clusters of densely connected nodes (Golino, Shi, et al., 2020). The Walktrap algorithm (Pons & Latapy, 2006) transforms the original correlation matrix into a matrix containing transition probabilities called a transition matrix. Transition probabilities refer to the probability of transitioning between nodes based on edge strength. Edge strength is defined by the strength of the relationship between nodes, in this case the partial correlation between variables. This is done using a series random walks, typically of length 4, to estimate a distance measure for each pair of nodes. The algorithm then seeks to minimize the sum of squared distances between each node and all other nodes in its cluster using Ward's hierarchical clustering method (Ward Jr, 1963).

More formally, the Walktrap algorithm begins with a weighted, undirected original input matrix, A, where A_{ij} is the strength between node i and node j. The algorithm reconstructs matrix A into a transition matrix, P, using a Markov chain random walk process defining the transition probability between node i and node j to be $P_{ij} = \frac{A_{ij}}{d(i)}$ and doing so in length t to be P^t_{ij}. Note that this probability will be influenced by the degree (number of connections) of node j such that there is a higher probability of transitioning to a node with a higher degree. P^t_{ij} will also be higher when i and j are in the same community, however a high P^t_{ij} does not necessarily mean that nodes i and j are in the same community. Using random walks, a distance d will be defined between nodes.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \frac{(P_{ik} - P_{jk})^2}{NS(k)}} \tag{1}$$

Where k refers to the cluster node i and j belong to. r should be smaller between node i and node j if they are in the same community and comparatively larger if they are not in the same community. This same logic can be applied to define the distance between node j and community C by

$$P^t_{Cj} = \frac{1}{|C|} \sum_{i \in C} P^t_{ij} \tag{2}$$

We can then define the distance between communities C_1 and C_2 to be

$$r_{C_1 C_2} = \sqrt{\sum_{k=1}^{n} \frac{(P^t_{C_1 k} - P^t_{C_2 k})^2}{d(k)}} \tag{3}$$

The Walktrap algorithm uses an agglomerative clustering approach beginning by defining the most general case where each node is its own cluster. The distance, r, is computed between each of the nodes. The algorithm then begins to iteratively merge nodes with edges between them into larger clusters. Per methodology proposed by Pons and Latapy (2006), this merging is done in such a way to approximately minimize the variation in squared distances between each node and its community (σ).

$$\Delta\sigma(C_1, C_2) = \frac{1}{n}(\sum_{i \in C_3} r^2_{iC_3} - \sum_{i \in C_1} r^2_{iC_1} - \sum_{i \in C_2} r^2_{iC_2}) \tag{4}$$

Where C_1 and C_2 are clusters being merged to form a third cluster, C_3. The resulting values will be stored in a dendrogram. From the probabilities given in P^t_{ij}, the length t of the random walks should be optimized to gather sufficient information to accurately partition the clusters.

Optimizing the number of steps in the Walktrap algorithm

As previously stated, the number of random walks (t) used in the Walktrap algorithm is generally set statically as an empirical compromise to computational efficiency, with $t = 4$ or $t = 5$ being recommended as the most computationally efficient approach (Pons & Latapy, 2006). In the network psychometrics literature, setting the number of steps as $t = 4$ has shown to be effective in recovering the number of simulated factors (Golino, Shi, et al., 2020) or communities of sparse count data (Gates et al., 2019). As Pons and Latapy (2006) states, t must be large enough to adequately capture the topography of the graph, but if t is too large, then the probability of transitioning from one node to another depends solely on the degree of the second node. For most applications in psychology in which the Walktrap algorithm is used to identify communities of densely connected nodes representing latent factors, as in the EGA approach, tuning the number of steps is highly desirable, since it can lead to improved factor estimation and can facilitate the interpretation of the factors due to a improved placement of variables per factor.

To tune the Walktrap hyperparameter (i.e. number of steps), we propose an iterative algorithm. First, a network is estimated (e.g., using the glasso or the TMFG network methods). Then, the Walktrap algorithm is applied with the number of steps set as 3. The fit of the resulting partition of the multidimensional space (in this case the partition of the network into communities) to the data is then computed using the *total entropy fit index* (**TEFI**: Golino, Moulder, et al., 2020). The TEFI index has shown to present the highest accuracy in detecting the correct dimensionality solution (i.e. number of factors and correct placement of variables per factor) in a Monte-Carlo simulation study (Golino, Moulder, et al., 2020) where traditional fit indices used in factor analysis and structural equation modeling were also used. The TEFI index assesses the degree of uncertainty of the partition of a multidimensional space into separate distinct categories (i.e., latent factors or clusters), where lower TEFI values indicate less uncertainty of the dimensionality solution. In other words, lower TEFI values indicates that a given dimensionality structure fits the data better than an alternative dimensionality solution with higher TEFI values, indicating that the former is more likely to represent the best organization of the variables than the latter. The TEFI index is calculated as follows:

$$TEFI = \left[\frac{\sum_{i=1}^{N_F} \mathcal{S}(\boldsymbol{\rho}_i)}{N_F} - \mathcal{S}(\boldsymbol{\rho})\right] + \left[\left(\mathcal{S}(\boldsymbol{\rho}) - \sum_{i=1}^{N_F} \mathcal{S}(\boldsymbol{\rho}_i)\right) \times \sqrt{N_F}\right] \tag{5}$$

Where N_F is the number of factors (or communities) estimated by the Walktrap algorithm, $\mathcal{S}(\boldsymbol{\rho}_i)$ is the Von Neumann entropy for each individual factor and $\mathcal{S}(\boldsymbol{\rho})$ is the total entropy of the system of variables. Golino, Moulder, et al. (2020) showed that the Von Neumann entropy can be approximately estimated in a correlation matrix by scaling it so that the trace of the matrix equals one (i.e. taking a correlation matrix and dividing all entries by the number of columns of the matrix). After scaling the correlation matrix, an entropy-like metric can be obtained by the negative of the trace of the product of the density matrix by the log of elements of the density matrix (see: Golino, Moulder, et al., 2020).

The TEFI index has two parts that can be separated as $TEFI = [A] + [B]$ (Golino, Moulder, et al., 2020). Element $[A]$ is similar to that of the total correlation of multiple variables (Watanabe, 1960), but instead of using the joint entropy for the partitions (factors or clusters), it uses the total entropy of the system (i.e. entropy calculated using all variables together). Additionally, the sum of the individual entropies (estimated per factor or cluster) is divided by the number of partitions (i.e. factors or clusters), yielding what Watanabe (2001) termed “K-function”. Element $[B]$ reduces the influence of $[A]$ by the number of factors used to describe a given data set. As Golino, Moulder, et al. (2020) note, while $[A]$ is expected to decrease monotonically as the number of factors increases, $[B]$ is expected to increase as the number of factors increase. $[B]$ represents the reduction in average entropy of a set of data conditional on a given factor or community structure. The square root of the number of factors was chosen in $[B]$ in order to control the expected growth trajectory of $[B]$ as the number of factors increases. Golino, Moulder, et al. (2020) argues that the expected decrease in total entropy going from 1 to 2 factors would be higher than the expected decrease in entropy going from 100 to 101 factors, and therefore the multiplication by the square root of the number of factors is used to model this behavior.

Methods

In order to better understand the impact of varying the number of steps taken by the random walks, t will be adjusted from 3 to 10 within multiple data structures and community structure accuracy will be compared. The next paragraphs will describe the data generation mechanism used (a two-step approach), the design of the Monte Carlo simulation implemented and how the results are analyzed.

Data will be generated using a Monte Carlo simulation manipulating various data properties. First, a four factor structure and resulting correlation matrix will be estimated varying the sample size, continuous or categorical variables (4 categories), the number of variables per factor, whether or not the factors have the same number of variables within them, factor loadings, and the correlation between factors. When the number of variables within a community are unequal, two factors are reduced by one variable and two factors are increased by one variable (i.e., 8 variable factors have four factors containing 7, 7, 9, and 9 variables). Relationships between the resulting variables will be estimated using either Pearson correlation, polychoric correlations (for categorical variables), or Louis-Guttman Image Structural Analysis (described below) and placed in a matrix.

Data Generation

The data generation mechanism used in the current paper follows a two-step approach. The first step follows the common factor model used by Golino, Shi, et al. (2020), that works as follows. First, the reproduced population correlation matrix (with communalities in the diagonal) is computed:

$$\mathbf{R_R} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}', \tag{6}$$

where $\mathbf{R_R}$ is the reproduced population correlation matrix, *lambda* ($\mathbf{\Lambda}$) is a $k \times r$ factor loading matrix for k variables and r factors, and *phi* ($\mathbf{\Phi}$) is the structure matrix of the latent variables (i.e., a $r \times r$ matrix of correlations among factors). This procedure implies that the generated data does not contain correlated residuals (minor factors) at the population level.

The population correlation matrix $\mathbf{R_P}$ is then obtained by inserting unities in the diagonal of $\mathbf{R_R}$, thereby raising the matrix to full rank. Next, a Cholesky decomposition of $\mathbf{R_P}$ is performed, such that:

$$\mathbf{R_P} = \mathbf{U}'\mathbf{U}. \tag{7}$$

If either $\mathbf{R_P}$ is not semi-positive definite (i.e., at least one eigenvalue is ≤ 0) or an item's communality is greater than 0.90, the $\mathbf{\Lambda}$ matrix is replaced and a new $\mathbf{R_P}$ matrix is computed following the same procedure. Subsequently, the sample data matrix of continuous variables is computed as:

$$\mathbf{X} = \mathbf{Z}\mathbf{U}, \tag{8}$$

where $\mathbf{Z}$ is a matrix of random standard normal deviates with rows equal to the sample size and columns equal to the number of variables.

Following Golino, Shi, et al. (2020) cross-loadings with magnitudes consistent to those commonly found in real data (Bollmann, Heene, Küchenhoff, & Bühner, 2015) are randomly drawn from a normal distribution (with mean zero and variance of .15) for all the items except for the first two in each factor, which were set as markers (i.e., all of their cross-loadings are fixed to zero). Of note, regarding the generation of the main loadings: The function generates the main loadings by drawing random values from a uniform distribution that has a range of $\pm.10$ from the specified value (so if the main loadings are set at 0.70, the function generates loading values between 0.60 and 0.80). The generated data is then used to compute an empirical correlation matrix $\mathbf{C_X}$.

After estimating the data following the procedure described above, a second step is implemented. As indicated by Figure 1, the resulting correlation matrix of the simulated data (from the factor model; $\mathbf{X}$) is then be multiplied by a predetermined undirected and unweighted network structure $\mathbf{N}$ (with number and composition of communities equal to the number and composition of factors as simulated in the first step above) where the probability of nodes being linked within a community and between communities will be varied from low to high. The networks are simulated following the framework of Girvan and Newman (2002) for generating networks with specific (i.e., known) community structures. Multiplying $\mathbf{C_X}$ by $\mathbf{N}$ generates a matrix of weights $\mathbf{W}$ with two important characteristics: the underlying factor structure is known (used to generate $\mathbf{C_X}$) and matches exactly the community structure of $\mathbf{N}$. The final sample data matrix of continuous variables is computed following a multivariate normal distribution with mean zero and variance-covariace matrix $\mathbf{W}$. The final sample data matrix contain continuous variables that can be discretized to generate polytomous data following the procedure described by Golino, Shi, et al. (2020).

This second step is necessary to control an important characteristic of networks that is

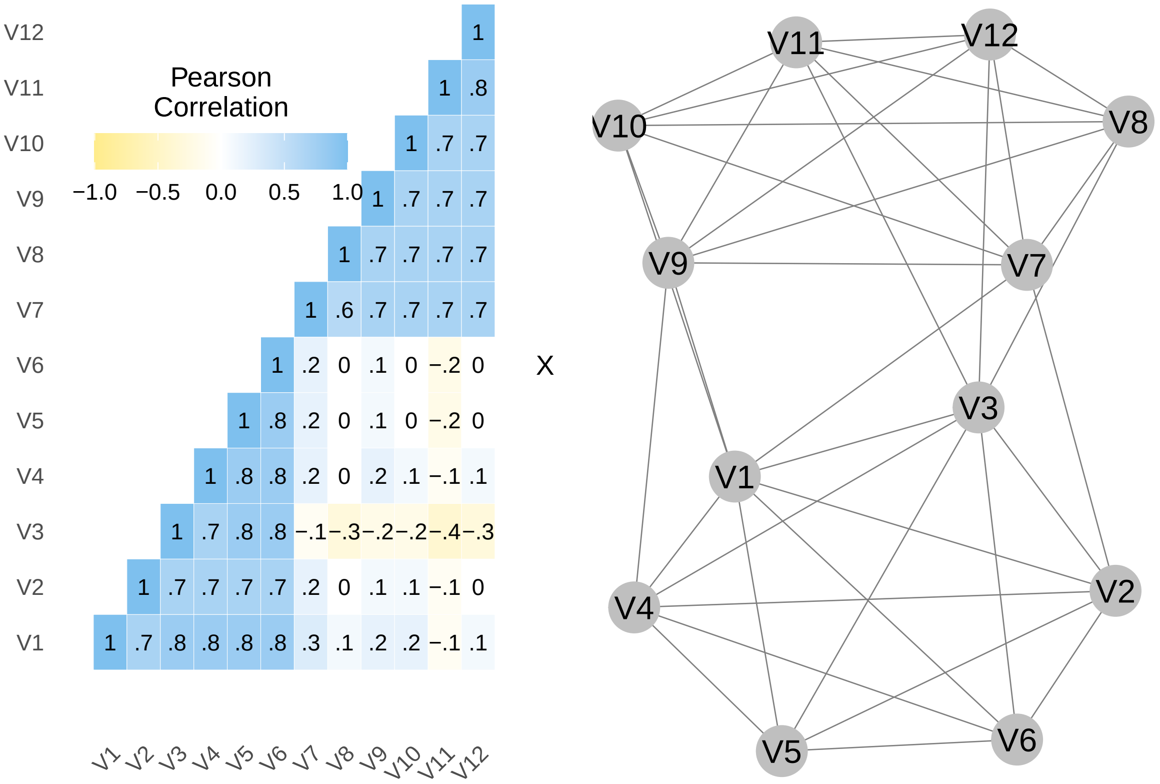

Figure 1. Network Data Generation

usually not controlled in the simulation studies employing the exploratory graph analysis technique (e.g., Golino, Shi, et al., 2020): the link probabilities. Using a two-step data generation approach is necessary to make the resulting data closer to real datasets.

In sum, data was generated in two steps: first, data is generated using a factor model and the sample correlation matrix is obtained. In the second step a unweighted, undirected network is generated and both matrices are multiplied to add a structural bias to the sample correlation matrix obtained in the first step. This second portion can be thought of as adding in spurious relationships at both the intra- and inter-community levels. We simulated a network with a known number of communities (matching the data-generation mechanism of the factor model), but with different levels of probabilities within and between communities in terms of edges. Therefore, even if the true factor model has low interfactor correlations, the structural bias will forcibly add edges between communities, or if the true factor model has low factor loadings, the structural bias will forcibly add edges within communities. Due to this methodology, we use the terms **factor** and **community** interchangeably.

All R code used in the current project are available in the Open Science Framework, as well as the R Markdown manuscript integrating code and text for data analysis.

Design

To investigate the suitability of our algorithm to optimize the number of steps of the Walktrap procedure, a Monte Carlo simulation was implemented and nine between-subject data factors were systematically manipulated. Within the factor structure, the sample size (500 and 1000), the equivalence in the number of variables per factor (i.e., whether or not all factors have the same number of variables), number of variables per factor (4, 8), factor loadings (.40 and .70), factor correlations (.30, .50 and .70), and type of variable (continuous or polytomous with four response categories). Within the network structure, the probability of links between communities (p-Out: .50 and .90), probability of links within communities (p-In: .50, .75), and network method (glasso and TMFG) were manipulated. The relationship between simulated variables was estimated either using traditional correlation coefficients (Pearson for the continuous data condition and polychoric for the polytomous data) or using the scaled covariance (or correlation) of images from Guttman's Image Structural Analysis (Guttman, 1953). In Guttman's image structural analysis, the covariance matrix of the *anti-images* ($\mathbf{\Gamma}$) for n variables is:

$$\mathbf{\Gamma} = \mathbf{S^2} \times \mathbf{R^{-1}} \times \mathbf{S^2}$$

where $\mathbf{S^2}$ is the diagonal matrix with the anti-norms ($\mathbf{S^2} = Diag\left(\frac{\Delta(R)}{\Delta(R_{ii})}\right)$, being $\Delta\mathbf{R}$ the determinant of $\mathbf{R}$ and $\Delta(R_{ii})$ the cofactor of R_{ii}), and $\mathbf{R^{-1}}$ is the inverse of the correlation matrix $\mathbf{R}$. The covariance matrix of the *images* ($\mathbf{G}$) is:

$$\mathbf{G} = \mathbf{R} + \mathbf{\Gamma} - 2\mathbf{S^2}$$

Guttman (1953) proposed a theorem in which any correlation coefficient can be regarded as the difference between two covariances, one for the common parts between the variables (images) and another by the alien parts (anti-images). This theorem leads to an important paradox, that the alien parts (i.e., the covariance of the partial anti-images) are more important to the structural analysis of a correlation matrix than the common parts (i.e., the covariance of the partial images), because a correlation matrix can be computed using only the partial anti-norms and the covariance of the anti-images (i.e., $\mathbf{R} = \mathbf{S^2} \times \mathbf{\Gamma^{-1}} \times \mathbf{S^2}$), but cannot be computed using the covariance of the images. Commoness in image structural analysis comes from the use of a multiple-regression approach in which correlations can be explained by means of the multiple regression of each variable on the remaining n-1 variables. Guttman's image structural analysis was linked to factor analysis in several classical works (Harman, 1976; Harris, 1962), and here the scaled covariance matrix of the images is also used, to be contrasted to the results obtained using traditional (Pearson or polychoric) correlation estimation techniques. The goal in using Guttman's image structural analysis is to investigate its effect in the accuracy of the Walktrap algorithm used in the EGA framework.

This design results in 1536 conditions to be compared. For each condition, 500 datasets were simulated.

Data Analysis

Assessing Accuracy of Cluster Partitions. For each simulated dataset, $TEFI$ values from $t = 3$ to $t = 10$ will be compared. The model with the lowest value of $TEFI$ will be identified and the structure and accuracy of the partition will be compared to $t = 4$ as seen in Figure 2.

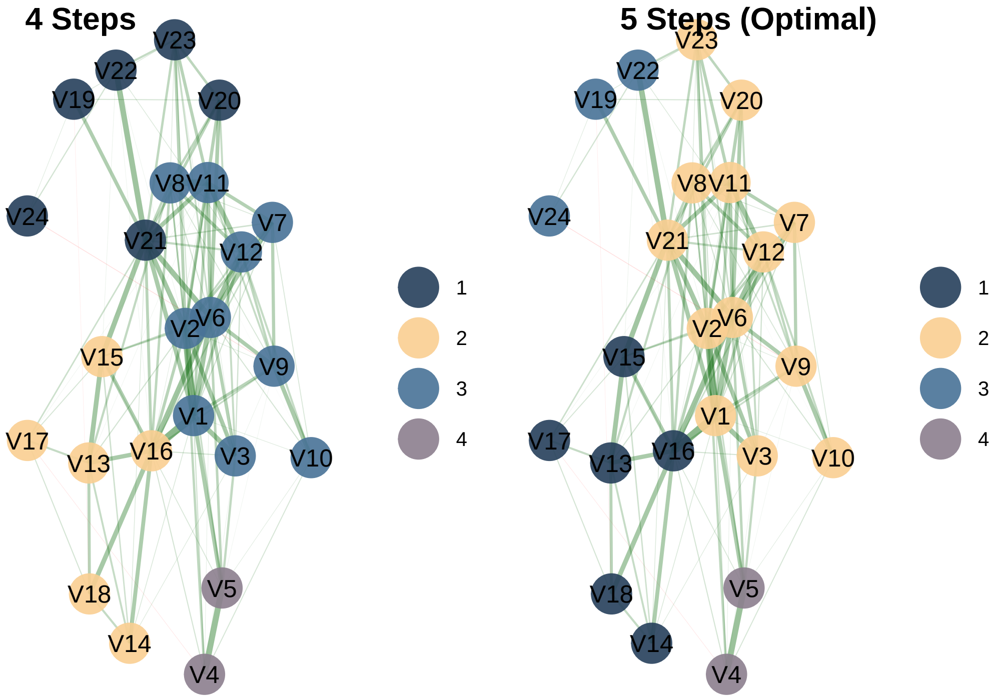

Figure 2. Default vs. Optimal Estimated Structure

The accuracy of a partition is considered to be higher when nodes sharing high edge weights are assigned to the same cluster while nodes sharing comparatively lower edge weights are assigned to separate clusters. The current paper will employ multiple measures of fit to assess the accuracy of the Walktraph algorithm: Majority Placement (MP), the Hubert-Arabie Adjusted Rand Index (ARI_{HA}), and Normalized Mutual Information (NMI). Additionally we use an overall measure of accuracy coded as 1 when the correct number of communities was discovered and 0 otherwise.

Majority Placement. Majority placement (MP) is a classification rate assessing the portion of nodes correctly classified (Gates et al., 2016; Girvan & Newman, 2002) If a node is placed in a community with more than 50% (the majority) of all other nodes from its true community then it is defined as being in its true community (Fortunato, 2010). More formally,

$$MP = \sum_{i=1}^{N} \frac{\tau_i}{N}, \begin{cases} 1 \text{ if node } i \text{ is placed with} \geq 50\% \text{ of nodes from its true community} \\ 0 \text{ otherwise} \end{cases} \tag{9}$$

where for each node i, τ_i is 1 if the node is in a community with 50% or more of other nodes from its true community. Note that this metric becomes unreliable if there are fewer communities identified than in the true structure (e.g., if only one community is detected, all nodes are placed with $\geq 50\%$ of the nodes from their true community).

Hubert-Arabie Adjusted Rand Index. Given the potential biases of relying solely on MP, we are additionally employing the *Hubert-Arabie Adjusted Rand Index* (ARI_{HA}; (Hubert & Arabie, 1985)). ARI_{HA} provides complementary information to the MP however it has more rigid constraints on what constitutes correct placement (Gates et al., 2016; Steinley, 2004). There are penalizations for pairing nodes in the same community if they are not paired in the true structure, and vice versa. In this way, ARI_{HA} penalizes the quality of fit for identifying fewer communities than exist in the true structure.

ARI_{HA} is formally defined as:

$$ARI_{HA} = \frac{\binom{N}{2}(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{N}{2} - [(a+b)(a+c) + (c+d)(b+d)]} \tag{10}$$

where a represents the number of paired nodes, in the same community, both in the true and recovered cluster solution; b represents the number of nodes paired in the same community in the true structure that were not paired in the same community in the recovered structure; c represents the number of nodes not paired in the same community in the true structure that were paired in the same community in the covered structure; and finally, d represents the number of node pairs that are not in the same community in both the true and recovered structure. $\mathrm{ARI_{HA}}$ was implemented using the clues package in R (Chang et al., 2010).

Normalized Mutual Information. Normalized mutual information (NMI) compares the true and recovered partitions by creating a confusion matrix where rows represent true communities and columns represent recovered communities (Danon, Diaz-Guilera, Duch, & Arenas, 2005). N_{ij} is the node in true community i that also appears in the recovered community j. This matrix is then used to assess the similarity of partitions. Formally, NMI is defined as:

$$I(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij} log(\frac{N_{ij}}{N_{i.}N_{.j}})}{\sum_{i=1}^{c_A} N_{i.} log(\frac{N_{i.}}{N}) + \sum_{j=1}^{c_B} N_{.j} log(\frac{N_{.j}}{N})} \tag{11}$$

where N_{ij} represents a matrix in which A represents a vector of true communities, B represents a vector of recovered communities, c_A and c_B denote the number of communities in either A or B, the sum of row i is denoted by $N_{i.}$, and the sum of column j is denoted by $N_{.j}$. NMI has a maximum of 1, the true and recovered communities are identical, and a minimum of 0, when no true communities are recovered.

Results

Comparing $TEFI$ across all values of t, the lowest $TEFI$ value within 49.7% of the simulated datasets were obtained by a value other than $t = 4$. Within all values of t other than 4, Figure 3 represents the proportion of datasets where each value of t provided the optimal fit (i.e., the lowest $TEFI$ value).

The question then becomes how does the optimal dimensionality structure (i.e., the structure with the lowest $TEFI$) compare to the default structure (i.e., the structure estimated with $t = 4$) when the optimal model is selected for a value of t other than 4. To

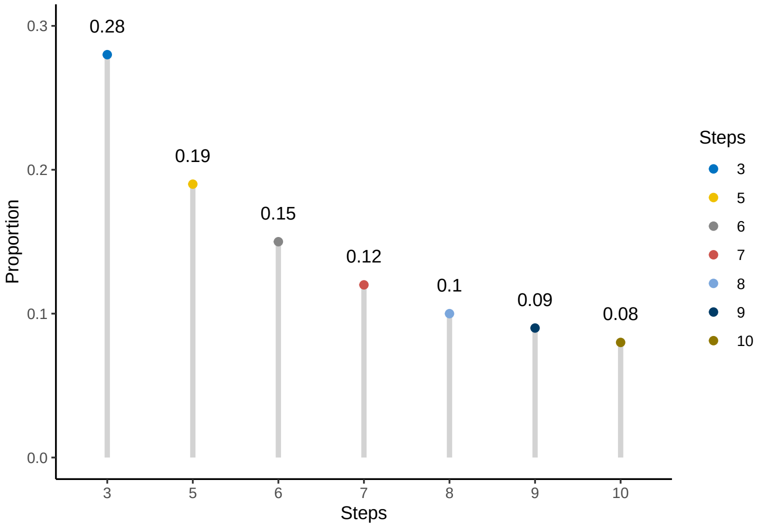

Figure 3. Proportion of Optimal Steps Within t Other Than 4

investigate this, we evaluated the interaction between each data factor. A 10-way ANOVA across all data factors was conducted using change in overall accuracy, MP, ARIHA, and NMI. Change in a given metric was computed by the metric value at the optimal number of steps minus the metric value at $t = 4$ for those datasets where the default structure estimated was not the structure estimated using the optimized number of steps. We recorded the effect size of each main effect and interaction using partial eta squared (η_p^2) following the guidelines of Cohen (2013) where values of 0.01 represent small effects, 0.06 medium effects, and 0.14 or more large effects.

For each metric, a greater positive difference between the optimal structure and the default structure is preferred. For instance, if the optimal structure for a given dataset had a MP value of 0.5 and the default structure had a MP value of 0.1, then the difference in MP would be $0.5 - 0.1 = 0.4$, indicating a gain in majority placement when the number of steps is optimized using the TEFI index. However, if the optimal structure had a MP value of 0.4 and the default structure had a MP value of 0.5, then the difference in MP would be $0.4 - 0.5 = -0.1$, indicating that by optimizing the number of steps the resulting structure has a lower majority placement value.

There were several data structures that exhibited no main effect or interaction after evaluating η_p^2. However, two such data conditions exhibited interesting results when visually inspecting change in accuracy metrics. 4 shows differences in the interaction between network estimation method and correlation type when split by variable type. When using Louis-Guttman Image Structural Analysis, we see slight improvement in accuracy for polytomous data regardless of network estimation method and improvement in MP when using *glasso*.

Additionally, we found small differences (as seen in Figure 5), a result that's in the opposite direction of what is obtained in the same conditions but for the traditional correlation techniques. For this data condition, we see improvements both in majority placement and accuracy.

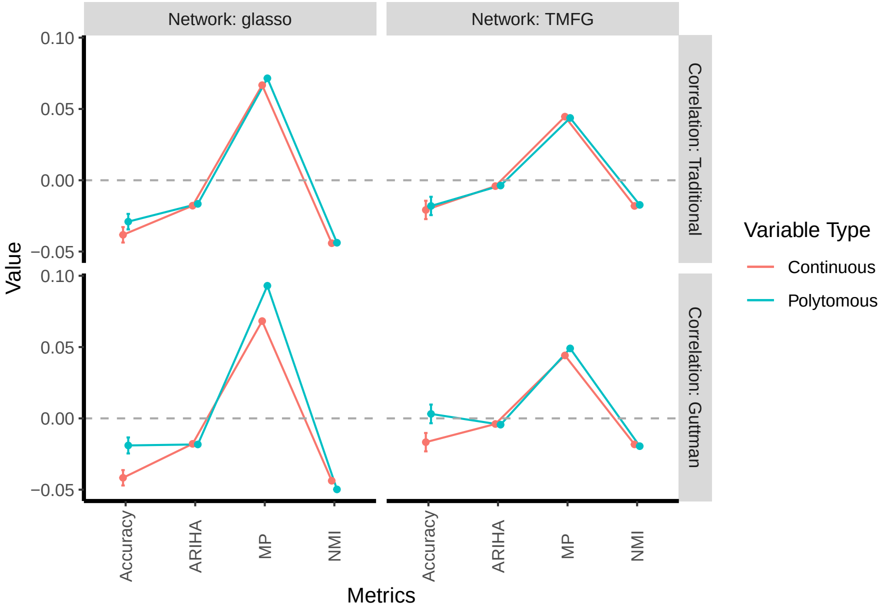

Figure 4. Interaction between Network Estimation and Correlation Type by Variable Type

Figure 6 shows the overall interaction of interfactor correlations, factor loadings, the probability of node connections within communities, and the probability of node connections between communities split by the number of variables per commmunity and the type of variable. In general, four variables per community do not see marked improvement in any

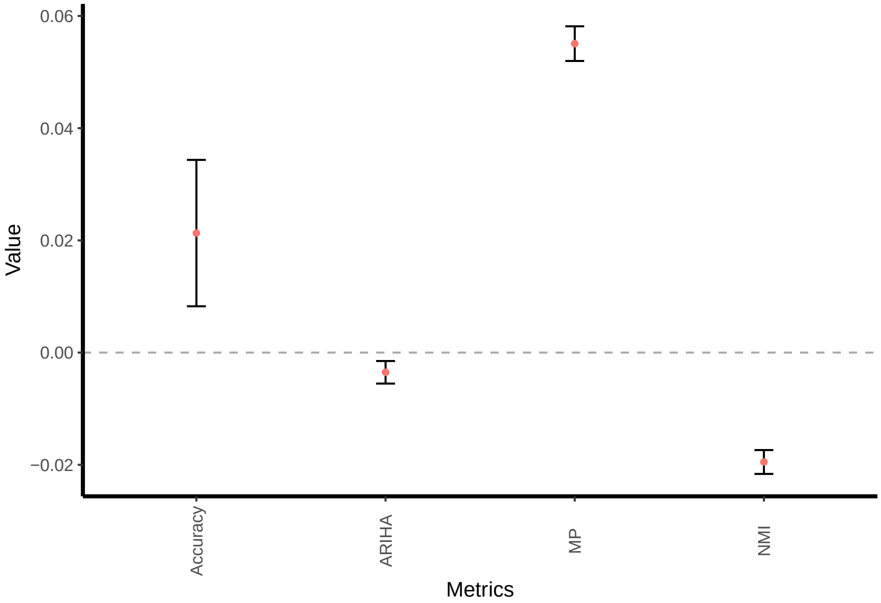

Figure 5. Effect with Sample Size of 1000, Using Louis-Guttman Image Structural Analysis, Number of Variables Is Not Equal Across Factors, TMFG Network Estimatioon, and Polytomous Variables

metric except for MP. Eight variables per community, however, is in general associated with greater improvement across each metric. This relationship is most notable and consistent within accuracy. When p-Out is lower, 0.5, number of variables per community is 8, as interfactor correlations increase so does the improvement in accuracy, regardless of factor loadings. However, when p-Out is higher, 0.9, the opposite is true. For 8 variables per community, as interfactor correlations increase, there is a decrease in the improvement of accuracy. In both scenarios the change in other metrics remain constant with the exception of MP. When p-Out = 0.90, there is a slight upward trend in the improvement of MP as interfactor correlations increase for both 4 and 8 variable per factor.

The relationship between interfactor correlations, factor loadings, and p-In and p-Out remain constant regardless of variable type. Interestingly, the trend in accuracy improvement as interfactor correlations increase seen when split by variable type for p-Out = 0.50 is no longer notable when split by variable type. However, when p-Out = 0.90, the improvement in accuracy still decreases as interfactor correlation increases and the improvement in MP still increases as interfactor correlation increases.

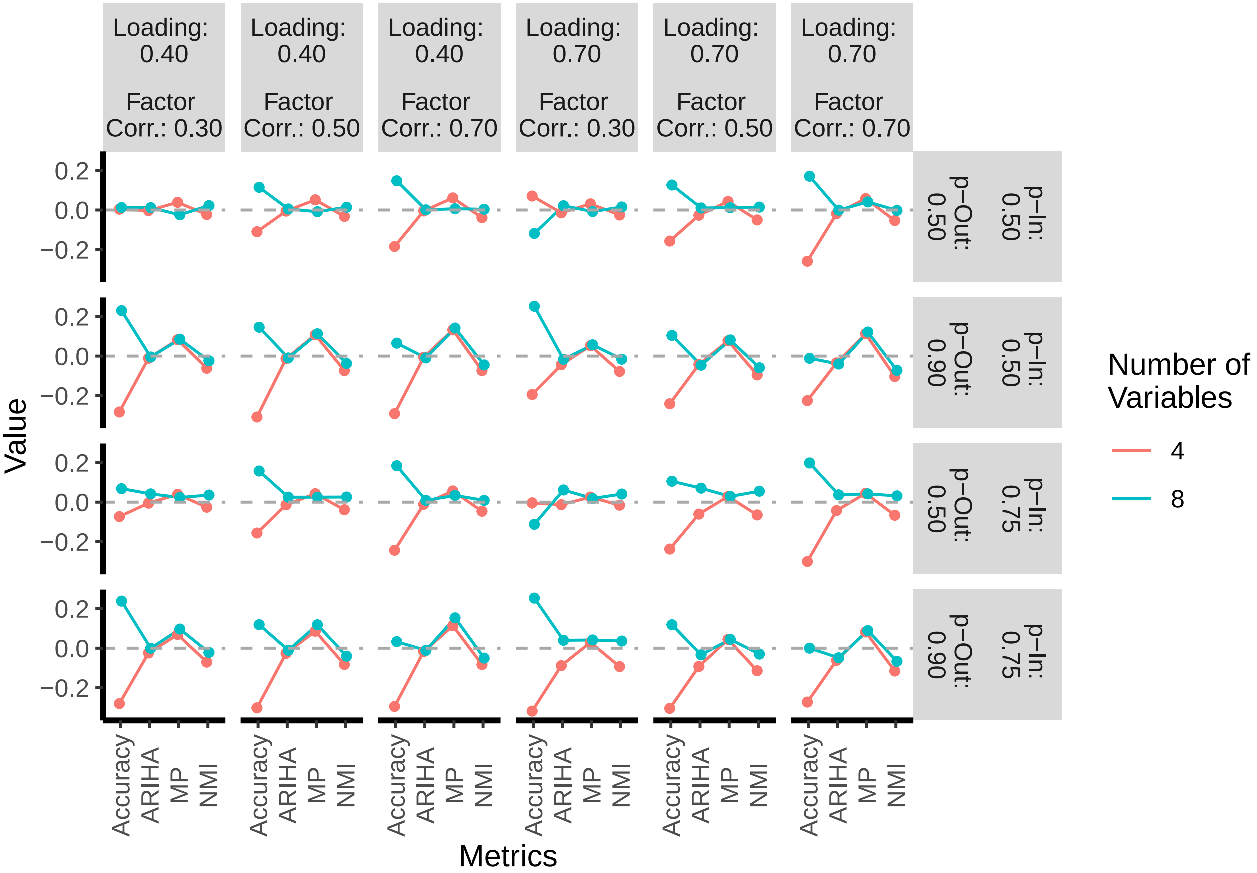

Figure 6. Interaction between Factor Loadings (Loading), Factor Correlations (Factor Corr.), p-In and p-Out by Number of Variables

Since most, if not all, of psychological measurement relies on polytomous response variable, and the larger effects of tuning the number of steps used in the Walktrap algorithm were seen in polytomous data conditions with more variables per factor (i.e., 8), the remainder of results will be reported on these conditions only and one 8-way ANOVA was conducted for these data structures. Table 1 shows the effect sizes for each effect from this model.

Figures 8, 9, and 10 show the 3 effects showing at least a small effect size ($\eta_p^2 > 0.01$). As seen in Figure 8, when p-In is greater (0.75), there is a slight improvement in each metric change compared to p-In at 0.50. Similarly, Figure 9 when p-Out is greater (0.90) and interfactor correlation is lower (0.30) there is greater improvement in accuracy and MP.

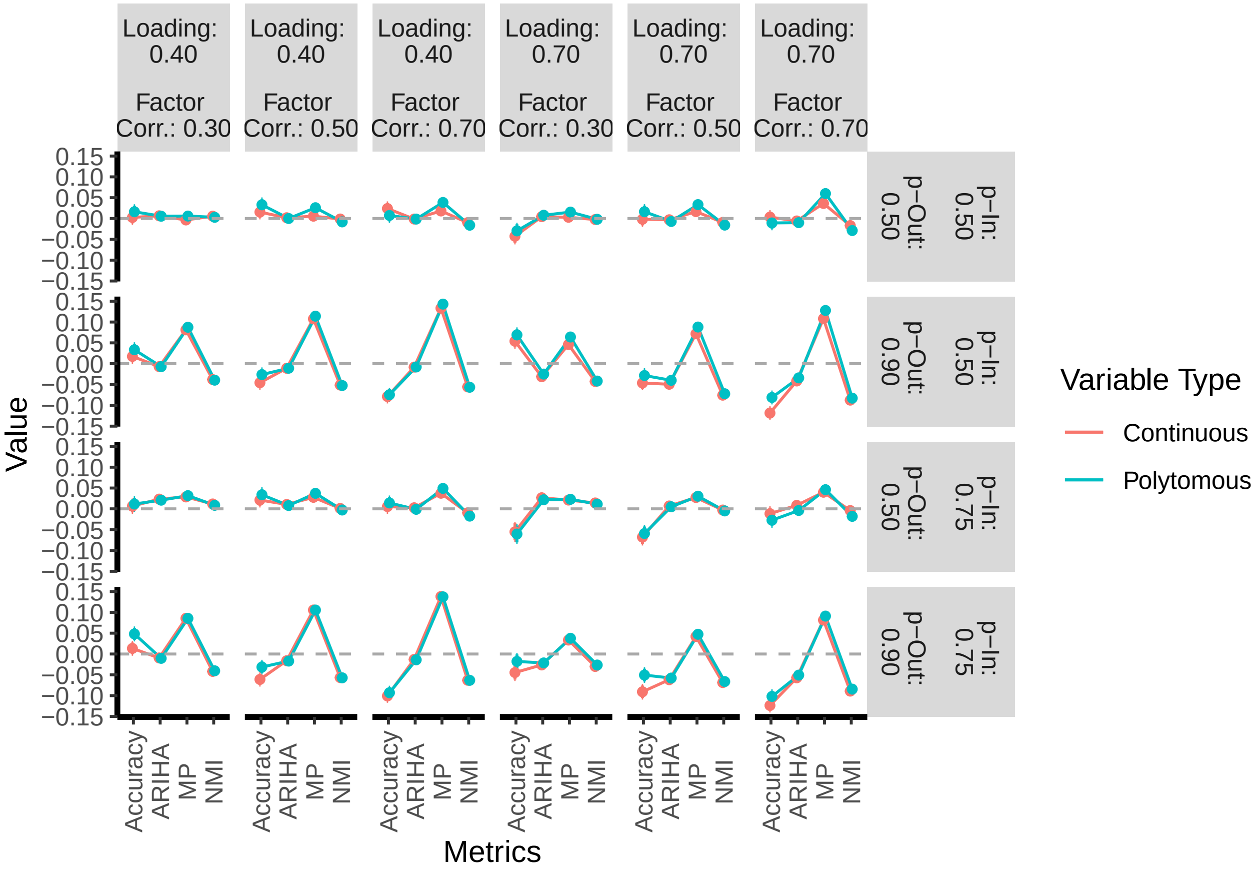

Figure 7. Interaction between Factor Loadings (Loading), Factor Correlations (Factor Corr.), p-In and p-Out by Variable Type

However, as interfactor correlation increases, this relationship is no longer consistent. Within p-Out at 0.50, as interfactor correlations increase there is an increase in the gain for accuracy but the other three metrics remain constant. Overall, the proposed method shows improvement in both accuracy and MP across levels of interfactor correlation, and presents the larger gains when there's more structural bias (i.e., larger interfactor correlations and lower p-Out, and lower interfactor correlations and high p-Out).

Finally, 10 is split by factor loadings where we see a similar relationship within p-Out at 0.90 where this is a more notable increase in accuracy and MP. When p-Out is lower (0.50) as factor loadings increase, there is a slight gain in each metric. Interestingly, across all three plots in Figures 8, 9, and 10 we see little to no improvement in NMI and $ARIH_{HA}$ and a majority of the improvement is notable in Accuracy and MP. Again, as happened with the

Table 1
Effect Size by Effect Tested for 8 Polytomous Variables per Factor

	Accuracy	ARIHA	MP	NMI
Network	0.000	0.001	**0.011**	0.006
CORF	0.001	**0.014**	**0.015**	**0.029**
P.OUT	0.000	**0.041**	**0.064**	**0.100**
CORF:P.OUT	**0.017**	0.002	0.001	0.005

p-Out link probability and interfactor correlation pairing, the larger the structural bias, the bigger the gain in accuracy and majority placement for the p-In and factor loadings pairing.

Empirical Example

To demonstrate the use of our approach to tune the number of steps used in the Walktrap algorithm used in exploratory graph analysis, we apply this method to the Developmental Coordination Disorder Questionnaire (DCDQ: Schoemaker et al., 2006). Data was provided to us through the Simons Foundation Powering Autism Research for Knowledge (SPARK) of the Simons Foundation Autism Research Initiative (SFARI), a large research initiative which has collected data from over 50,000 individuals with autism and their families (Feliciano et al., 2018). The DCDQ is a questionnaire given to parents of children (aged 5 to 15) to assess Developmental Coordination Disorder (DCD) commonly seen in individuals with autism spectrum disorders. DCD manifests as subtle motor skill impairment which affect things such as handwriting, clumsiness, energy levels, and athletic ability.

A grid search was conducted across values of t using EGA. Relationships between variables were estimated using partial correlation and the network was estimated using glasso. Figure 11 shows the $TEFI$ values across each level of t. When using the default $t = 4$, $TEFI = -7.64$. The lowest value of $TEFI$ (-9.72) occurs when $t = 9$ These results indicate that $t = 9$ provides the optimal model for this dataset. Figure 12 shows the difference in estimated community structure across $t = 4$ and $t = 9$.

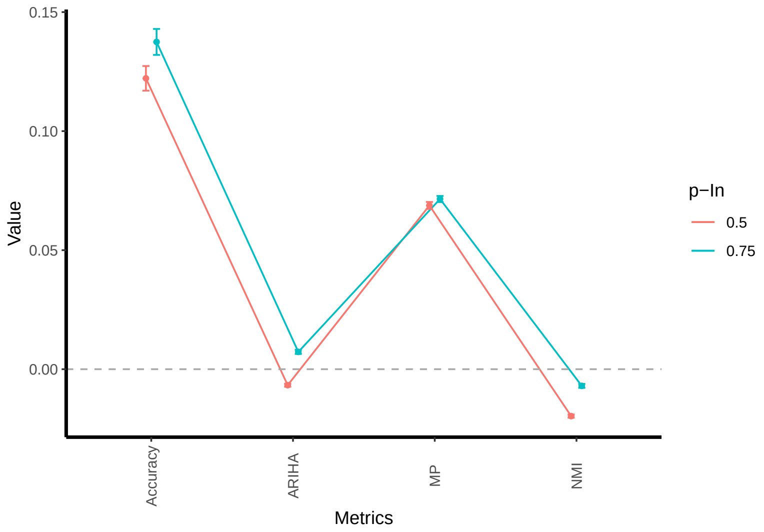

Figure 8. p-In Effect for 8 Polytomous Variables per Factor

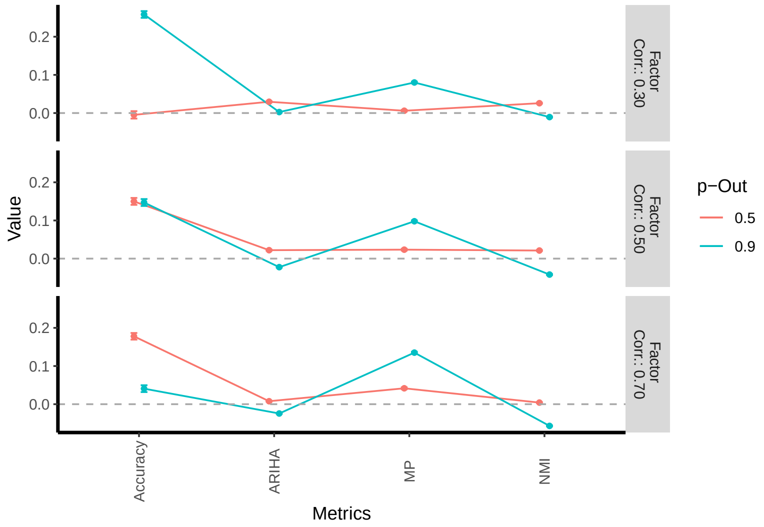

Figure 9. Interaction between p-Out and Factor Correlations (Factor Corr.) for 8 Polytomous Variables per Factor

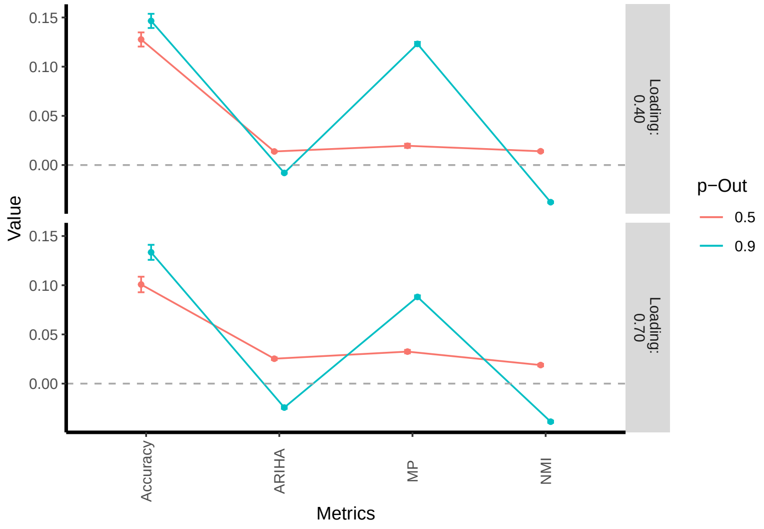

Figure 10. Interaction between p-Out and Factor Loadings (Loading) for 8 Polytomous Variables per Factor

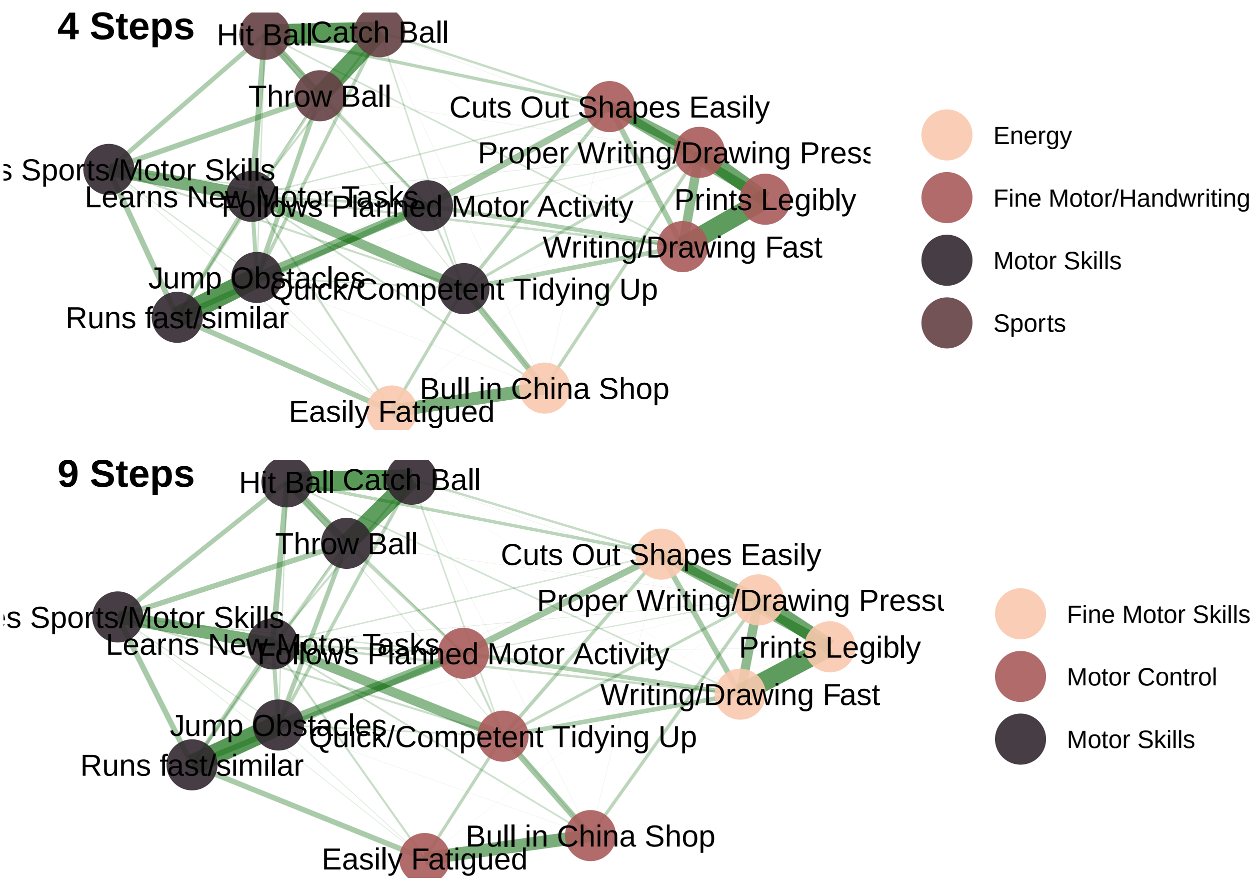

Figure 12. DCDQ Graph Estimations

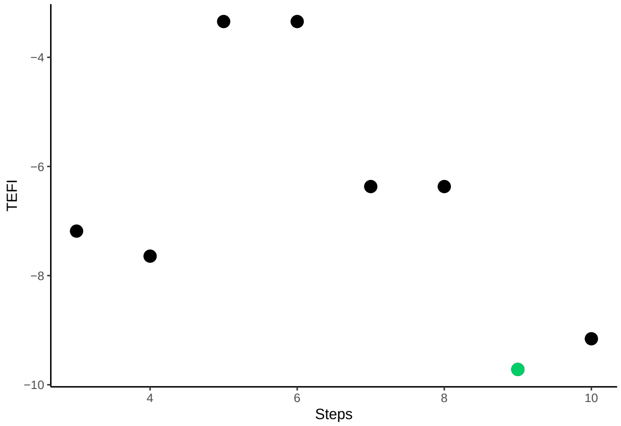

Figure 11. DCDQ TEFI Across Values of t

Table 2 shows the items of the DCDQ along with which scales they loaded onto from the original scale validation compared to the dimensions identified by EGA with an optimized t value. The $TEFI$ value obtained by the optimal EGA model (-9.72) is lower than the $TEFI$ value obtained by the original factor structure (-9.25). Both analyses revealed three very similar dimensions. However, the dimensionality uncovered by EGA using the proposed method moved items into slightly different positions within the community structure which in turn adjusts the interpretation of these communities.

Table 2

Comparing DCDQ Dimensionality Assessments

Item	Original Factor Analysis: TEFI = -9.25	Optimal EGA: TEFI = -9.72
Throws ball in a controlled and accurate fashion.	1. Control During Movement	1. Motor Skills
Catches a small ball from a distance.	1. Control During Movement	1. Motor Skills
Hits an approaching ball or birdie with a bat or racquet accurately.	1. Control During Movement	1. Motor Skills
Jumps easily over obstacles found in garden or play environment.	1. Control During Movement	1. Motor Skills
Runs as fast and in a similar way to other children of the same gender and age.	1. Control During Movement	1. Motor Skills
Is interested in and likes participating in sports or active games requiring good motor skills.	3. General Coordination	1. Motor Skills
Learns new motor tasks easily and does not require more practice or time than other children to achieve the same level of skill.	3. General Coordination	1. Motor Skills

Table 2

Comparing DCDQ Dimensionality Assessments (continued)

Item	Original Factor Analysis: TEFI = -9.25	Optimal EGA: TEFI = -9.72
Cuts out pictures and shapes accurately and easily.	2. Fine Motor/Handwriting	2. Fine Motor Skills
Printing, writing, or drawing is fast enough to keep up with the rest of the children.	2. Fine Motor/Handwriting	2. Fine Motor Skills
Printing or writing of letters, numbers, and words is legible, precise, or accurate.	2. Fine Motor/Handwriting	2. Fine Motor Skills
Uses appropriate effort or tension when printing, writing, or drawing.	2. Fine Motor/Handwriting	2. Fine Motor Skills
Can follow their plan of motor activity and organize their body to effectively complete the task.	1. Control During Movement	3. Motor Control
Child is quick and competent in tidying up, putting on shoes, dressing, etc.	3. General Coordination	3. Motor Control
Could be described as a ’bull in a china shop’ (appears clumsy, might break fragile things in a small room)	3. General Coordination	3. Motor Control

Table 2
Comparing DCDQ Dimensionality Assessments (continued)

Item	Original Factor Analysis: TEFI = -9.25	Optimal EGA: TEFI = -9.72
Fatiues easily, appears to slouch and ’fall out’ of the chair if required to sit for long periods.	3. General Coordination	3. Motor Control

All items related to fine motor skills and writing were identified in both analyses as a complete factor. Items related to sports and the enjoyment/proficiency of motor skills were assigned to the same community by EGA. In the original factor structure, whether or not the child enjoyed sports or enjoyed learning new motor skills did not load on to the same factor as items related to their abilities (e.g., throwing, catching, or hitting a ball). The item relating to whether or not the child is interested in participating in sports loaded onto a factor labeled “general coordination” with other items related to clumsiness, ability to clean up, and energy level. From a face valid standpoint, while these items are related generally to motor skills and coordination, it does not seem that they account for a similar type of variance in the overall construct. Rather the third community identified by EGA in this analysis appears to be more cohesive containing items related to ability to plan and accurately execute a task, ability to complete a task such as tidying up, and levels of clumsiness and fatigue.

Discussion

The Walktrap algorithm is a widely used community detection algorithm within network psychometrics particularly for estimating latent factors. However, the Walktrap algorithm contains a hyperparameter (t) the properties of which have not been fully researched. The present study tested a grid search approach for tuning t, identifying the optimal model with $TEFI$. Using synthetic data following data structures commonly found in psychological research, the benefits in model accuracy using this approach were investigated.

Data was simulated by multiplying a matrix of variables following a common factor model by a unweighted, undirected network to add a structural bias to the sample correlation matrix. 500 datasets were simulated across 1536 varied data structures, similar to the wide variety of structures found in substantive psychometric research. EGA was implemented varying the number of steps used by the Walktrap algorithm from 3 to 10. The optimal model with the lowest $TEFI$ value was identified and compared to the model at $t = 4$. Using overall accuracy, MP, ARIHA, and NMI as measures of partition accuracy, an analysis was conducted to identify whether or not and which kind of data structures benefit from varying t.

The results indicate that especially as sampling error is introduced into data, varying the number of steps within the Walktrap algorithm is beneficial. Importantly, it was demonstrated that the proposed method functioned similarly for both continuous and polytomous data. In line with previous dimensionality assessment research, the proposed method was particularly effective with a higher number of variables per factor (Garrido, Abad, & Ponsoda, 2011) as well as work in factor analysis positing that the increase in indicators also increases model error (MacCallum, Widaman, Preacher, & Hong, 2001).

As a higher probability of spurious intercommunity connections is introduced, the proposed method showed improvement over the traditional method both in estimating the correct number of communities but also the probability that nodes will be placed with other nodes from their true community. Spurious intercommunity connections not only interact with interfactor correlations, but also factor loadings. As the probability of spurious intercommunity connections increases, the proposed method provides improved model estimation. These findings are also in line with prior research indicating that higher interfactor correlations and lower loadings present particular challenges in accurate dimensionality assessment (Garrido et al., 2011; Garrido, Abad, & Ponsoda, 2013; Lubbe, 2019). Finally, when using Louis-Guttman Image Structural analysis as opposed to traditional correlations, there was a greater improvement in MP for polytomous data. The greatest metric improvement provided by the proposed method was seen in Accuracy and MP. These both relate directly to important aspects of how dimensionality assessment influences substantive research.

When substantive researchers validate a new measure, dimensionality assessment is often one of the first steps taken. As scales are broken down into further subscales, the facets of the larger latent structure being measured become clearer. In terms of structural validity, researchers and clinicians rely on the theory that scores and their variation directly relate to the structure of the scale and its subscales (Borsboom, Mellenbergh, & Van Heerden, 2004; Steger, 2006). As such, it is vital that any method applied to assess the relationship among items and the dimensionality of a scale is optimized to estimate the correct number of dimensions as well as place items together that assess the same dimension.

Flores-Kanter, Garrido, Moretti, and Medrano (2021) provide a great example of this using the Positive and Negative Affective Scale (PANAS; Watson, Clark, & Tellegen, 1988) where they review the multitude of studies evaluating its structure and discuss the implication of inconsistent structures estimated using traditional factor analytic techniques. In its validation, the PANAS was first identified as a three factor model: Positive Affect, Afraid, and Upset. These two negative affect scales (Afraid and Upset) represent orthogonal structures that many studies lump together as unidimensional. Flores-Kanter et al. (2021) evaluate the PANAS with EGA to assess the dimensionality (and stability thereof) and reveal a structure almost identical to the original three factor model. Similarly, the empirical example outlined in the current paper demonstrates the differences in scale interpretation based on the estimated factor structure. For both scales, the application of advanced and optimized methodology provided a clearer and more interpretable structure than traditional methods.

For the current study, it should be noted that in the data generation method, we introduced structural bias to better imitate empirical datasets. Therefore, error is introduced into model results not just from the model estimation process, but also from simulated sampling error. As a result, the magnitude of η_p^2 was impacted and no large effect sizes were found. Nonetheless, small and medium effect sizes revealed interesting relationships.

While the current paper reports rigorous testing of the proposed method in over 1500 combinations of common data structures, there are still several conditions not manipulated. For example, the proposed method was only tested for a four factor structure and also did not investigate how the method performs in very large networks (e.g., over 100 nodes). Additionally, the data generation method did not incorporate population error as it is traditionally implemented in psychometric literatur (Montoya & Edwards, 2020).

Future expansion on the current research should investigate the same application with different fit indices beyond $TEFI$ (e.g., AIC and BIC). While the current method has been shown to work well in cross-sectional factor designs, additional research should be conducted expanding into dynamical systems. Finally, particularly within polytomous data, additional work should be conducted investigating how this method functions when variables are skewed.

Conclusion

Proper latent trait modeling is the crux of almost every portion of psychological research. Many theories and statistical methods have been developed to assess the dimensionality of latent variables, each with its own strengths and weaknesses. EGA has been shown to perform well (and above and beyond similar methods) across data structures commonly found in psychological research. We aim to improve EGA even further by introducing a new technique when using the Walktrap algorithm for community detection. Instead of following standard guidelines statically setting t, a grid search can be conducted to optimize select the optimal value of t. In order to select the optimal value of t, we recommend using $TEFI$ due to its advantages in detecting the correct dimensionality solution.

The proposed method was tested across a variety of data structures commonly found in psychological research (e.g., highly correlated factors with spurious connections collected with polytomous response data). It was found to provide improvement above and beyond traditional methodology for the Walktrap algorithm in identifying the dimensionality and specific item-community organization. Additionally, the method was applied to a substantive dataset and shown to provide a clearer and more cohesive structure than both the original factor structure and the dimensionality structure identified by the traditional Walktrap application.

References

Bollmann, S., Heene, M., Küchenhoff, H., & Bühner, M. (2015). *What can the real world do for simulation studies? A comparison of exploratory methods.* LMU. Retrieved from Department%20of%20Statistics,%20University%20of%20Munich:%20https://epub.ub.uni-muenchen.de/24518/

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061.

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., … Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PloS One*, *8*(4), e60188.

Chang, F., Qiu, W., Zamar, R. H., Lazarus, R., Wang, X., & others. (2010). Clues: An r package for nonparametric clustering based on local shrinking. *Journal of Statistical Software*, *33*(4), 1–16.

Chen, J., & Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, 555–574.

Christensen, A. P., Cotter, K. N., & Silvia, P. J. (2019a). Reopening openness to experience: A network analysis of four openness to experience inventories. *Journal of Personality Assessment*, *101*(6), 574–588.

Christensen, A. P., Garrido, L. E., & Golino, H. (2021). Comparing community detection algorithms in psychological data: A monte carlo simulation. *PsyArXiv*. https://doi.org/10.31234/osf.io/hz89e

Christensen, A. P., Golino, H., & Silvia, P. J. (2019b). A psychometric network perspective on the measurement and assessment of personality traits. *Preprint*.

Christensen, A. P., Gross, G. M., Golino, H. F., Silvia, P. J., & Kwapil, T. R. (2019c). Exploratory graph analysis of the multidimensional schizotypy scale. *Schizophrenia Research*, *206*, 43–51.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Costantini, G., Richetin, J., Preti, E., Casini, E., Epskamp, S., & Perugini, M. (2019). Stability and variability of personality networks. A tutorial on recent developments in network psychometrics. *Personality and Individual Differences*, *136*, 68–78.

Danon, L., Diaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, *2005*(09), P09008.

Dijkstra, J. K., Cillessen, A. H., & Borch, C. (2013). Popularity and adolescent friendship networks: Selection and influence dynamics. *Developmental Psychology*, *49*(7), 1242.

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212.

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617.

Feliciano, P., Daniels, A. M., Snyder, L. G., Beaumont, A., Camba, A., Esler, A., … others. (2018). SPARK: A us cohort of 50,000 families to accelerate autism research. *Neuron*, *97*(3), 488–493.

Flores-Kanter, P. E., Garrido, L. E., Moretti, L. S., & Medrano, L. A. (2021). A modern network approach to revisiting the positive and negative affective schedule (panas) construct validity.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3-5), 75–174.

Foygel, R., & Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. *arXiv Preprint arXiv:1011.6640*.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of velicer's minimum average partial factor retention method with categorical variables. *Educational and Psychological Measurement*, *71*(3), 551–570.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at horn's parallel analysis with ordinal variables. *Psychological Methods*, *18*(4), 454.

Gates, K. M., Fisher, Z. F., Arizmendi, C., Henry, T. R., Duffy, K. A., & Mucha, P. J. (2019). Assessing the robustness of cluster solutions obtained from sparse count matrices. *Psychological Methods*.

Gates, K. M., Henry, T., Steinley, D., & Fair, D. A. (2016). A monte carlo evaluation of weighted community detection algorithms. *Frontiers in Neuroinformatics*, *10*, 45.

Gates, K. M., & Molenaar, P. C. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, *63*(1), 310–319.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826.

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS One*, *12*(6).

Golino, H., Moulder, R., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., … Boker, S. M. (2020). Entropy fit indices: New fit measures for assessing the structure and dimensionality of multiple latent variables. *Multivariate Behavioral Research*, 1–29.

Golino, H., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., … Martinez-Molina, A. (2020). Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods*.

Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika*, *18*(4), 277–296.

Harman, H. H. (1976). *Modern factor analysis*. University of Chicago press.

Harris, C. W. (1962). Some rao-guttman relationships. *Psychometrika*, *27*(3), 247–263.

Hoffman, M., Steinley, D., Gates, K. M., Prinstein, M. J., & Brusco, M. J. (2018). Detecting clusters/communities in social networks. *Multivariate Behavioral Research*, *53*(1), 57–73.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

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

Lubbe, D. (2019). Parallel analysis with categorical variables: Impact of category probability proportions on dimensionality assessment accuracy. *Psychological Methods*, *24*(3), 339.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, *36*(4), 611–637.

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., Bork, R. van, Waldorp, L., … Maris, G. (2018). An introduction to network psychometrics: Relating ising network models to item response theory models. *Multivariate Behavioral Research*, *53*(1), 15–35.

Massara, G. P., Di Matteo, T., & Aste, T. (2016). Network filtering for big data: Triangulated maximally filtered graph. *Journal of Complex Networks*, *5*(2), 161–178.

McNally, R. J. (2016). Can network analysis transform psychopathology? *Behaviour Research and Therapy*, *86*, 95–104.

Montoya, A. K., & Edwards, M. C. (2020). The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement*, 0013164420942899.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582.

Orman, G. K., & Labatut, V. (2009). A comparison of community detection algorithms on artificial networks. In *International conference on discovery science* (pp. 242–256). Springer.

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, *10*(2), 191–218.

Schoemaker, M. M., Flapper, B., Verheij, N. P., Wilson, B. N., Reinders-Messelink, H. A., & Kloet, A. de. (2006). Evaluation of the developmental coordination disorder questionnaire as a screening instrument. *Developmental Medicine and Child Neurology*, *48*(8), 668–673.

Steger, M. F. (2006). An illustration of issues in factor extraction and identification of dimensionality in psychological assessment data. *Journal of Personality Assessment*, *86*(3), 263–272.

Steinley, D. (2004). Properties of the hubert-arable adjusted rand index. *Psychological Methods*, *9*(3), 386.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.

Watanabe, H. (2001). Clustering as average entropy minimization and its application to structure analysis of complex systems. In *"2001 ieee international conference on systems, man and cybernetics": "E-systems and e-man for cybernetics in cyberspace"* (Vol. 4, pp. 2408–2414). https://doi.org/10.1109/ICSMC.2001.972918

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, *4*(1), 66–82.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, *54*(6), 1063.

Williams, D. R., Rhemtulla, M., Wysocki, A. C., & Rast, P. (2019). On nonregularized estimation of psychological networks. *Multivariate Behavioral Research*, *54*(5), 719–750.

Yang, Z., Algesheimer, R., & Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, *6*, 30750.