Structural Equation Modeling of Social Networks: Specification, Estimation, and Application

Haiyan Liu
University of California, Merced

Ick Hoon Jin
University of Notre Dame

Zhiyong Zhang
University of Notre Dame

2018

Citation: Liu, H., Jin, I. K., & Zhang, Z.(2018). Structural Equation Modeling of Social Networks: Specification, Estimation, and Application. Multivariate Behavioral Research

Author Note

Haiyan Liu, Psychological Science, University of California, Merced; Ick Hoon Jin, Department of Applied and Computational Mathematics and Statistics (ACMS), University of Notre Dame; Zhiyong Zhang, Department of Psychology, University of Notre Dame.

This study was partially supported by the Institute for Scholarship in the Liberal Arts, College of Arts and Letters, University of Notre Dame. Zhang was supported by grants from Institute of Education Sciences (R305D140037) and National Science Foundation (1461355).

Correspondence concerning this article should be addressed to Haiyan Liu, Psychological Science, University of California, Merced, 5200, N. Lake Road, Merced, CA 95343. Email: hliu62@ucmerced.edu

Abstract

Psychologists are interested in whether friends and couples share similar personalities or not. However, no statistical models are readily available to test the association between personalities and social relations in the literature. In this study, we develop a statistical model for analyzing social network data with the latent personality traits as covariates. Because the model contains a measurement model for the latent traits and a structural model for the relationship between the network and latent traits, we discuss it under the general framework of structural equation modeling (SEM). In our model, the structural relation between the latent variable(s) and the outcome variable is no longer linear or generalized linear. To obtain model parameter estimates, we propose to use a two-stage maximum likelihood (ML) procedure. This modeling framework is evaluated through a simulation study under representative conditions that would be found in social network data. Its usefulness is then demonstrated through an empirical application to a college friendship network.

Keywords: social network analysis, personality, latent space model, confirmatory factor model, nonlinear structural equation modeling, two-stage maximum likelihood approach

Structural Equation Modeling of Social Networks: Specification, Estimation, and Application

Introduction

Social network analysis is a popular interdisciplinary research topic in statistics, sociology, political science, and recently psychology (e.g., Hoff et al., 2002; Saul & Filkov, 2007; Schaefer et al., 2013; Wasserman & Faust, 1994). It is used to assess social structures through connections between entities/nodes/subjects in a bounded network (e.g., Otte & Rousseau, 2002) and has substantive applications in many fields. For instance, economists employ network techniques to address how the social, economic, and technological worlds are connected (Easley & Kleinberg, 2010). Epidemiologists use them to analyze the emergence of infectious diseases, such as the severe acute respiratory syndrome (SARS; Berger et al., 2004). Political scientists focus on how social networks influence individual's political preference (Lazer et al., 2010; Ryan, 2011). Sociologists are interested in the factors that explain the patterns in a social network. For example, Deana et al. (2017) studied the factors leading to friendship dissolution within a social network. For psychologists and behavioral scientists, a common goal of a social network analysis is to conduct behavioral intervention using the information from a network (Maya-Jariego & Holgado, 2015).

Different methods have been used for analyzing social network data (Hoff et al., 2002; Hunter & Handcock, 2006; Deana et al., 2017). The exponential random graph model (ERGM), for instance, is one well-known models for static networks (e.g., Anderson et al., 1999; Frank & Strauss, 1986), which treats the entire social network as a random variable. It intends to explain the probability of networks using their local features such as triangle counts and node degrees (Robins et al., 2007; Snijders et al., 2006; Wasserman & Pattison, 1996). The stochastic actor-oriented model (SAOM) was proposed for dynamic network analysis (Snijders et al., 2006). Like an ERGM, it only uses the information of the internal network structures to model the evolution of a network.

Latent space modeling is another widely used technique for a static network analysis (Hoff et al., 2002). It investigates pairwise edges in a network by including the latent Euclidean space for actors in logistic regression models. The existence of edges are independent conditional on the latent positions. The probability of an edge is a function of both manifest covariates (such as individual's characteristics) and the distance between actors in the latent Euclidean space. In the latent space modeling, the probability of a tie is constrained to decrease as the latent distance increases, which is based on the principle that similarity drives the formation of a network. Handcock et al. (2007) extended the model of Hoff et al. (2002) to account for clusters in a network by allowing for a mixture of normal distributions on latent positions. Sewell & Chen

(2015) expanded the model to longitudinal networks by linking the dynamics of networks with the trajectories of latent positions in a Euclidean space. Overall, a latent space model explains the dependence between actors using the closeness of latent positions and covariates of interests.

In recent years, the network analysis technique is also used in psychometrics to study the correlations/covariance among variables (Borsboom & Cramer, 2013; Schmittmann et al., 2013), which is termed as "network psychometrics" in the literature. In network psychometrics, each variable is treated as a node and the pairwise interactions among variables are represented by edges (Epskamp et al., 2017). An edge in a variable network is the relation between two nodes after controlling all other nodes and the edge weight stands for the degree of such relations (e.g., Cramer et al., 2010). For instance, an edge could be the partial correlation coefficients between two variables.

An edge in a social network represents certain social relations such as friendships, common hobbies, and marriage. These social relations influence individuals' social development and adjustment (e.g., Veenstra et al., 2013). Understanding their formation, therefore, can be particularly important to researchers (Deana et al., 2017). A broad range of actor attributes are relevant to his/her relationship with others (McPherson et al., 2001). Demographic variables, such as race, age, gender, and education level, may affect the likelihood of two actors being friends or romantic partners (McPherson et al., 2001; Rushton & Bons, 2005). In addition, different types of social spaces have been discussed to explain the existence of clusters in social networks (Faust, 1988; McFarland & Brown, 1973).

In this study, we focus on a personality factor space formed by individuals' personality factor scores. Personality has been shown to affect social relations (e.g., Asendorpf & Wilpers, 1998; Harris & Vazire, 2016). Some studies claimed weak personality similarity between partners and friends (Watson et al., 2014; Rushton & Bons, 2005). Some others found moderate similarity between romantic partners in some of the Big Five factors of personality such as "Openness" and "Conscientiousness" (McCrae et al., 2008; Watson et al., 2000). Asendorpf & Wilpers (1998) showed that the Big Five personality factors "Extraversion", "Agreeableness", and "Conscientiousness" predict the number of close relations with peers. A recent study by Youyou et al. (2017) revealed medium to large personality similarities between friends and couples. Although there are many studies indicating personality similarity between connected dyads, techniques for statistical inference are lacking for the extent to which they are related to each other. In addition, previous studies usually focused on dyads with social relations, but information on dyads without social relations is lacking. Social network data, however, contain information on both types of dyads, which allow us to conduct better hypothesis testing.

To link social network analysis to psychological research, we developed a new model under

the framework of structural equation modeling. The model contains both a measurement model and a structural model. A confirmatory factor model is used to estimate the latent personality traits (Cattell, 1952), for which data on multiple indicators are collected using the mini-IPIP scales (Donnellan et al., 2006). A logistic regression model is used to predict the probability of a connection between pairs of nodes. We consider the new model as an extension of structural equation modeling. As in most statistical models, conditional independence is assumed for our model. Specifically, the edges in a social network are conditionally independent given levels of covariates and latent personality traits.

Since outcomes of the logistic regression are from a network, the predictors should also be in the form of networks to match their dimensions. To have network predictors, we need to transform the covariates to nodal covariates measuring characteristics of paired actors. In the literature, many nodal covariates are computed based on the homophily principle that drives the formation of a network (McPherson et al., 2001). Hunter et al. (2008) introduced several nodal covariates for a uni-dimensional manifest variable. For example, for a categorical covariate, one may define a nodal covariate based on whether two nodal actors belong to the same category or not. For a continuous covariate, one may make use of a covariate based on objects of analysis. In our study, we define a new nodal covariate based on (dis)similarity of latent personality traits by calculating the Mahalanobis distance (Mahalanobis, 1936).

Because we compute one integrated latent personality score for each pair of actors, the structural relation between the latent variable(s) and the outcome variable (a network) is no longer linear or generalized linear. Therefore, we cannot directly apply the widely used estimation methods, such as the maximum likelihood (ML) and generalized least square estimation (GLS) methods, for structural equation modeling (SEM) to the newly developed model. To estimate the model, we propose a two-stage ML estimation procedure. At the first stage, we fit a confirmatory factor model to extract Thurstone-Thomson "regression" factor scores. At the second stage, we fit the structural model using the extracted factor scores at the first stage and obtain the regression coefficient estimates.

The remainder of this article is organized as follows. First, we will briefly describe social networks and latent space modeling. The aim is to introduce terminologies and notations specific to the field of social network analysis. Then, we will present a latent space model with a factor structure for social network data and describe a two-stage ML estimation procedure. After that, a simulation study is conducted to evaluate the performance and an empirical example is used to demonstrate the application of this model. Finally, we summarize our study and discuss its limitations and future directions.

A Brief Introduction to Social Network Analysis

In this section, we will describe some basic terminologies and notations used in social network analysis (SNA). We will also introduce the latent space model used for network analysis to provide readers some fundamental knowledge of SNA.

Social networks

In the literature, there are different types of social networks based on the substantive meaning of ties including a smoking social network (e.g., Schaefer et al., 2013), a friendship social network (e.g., Miething et al., 2016), and a social media network (e.g., Facebook or Twitter). In the present study, we will discuss our model in the context of a friendship network. Specifically, a tie/link/connection in a friendship network represents that the two nodal actors are friends. An un-directional tie between actors i and j means that they are mutually friends. The directional tie from $i \to j$ indicates that actor i nominates actor j as his/her friend and actor j does not choose actor i as his/her friend. In the remainder of this article, we will consider only the un-directional ties. Extension of our approach to a network with directional ties is straightforward.

In an un-directional friendship network with N actors, subjects, or participants, let $\mathbf{Y} = (y_{ij}), i = 1, \ldots\ldots, N, j = 1, \ldots\ldots, N$ be an adjacency matrix. In a dichotomous social network, $y_{ij} = 1$ if actors i and j are connected, and $y_{ij} = 0$ if they are not, for all $i \neq j$. Since we do not allow self-connection, we let $y_{ii} = 0$, for all $i = 1, 2, \cdots, N$. As a consequence, $\mathbf{Y}$ is a symmetric matrix with elements being either 1 or 0. For each actor, we define the *degree* as the number of edges that end up on that actor. An actors with larger degree typically is more important for a social network because the interaction among actors could pass by him/her. Figure 1 is the adjacency matrix and the graph representation of a network with 10 actors. In the plot, a gray line represents that the two ending nodes are connected/friends, which is indicated by "1" in the adjacency matrix. The size of a node corresponds to its degree. The bigger a node, the larger its degree. From the plot, we notice that the node with ID 9 has the largest degree, which is 6 based on the adjacency matrix.

One purpose of a social network analysis is to predict how likely two actors are to be connected based on some external information, i.e., finding variables related to the network. For a friendship network, demographic variables, such as gender, education, age, and race, might influence the existence of friendship between actors. For example, if two people are of the same gender, it may be more likely for them to be friends than if they are not.

Nodal covariates

In a social network with N actors, a covariate $\mathbf{x} = (x_1, x_2, \cdots, x_N)'$ is a N by 1 vector of an attribute observed for each of the nodes in the network. Because we focus on the dyadic ties in our present study, we need to define a nodal covariate by integrating the feature of a pair of actors through a suitably chosen function $h(x_i, x_j)$ (Hunter et al., 2008). There are different types of nodal covariates based on both the property of a variable (continuous vs. categorical) and the specific features of interests.

For a continuous covariate $\mathbf{x}$, the *main factor effect* is the average of the covariate of two actors i and j,

$$h_{ij} = h(x_i, x_j) = \frac{x_i + x_j}{2}, \tag{1}$$

which is used when one wants to investigate the effect of average attributes of two actors on their friendship. For instance, it could be the average depression levels of two actors.

Another widely used nodal covariate for a continuous covariate is termed as the *absolute difference factor effect*,

$$h_{ij} = h(x_i, x_j) = |x_i - x_j|, \tag{2}$$

which describes the similarity or dissimilarity of two actors through the Euclidean distance. Because the feature of homophily is common in relational data (McPherson et al., 2001), it is very natural to use this type of nodal covariate in a network analysis.

For a categorical covariate, the *nodal factor effect* is the effect when both actors forming a tie are in a particular level of a categorical factor,

$$h_{ij} = h(x_i, x_j) = \begin{cases} 2 & \text{if both nodes } i \text{ and } j \text{ have the specific factor level} \\ 1 & \text{if exactly one of } i, j \text{ has the specified factor level} \\ 0 & \text{if neither } i \text{ and } j \text{ has the specified factor level.} \end{cases} \tag{3}$$

For example, a nodal factor effect may be a female-female effect or a treatment-treatment effect.

Other types of nodal covariates for categorical data are based are the *homophily* of covariates. For *uniform* homophily effect, the statistic is as follows

$$h_{ij} = h(x_i, x_j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ have the same factor level} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

which is used when the effects of two actors belonging to the same group are the same for all

groups. In contrast, for the *differential* homophily, the statistic is defined for specific levels such as

$$h_{ij} = h(x_i, x_j) = \begin{cases} 1 & \text{if both } i \text{ and } j \text{ have the specific factor level} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

For instance, if both actor i and j are males, or in treatment groups, then h_{ij} is 1, otherwise, it is 0.

For multi-dimensional latent traits, there are no nodal covariates available yet. Thus, we would like to introduce the *latent factor similarity effect*, which measures the overall high-dimensional nodal similarity using the Mahalanobis distance (Mahalanobis, 1936),

$$d_{ij} = h(\boldsymbol{\eta}_i, \boldsymbol{\eta}_j) = \sqrt{(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)' \boldsymbol{\Phi}^{-1} (\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)}, \tag{6}$$

where $\boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_j$ are the vectors of factor scores of actors i and j, and $\boldsymbol{\Phi}$ is the population covariance matrix of latent factors. The Mahalanobis distance is the standardized distance of two vectors.

Latent space modeling

The latent space model was introduced by Hoff et al. (2002) to model the probability of dyadic ties in a dichotomous network. Under the latent space model, each individual j is projected into an unknown position $\mathbf{w}_j = (w_{j,1}, w_{j,2}, \cdots, w_{j,D})'$ in the Euclidean space $\mathbb{R}^D$. An intuitive and widely used distance measure for Euclidean space is

$$d_{ij} = |\mathbf{w}_i - \mathbf{w}_j| = \sqrt{(\mathbf{w}_i - \mathbf{w}_j)'(\mathbf{w}_i - \mathbf{w}_j)} = \sqrt{\sum_{d=1}^{D} (\mathbf{w}_{i,k} - \mathbf{w}_{j,k})^2}$$

for any pair of actors i and j. Given a vector of nodal statistics of interests (for instance, those defined in Equation (1)-(5)), a latent space model has the following general form,

$$\begin{cases} y_{ij} & \sim \text{Bernoulli}(p_{ij}) \\ \text{logit}(p_{ij}) & = \alpha + \boldsymbol{\beta}' \mathbf{h}_{ij} + \gamma d_{ij} \\ d_{ij} & = |\mathbf{w}_i - \mathbf{w}_j| \end{cases} \tag{7}$$

where $\boldsymbol{\beta}$ and γ are the coefficients of the manifest nodal covariates and the latent distance, respectively. With P manifest covariates, the symbol $\mathbf{h}_{ij} = (h_{1,ij}, h_{2,ij}, \cdots, h_{P,ij})'$ represents the $P \times 1$ vector of manifest nodal covariates observed on dyads (i, j). When $\gamma > 0$, two actors that are far away from each other in the latent space are more likely to be friends; while with $\gamma < 0$,

the closer of two actors in the latent space, the higher probability for them to be friends while controlling other nodal covariates.

Because the Euclidean distance is not scaling-invariant by constants, there are infinitely many pairs of $(\gamma, d_{ij}, i, j = 1, \cdots, N)$ with the same quantity γd_{ij} and the model is thus not identifiable. To solve this problem, the latent distance and the log odds are often restricted to have the same scale. Furthermore, many social networks show the property of homophily, which states that a network formation is driven by the similarity in actor attributes. Hence, the coefficient of the latent distance γ is usually set to be -1 as in Hoff et al. (2002). The other unknowns to be estimated for the latent space model include α, β, and $\mathbf{w}_i$'s.

Factor Space as the Latent Space

The purpose of latent space modeling is to model the dependence between actors using a space of unobserved characteristics. However, the latent space itself is arbitrary and cannot be interpreted substantively. In the present study, we make the latent space interpretable by proposing to use the latent personality factor space (Harris & Vazire, 2016). The resulting model contains two components: a measurement model for the latent traits and a structural model for the relationship between a network and latent traits. More specifically, the measurement model is a confirmatory factor model with correlations among factors, which provides information on latent traits. The structural model is a logistic regression model with both manifest and latent nodal covariates.

In our social network analysis, three parts of the data are available: $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$. With N actors, the outcome variable $\mathbf{Y}$ is an N by N matrix identifying all the edges and actors in a network. Each column of $\mathbf{X}$ contains the data on one covariate. With P covariates, $\mathbf{X}$ is a N by P matrix. The part of data $\mathbf{Z}$ are scores of indicators of latent traits. With J indicators, $\mathbf{Z}$ is a N by J matrix and it provides information on the latent traits of interests.

Measurement model (factor model)

Let $\boldsymbol{\eta}_i = (\eta_{i1}, \cdots, \eta_{iD})'$ be a vector of D latent variables and $\mathbf{z}_i = (z_{i1}, \cdots, z_{iJ})'$ be a vector of J indicators. A confirmatory factor model can be expressed as (Cattell, 1952),

$$\begin{cases} \mathbf{z}_i & = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\eta}_i & \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Phi}) \\ \boldsymbol{\varepsilon}_i & \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Psi}), \end{cases} \tag{8}$$

where $\boldsymbol{\varepsilon}_i$ is a $J \times 1$ vector of unique factors associated with case i and it follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Psi}$. The factor loading matrix $\boldsymbol{\Lambda}$ is a $J \times D$ matrix. $\boldsymbol{\Phi}$ is the factor covariance matrix to be estimated. In this model, the unknowns include individual factor scores $\{\boldsymbol{\eta}_i\}_{i=1}^N$ and model parameters $\{\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Psi}\}$. To identify the model, restrictions are usually posted (e.g., Mulaik, 2009). For instance, as in a standard structural equation model, it is necessary to set either the first factor loading of each factor or the factor variances to be 1. In this article, we restrict the first factor loading to be 1 in both simulation and empirical studies. The purpose of the confirmatory factor analysis is to obtain both model parameter estimates and the predicted factor scores of each actor, denoted by $\hat{\boldsymbol{\eta}}_i$, which will be used in the estimation of the structural model.

Structural model

As in latent space models, the latent personality traits can be used to predict the network. Since the latent factors are not necessarily orthogonal to each other, we use the Mahalanobis distance (Equation (6)) as a nodal covariate,

$$\begin{cases} y_{ij} & \sim \text{Bernoulli}(p_{ij}) \\ \text{logit}(p_{ij}) & = \alpha + \boldsymbol{\beta}'\mathbf{h}_{ij} + \gamma\sqrt{(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)'\boldsymbol{\Phi}^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)} \end{cases} \tag{9}$$

where $\mathbf{h}_{ij}$ is the vector of manifest nodal covariates whose elements are defined by Equation (1) to Equation (5), and $\sqrt{(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)'\boldsymbol{\Phi}^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)}$ is Mahalanobis distance formed by factor scores. The association of the Mahalanobis distance and the existence of an edge between actors is represented by the parameter γ. More generally, we can construct multiple nodal covariates using the latent factors, denoted as $\boldsymbol{\Theta}_{ij}$. Then the model can be written in a general form as

$$\text{logit}(p_{ij}) = \alpha + \boldsymbol{\beta}'\mathbf{h}_{ij} + \boldsymbol{\gamma}'\boldsymbol{\Theta}_{ij}$$

where $\boldsymbol{\gamma}$ is a vector of coefficients. In addition, coefficients of the observed covariates are represented by $\boldsymbol{\beta}$, and the intercept is α. Because of the involvement of the factor model, we call our model in Equation (8) and (9) a latent space model with a factor structure.

It is interesting to note that the model is flexible to include various nodal covariates and latent nodal covariates according to the objects of analysis. As just discussed, $\boldsymbol{\Theta}_{ij}$ could be $\sqrt{(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)'\boldsymbol{\Phi}^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)}$, i.e., the Mahalanobis distance of the factor scores of actors i and j, when one intends to test the *overall* latent factor similarity effect. A latent nodal covariate could also be defined for each individual latent factor as in Equation (2). A latent nodal covariate could

also be $\frac{\boldsymbol{\eta}_i+\boldsymbol{\eta}_j}{2}$, if the main factor effect of a latent factor (Equation (1)) is of interest. Researchers are free to test any type of nodal effects by using the corresponding forms of nodal covariates.

Our model is different from a latent space model in several ways. In latent space modeling (Hoff et al., 2002), the latent nodal covariate d_{ij} has a very specific form (Euclidean distance) and the coefficient γ is fixed to be -1 to make the model identifiable. Both the direction and magnitude of the effect of the latent distance on the network is fixed. The coefficient γ is, however, freely estimated in our model, which is more flexible. In addition, the "latent space" in our model has its substantive meaning. It represents a space formed by latent traits such as the Big Five personality traits. Figure 2 is a graphical portray of our model with a factor structure. Note that the latent nodal covariates $\boldsymbol{\Theta}$ are defined by latent factors based on research questions.

Same as in latent space modeling, conditional independence is assumed in our newly developed model. Specifically, the existence of edges between actors are independent with each other for given levels of nodal covariates. The conditional independence assumption is theoretically reasonable, because the nodal covariates are actually networks with weighted edges. The dependence structure nested in the outcome network would be explained by that in the covariate networks.

Our model consists of both a measurement (i.e., CFA model) and a structural model (i.e., logistic model) as in a general structural equation model. However, it also has distinct features from a traditional SEM model. First, all the variables involved in the model are dyadic. Second, because a latent nodal covariate $\boldsymbol{\Theta}$ is not a linear function of a single latent variable, the structural model is neither a linear nor a generalized linear function of latent variables.

Model estimation

In SEM, the model implied covariance matrix $\Sigma(\boldsymbol{\theta})$, a matrix function of model parameters $\boldsymbol{\theta}$, is usually computed in order to estimate the model. And the model parameter estimates are obtained by minimizing the discrepancy between $\Sigma(\boldsymbol{\theta})$ and the sample covariance matrix $\mathbf{S}$.

Since the network is not linearly or generalized linearly related to the latent variable in our model, it is impossible to derive the model implied covariance matrix as is done for a traditional SEM model. As a consequence, the widely used ML and generalized least square (GLS) estimation methods are no longer directly applicable in estimating our model. Since there are latent variables involved in our model, an expectation–maximization (EM) algorithm is seemingly useful in estimating our model (Bilmes et al., 1998; Bock & Aitkin, 1981; Ghahramani et al., 1996). However, the nodal latent covariates in our model are either the sum or difference of two factor scores. Thus, we are unable to split them using an EM procedure. We, therefore, turn to a two-stage maximum likelihood (ML) procedure to estimate the model.

At the first stage, we fit a confirmatory factor analysis (CFA) model. The ML estimates of model parameters of the CFA model (Equation (8)) is obtained by minimizing the following discrepancy function as in most SEM studies (e.g., Jöreskog, 1967),

$$F_{ML}(\boldsymbol{\theta}) = \log|\Sigma(\boldsymbol{\theta})| + \mathrm{tr}(\mathbf{S}\Sigma^{-1}(\boldsymbol{\theta})) - \log|\mathbf{S}| - p \tag{10}$$

where $\boldsymbol{\theta}$ is the collection of the parameters of the CFA model including factor loading matrix $\boldsymbol{\Lambda}$, factor covariance matrix $\boldsymbol{\Phi}$, and residual variances $\boldsymbol{\Psi}$, $\Sigma(\boldsymbol{\theta})$ is the model implied covariance matrix based on a CFA model which is a matrix function of the model parameters, and $|\cdot|$ indicates the determinant of a matrix. In our current analysis, we consider only common factor models with no specific factors, and the first factor loading of each factor is restricted to be 1.

In this study, we are interested in how the similarities in individual attributes predict the existence of social relations. We thus also obtain the predicted factor scores using the confirmatory factor analysis. In the literature, there are many discussions on the issue of factor score indeterminacy (Guttman, 1955; Grice, 2001; Mulaik, 1976, 2009). Mathematically, there are infinitely many groups of $(\boldsymbol{\eta}, \boldsymbol{\varepsilon})$ satisfying the factor model in Equation (8), depending on which estimator is used. Among so many estimators, the Thurstone-Thomson "regression" (Thurstone, 1935) estimator and the Bartlett estimator (Bartlett, 1937) are the two most commonly used predictors of factor scores. The Thurstone-Thomson "regression" predictor is obtained by minimizing the following function,

$$\mathrm{Loss}(\boldsymbol{\eta}) = \mathrm{tr}[\mathrm{E}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})'(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})] \tag{11}$$

and the resulted "regression" estimator is

$$\hat{\boldsymbol{\eta}} = \boldsymbol{\Phi}\boldsymbol{\Lambda}'\Sigma^{-1}\mathbf{Z} \tag{12}$$

where $\mathbf{Z}$ is the observed data. And the Bartlett factor score estimator (Bartlett, 1937) obtained by minimizing the loss function for given factor loading matrix $\boldsymbol{\Lambda}$ and residual variances $\boldsymbol{\Psi}$,

$$\mathrm{Loss}(\boldsymbol{\eta}) = (\mathbf{Z} - \boldsymbol{\Lambda}\boldsymbol{\eta})'\boldsymbol{\Psi}^{-1}(\mathbf{Z} - \boldsymbol{\Lambda}\boldsymbol{\eta}) \tag{13}$$

and the resulted weighted least square (WLS) estimator is

$$\hat{\boldsymbol{\eta}} = (\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\mathbf{Z} \tag{14}$$

where $\mathbf{Z}$ is the observed data.

When compared against each other, the Bartlett factor score is conditionally unbiased (Bentler & Yuan, 1997) in the sense that $E(\hat{\boldsymbol{\eta}}|\boldsymbol{\eta}) = \boldsymbol{\eta}$, and Thurstone-Thomson "regression" estimator has least deviation from the true score. When the predicted factor score used for independent variables in linear regression analyses with manifest outcome variables, the predicted factor score using the "regression" estimator is recommended with unbiased parameter estimates (Devlieger et al., 2016).

Since our model is not a linear model and an absolute difference score is used for independent variables, it is not clear which factor score estimation method works better. We, therefore, conducted a simulation study in which both types of factor scores are used. Based on the simulation results, the use of the Thurstone-Thomson "regression" estimator results in less biased parameter estimates in the second-stage estimation. Therefore, it is also recommended in our analysis. In the remaining part of this article, we will focus on the "regression" factor score, and it is used in both the simulation and the empirical studies.

At the second stage, we fit a logistic regression model using the predicted factor scores and estimated factor covariance matrix obtained at the first stage,

$$\begin{cases} y_{ij} & \sim \text{Bernoulli}(p_{ij}) \\ \text{logit}(p_{ij}) & = \alpha + \boldsymbol{\beta}'\mathbf{h}_{ij} + \gamma\sqrt{(\hat{\boldsymbol{\eta}}_i - \hat{\boldsymbol{\eta}}_j)'\hat{\boldsymbol{\Phi}}^{-1}(\hat{\boldsymbol{\eta}}_i - \hat{\boldsymbol{\eta}}_j)} \end{cases}. \tag{15}$$

The two-stage estimation procedure is conducted in the free statistical software R (R Core Team, 2016).

Simulation Study

In this section, we conduct a simulation study to evaluate the performance of the two-stage estimation procedure. We focus on how the performance of the two-stage procedure is impacted by the number of factors, number of indicators per factor, correlations between factors, and sample sizes. The evaluation focuses on the parameter and standard error estimates.

Simulation design

The latent space model with factor structures defined by Equations (8) and (9) is the data generating model, where the Mahalanobis distance is used as the latent nodal covariate $\boldsymbol{\Theta}$.

Measurement model. The measurement model is a confirmatory factor model as described by Equation (8). We consider 2 and 4 factors in the factor model. The number of indicators per factor is chosen to be 3, 5, 10, and 20, respectively.

CFA2. Let CFA2 denote a confirmatory factor model with 2 factors. The factor loading matrix is $\mathbf{\Lambda} = \begin{bmatrix} \boldsymbol{\lambda} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\lambda} \end{bmatrix}$ with $\boldsymbol{\lambda} = (1, .9, .8)'$, $\boldsymbol{\lambda} = (1, .9, .8, .9, 8)'$, $\boldsymbol{\lambda} = (1, .9, .8, .9, .8, .9, .8, .9, .8, .9)'$, and $\boldsymbol{\lambda} = (1, .9, .8, .9, .8, .9, .8, .9, .8, .9, .8, .9, .8, .9, .8, .9, .8, .9, .8, .9)'$ for models with 3, 5, 10, and 20 indicators per factor, respectively. The factor covariance matrix is set at $\mathbf{\Phi} = .9 \times \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$ and the correlation coefficient r takes value 0, .2, and .5, which ranges from no correlation to large correlation. The reliability of each item, defined as the ratio of the factor explained variance to the total variance, falls between 0.5 to 0.9.

CFA4. Let CFA4 denote a confirmatory factor model with 4 factors and its factor loading matrix is $\mathbf{\Lambda} = \begin{bmatrix} \boldsymbol{\lambda} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\lambda} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\lambda} \end{bmatrix}$ with $\boldsymbol{\lambda} = (1, .9, .8)'$, $\boldsymbol{\lambda} = (1, .9, .8, .9, 8)'$, $\boldsymbol{\lambda} = (1, .9, .8, .9, .8, .9, .8, .9, .8, .9)'$, and $\boldsymbol{\lambda} = (1, .9, .8, .9, .8, .9, .8, .9, .8, .9, .8, .9, .8, .9, .8, .9, .8, .9, .8, .9)'$, respectively for models with 3, 5, 10, and 20 indicators per factor. The factor covariance matrix is set at $\mathbf{\Phi} = .9 \times \begin{bmatrix} 1 & 0 & .2 & .5 \\ 0 & 1 & .5 & .2 \\ .2 & .5 & 1 & 0 \\ .5 & .2 & 0 & 1 \end{bmatrix}$. Therefore, we allow the factors to be correlated with each other and the correlation coefficients range from no correlation to large correlation.

In both CFA2 and CFA4, the covariance matrix of the unique factors is denoted $\mathbf{\Psi}$, which is a diagonal matrix with diagonal elements $\mathbf{1} - \text{diag}(\mathbf{\Lambda\Phi\Lambda}')$. Thus, the covariance matrix of the manifest indicators, which is $\Sigma = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Psi}$, becomes a correlation matrix. By setting the population values at those values, the ratio of the error variance and total variances of each indicator, $\frac{\sigma_\epsilon^2}{\sigma_z^2}$, is from 0.1 to 0.5. Equivalently, the reliability of each indicator is in the range [0.5,0.9].

Structural model. The structural model is a logistic regression model, and we consider both manifest and latent nodal covariates in the structural model.

Besides the latent traits, we consider two other predictors, which are simulated from a Bernoulli and a standard normal distribution, respectively. The latent space model with factor structures aims to predict the probability of $y_{ij} = 1$ from the characteristics of actors. Since the network data are dyadic, we need to transform the covariates to dyads to match the dimension of covariates and networks, which is usually based on the specific covariate effects to test. For the latent nodal covariates $\mathbf{\Theta}_{ij}$, we compute the Mahalanobis distance

$d_{ij} = \sqrt{(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)'\boldsymbol{\Phi}^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)}$, which is the standardized distance of latent vectors and measures the "closeness" of two actors in the factor space.

The categorical covariate X_1 is generated from the Bernoulli distribution-Bernoulli(0.5). To simulate its effect, we define the *uniform homophily statistic* based on Equation (4), termed as $h_{1,ij}$ for actors i and j,

$$h_{1,ij} = \begin{cases} 1 & \text{if } x_{1,i} = x_{1,j} \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

The continuous covariate X_2 is simulated from the standard normal distribution $N(0, 1)$. To simulate its effect, we compute the *absolute difference factor effect* as the nodal covariate according to Equation (2),

$$h_{2,ij} = |x_{2,i} - x_{2,j}|, \tag{17}$$

which is a measure of how similar two actors are in terms of the attribute measured by X_2.

The structural model expressed as a logistic model is

$$\text{logit}(p_{ij}) = \alpha + \beta_1 h_{1,ij} + \beta_2 h_{2,ij} + \gamma\sqrt{(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)'\boldsymbol{\Phi}^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)} \tag{18}$$

where α, β_1, β_2, and γ are the coefficients of the model.

The regression coefficients considered here are $(\alpha, \beta_1, \beta_2, \gamma) = (-1, 0.6, -0.2, -0.1)$, which are similar to those obtained in the empirical study in the next section.

Sample size . In most social network analysis, the size of networks (i.e., the number of actors in a network) is not very large (Kolaczyk & Krivitsky, 2015). We evaluate sample sizes of 50, 100, 150, 200, 250, and 300 in the simulation.

In total, there are $4 \times 3 \times 6(\text{CFA2}) + 4 \times 1 \times 6(\text{CFA4}) = 96$ conditions to evaluate in the simulation. Under each condition, we generate 1000 data sets and fit the latent space model with factor structures to them.

Evaluation criteria

We focus on how accurate the parameter estimates and their standard errors are. Three statistics will be computed: relative bias (Bias), standard error estimates ($\bar{se}$), and empirical standard errors ($e.se$).

Let θ represent a parameter or its true value. The *relative bias* is defined as the percent ratio of the discrepancy between the estimate and the true value with respect to the true value of

a parameter:

$$\text{relative bias}_\theta = \begin{cases} \frac{\bar{\theta}-\theta}{|\theta|} \times 100\% & \text{if } \theta \neq 0 \\ (\bar{\theta}-\theta) \times 100\% & \text{otherwise} \end{cases}, \tag{19}$$

where $\bar{\theta}$ is the average of the estimates in R, which is 1000 in our simulation, successful replications,

$$\bar{\theta} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r$$

with $\hat{\theta}_r$ denoting the parameter estimate in the rth replication.

The *standard error* estimate is defined as the average of all replicated standard errors,

$$\bar{se}_\theta = \frac{1}{1000}\sum_{r=1}^{1000} se_r(\theta)$$

where $se_r(\theta)$ is the standard error of θ in the rth replication.

With 1000 replications, the *empirical standard error* ($e.se$) is defined as

$$e.se_\theta = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_r - \bar{\theta})^2}$$

with $\bar{\theta}$ and $\hat{\theta}_r$ being the same as before. To quantify the accuracy of the standard error estimate of a parameter θ, its relative discrepancy from the empirical standard error ($e.se$) is calculated as follows

$$\text{Diff}_\theta = \frac{\bar{se}_\theta - e.se_\theta}{e.se_\theta} \times 100$$

with $\bar{se}_\theta$ and $e.se_\theta$ being standard error estimates and empirical standard error, respectively. A Diff_θ less than 10% indicates the standard error is estimated accurately (Hoogland & Boomsma, 1998).

Performance of the two-stage estimation procedure

We now present the results on the relative bias, standard error estimates, and empirical standard errors. Using the two-stage method, the factor model is fit using the ML estimation method, who can provide in the literature. We thus focus on the parameter estimates of the structural model.

For each generated data sets, we fit our model to it and parameter estimates were obtained

using the two-stage procedure described in the previous section. For comparison, we extracted both the Bartlett factor score and the "regression" factor score at the first stage. Two sets of parameter estimates were obtained in the second stage using the two types of predicted factor scores, respectively. Our simulation results show that when the "regression" factor score was used, the parameter estimates of the logistic model was less biased. In the following, we will explain the simulation results using the "regression" factor score, which are summarized in Tables 1-5. The results using Bartlett factor scores are provided in the supplementary materials.

Accuracy of parameter estimates. Based on the relative bias shown in Table 1 to Table 4, the two-stage procedure worked well in estimating the model parameters when the CFA model had two latent factors. First, all relative biases in the coefficients estimates of the manifest nodal covariates (i.e.,β_1 and β_2), were less than 5% across all conditions even with the size of networks as small as 50. The relative biases of γ were relatively larger. However, they decreased as the number of indicators per factor increased from 3 to 20. With 3 indicators per factor, the relative bias was around $5 - 10\%$ with only one exception bolded in the table. With 5 indicators per factor, it became smaller and was around 5%. When each factor was measured by 10 or 20 indicators, the relative bias of all regression coefficients estimates was less than 5% regardless of sample sizes.

Second, the relative bias decreased when the sample size increased. However, the degree of the improvement in parameter estimates is also related to the number of indicators per factor. In general, the influence of sample size is less when the number of indicators per factor is larger. In addition, the factor correlations have little influence on parameter estimates. With a correlation being 0 and a correlation being .5, the results are quite similar.

With 4 latent variables (see Table (5)), the results were similar.

Precision of parameter estimates. The precision of the parameter estimates is measured by its standard error, whose estimate is represented by $\bar{se}$. For comparison, we also obtained the empirical standard error ($e.se$), which approximates the true standard error. From Table 1 to Table 5, we noticed that the $\bar{se}$s were generally very close to $e.se$s with the relative discrepancy (Diff) less than 5%, which indicated that the two-stage procedure can estimate the standard error well. In addition, the $\bar{se}$s of the latent nodal covariate coefficient γ were mostly less than their corresponding $e.se$s. The reason was that at the second stage, the predicted factor scores were used and the variation of factor scores cannot be fully accounted in estimating the regression coefficients. Hence, the standard error estimates obtained by the two-stage procedure did not account for the uncertainty of factor scores. Consequently, they were smaller than they should be, which was approximated by the empirical standard errors. However, the discrepancy between the standard error estimate ($\bar{se}$) and the empirical standard errors ($e.se$) was less than

5% based on our simulation study.

In summary, the two-stage procedure using the predicted factor score from the "regression" estimator works well and it provides both accurate parameter estimates and reliable standard error estimates.

Empirical Study

We now illustrate our model using a set of real data. The data on friendship were collected from a group of college students from the same major in the same college. In the current study, data are available from 129 participants, among which 55.04% were females and 44.96% were males. During the data collection, each student was given a roster of all the students and was asked to report whether every other student was his/her friends or not. In addition, information on academic performance was also available, with scores ranging from 18 to 87. The average academic performance score was 54.99 with the standard deviation 10.94. In the analysis, we standardized the academic performance scores. Finally, a short form of personality test was used to measure two personality factors "Extraversion" and "Imagination", each of which was measured by 4 items.

Using the data, we investigate the association between covariates –gender, academic performance, and personalities– and friendships. Specifically, we evaluate whether similarity in student attributes (such as gender, academic performance, and personalities) can predict the formation of friendships. To test the gender effect, the *uniform homophily statistic* is defined as

$$h_{1,ij} = \begin{cases} 1 & \text{if actors } i \text{ and } j \text{ are of the same gender} \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

For the standardized academic performance score, we test the *absolute difference factor effect* with the following nodal statistic,

$$h_{2,ij} = |\text{score}_i - \text{score}_j|. \tag{21}$$

Finally, the personality effect is defined using the Mahalanobis distance of latent factors–Extraversion and Imagination.

We first obtain the predicted factor scores for latent factors –Extraversion and Imagination– through a confirmatory factor analysis. The model shown in Figure 3 was found to fit the personality data well with the chi-squared statistic 22.72 and p-value 0.159 (CFI=0.952 and RMSEA=0.051). The predicted factor scores from the Thurstone-Thomson "regression"

estimator were then obtained based on the model.

Given the predicted factor scores, the following structural model was estimated,

$$\text{logit}(y_{ij} = 1) = \beta_0 + \beta_1 h_{1,ij} + \beta_2 h_{2,ij} + \gamma\sqrt{(\hat{\boldsymbol{\eta}}_i - \hat{\boldsymbol{\eta}}_j)'\hat{\boldsymbol{\Phi}}^{-1}(\hat{\boldsymbol{\eta}}_i - \hat{\boldsymbol{\eta}}_j)},$$

where $\hat{\boldsymbol{\eta}}_i$ are the predicted factor scores obtained using the Thurstone-Thomson "regression" estimator for participant i. The parameter estimates were summarized in Table 6. First, the coefficient β_1 for gender was 0.588 with p-value less than 0.001, which was statistically significant at the significance level 0.05. This result indicates that two students with the same gender were more likely to be friends when controlling the effect of the academic performance and latent personality distance. Second, the coefficient β_2 for the academic performance score was -0.222 with p-value less than 0.001, indicating a significant effect. Therefore, two students with more similar academic performances were more likely to become friends when controlling the effect of gender and the distance in the personality space. The coefficient of the latent distance was 0.004 with the p-value 0.919. Since the coefficient is not significant, it indicates that whether two students are friends or not does not depend on their personality similarity after controlling the effect of gender and academic performance.

Discussion and Conclusion

Social network analysis is a technique used to understand the formation of associations among actors within a network. It is becoming increasingly popular in psychology and social sciences. Latent space modeling is one of the representative methods for social network analysis. Latent space modeling is based on the assumption that each actor holds an unknown position in a latent space. The relation between actors does not only depend on actor's characteristics, but also on their relative positions in the latent space. The closer two actors in the latent space, the more likely for them to be connected in a social network. The latent space itself is arbitrary with no substantive meaning. Individuals are naturally nested in different types of manifest and latent spaces including height-weight space, personality space, and intelligence space.

In psychological research, factor models are widely used to study the latent structure of observed variables and the latent factors usually have substantive meaning. For instance, the widely used Big Five factor model defines five correlated dimensions for personalities. Each dimension describes one personality trait. The factor scores represent the strength of personality in different dimensions, which naturally form a meaningful latent space. Therefore, we propose to use the latent factor space in our analysis. A confirmatory factor model is adopted to model the latent factors using data on multiple indicators. Consequently, the resulting model consists of

both a measurement model and a structural model, similar to structural equation models.

To estimate the model, we proposed a two-stage procedure. At the first stage, we fit the CFA model. Both the model parameters and the Thurstone-Thomson "regression" factor scores were extracted. At the second stage, the regression coefficients were obtained. Results from our simulation study showed that the coefficients of manifest nodal covariates were estimated very well with ignorable relative bias ($< 5\%$) across all conditions. The estimates of the coefficient of latent nodal covariates had slightly larger relative bias, but the bias decreased when the number of indicators per factor and the network size increased. With 5 indicators per factor, the relative bias was around 5% and with 10 or more indicators per factor, the relative bias was ignorable ($< 5\%$).

We applied our model in a friendship network analysis. The data on friendships were collected from a group of college students from the same major in the same college. We used the information on gender, academic performance scores, and two personality factors to predict the existence of friendships among students. According to our analysis, students of the same gender are more likely to be friends than those of different genders, while controlling the academic performance and personality factors. For college students, similar academic performance increases the possibility of two students to be friends while both the gender and personality factor scores are controlled. Similar personalities were not found to be related to the chance for two students to be friends.

From the simulation results, we observed that the coefficient estimates of the latent nodal covariates always biased towards zero, which indicated that the coefficient of the latent nodal covariate was under-estimated. The attenuation is ascribed to use of predicted factor scores in the second-stage of the analysis. However, the attenuation becomes less severe when the number of indicators per factor and/or the sample size increases. This is because that when the number of indicators per factor increases, we have more information on the factor score and less uncertainty involved in the predicted factor score. We also found that the standard error of the coefficient estimates of the latent nodal covariate was less than it was supposed to. The reason is that we used the predicted factor scores at the second stage and the uncertainty of latent factors was thus ignored in estimating the regression coefficients. Consequently, the obtained standard error estimates were smaller than if the true factor scores were used. However, the relative discrepancy of the estimated standard error and the empirical standard error is ignorable ($< 5\%$) when a factor has five or more indicators.

To estimate our model, Bayesian estimation methods could be useful and potentially work better especially when a factor is measured by a small number of indicators, for instance less than 3 indicators per factor. Because of the model complexity, it is very challenging to obtain

ML estimates using an ordinary EM algorithm. In the literature, variational EM (Han et al., 2015; Paul & Chen, 2016) and variational Bayesian (Aicher et al., 2014; Airoldi et al., 2008; Latouche et al., 2011) approaches have been developed for complex network models and they can provide model parameter estimates using a single optimization procedure. Estimating our model using variational approaches is thus something worthy investigation in future.

Although a one-stage method can potentially work better than the two-stage method, conducting a one-stage analysis can be more difficult for applied researchers. Compared to Bayesian estimation methods, the two-stage ML method is fast and simple, and still provides consistent parameter estimates. Therefore, it is more pragmatic to be used by applied researchers without much experience in Bayesian modeling.

Although both the two-stage procedure and potential one-stage procedure are able to provide consistent parameter estimates, the two types of estimation procedures are distinct from each other in some aspects. If a one-stage procedure is used, the network information contributes to the factor score and the factor covariance structure estimates, and the network is actually a manifest "variable" of latent factors. Using a one-stage estimation method, similarity in personalities causes social relations. Using the two-stage ML procedure, the network information is not used in the estimation of the factor model. The similarity in personalities is purely a predictor of the presence of social relations and a network is the outcome variable. In the substantive literature, it has been found that similar personalities may make two individuals feel comfortable and become friends (Youyou et al., 2017). On the other hand, friends may influence each other in many aspects including personalities. From these two aspects, we would think both the one-stage and two-stage estimation procedures have their applications in substantive research and the choice between the two is up to the interpretation of the relation between social relations and personality similarities.

In the near future, we would like to extend our model in different ways. First, because there are many factors that may influence the formation of social relations, it is impossible to exhaust all of them and the conditional independence assumption is thus easily violated in practice. In order to find an adequate model for given data, it is necessary to develop model selection procedures and fit indices to aid researcher to do model selection. Second, the current logistic model is not a causal model and no causal inference could be drawn from it. Theoretically, the relation between social network and personality traits may be reciprocal. To better understand their relation, we would like to extend the model to longitudinal social network analysis. Third, since humans are social actors and they belong to multiple social networks simultaneously (such as friendship network and social median network), we intend to expand the model for multiplex network analysis. Fourth, it is worth developing one-stage procedures to estimate the model.

References

Aicher, C., Jacobs, A. Z., & Clauset, A. (2014). Learning latent block structure in weighted networks. *Journal of Complex Networks*, *3*(2), 221–248. doi: 10.1093/comnet/cnu026

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, *9*(Sep), 1981–2014.

Anderson, C. J., Wasserman, S., & Crouch, B. (1999). A p* primer: Logit models for social networks. *Social networks*, *21*(1), 37–66. doi: 10.1016/S0378-8733(98)00012-4

Asendorpf, J. B., & Wilpers, S. (1998). Personality effects on social relationships. *Journal of Personality and Social Psychology*, *74*(6), 1531–1544. doi: 10.1037/0022-3514.74.6.1531

Bartlett, M. S. (1937). The statistical conception of mental factors. *British journal of Psychology*, *28*(1), 97–104. doi: 10.1111/j.2044-8295.1937.tb00863.x

Bentler, P. M., & Yuan, K.-H. (1997). Optimal conditionally unbiased equivariant factor score estimators. *Latent Variable Modeling and Applications to Causality*, *120*, 259–281.

Berger, A., Drosten, C., Doerr, H., Stürmer, M., & Preiser, W. (2004). Severe acute respiratory syndrome (sars)-paradigm of an emerging viral infection. *Journal of Clinical Virology*, *29*(1), 13–22. doi: 10.1016/j.jcv.2003.09.011

Bilmes, J. A., et al. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, *4*(510), 126.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*(4), 443–459. doi: 10.1007/BF02293801

Borsboom, D., & Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology*, *9*, 91–121. doi: 10.1146/annurev-clinpsy-050212-185608

Cattell, R. B. (1952). *Factor analysis: an introduction and manual for the psychologist and social scientist*. New York: Harper.

Cramer, A., Waldorp, L., van der Maas, H., & Borsboom, D. (2010). Comorbidity: a network perspective. *Behavioral and Brain Sciences*, *33*(2-3), 137–150. doi: 10.1017/S0140525X09991567

Deana, D. O., Bauerb, D. J., & Prinsteinc, M. J. (2017). Friendship dissolution within social networks modeled through multilevel event history analysis. *Multivariate Behavioral Research*, *52*(3), 271–289. doi: 10.1080/00273171.2016.1267605

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, *76*(5), 741–770. doi: 10.1177/0013164415607618

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, *18*(2), 192. doi: 10.1037/1040-3590.18.2.192

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. New York: Cambridge University Press.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network pschometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904–927. doi: 10.1007/s11336-017-9557-x

Faust, K. (1988). Comparison of methods for positional analysis: Structural and general equivalences. *Social networks*, *10*(4), 313–341. doi: 10.1016/0378-8733(88)90002-0

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the american Statistical association*, *81*(395), 832–842. doi: 10.1080/01621459.1986.10478342

Ghahramani, Z., Hinton, G. E., et al. (1996). *The em algorithm for mixtures of factor analyzers* (Tech. Rep.). Technical Report CRG-TR-96-1, University of Toronto.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological methods*, *6*(4), 430–450. doi: 10.1037/1082-989X.6.4.430

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Mathematical and Statistical Psychology*, *8*(2), 65–81. doi: 10.1111/j.2044-8317.1955.tb00321.x

Han, Q., Xu, K. S., & Airoldi, E. M. (2015). Consistent estimation of dynamic and multi-layer block models. *Proceedings of the 32nd International Conference on Machine Learning*, 1511-1520.

Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(2), 301–354. doi: 10.1111/j.1467-985X.2007.00471.x

Harris, K., & Vazire, S. (2016). On friendship development and the big five personality traits. *Social and Personality Psychology Compass*, *10*(11), 647-667. doi: 10.1111/spc3.12287

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association*, *97*(460), 1090–1098. doi: 10.1198/016214502388618906

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329–367. doi: 10.1177/0049124198026003003

Hunter, D. R., Goodreau, S. M., & Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, *103*(481), 248–258. doi: 10.1198/016214507000000446

Hunter, D. R., & Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, *15*(3), 565–583. doi: 10.1198/106186006X133069

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*(4), 443–482. doi: 10.1007/BF02289658

Kolaczyk, E. D., & Krivitsky, P. N. (2015). On the question of effective sample size in network modeling: An asymptotic inquiry. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *30*(2), 184–198. doi: 10.1214/14-STS502

Latouche, P., Birmele, E., & Ambroise, C. (2011). Overlapping stochastic block models with application to the french political blogosphere. *The Annuals of Applied Statistics*, *5*(1), 309-336. doi: 10.1214/10-AOAS382

Lazer, D., Rubineau, B., Chetkovich, C., Katz, N., & Neblo, M. (2010). The coevolution of networks and political attitudes. *Political Communication*, *27*(3), 248–274. doi: 10.1080/10584609.2010.500187

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (India)*, *2*(1), 49–55.

Maya-Jariego, I., & Holgado, D. (2015). Network analysis for social and community interventions. *Psychosocial Intervention*, *24*(3), 121–124. doi: 10.1016/j.psi.2015.10.001

McCrae, R. R., Martin, T. A., Hrebickova, M., Urbánek, T., Boomsma, D. I., Willemsen, G., & Costa, P. T. (2008). Personality trait similarity between spouses in four cultures. *Journal of personality*, *76*(5), 1137–1164. doi: 10.1111/j.1467-6494.2008.00517.x

McFarland, D. D., & Brown, D. J. (1973). Social distance as a metric: a systematic introduction to smallest space analysis. *Bonds of pluralism: The form and substance of urban social networks*, *6*, 213–252.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415-444. doi: 10.1146/annurev.soc.27.1.415

Miething, A., Almquist, Y. B., Östberg, V., Rostila, M., Edling, C., & Rydgren, J. (2016). Friendship networks and psychological well-being from late adolescence to young adulthood: a gender-specific structural equation modeling approach. *BMC psychology*, *4*(1), 34. doi: 10.1186/s40359-016-0143-2

Mulaik, S. A. (1976). Comments on "the measurement of factorial indeterminacy". *Psychometrika*, *41*(2), 249–262. doi: 10.1007/BF02291842

Mulaik, S. A. (2009). *Foundations of factor analysis* (2nd ed.). New York: Chapman and Hall/CRC.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, *28*(6), 441–453. doi: 10.1177/016555150202800601

Paul, S., & Chen, Y. (2016). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*(2), 3807-3870. doi: 10.1214/16-EJS1211

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social networks*, *29*(2), 173–191. doi: 10.1016/j.socnet.2006.08.002

Rushton, J. P., & Bons, T. A. (2005). Mate choice and friendship in twins: evidence for genetic similarity. *Psychological Science*, *16*(7), 555–559. doi: 10.1111/j.0956-7976.2005.01574.x

Ryan, J. B. (2011). Social networks as a shortcut to correct voting. *American Journal of Political Science*, *55*(4), 753–766. doi: 10.1111/j.1540-5907.2011.00528.x

Saul, Z. M., & Filkov, V. (2007). Exploring biological network structure using exponential random graph models. *Bioinformatics*, *23*(19), 2604–2611. doi: 10.1093/bioinformatics/btm370

Schaefer, D. R., Adams, J., & Haas, S. A. (2013). Social networks and smoking: exploring the effects of peer influence and smoker popularity through simulations. *Health Education & Behavior*, *40*(1), 24S–32S. doi: 10.1177/1090198113493091

Schmittmann, V. D., Cramer, A. O., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New ideas in psychology*, *31*(1), 43–53. doi: 10.1016/j.newideapsych.2011.02.007

Sewell, D. K., & Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association*, *110*(512), 1646–1657. doi: 10.1080/01621459.2014.988214

Snijders, T. A., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological methodology*, *36*(1), 99–153. doi: 10.1016/j.socnet.2006.08.003

Team, R. C. (2016). R: A language and environment for statistical computing. r foundation for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. C: University of Chicago Press.

Veenstra, R., Dijkstra, J. K., Steglich, C., & Van Zalk, M. H. (2013). Network–behavior dynamics. *Journal of Research on Adolescence*, *23*(3), 399–412. doi: 10.1111/jora.12070

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp. *Psychometrika*, *61*(3), 401–425. doi: 10.1007/BF02294547

Watson, D., Beer, A., & McDade-Montez, E. (2014). The role of active assortment in spousal similarity. *Journal of Personality*, *82*(2), 116–129. doi: 10.1111/jopy.12039

Watson, D., Hubbard, B., & Wiese, D. (2000). Self–other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of personality and social psychology*, *78*(3), 546. doi: 10.1037/0022-3514.78.3.546

Youyou, W., Stillwell, D., Schwartz, H. A., & Kosinski, M. (2017). Corrigendum: Birds of a feather do flock together: Behavior-based personality-assessment method reveals personality similarity among couples and friends. *Psychological Science*, *28*(3), 276-284. doi: 10.1177/0956797616678187

Table 1

Parameter and standard error estimates of logistic regressions with 2 factors in the CFA model and each factor has 3 indicators

	Parameter		$r = 0$		$r = .2$		$r = .5$	
N	Par	True	Bias(%)	Diff(%)	Bias(%)	Diff(%)	Bias(%)	Diff(%)
50	β_1	.6	-0.87	2.54	0.00	0.03	-0.49	-0.16
	β_2	-.2	-1.18	-2.72	-3.01	-3.67	0.39	1.44
	γ	-.1	8.81	-0.41	8.55	2.11	**12.47**	-2.31
100	β_1	.6	-0.57	1.67	0.33	-4.21	0.19	4.30
	β_2	-.2	-1.81	0.94	0.07	0.63	-0.12	-0.08
	γ	-.1	6.00	0.32	6.18	1.04	7.24	-2.47
150	β_1	.6	-0.05	1.67	0.26	1.40	-0.42	-1.89
	β_2	-.2	0.41	-2.27	0.80	-2.26	0.54	0.35
	γ	-.1	6.14	-2.83	6.13	-2.25	7.75	-0.81
200	β_1	.6	0.10	2.31	0.07	-1.62	-0.04	0.80
	β_2	-.2	-0.20	-0.23	0.15	-2.20	-0.14	1.34
	γ	-.1	6.67	-6.42	6.25	3.67	6.60	-2.23
250	β_1	.6	-0.09	1.22	0.08	-0.84	-0.20	-3.24
	β_2	-.2	-0.04	-3.10	-0.17	-2.24	-0.14	1.20
	γ	-.1	5.40	-3.72	5.89	-6.07	7.51	-5.57
300	β_1	.6	-0.18	0.77	-0.03	-0.52	0.05	2.31
	β_2	-.2	-0.39	0.76	0.13	0.93	-0.06	-1.96
	γ	-.1	5.37	-5.18	5.41	-5.13	6.14	-4.97

Bias means the relative bias, $\bar{se}$ is the average of standard errors across all replications, $e.se$ is the empirical standard error, Diff is $\frac{(\bar{se}-e.se)}{e.se} \times 100$; a bold number is a relative bias larger above 10%.

Table 2

Parameter estimates and standard error estimates of logistic regressions with 2 factors in the CFA model and each factor has 5 indicators

	Parameter		$r = 0$		$r = .2$		$r = .5$	
N	Par	True	Bias(%)	Diff(%)	Bias(%)	Diff(%)	Bias(%)	Diff(%)
50	β_1	.6	1.13	1.46	-0.69	-3.71	-0.58	0.56
	β_2	-.2	-1.67	-4.54	-3.34	1.25	0.55	-2.75
	γ	-.1	4.56	3.42	7.07	-6.73	6.50	1.85
100	β_1	.6	0.32	3.38	-0.02	-0.34	-0.35	-0.80
	β_2	-.2	-0.10	-1.60	0.30	1.62	-0.36	-1.19
	γ	-.1	5.96	-0.12	4.60	-0.45	3.67	-1.71
150	β_1	.6	0.52	-0.30	0.49	-3.39	0.33	-2.35
	β_2	-.2	0.21	4.96	0.14	4.39	0.46	-1.11
	γ	-.1	5.73	-1.11	5.27	-3.77	5.04	-4.94
200	β_1	.6	0.06	-1.19	0.02	-0.42	-0.03	4.02
	β_2	-.2	-0.44	-1.54	-0.18	-0.47	0.23	2.10
	γ	-.1	3.77	-2.48	4.84	-3.40	5.02	-8.14
250	β_1	.6	-0.14	-1.12	0.07	3.31	0.16	-3.84
	β_2	-.2	-0.01	-4.03	0.32	0.39	-0.19	-3.25
	γ	-.1	4.21	-2.39	4.61	-1.71	4.98	-1.90
300	β_1	.6	0.11	1.53	-0.12	-1.53	-0.08	1.36
	β_2	-.2	0.09	0.55	0.09	0.71	0.33	1.46
	γ	-.1	4.37	-2.86	3.96	-2.67	5.31	-3.52

Bias means the relative bias, $\bar{se}$ is the average of standard errors across all replications, $e.se$ is the empirical standard error, Diff is $\frac{(\bar{se}-e.se)}{e.se} \times 100$; a bold number is a relative bias larger above 10%.

Table 3

Parameter estimates and standard error estimates of logistic regressions with 2 factors in the CFA model and each factor has 10 indicators.

	Parameter		$r = 0$		$r = .2$		$r = .5$	
N	Par	True	Bias(%)	Diff(%)	Bias(%)	Diff(%)	Bias(%)	Diff(%)
50	β_1	.6	0.30	0.15	1.01	-1.34	0.41	-5.24
	β_2	-.2	0.05	-1.29	-0.29	2.40	-0.97	3.08
	γ	-.1	6.53	1.60	3.39	-2.26	4.48	-1.01
100	β_1	.6	-0.12	2.84	-0.28	2.60	-0.29	0.73
	β_2	-.2	-1.86	2.37	-0.41	-1.45	-1.07	-1.34
	γ	-.1	2.61	1.56	2.79	-2.30	3.05	-1.73
150	β_1	.6	-0.12	0.29	-0.07	0.50	-0.03	-6.74
	β_2	-.2	0.42	0.11	-0.45	-2.42	-0.29	-2.45
	γ	-.1	2.51	-0.32	2.58	-0.15	3.24	-3.24
200	β_1	.6	0.17	1.88	0.04	0.34	0.05	-1.76
	β_2	-.2	0.59	-0.01	-0.01	-1.33	-0.10	1.76
	γ	-.1	3.31	-1.61	1.18	-1.31	2.82	-1.25
250	β_1	.6	0.10	-1.18	-0.12	0.47	-0.08	-0.60
	β_2	-.2	0.18	-1.62	-0.41	-3.37	0.09	-0.38
	γ	-.1	1.43	-3.59	3.18	-2.95	3.97	-1.99
300	β_1	.6	-0.12	0.81	-0.06	-0.85	0.08	0.93
	β_2	-.2	-0.26	-2.16	-0.16	-1.02	0.21	0.27
	γ	-.1	3.16	-2.42	2.56	-0.47	3.48	-1.24

Bias the relative bias, $\bar{se}$ is the average of standard errors across all replications, $e.se$ is the empirical standard error, Diff is the $\frac{(\bar{se}-e.se)}{e.se} \times 100$; a bold number is a relative bias larger above 10%.

Table 4
Parameter estimates and standard error estimates of logistic regression with 2 factors in the CFA model and each factor has 20 indicators.

	Parameter		$r = 0$		$r = .2$		$r = .5$	
N	Par	True	Bias(%)	Diff(%)	Bias(%)	Diff(%)	Bias(%)	Diff(%)
50	β_1	.6	1.23	0.07	1.69	-1.91	1.20	-1.12
	β_2	-.2	-1.29	2.91	1.26	-3.05	-2.13	-1.29
	γ	-.1	3.40	0.26	1.30	-0.39	3.65	-5.65
100	β_1	.6	0.24	1.60	-0.36	1.04	-0.07	-0.86
	β_2	-.2	-1.12	1.66	0.46	-2.97	0.32	3.60
	γ	-.1	1.78	0.60	3.22	-2.08	2.53	0.78
150	β_1	.6	-0.43	0.02	-0.05	-0.79	0.08	-0.30
	β_2	-.2	-0.02	2.36	-0.22	0.91	0.70	0.83
	γ	-.1	2.18	0.39	2.17	0.18	1.06	-5.56
200	β_1	.6	0.02	0.40	-0.01	-0.90	0.15	0.44
	β_2	-.2	-0.24	5.43	0.20	2.47	-0.04	-2.05
	γ	-.1	1.11	-2.03	2.55	-2.96	2.68	-4.59
250	β_1	.6	0.06	-2.46	0.21	0.85	0.18	0.70
	β_2	-.2	0.06	2.69	-0.29	-1.29	0.02	-1.42
	γ	-.1	1.78	-4.66	2.13	-2.67	2.24	-2.20
300	β_1	.6	0.02	5.56	-0.23	-1.28	0.10	3.08
	β_2	-.2	0.26	-1.09	-0.28	-0.07	-0.05	-3.72
	γ	-.1	1.33	-5.62	1.88	-4.17	2.32	-0.72

Bias means the relative bias, $\bar{se}$ is the average of standard errors across all replications, *e.se* is the empirical standard error, Diff is $\frac{(\bar{se}-e.se)}{e.se} \times 100$; a bold number is a relative bias larger above 10%.

Table 5

Parameter estimates and standard error estimates of logistic regression using "regression" factor scores with 4 factors in the CFA model.

	#Indicator		3		5		10		20	
N	Par	True	Bias(%)	Diff(%)	Bias(%)	Diff(%)	Bias(%)	Diff(%)	Bias(%)	Diff(%)
50	β_1	.6	1.24	4.68	0.24	-2.40	-0.77	-2.22	-	-
	β_2	-.2	-1.43	-0.66	-1.51	-0.20	-0.58	-0.43	-	-
	γ	-.1	5.68	0.45	4.42	-1.35	1.25	2.53	-	-
100	β_1	.6	-0.25	0.80	-0.51	-0.39	0.58	8.72	-0.27	-0.68
	β_2	-.2	-0.51	-0.70	-0.60	1.67	1.13	-2.44	-0.61	-1.07
	γ	-.1	8.12	-1.25	5.54	-1.93	4.02	-1.15	1.88	-2.13
150	β_1	.6	-0.23	-0.81	0.07	-0.14	0.14	0.95	0.33	4.02
	β_2	-.2	-0.50	2.57	0.02	-0.79	0.11	-2.40	-0.06	0.79
	γ	-.1	5.91	0.18	5.38	-2.17	3.85	-4.00	2.69	-3.38
200	β_1	.6	-0.20	2.09	0.14	1.54	-0.37	-5.20	-0.02	-0.53
	β_2	-.2	0.06	2.19	0.33	0.57	-0.28	-5.69	-0.29	-0.98
	γ	-.1	6.27	0.17	4.65	0.34	3.34	-4.14	2.70	-3.42
250	β_1	.6	-0.06	-2.98	0.13	1.84	-0.16	-0.36	0.06	-1.10
	β_2	-.2	-0.25	0.79	0.40	0.28	0.03	-2.64	-0.38	-3.30
	γ	-.1	7.01	-0.81	4.88	-2.93	3.31	-1.37	2.62	-3.95
300	β_1	.6	-0.04	-3.47	0.14	3.39	-0.14	-2.32	0.02	0.07
	β_2	-.2	-0.10	-2.01	0.31	-2.32	-0.07	-4.31	0.00	0.70
	γ	-.1	5.87	-4.25	4.24	-4.66	3.27	-1.54	2.00	-1.82

Bias means the relative bias, $\bar{se}$ is the average of standard errors across all replications, *e.se* is the empirical standard error, Diff is $\frac{(\bar{se}-e.se)}{e.se} \times 100$; a bold number is a relative bias larger above 10%; The block with "-" means that the factor model is not estimable because of the too small sample size.

Table 6
Parameter estimates of the relation between friendship and nodal covariates.

	Estimate	Std. Error	z value	Pr(>|z|)
β_0	-1.240	0.0774	-16.704	< 0.001
β_1(gender)	0.588	0.053	11.035	< 0.001
β_2(score)	-0.222	0.030	-6.751	<0.001
γ	0.004	0.035	0.102	0.919
AIC	8879.1			
Residual deviance		8871.1	df	8252

ID	1	2	3	4	5	6	7	8	9	10
1	0	0	1	0	1	0	1	0	1	0
2		0	0	0	0	0	0	1	1	0
3			0	1	1	0	0	0	0	0
4				0	1	1	0	0	1	1
5					0	0	0	1	0	1
6						0	1	0	1	0
7							0	0	0	0
8								0	1	0
9									0	1
10										0

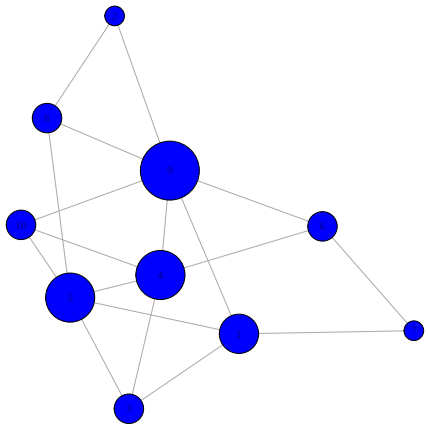

Figure 1. The top is the adjacency matrix with 10 actors and the bottom is the plot of them. Each blue node represents an actor and the a gray line means the corresponding two nodes are connected, which corresponding to "1" in the adjacency matrix. The node is labeled with its corresponding ID number. The size of a node corresponds to its degree. A bigger size means a larger degree.

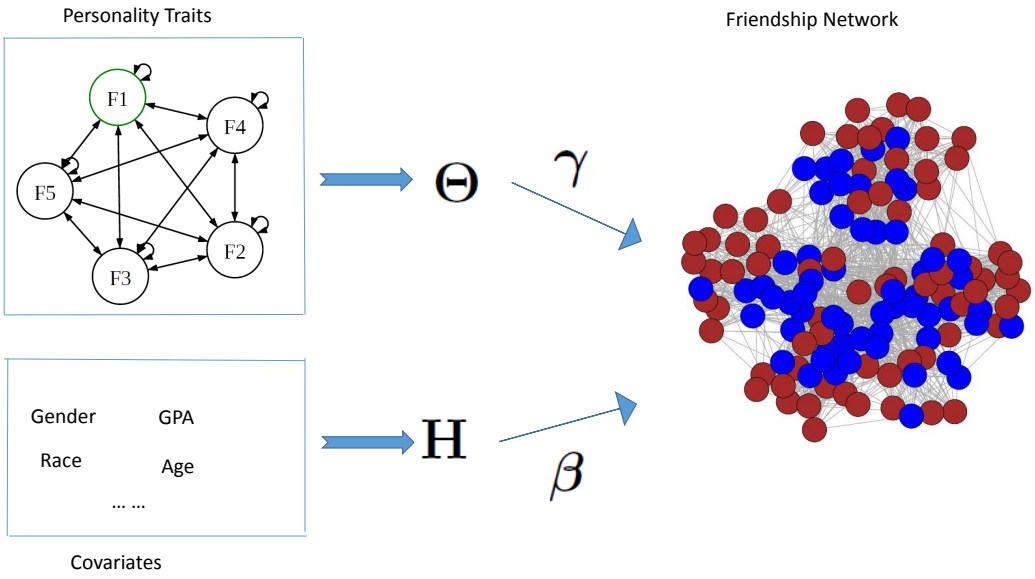

Figure 2. Path diagram of our model with factor structures. The latent personality traits contain five correlated latent factors. The manifest covariates include gender, race, GPA, age, etc. The nodal observed covariates and nodal latent covariates are $\mathbf{H}$ and $\mathbf{\Theta}$ with coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, respectively.

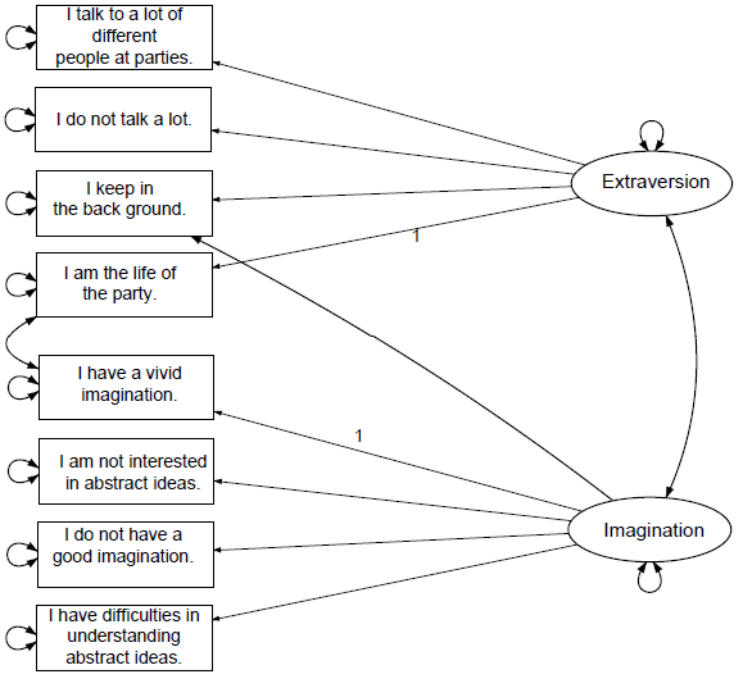

Figure 3. Path diagram of the confirmatory factor model used to extract factor scores. The first factor loading of each factor is restricted to be 1.