Psychometric Network construction, edge embedding and community detection

Tomáš Vojtíšek  
Interdisciplinary Research Team on Internet and Society  
Masaryk UniversityBrno, Czech republic  
vojtisektomas@mail.muni.cz

*Abstract*—The research project analyses a large psychometric dataset based around 2,000 fictional characters. More than 3 million respondents have rated their own personality and the personality characteristics of the fictional characters they were familiar with on a set of four hundred bipolar adjective pairs items. Two networks of statistical relationships between the self-report personality ratings and fictional character personality rating are estimated. The networks are then analyzed using network descriptives, community detection algorithms. Node embedding is used to compute pairwise node similarities within the networks and the correlation between the pairwise similarities across the two networks is estimated as a measure of overall similarity of the network structures. The implications of these results and their usefulness for personality psychology and psychometrics are discussed.

Keywords—Psychometrics, Network Science, Personality Psychology,

# Introduction

Personality measurement and assessment is a continuously growing field of research. The space of psychometrics is dominated by common cause models where an observed set of relationships between behaviors or experiences is explained by an unobserved higher order cause. For example, the behaviors of going to parties and the experience of being excited by having visitors at one’s house is, in the common cause paradigm explained by a personality trait: extraversion. Methods of factor analysis used to analyze data in this paradigm. However, in many areas, including personality psychology, the explanation via common cause contains unresolved issues. The fundamental issue is the inability to explain the nature of the common cause, as physiological and neurological structures in many cases do not correspond to the sets of relationships explained by the psychological common cause. It is therefore valuable to explore alternative explanations to the patterns of relationships that do not invoke an unobserved common cause.

Network models and theories of mutualistic relationships between psychological and behavioral phenomena are a growing field of alternative explanations for the phenomena studied by personality psychology. Instead of positing a latent cause, these theories state that the positive statistical relationship between behavioral and experiential instances (eg. 1) Going to parties and 2) looking forward to having friends visit you) are caused by direct relationships between these instances. A cultural explanation may be that the idea that a person who does 1) would also enjoy 2) may as well be the causal explanation for the observed positive relationship, and this explanation does not require a higher order cause, rather a system of connections between these instances that may be highly specific to each individual, a proposition that is completely unavailable to the common cause paradigm and its methods of factor analysis, and instead requires the study of the networks of pairwise relationships between behaviors, attitudes, or other relevant psychological constructs.

## Related works

The methods of Network Science are growing in relevance to the study of human experience, behavior and pathology in the field of psychology. A set of methods for the construction of networks of statistical conditional association in multivariate psychological data has been developed and quickly gained popularity in the past ten years [1]. These methods have been used to criticize dominant theories of intelligence [2], personality [3], personality pathology [4], and affective psychopathology [5].

In the simplest terms, two basic kinds of graphs (or network models) are used depending on the type of data analyzed. Between-person networks are graphs of relationships between analyzed variables obtained on data collected across many individuals at a single time point. Nodes of the graph are usually observed variables - individual items on a test or questionnaire. The strength of association between variables corresponds to edge weights, and the graphs are undirected. Within-person networks are estimated from data collected across multiple time instances from a single individual, and are weighted and directed, as the edges correspond to associations between variables through time. The directedness of the graphed edges has causal implications. For a more detailed overview of the psychometric network models and their estimations, see [6].

Given the nature of the psychometric graphs, the appropriate interpretation and application of network analytic methods in psychology is not straightforward. In comparison to other areas, where network science is used, in psychology, the edges are not taken from sets of observed connections but are estimated from statistical associations (most often correlations or partial correlations). Even recent works on network psychometrics ask more questions than they answer [1]. There is a growing body of literature showing the limits of generalizability of network psychometric findings using the most basic metrics of descriptive analysis [7]. Due to these challenges, it is critical to look for other methods from the field of Network science that may be applicable to the problems of psychometric network analysis.

## The current goal

The aim of this paper is to examine possible extensions of network psychometrics from network descriptive analysis to more complex methods of community detection and the use of node embeddings for overall network comparison.

# Method

## Data

### Measures

The questionnaire used to generate data consists of 400 items. Each item presents an adjective pair and a sliding answer scale. In the first part of the questionnaire, people answer a random subset of items about themselves, using the sliding answer scale to indicate where they fall on the spectrum between the two adjective opposites (eg. 20 % optimistic, 80 % pessimistic). In the second part of the questionnaire, they are asked what fictional universes they know (eg: Harry Potter, Mulan, etc.), and are then randomly assigned characters from these fictional works. For each fictional character, they are randomly assigned a subset of the 400 items and are asked to rate where does this fictional character fall on the spectrum between the two adjectives.

### Network estimation from raw data

The questionnaire was filled out by over 3.2 million people. The raw self-report dataset consists of their answers on randomly drawn subsets (ranging from 8 to 32) of questions about themselves, and there is therefore a large amount of data missing completely at random for each respondent. The raw character rating dataset has the same structure but is simplified into complete mean ratings of each of the 2000 fictional characters. The complete raw dataset can be downloaded at openpsychometrics.org [8].

The size of the raw self-report dataset alongside the missingness structure made the more commonly used methods for network estimation via partial correlations unfeasible, and I therefore opted for the simpler estimation of Pearson correlation coefficients instead. For consistency, this method was used both for the self-report and for the fictional character rating data. The resulting 400 by 400 correlation matrices as well as the R script used for data cleaning and correlation estimation can be found in the project repository [9].

The diagonal values of the correlation matrices are set to 0 as self-loops carry no meaning in this between-person network and are then used as graph adjacency matrices. To obtain a network that is not completely saturated, edge weights (correlations) of absolute value lower than 0.15 were set to 0 a practice known as network thresholding [10].

## Descriptive network analysis

The two networks are briefly described along their node degree, density, and clustering coefficient.

## Community membership detection

Communities are estimated through three methods: Label propagation, where each node is assigned a label randomly and then the labels are replaced so that each node shares its label with the maximum amount of neighbours; Fast community unfolding, where each node is assigned to a community so that the node modularity is maximized; and Walktrap algorithm, where each node is assigned to its own community and communities are then repeatedly merged so that distance between communities based on random walks is minimized.

## Network comparison through node embedding

The node2vec algorithm was used to generate node embeddings for both the self-report personality questionnaire network and the fictional character rating personality questionnaire network. 15-dimensional embeddings were generated with 100 walks of length 5. Then, similarity was computed for all pairs of nodes within each network, and the correlation between these similarity scores was used as an overall indication of similarity of the two network structures.

For comparison, the same procedure was used on two networks estimated from 2 subsamples of the real personality data, and correlation was computed for the pairwise embedding similarities. This procedure was used because randomly drawn subsamples from the large sample of 3 million observations can be expected to have the same network structure and can therefore serve as a scaling parameter of sorts to put the results of the above used procedure into perspective.

# Results

First, I describe the two obtained networks. Then I move on to present the results of the different community detection methods and briefly compare them. Finally, I present the results of network embedding similarity ratings strength of association between the two networks.

## Descriptives

### Self-report personality ratings

The network contains 400 nodes. Each node represents an item on the personality questionnaire, and each edge corresponds to the correlation of peoples answers on the two connected items (nodes). As can be seen in the following plot, the node degree does not follow a simple distribution.

Obsah obrázku text, snímek obrazovky, diagram, Vykreslený graf

Popis byl vytvořen automaticky

The network is relatively sparse (density = 0.25, transitivity = 0.47) and quite clustered (mean clustering = 0.52). The nodes are relatively varied in their degree of clustering, as can be seen in the following plot.

Obsah obrázku text, snímek obrazovky, diagram, Vykreslený graf

Popis byl vytvořen automaticky

### Fictional characters personality ratings

The network consists of 400 nodes. Each node represents an item on the personality questionnaire, but each edge corresponds to the correlation of mean rating of a given fictional character by a large number of people on the given item to the mean rating of the fictional character on a second item. The following plot shows that the node degree follows a unimodal left-skewed distribution.

Obsah obrázku text, diagram, snímek obrazovky, Vykreslený graf

Popis byl vytvořen automaticky

The fictional character personality network is relatively denser than the real personality network (density = 0.57, transitivity = 0.65) and marginally more clustered (mean clustering = 0.66). The following plot shows the range of node clustering is mostly limited between 0.5 and 0.8.

Obsah obrázku text, snímek obrazovky, diagram, Vykreslený graf

Popis byl vytvořen automaticky

## Community detection

Fast community unfolding converged on 2 communities for the fictional character rating network and 4 communities for the self-report personality ratings. Two step walktrap algorithm converged on 4 communities for the fictional character rating network and 43 communities for the self-report network. Finally, label propagation arrives at a single community in the fictional character network and two communities in the self-report network, although here the second smaller community contains only four nodes.

## Node embedding similarity across networks

Node2Vec embedding node similarities within networks had very low, but statistically significant correlation (r= 0.021, p < 0.001). In comparison, two networks estimated from two non-overlapping subsamples of the self-report personality dataset with nodes embedded through the same procedure had a correlation of 0.84 between the pairwise embedding similarities.

# Conclusions

Self-report real personality responses are of a distinctly different structure than 2nd person ratings of fictional characters.

Self-report network is relatively less dense and less tightly connected: answers on the personality questions (nodes) are less strongly connected than for fictional characters.

Explored algorithms for community detection produce differing results across the two networks and between individual methods, the usefulness of each may therefore be strictly limited to the nature of the psychometric network under examination and general statements or recommendations can’t be easily produced.

Finally, node2vec node embedding and the node pairwise similarities show very small correlation (although statistically significant r = 0.021, p < 0.001) corroborating the difference of structure of the two networks despite the substantive similarity of the data-generating process. In comparison, node2vec node embedding similarities from networks that are expected to be highly similar show high degree of association (r=0.84). This indicates that the outlined procedure may be usefully deployed to compare networks estimated from different samples of individuals on the same questionnaire items or measurement units (nodes), to establish a measure of similarity between the network structures.

##### References

1. D. Borsboom, “Possible Futures for Network Psychometrics,” Psychometrika, vol. 87, no. 1, pp. 253–265, Mar. 2022, doi: 10.1007/s11336-022-09851-z.
2. K.-J. Kan, H. de Jonge, H. L. J. van der Maas, S. Z. Levine, and S. Epskamp, “How to Compare Psychometric Factor and Network Models,” Journal of Intelligence, vol. 8, no. 4, p. 35, Oct. 2020, doi: 10.3390/jintelligence8040035.
3. G. Costantini et al., “State of the aRt personality research: A tutorial on network analysis of personality data in R,” Journal of Research in Personality, vol. 54, pp. 13–29, Feb. 2015, doi: 10.1016/j.jrp.2014.07.003.
4. A. Y. See, T. A. Klimstra, A. O. J. Cramer, and J. J. A. Denissen, “The Network Structure of Personality Pathology in Adolescence With the 100-Item Personality Inventory for DSM-5 Short-Form (PID-5-SF),” Frontiers in Psychology, vol. 11, 2020.
5. E. I. Fried, S. Epskamp, R. M. Nesse, F. Tuerlinckx, and D. Borsboom, “What are ’good’ depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis,” Journal of Affective Disorders, vol. 189, pp. 314–320, Jan. 2016, doi: 10.1016/j.jad.2015.09.005.
6. D. Borsboom et al., “Network analysis of multivariate data in psychological science,” Nature Reviews Methods Primers, vol. 1, no. 1, pp. 1–18, Aug. 2021, doi: 10.1038/s43586-021-00055-w.
7. T. L. Rodebaugh et al., “Does centrality in a cross-sectional network suggest intervention targets for social anxiety disorder?” Journal of Consulting and Clinical Psychology, vol. 86, no. 10, pp. 831–844, Oct. 2018, doi: 10.1037/ccp0000336.
8. „Data from the statistical ‚which character´ personality quiz". https://openpsychometrics.org/tests/characters/data/ (retrieved in April 2023).
9. Analytical script and data available at <https://github.com/tomvojtisek/Network-Science-Project>
10. A.-M. Isvoranu, S. Epskamp, L. J. Waldorp, a D. Borsboom, Ed., *Network psychometrics with R: a guide for behavioral and social scientists*. in Research methods and statistics. Abingdon: New York, NY, 2022.