A Standardized Approach to Social Network Data Validation

Neil Yetz, M.P.H., M.S. candidate

Colorado State University

Thesis Proposal

One of the most important concepts in the social sciences is the notion that individuals are part of a larger set of related systems. The social ecological model, common in the social science literature, emphasizes the importance of the social environment as a pertinent level to predicting determinants of health. One way to approach this set of social systems is through social network analysis (SNA). Social network analysis is the process of understanding structures through networks and graph theory through quantitative approaches.

The concept of social networks has been around since the time of Plato in Socrates, when they analyzed the influences of social classes on people. The year 1932 marks the first-time social network theory was used in an applied setting (Borgatti, 2009). Since the early 2000’s and the creation of social network sites such as Facebook and Twitter have provided a path for large quantities of social network data. However, modern utilization of SNA goes well beyond the of social networking websites. Recent literature has social network methods to analyze terrorist networks, tobacco smoking incidence, and social structures in obesity.

The collection of social network data is subject has strict methodological practices. Social network data collection from online social networks (OSN) are often times collected via Application Programing Interface provided by OSN providers (i.e. Facebook, Twitter). These OSN websites automatically collect a wealth of data from their users (Abdesslem, Parris & Henderson, 2012).

Although there are archival methods to social network data collection, survey instruments are often used to assess networks in real time. Survey research requires different approaches to network data collection. There are two common approaches to collecting social network data in survey research:

1. *Rosters:* This method shows survey responders a list of names of individuals within the network. Roster methods require the use of a stem question such as, “To whom do you report to at work?” or “Please select individuals you have a friendship with…”. However, the roster method may only be utilized in networks where the set of potential alters in known. (Butts, 2008). Roster methods are exclusive to a closed network and do not allow the option of anyone entering or leaving the network.
2. *Name Generator/Nomination:* This method gives participants to name anyone or several individuals within a network. The name that may be generated are limitless. A common prompt a participant may see is, “Please indicate five individuals that you would seek advice from within your office…”. Several limitations exist with this method including, false negatives due to subject forgetting or fatigue (Butts, 2008). Errors especially occur in instances where the ego has a large number of connections (Brewer, 2000).

There has been an increase in utilizing SNA as an applied method for research during recent years. Advances in statistical programming have provided researchers the ability to run studies with a strong social network component. Advanced Bayesian modeling techniques such as Exponential Random Graphing Models (ERGM) and *S*imulation Investigation for Empirical Network Analysis (SIENA) models, two common techniques in the SNA literature, require intensive processing power. These techniques are now feasible thanks to advances in advanced statistical programming languages (i.e. R, SAS, Python). Additionally, several research studies focus on collecting social network data because of its implications on health outcomes.

As the popularity of SNA grows, so does its implications for interventions, communities & policy. Interventions utilizing a social network method include enhancing relationships with moderate and severe disabilities (Haring & Breen, 1992), opiate treatments (Day et al, 2018) adolescent drug use (Valente, 2003).

Despite the popularity increase in SN data collection and modeling techniques, no standardized validation techniques for this data type exists. Cole and colleagues (2011) have proposed latent variable approaches to validating social network data. However, a standardized process to validate collected social network data remains unclear.

Fields focused on testing and measurement incorporate a numerous number of methods to ensure validity and reliability exist within the measures they use. Social network data is a measure of relationships and bonds and there is no excuse to ensure the quality. Furthermore, data quality measures exist in many other fields. This thesis project proclaims the assumption that social network data needs to be validated via standardized methods.

The validation of network data comes with its challenges. There are a variety of methods for collecting friendships and bonds. Despite these differing methods, the way social network data is processed is quite similar. Social network data may be represented in terms of an edge list or adjacency table. However, despite the standardized way of representing social network data, there is no standard process in existence to ensure the quality of social network data.

Social science’s dependability on reporting of standardized validity and reliability measures helps to ensure the credentials of the field. Social Network Analysis, an expanding part of the social science field does not follow these same rules. It is naïve to assume that the network data collected by researchers is valid and appropriate for analyses. Even more so, populations of interested may differ in the validity of network data collected. For example, research on adolescent survey self-report is mixed and no analogous studies exist on their ability to self-report social network data. Therefore, because there is no standardized methodology to evaluating social network data, the legitimacy of the field is at risk.

Past research has been performed the validation of the sociometric properties of social networks (Dunn & Westbrook, 2011). Often sociometric techniques are focused on comparing observed network properties to that of randomly generated network.

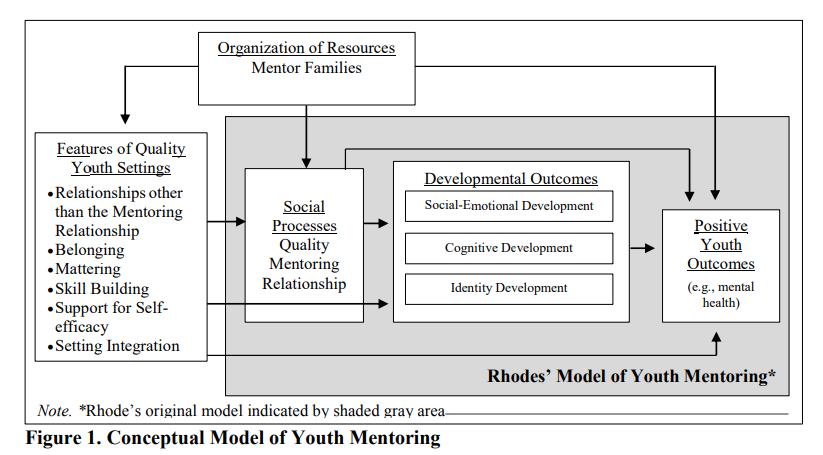
I propose creating a standardized method for validating social network data. My proposal serves as a set of analogous measures to those used in psychometric testing and measurement. Several analogous measures will be tested such as reliability, trustworthiness, and validity levels.

At the end of the project, a series of “checks” will be created to create an understanding of the validity of social network data. This is an essential step to determine the true effects of the network on health outcomes.

**Data Collection**

Data for this thesis project were collected via Campus Connections (CC), a youth mentorship service-learning program at Colorado State University (CSU). Campus Connections trains undergraduate students to mentor at-risk adolescents (i.e., those deemed at risk for not 2 E reaching their full potential due to significant individual and/or environmental risk) aged 11-18 within the local community. Youth and mentors pair up one time a week for four hours per week for 12 weeks. To date, more than 1,500 mentee-mentor pairs have participated in this program to date. Past work indicates that CC is practical, feasible, and in high demand.

Campus Connections utilizes evidence-based practice to promote positive youth outcomes. Along with CC’s desire for evidence-based practice, it utilizes and extension of Rhode’s model of youth mentoring as shown in Figure 1(Rhodes, 2002, 2005).



Results from past research indicate that Campus Connections adolescent outcomes rank similarly to other youth service programs. These effects on positive youth development tends to be relatively modest with some key outcomes not being substantially improved. A unique aspect of CC is the establishment of a *Mentor Family* component. Mentor families involve nesting three to four mentor-mentee pairs into a “family group” to experience the program in unison. These mentor families consist of multiple levels of support. Level one consists of the mentor-mentee dyad. Mentors and mentees establish a closely-knit bonding experience between the mentor and mentee. Level two is the establishment of the mentor family. The mentor family component of CC is based on a

References

Abdesslem, F. B., Parris, I., & Henderson, T. (2012). Reliable online social network data collection. In *Computational Social Networks* (pp. 183-210). Springer, London.

Butts, C. T. (2008). Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, *11*(1), 13-41.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *science*, *323*(5916), 892-895.

Cole, V. T., Hussong, A. M., Faris, R. W., Rothenberg, W. A., Gottfredson, N. C., & Ennett, S. T. (2018). A Latent Variable Approach to Measuring Social Dynamics in Adolescence. *Journal of Research on Adolescence*.

Day, E., Copello, A., Seddon, J. L., Christie, M., Bamber, D., Powell, C., ... & Frew, E. (2018). A pilot feasibility randomised controlled trial of an adjunct brief social network intervention in opiate substitution treatment services. *BMC psychiatry*, *18*(1), 8.

Haring, T. G., & Breen, C. G. (1992). A peer‐mediated social network intervention to enhance the social integration of persons with moderate and severe disabilities. *Journal of applied behavior analysis*, *25*(2), 319-333.

Rhodes, J. E. (2002). Stand by me: The risks and rewards of youth mentoring today’s youth. Cambridge, MA: Harvard University Press.

Rhodes, J. E. (2005). A model of youth mentoring. In D.L. Dubois & M.J. Karcher (Eds.), Handbook of youth mentoring, (pp. 30-43). Thousand Oaks, CA: Sage.

Valente, T. W. (2003). Social network influences on adolescent substance use: An introduction. *Connections*, *25*(2), 11-16.