An assessment of social network data trustworthiness

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

Colorado State University

Thesis Proposal

“One of the most potent ideas in in the social sciences is the notion that individuals are embedded in thick webs of social relations and interactions” (Borgatti et al, 2009). Social science as a field could not exist without social relations. This idea relates directly to the social ecological model (Bronfenbrenner, 1977). The social ecological model is arguably the most heavily utilized model in public health and social science. The social ecological model emphasizes the integration of the social environment and an individual predicting determinant of health. We can approach this integration of social systems and the individual via social network analysis (SNA). Social networks analysis helps to define and measure the connections among people, organizations, and/or other individual units (Valente, 2010).

More specifically, SNA is the process of understanding social structures through networks and graph theory through quantitative approaches (Butts, 2008). The entire realm of SNA is built off social network theory. Social network theory can be applied to a wide variety of levels spanning from the simple connection to two people, up to a collection of people and how those people are integrated in a set of systems (Kadushin, 2012). Social network analysis takes a quantitative aim to explain the relationships between these systems at a micro level (connections of individual units) and macro level (connection of systems) that is emphasized in social network theory.

It is important that we define another term before proceeding: *Network*. What is a network? A network is defined as a set of relationships between objects (otherwise known as nodes) and how they can be mapped (Kadushini, 2012). In social sciences, the term is most commonly referring to people, but the social networks to any set of related objects. Regardless of whether we are talking about people, neuronal connections, or business relationships, the mathematical properties remain the same.

Historically, the understanding of social connection has been around since the time of Socrates, whom theorized about the influences of social classes on people (Borgatti, 2009). However, the quantitative approach to understanding them is relatively new. It was not until the 1960’s that researchers began to theorize how the connections between people may impact the human experience. Milgram (1967) used social network methodology to determine everyone connected via six degrees of separation. The scarcity of social network research in past decades is due to the overall complexity of social networks. Analyzing large scale social networks in early scientific studies was nearly impossible during this time.

Thankfully the number of studies with social network methodologies have increased since the turn of the century. Advances in statistical programming have provided social science researchers an avenue to include strong social network components within their research studies (Luke, 2015). Several cutting-edge social network analysis methodologies require intensive processing power to be performed. Advanced Bayesian modeling techniques such as Exponential Random Graphing Models (ERGM) and *S*imulation Investigation for Empirical Network Analysis (SIENA) models (Ripley et al., 2018; Snijders, 2005) are examples of modeling techniques that require heavy processing power. These techniques are now feasible thanks to advances social computing that increase processing speeds and allow for large datasets (Parameswaran & Whinston, 2007). Social network data is notorious for producing these large datasets.

Even social network data collection, the process prior to analysis, has become more feasible over time. Social network sites such as Facebook and Twitter have provided a path for large quantities of social network data. Online social networks (OSN) are often collected via an application programing interface (API) provided by OSN providers. These OSN websites automatically collect a wealth of data from their users (Abdesslem, Parris & Henderson, 2012). However, modern utilization of SNA goes well beyond the of social networking websites. Research implementing social network methods include identifying characteristics of terrorist networks (Ressler, 2006), obesity (Cohen-Cole & Fletcher, 2008) and adolescent drug use (Valente, 2003).

Social network analysis has additionally been shown to have a useful place in community interventions. For example, Klovdahl (1985) created a social network intervention to identify and prevent HIV outbreaks within a homosexual population. DeLay and colleagues (2015) have used adolescent friendship networks to evaluate the Family Check-up model within adolescent populations. Lastly, social network interventions have been utilized in schools to reduce antisocial and violent behaviors within adolescent populations (Kornienko, Dishion & Ha, 2018). Measures of social network can inform and improve interventions directly. Experimental research by Valente (2003) found differences in tobacco intervention programs that identified group leaders in network analysis.

In summary, the dissemination of SNA information has real world implications that may influence communities. Based on the real-world implications SNA has, researchers need to be held accountable that the data they collect for analytical purposes are valid and trustworthy. However, this is not the case. There is light research concerning this issue of social network data validity. Other social science fields place a strong emphasis on ensuring construct measures are both reliable and valid.

Testing and measurement practices emphasize extreme precautions to ensure their measures are both reliable and valid. Because social science relies heavily on self-report measures it is a necessary component of the literature. Measurements of reliability is among the most emphasized subjects in the social science domain. As with social network survey methodology, many psychological measures depend on self-report. These self-report questionnaires need to be assessed for their reliability through many statistical methods. Internal consistency is mathematically calculated via Cronbach’s Alpha (Cronbach, 1951). Other measures such as test-retest reliability and interrater reliability are commonly reported in the literature as well. Without these measures of

Without these measures of reliability validity of psychometrics metric would cease to exist. Content, criterion, and construct validity all are necessary components to ensuring that self-report measures are capable of measuring the construct of choice. Devellis

Research on testing in employment processes has shown that anywhere from 15% to 40% of provide fake responses (Arthur, Glaze, Villado, & Taylor, 2010; Griffith, Chmielowski & Yoshita, 2007). Novel faker detection methods such as unlikely response patterns detection (Holden et al., 2017) and social desirability scales (Crowne & Marlowe, 1960; Paulhaus 1999) have been utilized in order to retain scale validity and trustworthiness.

Social network research has no excuse to pursue similar practices to ensure the field is taken seriously. Without assessing the trustworthiness of collected data, there is no understanding how valid our conclusions are. Cole and colleagues (2011) have proposed latent variable approaches to validating social network data. Sociometric reliability has also been a topic of discussion. Pepinsky (1949) mentions that there is that there is no guarantee of consistency regarding sociometric data. Contrarily, Harmon (1949) made the exact opposite argument, stating that reliability in sociometric data collection is not necessary because it serves as a description of the real world itself. Terry (2000) later indicated the validity of sociometric methods are of concern because the there is no consistency in question stems (i.e. “Who are your best friends” vs. “Name your friends”).

All the past research takes a focus on understanding data validity by use of the structural components of the network. However, the assumption that actors in a social network study can properly assess their own network is not nearly as researched. Often it is merely assumed that an actor assessing their own network through survey questions may accurately report. Ideally there needs to be systematic checks to determine the quality and trustworthiness of the data that has been collected. However, a standardized process to validate collected social network data remains unclear.

Survey research with a social network component consists of questionnaires that ask about relationships among a specified target group (Serrat, 2017). The questionnaires focused on identifying a network requires careful thought about their methods. There are two common approaches to collecting social network data in survey research:

1. *Social Cognitive Mapping/Roster:* Originally developed by Cairns and colleagues (1997), 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 method:* 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 (an individual node in the network) has many connections (Brewer, 2000).

Both methodologies are notorious for creating enormous datasets. Datasets that are hard to sift through without a systematic and methodical approach.

Using these methods to collect data requires a large assumption: *That respondents can properly assess their own networks*. This assumption is especially important depending on the population being assessed. For example, are adolescents able to properly assess their friendships with their peers? This overarching question is important and scarce in the literature.

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. When social network data is collected, it

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.

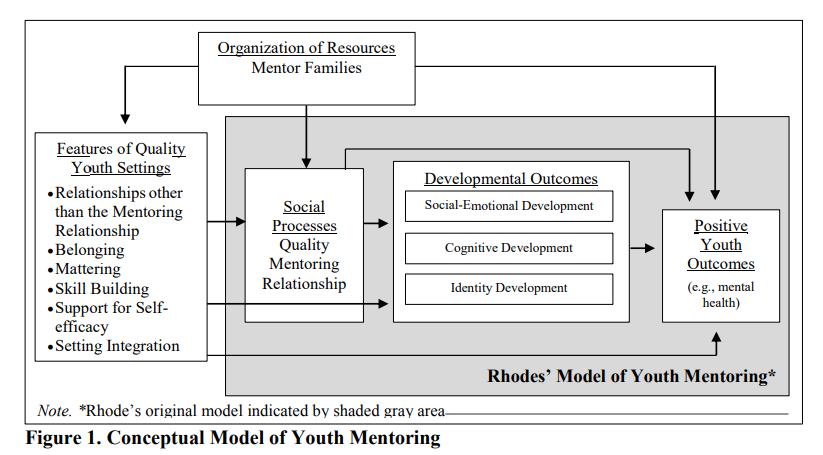
At the end of the project, a series of “checks” will be created to create an understanding of the validity of social network data. This is an essential step to determine the true effects of the network on health outcomes.

Due to the quantitative properties of SNA, a series of quantitative terms are used to represent relationships between people and systems. When these terms are used, we are establishing a mathematical representation of the network as a whole or the individual components of the network. Individual units in the network are referred to as nodes. In SNA, we study the connection and relationships between nodes (Marin & Wellman, 2008). The connecting points representing edges within a network are known as edges.

**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.

Arthur Jr, W., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet‐based tests of cognitive ability and personality. *International Journal of Selection and Assessment*, *18*(1), 1-16.

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.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American psychologist*, *32*(7), 513.

Cairns, R. B., Perrin, J. E., & Cairns, B. D. (1985). Social structure and social cognition in early adolescence: Affiliative patterns. *The Journal of Early Adolescence*, *5*(3), 339-355.

Cohen-Cole, E., & Fletcher, J. M. (2008). Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. *Journal of health economics*, *27*(5), 1382-1387.

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*.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297-334.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of consulting psychology*, *24*(4), 349.

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.

DeLay, D., Ha, T., Van Ryzin, M., Winter, C., & Dishion, T. J. (2016). Changing friend selection in middle school: A social network analysis of a randomized intervention study designed to prevent adolescent problem behavior. *Prevention Science*, *17*(3), 285-294.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review*, *36*(3), 341-355.

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.

Harmon, L. R. (1949). A Note on Pepinsky's Analysis of “Validity” and “Reliability” of Sociometric Data. *Educational and Psychological Measurement*, *9*(4), 747-747.

Kadushin, C. (2012). *Understanding social networks: Theories, concepts, and findings*. Oup Usa.

Klovdahl, A. S. (1985). Social networks and the spread of infectious diseases: the AIDS example. *Social science & medicine*, *21*(11), 1203-1216.

Kornienko, O., Dishion, T. J., & Ha, T. (2018). Peer network dynamics and the amplification of antisocial to violent behavior among young adolescents in public middle schools. *Journal of Emotional and Behavioral Disorders*, *26*(1), 21-30.

Luke, D. A. (2015). *A user's guide to network analysis in R*. London, England: Springer.

Marin, A., & Wellman, B. (2011). Social network analysis: An introduction. *The SAGE handbook of social network analysis*, *11*.

Milgram, S. (1967). The small world problem. *Psychology today*, *2*(1), 60-67.

Paulhus, D. L. (1998). *Paulhus deception scales (PDS): the balanced inventory of desirable responding-7: user's manual*. North Tanawanda, NY: Multi-Health Systems.

Parameswaran, M., & Whinston, A. B. (2007). Research issues in social computing. *Journal of the Association for Information Systems*, *8*(6), 22.

Pepinsky, P. N. (1949). The meaning of" validity" and" reliability" as applied to sociometric tests. *Educational and psychological measurement*, *9*(1), 39-49.

Ressler, S. (2006). Social network analysis as an approach to combat terrorism: Past, present, and future research. *Homeland Security Affairs*, *2*(2).

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.

Ruth M. Ripley, Tom A. B. Snijders, Zsofia Boda, Andras Voros, and Paulina Preciado (2018). Manual for Siena version 4.0. R package version 1.2-12. <https://www.cran.r-project.org/web/packages/RSiena/>.

Serrat, O. (2017). Social network analysis. In *Knowledge solutions* (pp. 39-43). Springer, Singapore.

Snijders, T. A. (2005). Models for longitudinal network data. *Models and methods in social network analysis*, *1*, 215-247.

Terry, R. (2000). Recent advances in measurement theory and the use of sociometric techniques. *New directions for child and adolescent development*, *2000*(88), 27-53.

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

Valente, T. W. (2010). *Social networks and health: Models, methods, and applications*. Oxford University Press.