Strategies for Gathering Social Network Data: Overview, Assessment and Ethics

jimi adams

Health & Behavioral Sciences – University of Colorado Denver

Tatiane Santos & Venice Ng Williams

Colorado School of Public Health – University of Colorado Anschutz Medical Campus

Abstract:

This chapter provides an overview of social network data collection strategies. We begin by outlining the primary principles of sampling and measurement design, then describing how those combine into what is labeled as the “boundary specification problem” for social network research. We accompany these definitions with examples of how these elements are applied across ego-, partial- and complete-network designs. Next, the chapter turns to the primary ways that network data have been evaluated, highlighting both the implications of those evaluations for their use in network analyses, and various strategies for how the identified limitations can be leveraged for optimal data and analytic quality. The chapter concludes by addressing some of the ethical considerations that are unique to the gathering and analyses of social network data.

Keywords: data collection, research design, ethics, data quality

Introduction

For decades, the field of social network analysis was dominated by a popular set of small-scale datasets.¹ Scholars reanalyzed these existing datasets as newly developing methods allowed for refined strategies to answer the original research questions. For example, one of the most widely used datasets—e.g. Zachary’s (1977) karate club data—are so well known that they are used to benchmark new approaches for estimating community detection (see Fortunato 2010).² Commonly available data were also re-analyzed to evaluate new research questions or theoretical reformulations of previous ideas. A prominent example of this can be found in the numerous re-analyses of the Coleman, Katz and Menzel (1957) data on physician prescribing practices. As scholars returned to these data, each subsequent paper reassessed the diffusion pattern of tetracycline among these physicians, with each new paper offering a modified account of the primary diffusion drivers that came in each study before (Burt 1987; Kilduff & Oh 2006; Marsden & Podolny 1990; Strang & Tuma 1993; Van den Bulte & Lilien 2003). As the remainder of this chapter will detail, many best practices for gathering social network data are expensive (in time, effort, and resources), explaining why this and other popular early social network datasets were so frequently re-analyzed.

Over the past few decades, scholars are incorporating network data collection into a rapidly expanding number and type of research projects. Alongside this growth, researchers have also moved to develop, and standardize methods employed in those studies. Fortunately, the field seems to have retained the strong norm of data sharing, meaning that many of these newly emerging datasets are

¹ This is not to suggest that other scholars were not also gathering their own data. Many were. It simply points out how few such scholars there were. And as with any burgeoning field, each employed their own newly developing strategies for gathering that data.

² This is so frequently the case that the NetSci annual meetings have adopted the norm of awarding a karate trophy to the first presentation each year that makes use of these data.

available to researchers beyond those directly engaged in their collection. This expansion has grown to the point that even some individual projects now serve as repositories for hundreds of unique network data sets (Bevc et al., 2015).

With the proliferation and availability of new social network data, the field has identified best practices for a more comprehensive set of dimensions that allow researchers to optimize social network data collection. This chapter provides an overview of three main dimensions—design strategies, assessment and implications of design strategies, and ethical considerations. However, before engaging in any discussion of how to gather social network data, it is imperative to identify what exactly it is that researchers aim to capture when soliciting information on social networks. This particularly matters for how we gather social network data, because different conceptualizations of the ties of interest can fundamentally alter strategies for measuring those relationships.

What is the Goal? Theory's Role in Gathering Network Data

Borgatti and colleagues (2009) provide a means for classifying the different types of network ties most frequently hypothesized to shape outcomes that interest social scientists. They differentiate among three primary types of ties: (1) social relationships, (2) interactions, and (3) flows.³ **Social relationships** capture particular relationship-based roles that people occupy (e.g., friendships, kinship ties, affective relationships, etc.).⁴ These types of ties are frequently thought to exhibit a number of characteristics that enhance our ability to measure them—including that members of the relationship can easily perceive and report about them, and are often enduring (rather than fleeting) in time, allowing for

³ In the paper, they also present a fourth type of "relationship" often modeled as a social network—similarities (e.g., being the same gender, from proximate locations, or having similar attitudes). We leave these out of the discussion here, because those are better conceptualized as similarities that are modeled as social relationships, not actually *representing* social relationships. That is, these can be captured as individual attributes with the represented relationships then estimated analytically—i.e., is outside the scope of this chapter. Moreover, these types of ideas require a different theoretical framework and subsequent analytic strategies that apply to the types of relationships addressed here.

⁴ Networks can also represent the relationships among entities other than people. Here, we discuss ties connecting people, but encourage readers to recognize that this is merely a shorthand to facilitate ease of writing.

stable measurement. **Interactions** capture the types of (frequently momentary) activities in which pairs of people jointly participate (having sex, sharing resources, sending and receiving messages, etc.).⁵

While social relationships differ from interactions conceptually, they also differ in the approaches taken for their analyses—e.g., between models that focus on more stable *relationships* (e.g., exponential random graph models or stochastic actor based models) versus those that focus on the sequential structural patterns among more momentary *interactions* (e.g., relational event models). Finally, **flows** capture the actual spreading or contagion of various things (ideas, diseases, resources, patient transfers) between the nodes in a network who are connected by relational or interaction ties.

These varied conceptualizations each carry quite different implications for what types of ties researchers aim to measure, and therefore how we should go about capturing each type. Even when a study's aims clearly aligns with one of these conceptualizations, and clearly indicates an appropriate strategy for gathering the appropriate data (e.g., the transmission of an STD), practical limitations of data collection often requires that researchers' theoretical aims can only gather empirical proxies that approximate the actual processes of interest (e.g., using self-reported sexual partnership data, perhaps combined with biomarkers of disease status at multiple time points). Moreover, when theory and methods can align on the pertinent tie types, the methods of data collection frequently cannot match the precision needed to actually capture all of the relevant dynamics from a population of interest. For example, even if researchers can accurately capture behaviors within sexual partnerships, it likely is not possible to do so with the temporal granularity to distinguish the discrete act in which transmission actually takes place from other non-transmitting interactions between the same partners. This requires researchers to carefully identify potential sources of conceptual slippage between the concepts of

⁵ While social relationships and interactions can involve groups, to gather and represent these collective experiences, we record the collection of dyadic relationships/interactions occurring within that group.

theoretical interest and their actual data and analytic capacities.⁶ In addition to these conceptual comparisons, later in the chapter we also describe existing strategies for—and common findings from—directly assessing the quality of network data.

Following longstanding norms in the literature, we orient much of our discussion below of data collection strategies around approaches developed for survey-based approaches, which primarily address gathering social relationship data (Marsden 2011). Most of these considerations carry over into gathering data on other tie types, and we describe the necessary adaptation of those concepts for other approaches where appropriate.

Design Strategies for Sampling and Measurement

Researchers are faced with three primary design considerations when gathering social network data, regardless of which type of tie(s) they are interested in capturing. First, which actors should be included? Second, which of those actors' ties will they aim to capture? Third, how will those relationships be measured?

The “Boundary Specification” Problem

Generally in social science, sampling is addressed solely as a question of which individuals within a population should be targeted for inclusion in a study. When the analytic focus shifts to relationships, which are the foundation of social network analysis (Berkowitz 1982), this sampling question must simultaneously bound which individuals *and* relationships will be included in the analysis—a consideration labeled the “boundary specification problem” (Lauman et al., 1994).

Broadly, there are three primary strategies employed to define the boundaries of network based study designs. **Complete** network designs begin by enumerating an entire population of interest, acquire a saturated sample from that population that represents a full census of respondents, then

⁶ Data proxies are often the coin of the realm across the social sciences, so this alone is not unique to social network data. However, the ramifications of such proxies can often be more pronounced in social network data (see for example the discussion of tie reliability and validity below).

identify all of the relationships that exist among members of that population. Complete network designs typically are most effective in populations that are easily bounded (e.g., formal organizational memberships), and provide salient boundaries that constrain the majority of relationships to exist among members of that bounded population (e.g., schools or classrooms). However, just because boundaries can be clearly defined does not necessarily mean that they provide the social boundaries that constrain social relationships. While some members can be clearly demarcated as within certain boundaries, their relationships may—or may not—be similarly contained within those boundaries (e.g., out-of-school friendships). This is where network scholars have repeatedly noted that the nature of the boundary specification problem is uniquely multilevel in social network studies; we must simultaneously consider how fully the population and relationships of interest are contained within the demarcated boundary (Laumann et al., 1994; Morris 2004).

On the other end of the spectrum, **ego network** designs investigate networks primarily as characteristics of the relationships surrounding particular individuals (see Smith chapter). Because of this focus on individuals and their relationships, for ego network designs, sampling considerations generally follow relatively standard individual level procedures. As a result, ego network designs can be included in other study designs, adding personal network information to the other individual-level data that are gathered, such as from nationally-representative, population-based samples. For example, the General Social Survey has periodically included a network module (Burt 1984).

In addition to the opposing extremes of the approaches described above for determining study boundaries, many studies take a more blended strategy labeled a “partial network” design (Morris 2004). This design relies on a **link tracing** approach, wherein the researcher first identifies a sample of “seed” respondents, then constructs a strategy to subsequently recruit some portion of each ego’s nominated alters to be recruited as respondents themselves. This pattern can then be repeated as many

or as few times as desired. While this strategy has similarities to general snowball sampling, the additional waves of recruitment in link tracing designs are typically more systematically determined (see McCormick's chapter). For example, epidemiologists developed many of the earliest, and most sophisticated, versions of these designs to trace infectious disease exposure and transmission (Klov Dahl 1985).⁷

Beyond these different conceptualizations for which people and ties to include, Laumann and colleagues (1994) also distinguish between strategies for who can make those boundary definitions. That is, even if researchers determine their study requires a "complete" network study of members within an organization, different people may have divergent perspectives on who are appropriately identified as members. The *realist* approach to boundary specification relies on people within the population of interest to subjectively determine who should be included or excluded from consideration as members of an organization (Laumann et al., 1994). A *nominalist* approach begins with some pre-determined identification of the population boundary—e.g., via a membership roster. It is important to point out that either of these boundary specification approaches merely identifies which members and ties *potentially* could be included in the study,⁸ and is a separate question from the question we turn to in the next section: which ties are actually reported on.

Each of these approaches come with their own potential benefits and drawbacks. The prominent National Longitudinal Study of Adolescent Health (Add Health) demonstrates some of the potential limitations that arise even in seemingly well-bounded networks. Bearman and colleagues' (2004) study of romantic and sexual networks found that even when they asked students to identify

⁷ Adaptations of such link tracing designs provide the backbone of study designs used to identify and study hard to reach populations for whom no appropriate sampling frame exists—e.g., respondent driven sampling (Heckathorn 1997; Salganik and Heckathorn 2004). However, while RDS studies sample over networks, they rarely actually include sufficient network data to analyze network structure within those populations (for an exception, see Khan et al., 2015).

⁸ As is frequently highlighted in SNA literature, compared to those that are present, ties that are absent—but possible—are frequently as important (if not more so) for understanding the network structure and its implications within a population.

their partners from a roster of other students attending their school,⁹ approximately half of the indicated partnerships were outside the school. Given these permeable boundaries, those ties outside the school had broad potential implications for the study's aims to understand how these networks shaped population-level diffusion potential and determinants of individual risk for sexually transmitted infections.

Name Generators - Which Relationships?

In the context of a social network survey, once researchers have a strategy to identify respondents, they must consider what number and type(s) of ties they will ask about, and what information they will gather about those relationships. In network studies, these distinctions typically breakdown into selecting the name generators and name interpreters that will be used. Simply, a **name generator** identifies which partners (often referred to as **alters**) will be elicited within a study, based on a specified relationship type. Traditionally, social network studies have often focused on a single name generator to elicit information on one relationship type at a time. In this case, researchers rely on theory to carefully identify the most appropriate relationship on which to elicit information.

Perhaps the single most frequently used name generator is the General Social Survey's "important matters" question.¹⁰ Prompts like these allow respondents to name partners with whom they share the specified relationship. Name generators like the "important matters" question leave substantial room for respondent interpretation, leading some to question the precision in its resulting data (Bearman and Parigi 2004). Research has repeatedly shown that more specific name generators solicit more reliable, precise, and accurate data (Brewer et al., 2005; adams and Moody 2007). However, others note that the GSS "important matters" question was used primarily to elicit respondents' salient

⁹ While the prompt focused on in-school partnerships, students had the option to name partners outside the school.

¹⁰ Specifically, its prompt asks respondents to name "the people with whom you discussed matters important to you" (Burt 1984).

relationships from which they could potentially draw social support and regular interaction. Given these aims, some research has shown that questions open to respondent interpretation can appropriately prompt responses of their salient personal relationships, even if relying on substantially different mental maps of what the question means while formulating their responses (Brashears 2014).

Practically, using a single name generator reduces both interviewer and respondent burden, but may reduce the specificity and/or the range of relationships successfully captured (Marin and Hampton 2006). These tradeoffs raise the important question of whether a single name generator is adequate for a study's aims. Ultimately, there is no blanket answer to this question; the recommended best practices have typically relied on a study's theoretical motivation—that is, the “ideal” network should include all the individuals or organizations who have a role in the process of interest (e.g., public policy decision-making, quality improvement, etc.); while excluding individuals who have no bearing on the process (Burt et al., 2012). An increasing number of studies show that for many questions, a single name generator cannot replace the same coverage provided by multiple name generators (Klofstad, McClurg, and Rolfe 2009; Marin and Hampton 2006). For example, a study of physicians sought to limit respondent burden by identifying the minimal set of name generators needed to sufficiently characterize networks involved in quality improvement (QI) intervention and dissemination (Burt et al., 2012). The researchers found that single name generators insufficiently enumerated the ties that are relevant to QI network processes. Other work highlights the importance of multiple name generators for spanning the different relational domains pertinent to individuals' lives (McCarty et al., 2007).

Whether employing one or multiple name generators, all studies must consider a number of additional factors regarding how to employ the selected name generator(s). A first consideration is whether respondents are provided with a roster of other members of the population from whom to identify potential alters, or are asked to free recall their names. Employing a free recall approach has the

potential to under-represent respondents' personal network size, due to limits on memory and recall (Brewer 2000). For example, in a longitudinal study of older adults, after respondents named confidant alters via free recall, researchers then prompted respondents with alters they named in previous waves of the study to follow-up on their exclusion from subsequent rounds (Cornwell et al., 2014). This study went on to show that these exclusions both represented real network change (i.e., former confidants who are no longer partners), and memory errors (i.e., people the respondent confirms as still being a confidant, but only after prompting).

A second consideration is how many alters about whom the researchers aim to elicit information for each name generator. Are they going to cap the number of alters about whom they elicit information? Many studies limit respondents to a small number of nominations (e.g., 3 to 5, see for example Merluzzi and Burt 2013), either because small numbers are thought sufficient to capture the theoretically salient relationships, or because of the burdens that each additional named alter implies for follow up questions (Paik and Sanchagrin 2013). Caps can either be implicit (i.e., only known to the researcher) or explicit (i.e., mentioned to respondents, perhaps in the question script). Research shows that explicitly identified caps can have a range of unintended consequences. Mentioning a number could introduce floor effects that artificially inflate the number of named partners,¹¹ which could alter a study's analytic findings. For example, when Add Health asked respondents to name their five best male *and* female friends (Bearman et al., 1997), it may have inadvertently reduced the amount of gender-based homophily (Kandel 1978; McPherson et al., 2001) observed in that population by encouraging respondents to name more opposite sex friends than they would have without such an explicit numerical anchor in the question prompt. If respondents or interviewers interpret a number

¹¹ This was the explicit protocol for the important matters question in the General Social Survey. If respondents named fewer than five alters, interviewers were instructed to probe respondents for any additional alters (Burt 1984).

included in the prompt as a cap on the number of alters to be elicited, the population degree distribution could be artificially truncated (by stopping reporting or recording nominations above the cap), thus reducing both the observed range and variability within the population.

Name Interpreters - Information about Identified Social Ties

Most studies follow name generator question(s) with **name interpreters**. Name interpreters ask respondents to provide additional information about each alter nominated in their responses to the name generator question(s). Typically, these focus on: (1) attributes of the nominated alters, (2) details of the relationship between the respondent and each alter (e.g., strength, type, or frequency of the relationship), and (3) the respondent's estimate of relationships that exist among their alters (e.g., "How well does Person A know Person B?" for each pair of alters).

For most network-based investigations, we are not only interested in the number of ties that individuals have, but also attributes of those alters to whom they are connected. These attributes can be used to examine the composition of networks (e.g., how many, what proportion, etc. of one's alters have particular characteristics), or the (dis)similarity between ego and alter attributes (e.g., the well-known network pattern of homophily, McPherson et al., 2001). This aim leads to most network studies including some strategy to gather information on alters' attributes.

Any time alters are included among a study's respondents (e.g., in complete network designs, and once recruited into a partial network design), their attributes can be gathered from these respondents themselves. In the case of ego network designs, respondents' reports on their alters' attributes are likely the only source for obtaining such information. Moreover, sometimes, even when those alters are included among a study's respondents, it can still be valuable to get a respondent's perception of their partners' attributes (e.g., as validation, or to model perception vs. reality). However, as with many aspects of network data collection, respondent burden rapidly increases: each additional name interpreter must be asked for each alter identified by the name generator. Therefore, when

gathering network data, it is often thought to be even more important to optimize the efficiency and appropriateness of name interpreter items (Young et al., 2016), by avoiding the collection of irrelevant, redundant, costly, and/or time-consuming data (McCarty et al., 2007).

In addition to alters' attributes, studies also frequently rely on egos to provide additional information about the relationships identified with the name generators. For example, in Add Health, respondents were asked to describe whether they had engaged in a selection of activities (talked on the phone, visited their house, etc.) for each of their named friends (Bearman et al., 2004). These tie-based interpreters can either be considered as separate (behavioral) network questions to be addressed independently (e.g., by analyzing the patterns of phone communication networks among identified friends), or more frequently as an opportunity to interpret how the networks elicited by the name generators should be used in analyses (e.g., by considering how phone communication fosters friendship ties). Alternatively, these relationship-based name interpreters can allow researchers to investigate how respondents potentially differentially interpret the same question prompts (see Bearman and Parigi 2004).

In summary, network data collection design strategies entail the three primary components described: (1) delineating which nodes and which ties will (and will not) be of interest to a study—frequently described as the "boundary specification problem," (2) determining which set(s) of relationships on which to gather information—referred to as "name generators" in survey approaches, and (3) attaching additional information about the nodes and ties under study—referred to as "name interpreters" in survey approaches.

While the descriptions above focus on survey approaches for eliciting network data (i.e., by asking some ego to report on themselves and the ties they have to a specified set of alters), data collection strategies have rapidly evolved to incorporate a variety of other strategies (for example, see

the chapters in this volume by Kitts and Hogan, or Dominguez and Hollstein 2014). These other strategies can range from relying on external sources of information on individuals (e.g., from archival sources) that can be attached as node attributes, to passive behavioral monitoring, e.g., recording behavioral interaction data from automated sensors (Eagle, Pentland and Lazer 2009; Salathe et al., 2010). While each of these strategies differ in the ways that data are actually collected, all of the considerations described above apply equally to any of these strategies. For example, boundary specification is an issue that applies to data collected by survey or by sensor. Moreover, since they were collected for other purposes, archival or passive data sources' inclusion criteria have the potential to differ in meaningful ways from the analytic aims of any study in which they are used, and researchers should carefully consider how the differences between their aims and the qualities of available data may alter how they interpret any subsequent analytic results.

Data Quality & Assessment: Did we Capture what we Intended to Capture?

Once gathered, social network data often provide unique capabilities for assessing their quality (Wald 2014). Here we describe some of the: (1) primary ways network data has been evaluated, and common descriptive patterns found across these assessments, (2) implications of these types of assessments for the analytic capabilities of network data in other studies, and (3) existing strategies to overcome the limitations identified in the first two questions.

Tie Reliability and Validity

Because social network ties—by definition—involve two different actors, frequently the same information can potentially be reported by both actors involved—and as noted above, sometimes even by their partners (see an additional perspective on this in the section on cognitive social structures below). This allows direct comparisons of how consistently the same ties are reported among these various reports, providing the capacity to assess data reliability—if not validity—of the gathered tie information. For example, Adams and Moody (2007) evaluated how consistently partners reported

sexual and needle-sharing partnerships in a project examining HIV risk (Potterat et al., 2004). This analysis found broad general agreement among partners when reporting on their own behaviors—which improved when properly accounting for the temporal specificity with which respondents were asked about their partnerships (adams and Moody, 2007).¹²

This type of analysis builds on a long research tradition, which raises a number of questions about how reliably people report partnership information (see e.g., Killworth and Bernard, 1976; Bernard, Killworth & Sailer, 1979). But as exhibited by the example above (adams and Moody 2007), tie reports that are more salient (e.g., strong vs. weak relationships) and more precisely elicited ties (e.g., with specific time-bounding of relationship reporting windows) are more likely to be reported consistently (see also Brewer, 2000; Brewer and Webster 1999).

Implications & Quality Assessment

Researchers are increasingly incorporating the types of evaluations described above into their data collection efforts (Phillips et al., 2017; An and Schramski 2015). But beyond simply describing patterns of data fidelity, researchers are increasingly demonstrating how any such imprecisions in relational data (e.g., mis-identifying characteristics of alters) can influence the interpretation of network patterns from the subsequent data (Young et al., 2016). Different strategies for handling partner disagreements on relationships can lead to altered estimates of a relationship's existence, duration, and content (adams and Moody, 2007; Phillips et al., 2017). Such variability between partners' reports has proven important in some contexts for estimating behavioral implications for population-level HIV-relevant risk behaviors (Helleringer et al., 2011), but other researchers have shown that estimates of an influenza

¹² That is, simply because two people disagree about whether they shared a particular relationship does not necessarily indicate disagreement. In studies where partners are asked to only report on partnerships that occurred within a particular timeframe (Colorado Springs respondents were asked about partnerships occurring in the previous six months; Potterat et al., 2004), if the interview & relationship windows are misaligned, seeming disagreements can both be accurate. For example, suppose I report today about a partnership I had six months ago that only lasted a few weeks. If that partner is interviewed a month from today, they could accurately leave me out of the partnerships they report within the six months prior to their interview.

epidemic can be quite robust to how this variability in reporting is incorporated into their models (Potter et al., 2015). As such, the utility of such reliability and validity assessments is best understood in the context of the data's intended applications rather than solely as a stand-alone question.

Beyond simply considering network reporting inconsistencies as a nuisance to accurate representation that hampers accurate analytic interpretations, other researchers have attempted to leverage these incongruences as theoretically informative. For example, in populations where preferential attachment processes are at work, we find high correlation between received nominations and most centrality indicators (Smith et al., 2017). However, because of recall limitations and/or alter list truncation, the most popular actors cannot reciprocate all of the nominations they receive. Grippa and Gloor (2009) highlight how this greater likelihood for high-degree actors to have unreciprocated nominations can be leveraged as an independent estimator of these individuals' centrality (on the underlying complete graph, even if it is not observed), or their reputation and informal leadership within the group.

Strategies for Optimizing Data Fidelity

A variety of strategies have emerged for resolving discrepancies between partners' reports on the same relationships. The most straightforward possibilities are either to use the intersection set, which only counts a relationship as existing when multiple reports *both* agree on the nature of the relationship, or using the union set, which counts a relationship as existing if *either* of the members of a relationship describe its presence. When choosing only between these two options, researchers more commonly rely on union-set logic (Brewer and Webster 1999). The rationale for this stems from studies often precluding one partner's potential to report on a tie's presence (e.g., through non-response), and findings from concordance comparisons in empirical literature that "false negatives" are a much more common reason for discrepancies. That is, researchers choose to include ties that are reported by either

partner, and do not take the other partner's failure to confirm the nomination as necessarily indicating it is not present (Brewer 2000).¹³

While such simple (and mostly a priori) decisions about tie inclusion/exclusion are common, recent work has suggested the probabilistic estimation of tie likelihoods as an alternative strategy for resolving inconsistencies between tie reports (Butts 2008). One example of this strategy, requires computing a score of each actor's "credibility," then applying this credibility as a weight to each respondent's respective report of a tie (An and Schramski 2015). This allows for model-based estimation of multiple possible graph representations for a single network, derived from a single set of survey responses.

Cognitive Social Structures

David Krackhardt (1987) introduced a different conceptualization for capturing social network data—known as "cognitive social structures" (CSS)—wherein instead of asking people to only report on the ties they are personally involved in, researchers ask *all* members of a group to estimate the patterns of relationships among the *entire* population. For example in Krackhardt's (1987) initial introduction of the idea, he shows how CSS can provide an improved prediction of individuals' performance in a small management firm. While generally thought to provide a different view of networks' capabilities in a population, it has also been used as a means to fill in missing data, or provide additional means for adjudicating between conflicting reports. Neal (2008) shows how aggregating across the complete set of "perceptual networks" provided by CSS, can provide a single aggregate consensus structure in a classroom setting. This consensus structure provides valued information on a tie's presence and salience. She shows how this consensus structure overcame some of the limitations of relying solely on the self-reported network, by more completely capturing the classroom's clustering into separate

¹³ This approach is also consistent with the assumptions underlying analyses of ego network data—that individuals can accurately report on their own behaviors, without the possibility for assessing corroboration by their partners.

groups. In part this improvement of fit for describing the global pattern arose from smoothing over some of the potential information loss derived from unreciprocated tie reports—particularly by reducing the likelihood of "false negative" reports (Neal, 2008).

Unique Ethical Considerations of Network Data

Ethics in Data Collection

The same general principles used in the social sciences for the protection of research participants also apply when gathering network data. However, to follow the *intent* of these principles and not just the prescriptions developed to uphold them in practice requires some additional considerations beyond standard protocols for network data. A special issue of *Social Networks* (Breiger, 2005) was devoted to identifying, and making recommendations for handling, these unique considerations. It should be noted that though this issue is now more than a decade old, and some of the pragmatic solutions to the issues raised in it have since changed. However, it still provides useful elaboration of the main set of unique ethical factors that network studies should consider. Here, we begin with those considerations that stem directly from the Belmont Report's (National Commission for the Protection of Human Subjects, 1979) primary ethical protections—those of minimized risk and informed consent.

Charged with minimizing the risks to their research participants, social scientists are frequently primarily concerned with providing respondents with assurances of either anonymity or confidentiality. Anonymity entails not gathering information from research participants that could potentially be personally identifying. In practice, for all but the simplest ego network designs, anonymity is often not possible for the analytic aims of social network research; to analyze the connections among members of a population, we need to know which members of the population the data represent (Borgatti and Molina 2003). Alternatively, confidentiality ensures that while researchers will know (or have the capacity to know) research participants' identity, they will report all data and analytic findings in a way

that does not convey any personally identifying information. Confidentiality is more viable for network studies, but either strategy is simpler in principle than they are in practice.

A burgeoning literature has identified the potential problems of **deductive disclosure**. This work shows that with enough information, individuals' identities can be deduced from de-identified data, even in some cases where researchers began with truly anonymized data. Computer scientists have devoted considerable attention to these questions in the domain of "big data." For example, Narayanan and Shmatikov (2009) gathered an anonymous graph of several thousand Twitter users, which they then aimed to (re-)identify using auxiliary information from a complementary dataset of Flickr users. Using only features of the network from the Flickr dataset, they were able to successfully re-identify 30.8% of sampled Twitter user pairs. The vast majority of the 12.1% rate of mis-identified false positives from their automated methods matched to a person in the same geographic location and or only one step removed from the true match, which could mostly be manually corrected.¹⁴ This example highlights how even strict data anonymization protocols do not sufficiently protect participant's identities, and often times network-based information (which may be readily, and publicly available) can facilitate this identification. In other words, while deductive disclosure is increasingly a concern for all human subjects research, network data may be particularly susceptible to the potential limitations of standard de-identification procedures.

The Belmont Report has also been used to ensure that human subjects research participation is voluntary and includes informed consent—requiring that participants know the potential risks involved before agreeing to be involved. This dimension of human subjects protections has occasionally proven particularly perplexing for institutional review boards (IRBs) when evaluating social network research.

¹⁴ The remaining 57% were not matched to any Twitter users from their sample. The methods used do not allow them to assess whether these were non-Twitter users, Twitter users outside their sample, or Twitter users inside the sample that were simply not matched.

Given that network designs often request participants to report on the identity, characteristics—and even the relationships—of their partners, IRBs have debated whether these alters should be classified as "secondary subjects" (Marsden 2011), and therefore whether their informed consent is required to gather and/or retain such data (Morris 2004; Sönmez et al., 2016).

Researchers have devised a number of strategies to address the privacy concerns for such secondary subjects. A first strategy is to assess the actual risks of their inclusion. If it can be established that the inclusion of these alters' information would provide no more potential risks than they encounter in their daily lives, some IRBs have concluded that it is not necessary to obtain their consent (Klov Dahl 2005). Another approach leverages the fact that many network studies attempt to collect and analyze data on entire populations, not just a sample from it. In this case, recruitment strategies can themselves directly address the consent for everyone in the population. For example, Add Health employed an active consent procedure whereby every student within the recruited schools were asked for their consent to be included in the study (Bearman et al., 1997). The research team then generated the rosters from which respondents nominated friends and romantic partners to only include those who gave their consent to be included.¹⁵ Some researchers have successfully argued that active consent is an undue burden on researchers when studies involve minimal risk, leading IRBs to allow passive consent—that is notifying the entire population of the aims and potential risks of the study, then only excluding information from being collected on those who opt out (Lorant et al., 2015).¹⁶

Network-based sampling approaches—like those that comprise the partial network designs described above—have devised strategies that allow for the anonymized matching of nominated

¹⁵ In fact, Add Health's consent procedure was a two-stage active consent, in that both the students themselves *and* their parents had to agree to the students' inclusion in the study.

¹⁶ However, passive—or even assumed—consent has been criticized as potentially inconsistent with standard interpretations of the "Common Rule." For example, see the discussion (e.g., Fiske and Hauser, 2014; Kahn et al., 2014) regarding a massive Facebook experiment on emotional contagion (Kramer et al., 2014).

partners. For example, respondent driven sampling (RDS) uses a staged procedure which: (1) asks respondents to report only on characteristics of their alters (rather than identifying them), (2) then provides respondents with uniquely-identified tokens that respondents give to those described partners, (3) who in turn opt-in to participating in the study by returning the token to the study team (Gile and Handcock 2010; Wejnert and Heckathorn 2008). These RDS recruitment chains can then be linked via the tokens' unique identifiers, and after the linkage is made, the identifying information can be removed prior to data storage and analysis (Sönmez et al., 2016).

Ethics in Data Analysis and Presentation of Results

Sharing research results with the population under study has become an increasingly common practice.

It can simultaneously allow the population to inform the interpretation of study results, and facilitate their ability to benefit from its findings (Wallerstein and Duran 2006). Kadushin (2005) adapts these ideas for social network data. He notes the challenges of potential participant identifiability (especially in smaller networks), therefore recommending that findings reports are generalized in a way that removes identifiable features. This de-identification may necessitate going beyond simply removing individual's characteristics. For example, presenting a network visualization in some cases could reveal *any* individuals as uniquely identifiable from their position, which could be extrapolated to identify other actors they are (directly or indirectly) connected to.

As with any research, it is also important to evaluate how findings could be used. Network studies present a number of unique usage considerations. For example, in surveillance of infectious diseases, the ability to track potential transmission through a population may require modest risks stemming from individuals' partner identification, but these may be outweighed by the potential population level benefits of slowing its spread (Klov Dahl 2005).

Management and business consulting has increasingly incorporated network analysis into evaluating and optimizing firm performance. These types of analyses have been used to determine vital

outcomes for employees in those organizations (e.g., promotion, salary and termination decisions). In these cases, while traditional informed consent procedures can allow participants' awareness of potential risks, assurances of confidentiality, and how voluntary their participation is (Borgatti and Molina 2003), they may not sufficiently protect participants from detrimental (or biases in beneficial) outcomes. As a result of these potential real-life harms from organizational network analyses, researchers have crafted careful agreements with these partner organizations, which control who will see the data, how it could be shared, and what ramifications may arise if someone else sees a respondent's information (Borgatti and Molina 2005). Some organization-based researchers even recommend explicitly executing contractual agreements with the studied organization's leadership to ensure that the researcher maintains control over all data ownership and presentation, so that individual participants are protected from potential harms (Borgatti and Molina, 2005; Kadushin, 2012).

In sum, appropriately protecting human subjects in social networks research designs frequently requires the researchers to take additional steps beyond those common to social science—even to achieve the same goals.

Summary

This chapter has overviewed social network data collection strategies, focusing particularly on study design, data assessment, and ethical considerations. As with most of the topics in this handbook, these and related questions could warrant a standalone course, whereas it's often a single module in larger courses on SNA.

Gathering social network data requires a different set of orienting principles than governs general strategies for data collection in the social sciences. For sampling, this requires considering relationships in addition to people. For measurement, this means carefully weighing the variety of theoretical and pragmatic tradeoffs that arise from the number of name generators and name

interpreters to be employed. Moreover, as other opportunities for gathering network data become increasingly common (e.g., from passive collection or organic data sources), these same considerations should be used to identify what datasets include. Simply, "big data" are not immune from the need to carefully factor in how these same choices shape what data represent.

Network data, and the design strategies used to gather it, uniquely provide a variety of possibilities for descriptively assessing the quality of data, and in turn how that quality alters the capacity of analyses to which network data are employed.

Simply applying the human subjects' protections from general social science protocols are not sufficient in network studies to maintain the aims of minimized risk and informed consent that motivate research protections. In particular, this must be carefully evaluated as research fields continue to embrace ideals of open science and data sharing, aims shared historically within the field of SNA.

References

- adams, jimi, and James Moody. 2007. "To Tell the Truth? Measuring Concordance in Multiply-Reported Network Data." *Social Networks* 29:44-58.
- An, Weihua, and Sam Schramski. 2015. "Analysis of Contested Reports in Exchange Networks based on Actors' Credibility." *Social Networks* 40:25-33.
- Bearman, Peter S., Jo Jones, and J. Richard Udry. 1997. "The National Longitudinal Study of Adolescent Health: Research Design." University of North Carolina.
- Bearman, Peter S., James Moody, Katherine Stovel, and Lisa Thalji. 2004. "Social and Sexual Networks: The National Longitudinal Study of Adolescent Health." Pp. 201-20 in *Network Epidemiology: A Handbook for Survey Design and Data Collection*, edited by Martina Morris. Oxford University Press.
- Bearman, Peter, and Paolo Parigi. 2004. "Cloning Headless Frogs and Other Important Matters: Conversation Topics and Network Structure." *Social Forces* 83(2):535-57.
- Berkowitz, Steven D. 1982. *An Introduction to Structural Analysis: The Network Approach to Social Research*. Butterworths.
- Bernard, H. Russell, Peter D. Killworth, and Lee Sailer. 1979. "Informant accuracy in social network data IV: a comparison of clique-level structure in behavioral and cognitive network data." *Social Networks* 2(3):191-218.
- Bevc, Christine A., Jessica H. Retrum, and Danielle M. Varda. 2015. "Patterns in PARTNERing across Public Health Collaboratives." *International Journal of Environmental Research and Public Health* 12(10):12412-25.
- Borgatti, Stephen P., Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. 2009. "Network Analysis in the Social Sciences." *Science* 323:892-95.
- Borgatti, Stephen P. & Molina, José Luis. 2003. "Ethical and strategic issues in organizational social network analysis." *Journal of Applied Behavioral Science*, 29(3), 337-349.
- Borgatti, Stephen P., and José-Luis Molina. 2005. "Toward ethical guidelines for network research in organizations." *Social Networks* 27(2):107-17.
- Brashears, Matthew E. 2014. "'Trivial' Topics and Rich Ties: The Relationship between Discussion Topic, Alter Role, and Resource Availability Using the 'Important Matters' Name Generator." *Sociological Science* 1:493-511.
- Breiger, Ronald L. (ed.) 2005. "Ethical Dilemmas in Social Network Research." Special Issue of *Social Networks* 27(2):89-168.
- Brewer, Devon D. 2000. "Forgetting in the Recall-Based Elicitation of Person and Social Networks." *Social Networks* 22:29-43.
- Brewer, Devon D., J. Poterat John, Stephen Q. Muth, Patricia Z. Malone, Pamela Montoya, David L. Green, Helen L. Rogers, and Patricia A. Cox. 2005. "Randomized Trial of Supplementary Interviewing Techniques to Enhance Recall of Sexual Partners in Contact Interviews." *Sexually Transmitted Diseases* 32(3):189-93.
- Brewer, Devon D., and C. M. Webster. 1999. "Forgetting of Friends and its Effects on Measuring Friendship Networks." *Social Networks* 21:361-73.
- Burt, Ronald S. 1984. "Network Items and the General Social Survey." *Social Networks* 6:293-339.
- Burt, Ronald S. 1987. "Social Contagion and Innovation: Cohesion versus Structural Equivalence." *American Journal of Sociology* 92:1287-1335.
- Burt, Ronald, Meltzer, David O., Seid, Michael, Borgert, Amy, Chung, Jeanette, Colletti, Richard B., Dellal, George, Kahn, Stacy A., Kaplan, Heather C., Peterson, Laura E., and Margolis, Peter. 2012. "What's in a name generator? Choosing the right name generators for social network surveys in healthcare quality and safety research?" *BMJ Quality and Safety* 21:992-1000.

- Coleman, James S., Elihu Katz, and Herbert Menzel. 1957. "The Diffusion of an Innovation Among Physicians." *Sociometry* 20(4):253-70.
- Cornwell, Benjamin, L. Phillip Schumm, Edward O. Laumann, Juyeon Kim, and Young-Jin Kim. 2014. "Assessment of Social Network Change in a National Longitudinal Survey." *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, 69(8):S75-S82.
- Dominguez, Silvia, and Betina Hollstein. 2014. *Mixed Methods Social Networks Research: Design and Applications*. Cambridge University Press.
- Eagle, Nathan, Alex (Sandy) Pentland, and David Lazer. 2009. "Inferring Friendship Network Structure by Using Mobile Phone Data." *Proceedings of the National Academy of Science* 106(36):15274-78.
- Fiske, Susan T., and Robert M. Hauser. 2014. "Protecting human research participants in the age of big data." *Proceedings of the National Academy of Sciences* 111(38):13675-76.
- Fortunato, Santo. 2010. "Community Detection in Graphs." *Physics Reports* 486:75-174.
- Gile, Krista J., and Mark S. Handcock. 2010. "Respondent Driven Sampling: An Assessment of Current Methodology." *Sociological Methodology* 40:285-327.
- Grippa, Francesca, and Peter A. Gloor. 2009. "You are Who Remembers You: Detecting Leadership through Accuracy of Recall." *Social Networks* 31(4):255-61.
- Heckathorn, Douglas D. 1997. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems* 44(2):174-99.
- Helleringer, S., Hans-Peter, K., Kalilani-Phiri, L., Mkandawire, J, & Benjamin, A. 2011. "The reliability of sexual partnership histories: Implications for the measurement of partnership concurrency during surveys." *AIDS* 25:503-511.
- Kandel, D. B. 1978. "Homophily, Selection, and Socialization in Adolescent Friendships." *American Journal of Sociology* 84:427-36.
- Kahn, Jeffrey P., Effy Vayena, and Anna C. Mastroianni. 2014. "Opinion: Learning as we go: Lessons from the publication of Facebook's social-computing research." *Proceedings of the National Academy of Sciences* 111(38):13677-79.
- Khan, Bilal, Kirk Dombrowski, Ric Curtis, and Travis Wendel. 2015. "Estimating Vertex Measures in Social Networks by Sampling Completions of RDS Trees." *Social Networking* 4(1):1-16.
- Kilduff, Martin and Hongseok Oh. 2006. "Deconstructing diffusion: An ethnostatistical examination of medical innovation network data reanalyses." *Organizational Research Methods* 9:432-455.
- Killworth, P. D., and H. R. Bernard. 1976. "Informant Accuracy in Social Network data." *Human Organizations* 35:269-86.
- Klofstad, Casey A., McClurg, Scott D., and Rolfe, Meredith. 2009. "Measurement of Political Discussion Networks: A Comparison of Two "Name Generator" Procedures." *American Association for Public Opinion Research* 73(3), 462-483.
- Klov Dahl, Alden S. 1985. "Social Networks and the Spread of Infectious Diseases: The AIDS Example." *Social Science Medicine* 21:1203-16.
- Klov Dahl, Alden S. 2005. "Social Network Research and Human Subjects Protection: Towards More Effective Infectious Disease Control." *Social Networks* 27:119-37.
- Krackhardt, David. 1987. "Cognitive Social Structures." *Social Networks* 9: 109-134.
- Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. "Experimental evidence of massive-scale emotional contagion through social networks." *Proceedings of the National Academy of Sciences* 111(24):8788-90.
- Laumann, Edward O., Peter V. Marsden, and David Prensky. 1994. "The Boundary Specification Problem in Network Analysis." in *Research Methods in Social Network Analysis*, edited by Linton C. Freeman, Douglas R. White, and A. Kimball Romney. Transaction Publishers.

- Lorant, Vincent, Victoria Eugenia Soto, Joana Alves, Bruno Federico, Jaana Kinnunen, Mirte Kuipers, Irene Moor, Julian Perelman, Matthias Richter, Arja Rimpelä, Pierre-Olivier Robert, Gaetano Roscillo, and Anton Kunst. 2015. "Smoking in school-aged adolescents: design of a social network survey in six European countries." *BMC Research Notes* 8(1):91.
- Marin, Alexandra and Hampton, Keith N. 2006. "Simplifying the Personal Network Name Generator Alternatives to Traditional Multiple and Single Name Generators." *Field Methods* 19(2), 163-193.
- Marsden, Peter V. 2011. "Survey Methods for Network Data." Chapter 25 in *The Sage Handbook of Social Network Analysis*, edited by John Scott and Peter J. Carrington. Sage.
- Marsden, Peter V. and Joel Podolny. 1990. "Dynamic analysis of network diffusion processes." In: J. Weesie and H. Flap (Eds.). Pp. 197-214 in *Social Networks through Time*. ISOR, Utrecht.
- McCarty, Christopher, Peter D. Killworth, and James Rennell. 2007. "Impact of Methods for reducing Respondent Burden on Personal Network Structural Measures." *Social Networks* 29:300-15.
- McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415-44.
- Merluzzi, Jennifer and Burt, Ronald S. 2013. "How many names are enough? Identifying network effects with the least set of listed contacts." *Social Networks* 35(3):331–337.
- Morris, Martina (ed.). 2004. *Network Epidemiology: A Handbook for survey design and Data Collection*. Oxford University Press.
- National Commission for the Protection of Human Subjects. 1979. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. US Government Printing Office.
- Narayanan, Arvind, & Vitaly Shmatikov. 2009. "De-anonymizing social networks." *IEEE Symposium on Security and Privacy* 30:173-187.
- Neal, Jennifer Watling. 2008. "“Kacking” the Missing Data Problem: Applying Krackhardt's Cognitive Social Structures to School-Based Social Networks." *Sociology of Education* 81(2): 140-162.
- Paik, A., & Sanchagrin, K. (2013). "Social isolation in America: An artifact." *American Sociological Review*, 78, 339–360
- Phillips, Gregory II, Patrick Janulis, Brian Mustanski, and Michelle Birkett. 2017. "Validation of Tie Corroboration and Reported Alter Characteristics among a Sample of Young Men who have Sex with Men." *Social Networks* 48:250-55.
- Potter, Gail E., Timo Smieszek, and Kerstin Sailer. 2015. "Modelling Workplace Contact Networks: The Effects of Organizational Structure, Architecture, and Reporting Errors on Epidemic Predictions." *Network Science* 3(3):298-325.
- Potterat, J. J., D. E. Woodhouse, S. Q. Muth, R. B. Rothenberg, W. W. Darrow, A. S. Klov Dahl, and J. B. Muth. 2004. "Network dynamism: history and lessons of the Colorado Springs study." in *Network Epidemiology: A Handbook for Survey Design and Data Collection*, edited by Martina Morris. Oxford University Press.
- Salathe, Marcel, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W. Feldman, and James H. Jones. 2010. "A High-Resolution Human Contact Network for Infectious Disease Transmission." *Proceedings of the National Academy of Science* 107(51):22020-25.
- Salganik, Matthew J., and Douglas D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations using Respondent Driven Sampling." *Sociological Methodology* 34:193-240.
- Smith, Jeffrey, James Moody, and L. Smith-Lovin. 2017. "Network sampling coverage II: The effect of non-random missing data on network measurement." *Social Networks* 48(1):78-99.
- Sönmez, Sevil, Apostolopoulos Yorghos, E. Tanner Amanda, Massengale Kelley, and Brown Margaret. 2016. "Ethno-epidemiological research challenges: Networks of long-haul truckers in the inner city." *Ethnography* 17(1):111-34.

- Strang, David, and Nancy Brandon Tuma. 1993. "Spatial and Temporal Heterogeneity in Diffusion." *American Journal of Sociology* 99:614-639.
- Van den Bulte, Cristophe, and Gary L. Lilien. 2001. "Medical innovation revisited: Social contagion versus marketing effort." *American Journal of Sociology* 106:1409-1435.
- Wald, Andreas. 2014. "Triangulation and Validity of Network Data." Pp. 65-89 in *Mixed Methods Networks Research: Design and Applications*, edited by Silvia Dominguez and Betina Hollstein: Cambridge University Press.
- Wallerstein, Nina B., and Bonnie Duran. 2006. "Using Community-Based Participatory Research to Address Health Disparities." *Health Promotion Practice* 7(3):312-23.
- Wejnert, Cyprian, and Douglas D. Heckathorn. 2008. "Web-Based Network Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research." *Sociological Methods & Research* 37(1):105-34.
- Young, April M., Abby E. Rudolph, Amanda E. Su, Lee King, Susan Jent, and Jennifer Havens. 2016. "Accuracy of name and age data provided about network members in a social network study of people who use drugs: Implications for constructing sociometric networks." *Annals of Epidemiology* 26(11):802-09.
- Zachary, Wayne W. 1977. "An Information Flow Model for Conflict and Fission in Small Groups." *Journal of Anthropological Research* 33(4):452-73.

Author Bios

jimi adams is an Associate Professor in the Department of Health and Behavioral Sciences at the University of Colorado Denver. His work focuses on examining social networks to understand how infectious diseases and novel ideas spread. This has included modeling HIV/AIDS risk in the US and sub-Saharan Africa, and the organizational dynamics of interdisciplinary fields. He is the author of a SAGE "little green book" on Gathering Social Network Data.

Tatiane Santos's research has focused on evaluating the impact of the Patient Protection and Affordable Care Act provisions on population health outcomes, healthcare utilization, and costs. She has evaluated Colorado's Medicaid reform efforts, as well as Colorado's State Innovation Model that seeks to integrate behavioral health within the primary care setting. She is interested in applying organization theory and social network methods to explore the role of community social capital in promoting public health. Aside from health policy, Tatiane has worked on immigration policy related to the Dream Act and in-state tuition eligibility for "dreamers".

Venice Ng Williams, MPH CHES is a doctoral candidate in Health Services Research at the University of Colorado and the lead mixed-methods research analyst at the Prevention Research Center for Family & Child Health. Trained in program planning, evaluation, and econometrics, her research focuses on mixed-methods, maternal-child health, organizational collaboration, and translating research into practice within the context of prevention programs. Her dissertation examines cross-sector collaboration in the Nurse-Family Partnership and implementation outcomes. She has previously worked in health promotion, tobacco prevention policies, systems change evaluation, health impact assessments, and hospital community-benefit research.