Social Network Data Collection: Principles and Modalities

jimi adams, University of Colorado Denver

Miranda J. Lubbers, Autonomous University of Barcelona

Introduction

Relational questions require relational data and methods. That may seem like a truism that is unnecessary to make explicit, but the social and behavioral sciences are replete with relational theories (Borgatti and Halgin, 2011) and questions that have repeatedly been examined with data that make essentializing assumptions (Emirbayer, 1997); that is, relying on *non*-relational data.

There are numerous potential reasons for this disconnect. First, disciplines carry strong methodological norms that do not readily incorporate relational theoretical perspectives, ranging from those that focus on desirable features of data (e.g., the "large nationally representative samples" that predominate in many corners of sociology; Martin, 2017) to those that prioritize particular analytic dimensions (e.g., the preoccupation with causal identification within some domains of economics and political science; Mogstad and Torgovitsky, 2018), and those that prioritize internal validity or demand experimental control (e.g., in many lab-driven fields; Falk and Heckman, 2009). Each of these cases carries (sets of) strong assumptions that require priorities that are incompatible with network approaches, leaving relational questions to be shoe-horned into available models. Second, some have critiqued social network analysis as being "only a method" and thus not having the conceptual and theoretical range that demands its development and use across the social sciences in the way that those who see *most* social science as foundationally relational might prefer (Borgatti et al., 2009). Third, as is often the case in scientific fields, theory, data, and analytic capabilities often develop in trajectories that

are not well synced with one another, such that the advances in one domain get out well ahead of the others (Lakatos, 1978). Said more positively, we may just be at an opportune point in the life course of social network research when rethinking the approaches to social network data collection is especially needed. Fourth is pragmatic; network data are often demanding to gather, and even more difficult to do well (adams, 2019; McCarty et al., 2019). Therefore, if researchers can get away without those added burdens, they will frequently opt to avoid them.

Despite these potential barriers, we take as given that network research offers a range of unique theoretical, empirical, and other contributions, which warrant well-designed strategies to collect and analyze social network data.[1] Therefore, our focus here is to overview the core principles that have governed many primary strategies for approaching that task[2] and illustrate their utility for a range of social and behavioral scientific questions. We will first discuss the general principles of collecting network data, and then elaborate on how those principles have been applied in four strategies of network data collection: Experiments, surveys and interviews, observation, and trace data.

General Principles of Network Data

As noted above, what makes network research unique is the focus on relationships. Contrasting most social sciences that commonly theorize at the levels of individuals, groups, or their aggregations (e.g., by targeting explanations for the associations between various attributes of whichever of those units are a study's focus), social network research shifts its primary gaze to the relationships *between* those units. This focus on relationships in network research implicates several dimensions that entail unique requirements for data collection. In this section,

[1] If you need more convincing, we point you to conceptual and theoretical chapters elsewhere in this volume (e.g., Chapter Prell/Hollway/Todo) or the application to topic(s) of your particular interest below.
[2] Coincidentally, one approach we discuss in detail below –ethnographic observation– is among the most primed to incorporate relational conceptualizations into 'standard' research design approaches.

we outline the importance and strategies for addressing those pertaining to: (1) the definition and interpretation of units involved (both the relationships and the nodes), (2) specifying "the rules of inclusion for different network elements," in what has been labeled as the network "boundary specification" problem (Laumann et al., 1983), along with (3) how each of these can vary across different relationship dimensions, nodesets (also termed 'modes'), levels, and timescales.

First, the definition and interpretation of units involve conceptualizing the *type of relationships* we are interested in, which impacts measurement. We often focus on positive relationships, such as collaboration or friendship, but sometimes negative ties (e.g., conflict, disliking)–or their absence–can be as/more influential than positive ties (e.g., support, liking). We can further distinguish: (1) relatively stable relationships ('states', e.g., dislike, knowing someone) from highly changeable and ephemeral ties ('events', e.g., physical contact, soccer passes, sexual encounters, political discussions), (2) objectively estimable relationships (e.g., financial exchanges, text messages) from those that are perceptually oriented (e.g., trust, 'cognitive social structures'); and (3) other *conceptual* distinctions of the types of relationships that matter (e.g., flows, potential/latent ties, etc.; see Borgatti et al., 2009). Shared across this variety is the *relational* nature of each possibility–turning the analytic lens from the units that comprise a population, and their various characteristics, to the spaces between them and the relationships that fill those spaces. Often our understanding of social dynamics via network ideas requires us to account for the patterning of when relationships do *not* form, as much as when they do.

We must also define the *type of nodes* involved in those relationships of interest, which affects sampling. Depending on the research questions, they can be any individual or collective social entity that has relationships with others, such as humans, animals, literary characters, teams, organizations, countries, or words.

Second, and related, we need to define the *boundary on the set of relationships and nodes*: where do we draw the line for which of these belong to the network and which do not? The answer affects both sampling and measurement strategies. Historically, network research has fallen into two groups for addressing this question based on the nodesets: networks in bounded settings such as classrooms, organizations, or neighborhoods (sociocentric networks, or "networks in a box," Kadushin, 2011:17)[3] versus networks surrounding focal nodes (egocentric or personal networks; see Chapter Small/Perry). In the first case, we study a 'whole' setting, assuming that everyone who belongs to the setting(s) is a network node, seeking to map the relationships among that presumed 'complete' population. In the second case, we focus on an individual, or more commonly, a sample of individuals, and examine the entities with which each focal node has the relationship(s) of interest. The focal node, its network members, their relationships, and potentially the relationships among those network members then form the node's (ego-)network.

Notably, the definition of the tie further delimits the set of nodes in an egocentric network. In theory, every person that individuals know could be identified as a network member, but in practice, researchers often focus on stronger ties (e.g., those they feel closest to or trust most) or more particular relationships governing the substantive research focus, thus generating further boundaries on the network (see Chapter Brashears/Money). Additional restrictions on the nodes –for instance, on the minimum age of network members– or the relationships –for instance, regarding their duration (e.g., support given within the past year)– can make the amount of data more manageable.

Apart from sociocentric and egocentric networks, a third strategy recognizes that some phenomena of interest entail blurred boundaries. The nodes are not neatly organized in an

[3] While these are frequently referred to as 'whole' or 'complete' networks, the boundary specification and/or data collection practicalities often enforce limitations on how complete they truly are.

institutional or geographical setting, as in sociocentric networks, but neither are they centered on individuals, as in ego-centered networks. These have been labeled 'partial networks' (Morris, 2004) or 'open system networks' (Kadushin, 2011). When there are no clear a priori boundaries, they are often delineated by using some type of network sampling (e.g., Mouw and Verdery, 2012), such that they start with the egocentric networks of a few 'seeds,' and then the contacts of these seeds are also invited to participate, and so on, until producing a representation of the overall network. Other times, when a partial network is embedded in systems for which digital trace data exist (for instance, Twitter networks), the network is created by limiting the geographical area of nodes, the time span of contents, and the topics discussed.

Third, networks can be 'multilayered': They can have more than one type of ties, one type of nodes, or one time point. Research can focus on a single type of relationship or study how different types of relationships interact (e.g., liking and disliking, or advice and information). These are multivariate (or multiplex) networks. Additionally, some studies include two nodesets (or node partitions), such as individuals and organizations, patients and doctors, or Wikipedia entries and editors. This makes the design more complex. If edges can only be formed between nodes of different partitions, the network is called bipartite or two-mode (see Chapter Jasny), such as networks of Wikipedia edits (Keegan and Fiesler, 2017). Even though we only directly observe the connections of people (authors) to entries in this network, we can derive single-mode networks from it, of people who have coauthored Wikipedia entries (people-to-people networks), or networks of entries edited by the same people (entry-to-entry networks). When one type is hierarchically embedded in the other and relationships can be formed between nodes of the same and different levels, we speak of multilevel networks (see Chapter Lazega/Wang).[4] For instance, we may study to what degree individuals working in a set of

[4] If the higher-level groups do not have relationships among them, they are also called multilevel networks, but in that case the higher-level groups are not nodes.

organizations collaborate with one another, within and between organizations, and whether the formal collaboration ties among organizations shape such interactions. Furthermore, researchers may study relationships at one time point, or longitudinally (in continuous or discrete time; see Chapter Snijders). Especially when investigating event-like relationships, collecting longitudinal data allows us to follow network changes over time, which inform us better about its functioning than a 'snapshot' of the network (see Chapter Contractor/Schecter).

Modes of data collection

Building on the common principles described above, we now introduce the main methods used for data collection in social network research: experiments, surveys and interviews, observation, and behavioral trace data. We highlight their unique contributions, subtypes, and design decisions, and give examples.

Experiments

Experiments are ideal for analyzing the causal mechanisms underlying social influence or selection processes and have been used for decades for network research (e.g., Bavelas and Barrett, 1951). Other data collection modes mostly show associations between networks and behaviors, but are less suitable for explicitly demonstrating causation.[5] Suppose we observe that highly central people in a collaboration network use more technological devices for their work, such as earphone translators or collaboration software. We may postulate that this is caused by social selection on the variable of interest: people are drawn toward technologically savvy users, which explains their centrality. However, alternative explanations (cf. Shalizi and

[5] Survey and observation researchers can take certain measures to strengthen their designs in this respect: they can collect additional data to control for the most plausible alternative explanations, longitudinal data, or add qualitative questions to have respondents explain the observed behavior in their words. While such measures make a stronger case for causation, they may still not provide definitive evidence.

Thomas, 2011) include social selection on other variables (people are drawn to higher-ranked individuals, who also have the money to buy more gadgets), diffusion (highly central people may receive much more information about the benefits of such devices than less central people, and thus adopt them more often) or adaptation to the same environment (people in similarly central network positions have to manage many social relationships, and doing so is easier with devices). Experiments help determine causal mechanisms and are important tools for theory formation.

The randomized controlled trial (RCT) is typically seen as the 'gold standard' experimental design. In RCTs, researchers randomly assign individuals to either the experimental or the control condition without telling them which condition they are assigned to, to control potentially confounding variables. The difference between these two conditions is how the independent (explanatory) variable is manipulated. The effect of the independent (predictor) on the dependent (outcome) variable is measured by comparing the treatment to the control groups (through surveys, observation, or trace data). In network experiments, networks can be the independent variable –when studying network effects– or the dependent variable –when studying network formation. Often, the dependent variable is measured at least once pre- and once post-intervention.

We can distinguish four types of network experiments based on their setting (see Table 1). First, network experiments have been conducted in semi-controlled, 'natural' environments (field experiments) and fully controlled environments (lab experiments). Field experiments focus on existing network structures and are, therefore, more realistic. Lab experiments tend to construct networks artificially, but they can control confounding variables better. Second, we can distinguish between offline and online experiments for each type. The emergence of online platforms such as Facebook or Amazon Mechanical Turk, and the possibilities of customizing

aspects of these environments and following behaviors in real time have incited new experimental research.

[Insert Table 1 near here]

Let us illustrate each quadrant of Table 1 with recent examples using RCT designs. An example of *offline field experiments* (quadrant 1) is Paluck et al.'s (2016) study of the effect of a peer-to-peer anti-conflict intervention in middle schools. The researchers hypothesized that the intervention would reduce conflict behavior, particularly if the selected peers were highly central in the networks ('social referents'). To test this, they randomly assigned 56 schools to experimental or control conditions. In both conditions, they measured peer networks through a survey at the beginning (before randomization) and the end of the school year. In the experimental condition, a small portion of students were selected to serve as 'seeds' for the intervention; they met biweekly with trained research assistants to identify conflict behaviors in their school and take a public stance against conflicts. The intervention reduced conflict behavior substantially, with social referent seeds being most effective.

An illustration of *online field experiments* (quadrant 2) is Bond and colleagues' (2012) study of how political behavior spreads on Facebook and whether strong ties influence the spread more than weak ties. On the US congressional election day in 2010, they conducted an experiment with all 61 million adult US Facebook users, assigning them randomly to one of three conditions. In the social transmission condition, users received a message shown at the top of their news feed, which encouraged them to vote, displayed a link to information about nearby polling places, presented a button users could click to indicate that they had voted, and showed the IDs of up to six Facebook friends who had already clicked the 'I voted' button alongside the total number of friends who had done so. The second condition similarly encouraged users to vote but did not show the profile pictures or the number of Facebook

friends who had voted. Participants in the control condition did not receive any targeted messages. The researchers captured digital trace data of the networks and behaviors, and validated voting behaviors with public voting records. As hypothesized, results showed that social transmission was more effective than encouragement alone, particularly through strong ties.

These first two cases captured 'naturally' occurring networks. In contrast, the following two examples address lab experiments in artificially created networks. An example of *offline lab experiments* (quadrant 3) is Traeger and colleagues' (2020) study of how social robots shape interactions among people (here, networks were the dependent variable). Participants came to an onsite lab and filled in a pre-test survey. Then, they were randomly assigned to groups of three accompanied by a humanoid social robot, the fourth participant. Each group (51 in total) engaged in 30 rounds of a collaborative game. In the vulnerable condition, the robot made vulnerable utterances at the end of each round and admitted its own mistakes; in the neutral condition, it made neutral utterances and did not admit mistakes. In the control condition, the robot was silent. Researchers observed speaking time and directionality, and also measured group perceptions. Participants spoke more with one another (for all three dyads), had more balanced speaking times, and perceived their group more positively in the vulnerable condition, showing that robots can effectively alter group behavior.

Centola's (2010) examination of how network structure affects the spread of behaviors provides an example of *online lab experiments* (quadrant 4). While small-world networks diffuse information or germs more rapidly than locally highly clustered networks, Centola hypothesized that this does not hold for complex contagion, when people need contact with multiple adopters to adopt a behavior. To test this, he constructed a health forum accessible only to invited users. 1,528 participants were recruited from health-interest websites and created profiles on the platform, then were each randomly assigned to one of two conditions. In one condition, the

network had a clustered-lattice structure; in the other, part of the clustered-lattice structure was randomly rewired based on the small-world model. All participants had the same number of contacts in the network ('network buddies'), whose behaviors were visible to them. They could not contact their buddies nor create new ties, so the structure was fixed. Then, a randomly selected participant in each condition was informed about an information-rich health forum website (which Centola also maintained) and encouraged to register. Centola followed the dissemination of this information in real-time and found that it spread faster and farther in the clustered-lattice structure than in the small world structure, confirming his hypothesis.

The range of these network experiments demonstrates the field's diversity. Apart from the setting, the manipulated network element varies across studies: Paluck et al. (2016) manipulated the behaviors of a randomly selected set of nodes (a 'peer encouragement design'; Aral, 2016), Traeger et al. (2020) the network context ('settings design'; Aral, 2016), Bond et al. (2012) the information that flows through ties ('mechanisms design'; Aral, 2016), and Centola (2010) the network structure ('structural design'; Aral, 2016). The manipulation creates the exogenous variation needed to establish causality.

Not all network experiments use full-fledged RCT designs, however. Some manipulate a variable to make it measurable or salient but do not use control groups or random assignment. Milgram's (1967) famous small-world experiment is a good example. Milgram asked participants to pass a letter to a target stranger, but only via acquaintances who might either know this person or a person who might know them, and counted the number of intermediaries required for letters that reached the target. Some researchers also cleverly use external events such as a natural disaster to study network formation, comparing it with similar settings where the event did not happen (Phan and Airoldi, 2015)–a natural experiment.

Surveys and Interviews

While common perceptions of science often *assume* experimental designs, other designs are more prevalent for studying social networks. Survey and interview-based approaches have a much longer and more detailed history in the social-scientific study of networks (adams, 2019; Marsden, 1990), even in psychology, where experiments are generally highly prioritized (Neal, 2020). Before digging into some of the practicalities of how surveys and interviews have been used to study social networks, it may be helpful to briefly examine why survey and interview methods predominate.

The Thomas and Thomas theorem ("If [people] define situations as real, they are real in their consequences") remains as central to the social sciences today as it was when initially stated in 1928. In part, this is because *perceptions* of social reality can be as important for explaining social and behavioral outcomes as their 'objective' reality. This often leads researchers to prioritize talking to research participants about their social relationships–whether in structured ways through surveys, or more flexible interview techniques. Moreover, some of the most concretely theorized relationships within social networks research (friendship, liking/disliking, esteem, social support) are fundamentally perceptions rather than externally identifiable (Fischer, 1982), thus *necessitating* data collection via surveys or interviews. Finally, even those relationships that have the potential for external validation (e.g., sexual contact, financial exchanges, migration flows) are not always empirically accessible or precisely recordable in the ways researchers desire, thus requiring the use of various proxies for estimating the presence/absence of such relationships (adams, 2019; Kitts, 2014).

As such, whether seeking to collect data on objectively observable ties or perceptions of more subjective relationships, asking respondents has been a foundational strategy for eliciting such data throughout the field's history (adams 2019; Marsden 2011; McCarty et al. 2019). Within survey and interview-based strategies, several design considerations require adaptations to the particular needs of *network* data.

In social network surveys, the type of relationship asked about is commonly labeled the type of ‘name generator’ question used (i.e., for what type of relationship are we recording reported partners?). Name generator questions are complemented with a battery of ‘name interpreter’ questions, which are follow-up questions about the characteristics of the named partners (especially in the case of egocentric networks where such information is not gained from the network partners themselves), such as their gender or race, and the relationships with these partners, such as tie strength or duration. In egocentric networks, ‘name interconnector’ questions (Borgatti et al., 2013) may be added that enquire about the perceived relationships *among* a respondent’s network partners.

Survey (and interview) researchers are typically quite well-attuned to the dual concerns of measurement and sampling considerations. Measurement evaluation and validation are primary concerns in many fields. Evaluations of network data have therefore often taken the form of assessing the agreement between self-reports and external sources of validation (Killworth and Bernard, 1976) or–given the potential for network data to be reported by both partners–comparisons between these multiple reports (adams and Moody, 2007; An, 2022). Generally, these studies find that more precision in questions and higher investment in the behavior/relationship examined reduces measurement errors.

Measurement considerations have also led studies to examine the overlaps (and divergences) that arise from using different question prompts to elicit network data. For example, Marin and Hampton (2007) showed that a range of commonly employed general name generators elicit different sets of confidants from respondents that overlap only partially. Using a complementary set of multiple name generators that measure the same relational dimension can therefore help reduce measurement bias. Furthermore, researchers have experimented with ways to lower the respondent burden caused by the repetitive questions of name interpreters and interconnectors, by either keeping such questions at a minimum or

asking some questions for a sample of ties rather than for all ties (cf. McCarty et al., 2019). Studies have further investigated survey mode and interviewer effects.

As noted above, because of the 'boundary specification problem' in networks research, the nomination of partners is not *only* a question of measurement, but can also determine the sampling strategy for a study. Particularly in link-tracing (or 'partial') network designs, nominated partners can, in turn, become a device for further sampling from the population of interest, by recruiting them as subsequent study participants. This approach has been formalized in strategies like respondent-driven sampling (Heckathorn and Cameron, 2017) and network sampling with memory (Mouw and Verdery, 2012). These strategies seek to optimize the capacity for statistical inference by specifying particular rules for determining which links are followed to generate a study sample.

Survey network research has a long history and has been conducted in many formats (evolving across time from face-to-face to pencil-and-paper, phone, internet, etc.), with particular adaptations to more secure and private options when the data sought are especially sensitive (Szinovacz and Egley, 1995). While technological advances play some role in shaping those options and preferences, the theoretical and practical enhancements that such developments have fostered are also worth noting.

Network data collection on digital platforms (whether on local laptops or tablets or administered via online portals) illustrates how these developments can accomplish multiple aims at once. Practically, via back-end programming, these approaches encode the data *while they are being collected* in formats that facilitate rapid analysis and minimize (but do not eliminate entirely) the need for data coding and cleaning. Moreover, the potential for interactive approaches within these strategies can reduce respondent burden and enhance respondents' survey experience, which improves the quality of the resulting data (McCarty et al., 2019). One

recent example of such digital platforms is Network Canvas, which flexibly elicits a range of respondent attributes, name generators, interpreters, and interconnectors–each interactively leveraging these empirically demonstrated benefits (Birkett et al., 2021). For example, in a study of Black men who have sex with men and transgender persons, Network Canvas was found to be intuitive and useful for eliciting highly sensitive relational data (Crawford et al., 2021).

Surveys that include network modules draw on many of the demonstrated benefits of survey research in general –including standardized prompts allowing for comparability across respondents and within respondents over time, and capacity for measurement validation and replication across studies. However, as noted earlier, this pursuit of objective measures as the 'gold standard' often belies the importance of social actors' understanding and interpretation of their own relationships to faithfully examining social phenomena. As such, there are both: (1) calls for survey and interview researchers to increasingly (re-)focus on the *meaning* (and not just the structuring) of social relationships (adams, 2019), and (2) other methods with long histories and recent developments that prioritize these aims in their approaches to gathering social network data.

Regarding the first point, social network researchers increasingly use more open, *qualitative interviewing* strategies to enquire about social networks, either alone or in combination with structured network elicitation. In addition to focusing on individuals' understanding and interpretation of their own networks, qualitative interviewing is also suitable for exploring network formation and individuals' agency in networks, and why networks matter in their lives (see Chapter Hollstein). Qualitative interviews can also help validate data obtained with structured interviewing. Regarding the second point, the following sections will elaborate on methods that prioritize both more 'objective' and interpretive elements of measurement.

Observation

While surveys and interviews are the most common modes of social network data collection, they are not always the best choice. For instance, they are not viable for research with subjects that cannot read or reflect on relationships, such as young children (e.g., Martin et al., 2005) or animals (e.g., Mann et al., 2012; Rushmore et al., 2013). Furthermore, surveys may not capture behaviors reliably. For instance, humans are typically good at recalling routinized interactions but easily forget more fleeting or less recent interactions (e.g., Killworth and Bernard, 1976). In addition, behaviors are situated in contexts, and when abstracted from these contexts, accounts of behavior are often inconsistent with behavior (Jerolmack and Khan, 2014). Observation may be a better alternative in these cases when the ties are observable interactions or transactions.

Like surveys, observation has been used since the origins of social network research. For instance, Bott (1928) and Hagman (1933) observed social interactions among preschool children. Each day, Bott chose one child in the group and recorded its interactions with the other children. Hagman compared her observation of children's interactions with the children's recall of the same interactions-a forerunner of informant accuracy studies (e.g., Killworth and Bernard, 1976).

As with surveys and interviews, we can distinguish between structured (or systematic) and unstructured (ethnographic) observation. For *systematic observation*, researchers face partially the same decisions regarding the network boundary problem as in surveys. First, they must define the ties they want to observe. For instance, what counts as an interaction among children, regarding content, time, or directionality? When a child plays alongside another but not *with* them, does it count as an interaction? Second, researchers must define who belongs to the network. For instance, they can observe interactions among preschool children in the classroom and thus sample whole classes, but if they are more interested in unstructured play, they could also sample children and study their ego-centered networks on playgrounds or playdates. For animals, researchers often delineate a geographical area or follow a community and adopt a

threshold for the minimal number of sightings needed before the animal counts as belonging to the focal population. Researchers must also decide what other aspects of the interactions to observe (e.g., type, contact duration). These decisions ultimately depend on the research questions, as well as on equipment, available time, and team size.

Researchers relying on observation need to design measurement procedures that lower the complexity of observing a group of people over a prolonged period. Given the sheer volume of simultaneously occurring interactions, they must make choices in what to record, often by sampling on nodes and/or time. For nodes, they may focus on either one subject at a time ('focal sampling'; see the example of Bott at the start of this section) or the entire group ('all occurrence sampling'). For time, they may record networks continuously for a specific duration ('sequence sampling') or observe the subject or group at set intervals (i.e., 'scan sampling'), among other possibilities. For instance, Rushmore et al. (2013) observed a wild chimpanzee community in Uganda for nine months. Each morning, they randomly selected a chimpanzee and followed them for ten hours, scanning their company every 15 minutes. Given earlier primate research, they conceptualized being at a distance of up to fifty meters as contact, and up to five meters as close contact. In research among preschoolers, Martin et al. (2005) combined two procedures. On the one hand, they cycled through a randomly ordered list of names, observing each child for 10 seconds before continuing with the next name on the list. They paused for five minutes at the end of the list and then started again, recording the focal child's company, behaviors, and affective expressions. On the other hand, they used focal observations centering 10 minutes on the same child to observe its actions and the responses of its interaction partners. Coding protocols are typically completed for each subject-time unit and are as structured as surveys.

Technological advances have dramatically augmented the possibilities of systematic observation. Wearable sensors and tracking devices are increasingly used to record

interactions, amplifying the details researchers can capture. While we elaborate digital trace data below, here we address how researchers have strategically deployed them in observational designs. For instance, Mastrandrea et al. (2015) had students of one French high school wear sensor-containing badges, which captured when pairs of students were within 1-1.5 meters of one another, facing one another for at least 20 seconds. From those observations, the researchers inferred contact, recording more than 67,000 contact events in one week. Comparing these data with contact diaries completed by the same students, they found that the sensors reflected short contacts much better than the diaries, although the latter accurately presented the network's backbone. Tracking devices log the exact location of individuals at all times, and when given to a bounded group, they capture who is within range of whom at what times. Animal social network researchers have similarly used GPS trackers, sometimes enhanced with accelerometers or heartbeat monitors, and radiofrequency identification detection systems (Smith and Pinter-Wollman, 2021).

Researchers examining human networks have also used mobile phone apps (e.g., Keil et al., 2020). These include professional conference apps (de Vaan and Wang, 2020), which track the location of each conference attendee and allow them to contact each other on the app. Some mobile phone apps allow researchers to ask participants at set times (using alarms) to enter the names of the people with whom they interact and further information about these interactions, such as type of contact, duration, and quality. They are a viable alternative to pen-and-paper contact diaries (Fu, 2007). While these technological advances augment the granularity of observation data, they can be costly and involve technological challenges (e.g., battery life), and not all capture the nature or context of interactions, although combining them with human observation can offset this last challenge (Smith and Pinter-Wollman, 2021).

Unstructured ('ad libitum') and *ethnographic observation* have also long been used to study social networks (e.g., Loomis, 1941), either alone or complementing interview and survey

data. Ethnographers tend to focus on communities, and relationships are integral to communities. Their long-term immersion in a field allows ethnographers to gain trust within the community and observe relationships and interactions naturally on-site, in real-time, and in changing conditions. In contrast to structured observation, ethnographers may not predefine who belongs to the network or what constitutes a tie. Furthermore, they do not use predetermined protocols but record what they believe is most relevant for their research questions, in a more free-flowing format. Nonetheless, as in structured observation, they cannot observe everything. Ethnographers' attention is also focused on their research questions, but more broadly: They pay attention to events or behaviors among individuals, across situations, and over time, behaviors that affirm their hypotheses but also those that challenge them. They observe how the behaviors or events are situated in the physical and cultural context, and how informants' perceptions of them ('emic') deviate from the ethnographer's interpretation ('etic'). This open, typically inductive focus engenders the element of surprise, allowing for discovering unforeseen network patterns and practices.

As a recent example, Torres (2019) interviewed older adults in a New York City neighborhood about their core discussion ties while also observing her participants in coffee shops and other neighborhood sites during five years of fieldwork. Ethnographic observation showed that many ties defied the strong tie-weak tie binary. Torres called these ties 'elastic' – they were both strong, providing social support, and weak, as the participants kept their distance by gossiping or not recalling their acquaintances' names. These novel insights could only be gained with an inductive, multi-pronged, exploratory approach.

Since encounters among informants increasingly occur online, ethnographers have also entered digital spaces, using virtual or digital ethnography or 'netnography', often complementing ethnography in physical sites. Digital ethnography is related to digital trace data (discussed below) but focuses more on how virtual communities emerge, the contents of

interactions, their meaning, and the formation of identities in such groups. In some cases, digital ethnography is complemented with network visualization. For instance, Cottica et al. (2020) used semantic networks (see Chapter González-Bailón/Shugars) to analyze the contents of discussions in an online forum, and Akemu and Abdelnour (2020) visualized email exchanges in an organization (trace data) among workers seated in different areas in the organization (ethnographic observation).

Ethnographic observation has received comparatively less attention in social network scholarship, but it has been vital to understanding networks (Lubbers and Molina, 2021). It has detected culturally salient relationship categories and their basic properties and explored individual agency in networks (i.e., networking practices or relational work), long-term network dynamics, and networks' embeddedness in larger contexts. Ethnography is, however, a highly intensive mode of data collection for both researchers and participants, in stark contrast to the mode we will discuss next.

Behavioral trace data

Many of the principles that underpin ethnographic observational approaches to gathering social network data have also motivated the recent proliferation of more passively collected 'bread crumb' varieties of network data, especially those available through digital platforms (see e.g., Salganik, 2019). Thus, the fourth mode of data collection we address is harvesting the traces of human behaviors, or as Paxton and Griffiths (2017:1631) called them, 'wild' data. With their proliferation, behavioral trace data have become a rich resource for network researchers. Many of these data are time-stamped (i.e., event-based, longitudinal data), and some are geolocated.

We can again distinguish between data about offline and online behaviors. *Trace data about offline behaviors* can be extracted from (offline or online) historical archive records (e.g., Bloch et al., 2022), population register data (Van der Laan et al., 2022), newspaper archives for

discourse and policy networks (Nagel and Satoh, 2019), epidemiological contact tracing data for COVID-19 contagion (Hâncean et al., 2020), Scopus or Web of Science records for coauthorship (Akbaritabar and Barbato, 2021), policy event descriptions (Pal and Spence, 2020), or national statistics for networks of flows among countries (Danchev and Porter, 2018), to name but a few options. Many of these data sources are digitized, facilitating research (see e.g., McLevey and McIlroy-Young, 2017). Furthermore, several authors have proposed automated extraction methods, for instance, using NLP (Bloch et al., 2022), decreasing some burdens of the task, though increasing others.

Trace data on online behaviors have become abundant. The website 'internet live stats',[6] curated by the Real Time Statistics Project, gives an instant idea of the thousands of tweets, Instagram and tumblr posts sent, skype calls made, and YouTube videos viewed, and the millions of emails sent globally *in just a single second*. Many of these data contain relational information, as people like, respond to, or retweet tweets, for instance. Other online communities (e.g., gaming communities, crowdfunding sites) also continuously produce digital trace data on networks, which researchers can study (Bainbridge, 2007).

When using trace data, it is crucial to know the system that has generated the data thoroughly to ensure construct validity and reliability: what does the system afford users to do (and what not), how do users adopt the system, have the algorithms or use changed over time, how does the system archive interactions? Furthermore, researchers must still specify the network boundaries, even if this occurs implicitly. For example, node and edge types are sometimes determined by the type of data, but researchers need to clearly determine boundaries on the set of nodes they include in their analysis. While trace data may be

[6] https://www.internetlivestats.com/

sociocentric, the networks are often so large that researchers limit the set by defining, for instance, geographical, temporal, or topic-related boundaries (González-Bailón et al., 2014).

Despite their abundance, these data also come with some caveats that must be addressed. A disadvantage of trace data is that important sociodemographic attributes and outcome variables are often unavailable. Furthermore, it may be difficult to gauge how representative data are of larger populations of ties or nodes. In addition, extracting unambiguous network data from records can be labor-intensive, depending on the type of data. The documents have typically not been created for research, and in many cases, relations are embedded in texts (with qualitative meaning), and the contents may be unclear, difficult to code systematically, prone to errors, or contain missing data. It may also be challenging to disambiguate nodes, as some users have multiple accounts, and some accounts are operated by bots. Thus, while behavioral trace data are promising for collecting data on large-scale, naturally occurring networks, their cleaning and validation need substantial work.

Final considerations

We focused on the main modes and types of social network data collection, ignoring less conventional modes such as contact diaries (Fu, 2007). We have not discussed mixed-methods designs, which combine multiple modes of data collection simultaneously or sequentially (e.g., quantitative methods followed by qualitative methods, or vice versa; e.g., Small, 2011). Promising possibilities for mixing data include ‘data linkage’, where, for instance, survey participants are asked to give their social media accounts and consent to their analysis, and combining ‘big’ and ‘thick’ data, i.e., digital trace data with in-depth interviews with selected nodes. This illustrates how data collection is a highly creative puzzle involving the choice of modes, their combination, and the many smaller design decisions with myriad possibilities. We have sought to highlight the principles governing choices among these possibilities.

A final consideration concerns the ethical challenges of collecting social network data (Tubaro et al., 2021). Like all social science data, social network data need to be collected with the highest care for ethics, ensuring respect for participants (e.g., confidentiality, informed consent, special protection of vulnerable research participants), beneficence (minimizing risks and maximizing benefits), and justice in the social distribution of research risks and benefits. Each data collection mode comes with its own challenges, each described at length in general methodology books. A particular though not unique[7] challenge for social network analysis is the collection of data about network members who have *not* given explicit consent to their collection. In sociocentric networks, individuals who did not give their consent are often not placed on rosters of names from which respondents pick their nominations. This omission introduces methodological problems, as the group's network structure may not be adequately represented if persons are missing from the analysis (see Chapter Krause/Huisman): whether they are highly central or peripheral in the network changes the group's network structure considerably. In egocentric networks, we are typically not interested in identifying network members, but rather in understanding their attributes, the characteristics of the relationships the respondent has with them, and the relationships they have with one another. In this case, researchers often reduce ethical problems by asking respondents to use nicknames or initials when nominating network members, rather than their full names. When network members are not personally identifiable, network data become attributes of the respondents (their perceptions of their personal networks), and anonymizing respondent data is then sufficient for ensuring confidentiality of network members.

Acknowledgments

[7] Surveys enquiring about household members, or students rating their teachers, involve similar problems.

Miranda Lubbers is grateful for the funding of the Catalan Institution for Research and Advanced Studies (ICREA Acadèmia).

References

adams j (2020) *Gathering Social Network Data*. Sage Publications.

adams j, and Moody J (2007) To tell the truth: Measuring concordance in multiply reported network data. *Social Networks* 29(1):44-58.

Akemu O, and Abdelnour S (2020) Confronting the digital: Doing ethnography in modern organizational settings. *Organizational Research Methods* 23(2):296–321.

Akbaritabar A, and Barbato G (2021) An internationalised Europe and regionally focused Americas: A network analysis of higher education studies. *European Journal of Education* 56(2):219–234.

An W (2022) You said, they said: A framework on informant accuracy with application to studying self-reports and peer-reports. *Social Networks* 70:187-197.

Aral S (2016) Networked experiments. In: Bramoullé Y, Galeotti A, and Rogers BW (Eds.), *The Oxford Handbook of the Economics of Networks* (pp. 375–411). Oxford University Press.

Bainbridge WS (2007) The Scientific Research Potential of Virtual Worlds. *Science* 317(5837):472–76.

Bavelas A, and Barrett D (1951) An experimental approach to organizational communication. *Personnel* 27:386–397.

Birkett M, Melville J, Janulis P, Phillips II G, Contractor N, and Hogan B (2021) Network Canvas: Key decisions in the design of an interviewer assisted network data collection software suite. *Social Networks* 66:114-124.

Bloch A, Vasques Filho D, and Bojanowski M (2022) Networks from archives: Reconstructing networks of official correspondence in the early modern Portuguese empire. *Social Networks* 69:123-135.

Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, Settle JE, and Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–298.

Borgatti SP, Everett MG, and Johnson JC (2013). *Analyzing Social Networks.* Sage Publications.

Borgatti SP, and Halgin DS (2011) On network theory. *Organization Science* 22(5):1168–1181.

Borgatti SP, Mehra A, Brass DJ, and Labianca G (2009) Network analysis in the social sciences. *Science* 323(5916):892–95.

Bott H (1928) Observation of play activities in a nursery school. *Genetic Psychology Monographs* 4:44–88.

Centola D (2010) The spread of behavior in an online social network experiment. *Science* 329(5996):1194–1197.

Cottica A, Hassoun A, Manca M, Vallet J, and Melançon G (2020) Semantic social networks: A mixed methods approach to digital ethnography. *Field Methods* 32(3):274–290.

Crawford ND, Josma D, Harrington KRV, Morris J, Quamina A, Birkett M, and Phillips II G (2021) Using the think-aloud method to assess the feasibility and acceptability of Network Canvas among Black men who have sex with men and transgender persons: Qualitative analysis. *JMIR Formative Research* 5(9):e30237.

Danchev V, and Porter MA (2018) Neither global nor local: Heterogeneous connectivity in spatial network structures of world migration. *Social Networks* 53:4–19.

de Vaan M, and Wang D (2020) Micro-structural foundations of network inequality: Evidence from a field experiment in professional networking. *Social Networks* 63:213–230.

Emirbayer, M (1997). Manifesto for a relational sociology. *American Journal of Sociology* 103(2):281–317.

Falk A, and Heckman JJ (2009) Lab experiments are a major source of knowledge in the social sciences. *Science* 326(5952):535-538.

Fischer CS (1982) *To Dwell among Friends: Personal Networks in Town and City.* University of Chicago Press.

Fu Y-C (2007) Contact diaries: Building archives of actual and comprehensive personal networks. *Field Methods* 19(2):194–217.

González-Bailón S, Wang N, Rivero A, Borge-Holthoefer J, and Moreno Y (2014) Assessing the bias in samples of large online networks. *Social Networks* 38:16–27.

Hagman EP (1933) The companionships of preschool children. *University of Iowa Studies in Child Welfare* 7:10-69.

Hâncean M-G, Perc M, and Lerner J (2020) Early spread of COVID-19 in Romania: Imported cases from Italy and human-to-human transmission networks. *Royal Society Open Science* 7:200780.

Heckathorn DD, and Cameron J (2017) Network sampling: From snowball and multiplicity to respondent-driven sampling. *Annual Review of Sociology* 43:101–119.

Jerolmack C, and Khan S (2014) Talk is cheap: Ethnography and the attitudinal fallacy. *Sociological Methods and Research* 43(2):178–209.

Kadushin C (2011) *Understanding Social Networks: Theories, Concepts, and Findings.* Oxford University Press.

Keegan B, and Fiesler C (2017) The Evolution and Consequences of Peer Producing Wikipedia's Rules. *Preprint SocArXiv* 10.31235/osf.io/28sgr.

Keil TF, Koschate M, and Levine, M (2020) Contact logger: Measuring everyday intergroup contact experiences in near-time. *Behavior Research Methods* 52(4):1568–1586.

Killworth PD, and Bernard HR (1976) Informant accuracy in social network data. *Human Organization* 35(3):269–286.

Kitts JA (2014) Beyond networks in structural theories of exchange: Promises from computational social science. *Advances in Group Processes* 31:263-298.

Lakatos I (1978) *The Methodology of Scientific Research Programmes*. Vol. 1. Cambridge University Press.

Laumann EO, Marsden PV, and Prensky D (1983) The boundary specification problem in network analysis. In: Burt RS and Minor M (Eds.), *Applied Network Analysis: A Methodological Introduction* (pp. 18–34). Beverly Hills: Sage Publications.

Loomis CP (1941) Informal groupings in a Spanish-American village. *Sociometry* 4(1):36.

Lubbers MJ, and Molina JL (2021). The ethnographic study of personal networks. *Etnografia e Ricerca Qualitativa* 14(2):185–200.

Mann J, Stanton MA, Patterson EM, Bienenstock EJ, and Singh LO (2012) Social networks reveal cultural behaviour in tool-using dolphins. *Nature Communications* 3(1):980.

Marin A, and Hampton KN (2007) Simplifying the personal network name generator: Alternatives to traditional multiple and single name generators. *Field Methods* 19(2):163–93.

Marsden PV (1990) Network data and measurement. *American Review of Sociology* 16:435–463.

Martin JL (2017) *Thinking through Methods: A Social Science Primer.* University of Chicago Press.

Martin CL, Fabes RA, Hanish LD, and Hollenstein T (2005). Social dynamics in the preschool. *Developmental Review* 25(3–4):299–327.

Mastrandrea R, Fournet J, and Barrat A (2015) Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE* 10(9):e0136497.

McCarty C, Lubbers MJ, Vacca R, and Molina JL (2019) *Conducting Personal Network Research: A Practical Guide*. Guilford Press.

McLevey, J, and McIlroy-Young, R (2017) Introducing metaknowledge: Software for computational research in information science, network analysis, and science of science. *Journal of Informetrics*, 11(1), 176–197.

Milgram S (1967) The small-world problem. *Psychology Today* 1(1):60–67.

Mogstad M, and Torgovitsky A (2018) Identification and extrapolation of causal effects with instrumental variables. *Annual Review of Economics* 10:577-613.

Morris M (2004) *Network Epidemiology: A Handbook for Survey Design and Data Collection.* Oxford University Press.

Mouw T, and Verdery AM (2012) Network sampling with memory: A proposal for more efficient sampling from social networks. *Sociological Methodology* 42(1):206-256.

Nagel M, and Satoh K (2019) Protesting iconic megaprojects. A discourse network analysis of the evolution of the conflict over Stuttgart 21. *Urban Studies* 56(8):1681–1700.

Neal JW (2020). A systematic review of social network methods in high impact developmental psychology journals. *Social Development* 29(4):923–944.

Pal LA, and Spence J (2020). Event-focused network analysis: a case study of anti-corruption networks. *Policy and Society* 39(1):91–112.

Paluck EL, Shepherd H, and Aronow PM (2016) Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences* 113(3):566–571.

Paxton A, and Griffiths TL (2017) Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods* 49(5):1630-1638.

Phan TQ, and Airoldi, EM (2015) A natural experiment of social network formation and dynamics. *Proceedings of the National Academy of Sciences* 112(21):6595–6600.

Rushmore J, Caillaud D, Matamba L, Stumpf RM, Borgatti SP, and Altizer S (2013) Social network analysis of wild chimpanzees provides insights for predicting infectious disease risk. *Journal of Animal Ecology* 82(5):976–986.

Salganik, MA (2019) *Bit by Bit: Social Research in the Digital Age.* Princeton University Press.

Shalizi CR, and Thomas AC (2011) Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research* 40(2):211–239.

Small ML (2011). How to conduct a mixed methods study: Recent trends in a rapidly growing literature. *Annual Review of Sociology* 37:57–68.

Smith JE, and Pinter-Wollman N (2021) Observing the unwatchable: Integrating automated sensing, naturalistic observations and animal social network analysis in the age of big data. *Journal of Animal Ecology* 90(1):62–75.

Szinovacz ME, and Egley LC (1995) Comparing one-partner and couple data on sensitive marital behaviors: The case of marital violence. *Journal of Marriage and Family* 57(4):995-1010.

Torres S (2019) On elastic ties: Distance and intimacy in social relationships. *Sociological Science* 6:235–263.

Traeger ML, Sebo SS, Jung M, Scassellati B, and Christakis NA (2020). Vulnerable robots positively shape human conversational dynamics in a human-robot team. *Proceedings of the National Academy of Sciences* 117(12):6370–6375.

Tubaro P, Ryan L, Casilli AA, and D'Angelo A (2021) Social network analysis: New ethical approaches through collective reflexivity. *Social Networks* 67:1–8.

van der Laan J, de Jonge E, Das M, Te Riele S, and Emery, T (2022) A whole population network and its application for the social sciences. *European Sociological Review*, online first. https://doi.org/10.1093/esr/jcac026