Measuring Networks in the Field

Jennifer M. Larson and Janet I. Lewis*

December 14, 2016

Abstract

Measuring networks in the field – usually by asking individuals systematically about their networks – entails complex design choices, with large consequences for the resulting data. Because observations in a network are interconnected, well-established practices from non-network survey settings can lead researchers astray; while small samples are informative about a population when observations are independent, small samples of nodes in a network often do not allow for meaningful inferences about the overall network. Despite the increasing focus on networks in political science, little guidance is available for researchers facing high-stakes decisions when designing a study to elicit networks. This article serves as a practical guide. It offers a simple framework for constructing a network theory, illuminates tradeoffs like measuring more nodes versus more ties per node or asking for names versus selections from a list, and suggests a new technique for cleaning relational data.

Social Networks | Fieldwork | Surveys | Research Design | Network Theory

9,695 Words

*We are grateful to Cesi Cruz, Guy Grossman, Franziska Keller, Lucy Martin, Matthew Simonson, and Josef Woldense for their thoughtful feedback and assistance.

1 Introduction

People's social relationships serve as crucial sources of information and influence, with important implications for politics. The presence, quality, and pattern of these relationships – features of social networks – have been found to affect outcomes like voting behavior (Sinclair, 2012), public opinion (McClurg, 2006), policymaking (Scholz, Berardo and Kile, 2008), social capital (Jackson, Rodriguez-Barraquer and Tan, 2012), protests (González-Bailón and Wang, 2016), and political violence (Parkinson, 2013; Larson and Lewis, 2016), as well as public health decisions (Rao, Mobius and Rosenblat, 2007) and technology adoption (Conley and Udry, 2010).

In the wake of promising findings like these, and because observational data on relevant networks is rarely available, scholars are increasingly undertaking fieldwork that entails eliciting original data on networks among individuals or small groups. Doing so has been central to advances in knowledge about, broadly speaking, the consequences of network position for individual behavior (Barr, Ensminger and Johnson, 2009; Baldassarri and Grossman, 2013; Baldassarri, 2015; Breza, Chandrasekhar and Larreguy, 2015); the spread of information through a group (Banerjee et al., 2014; Alatas et al., 2015; Larson and Lewis, Forthcoming); the presence and magnitude of peer influence (Conley and Udry, 2010; Banerjee et al., 2013; Paluck, Shepherd and Aronow, 2016); and the way social networks change in response to shocks (Fafchamps and Gubert, 2007; Jackson, Rodriguez-Barraquer and Tan, 2012). The contexts for these studies vary widely, for example from villages in the Philippines and India to high schools in the U.S.

In any context, however, measuring networks in the field is not a straightforward task. Since respondents in a network are interrelated – and thus cannot be taken as independent and identically distributed observations – problems that would be minor under "the usual" independent sampling procedures can compound into large problems that generate misleading results in a network context. To measure any one network, the way nodes and links are sampled, and the way survey questions are worded and ordered have large consequences for the resulting data. Moreover, for a given group of people, numerous different networks could be measured. Because resources, time in the field, and the attention span of respondents are all finite, researchers measuring networks face a large set of high-stakes tradeoffs that can be difficult to evaluate prior to collecting data.

Yet little practical guidance is available for the measurement of networks in the field. While researchers have ready access to guides to formal network theory (Wasserman and Faust, 1994; Jackson, 2010; Easley and Kleinberg, 2010; Siegel, 2011), statistical methods using network data that already exist (Butts, 2008; Kolaczyk, 2009), and causal inference on networks (Fowler et al., 2011), such works do not aim to assist researchers with the myriad choices involved when designing a study that collects original network data from people.[1]

This article's aim is to help researchers design a network elicitation study. Naturally, there is no one-size-fits-all design; the optimal design will vary study by study. Consequently, we stress that carefully first specifying a network-related theory is crucial in order to guide complex design decisions. Simply put, one must know which network(s) she is trying to capture and why before designing a study to do so. We begin by presenting a simple framework for constructing such a theory.

We then systematically address a battery of dimensions on which network studies can vary, and hence about which a researcher faces choices. We highlight the often subtle tradeoffs entailed in each. The first set of tradeoffs we consider pertains to the method of eliciting

[1] An exception is Marsden (2005), which reviews the findings from studies testing the consequences and reliability of design decisions in sociology and psychology. This is a useful starting point, but leaves many of the issues faced by political scientists that are covered in this article untouched.

the network: how exactly should the study identify the presence of a link between two people? The second set pertains to precision: how precisely must the true network be measured, and what are the consequences of a variety of options for measuring it more coarsely? The third set pertains to cleaning the data after collection: how should problems like spelling discrepancies across names or ties to people outside the sample be handled? The third set is easy to ignore ex ante, but anticipating these problems allows researchers to build in straightforward safeguards that vastly simplify the cleaning ex post. We also introduce a new technique for cleaning relational name data. Throughout, we emphasize that many of the tradeoffs can only be adjudicated by theory.

Finally, we emphasize the need to determine the importance of network *structure* to the research question, and offer a classification to help with this assessment. The more the research question depends on structure – identifying how central people are, how socially distant any two people are, and so on – the more important it is to measure a network as thoroughly as possible. Since most network studies seek to uncover network structure and small samples of nodes recover structural properties poorly (Chandrasekhar and Lewis, 2011), a key implication is that *most network studies would benefit from prioritizing surveying as many nodes as possible.* This consideration is particularly important for researchers planning to collect data in, and contrast network structures between, multiple locations. While measuring networks in many locations with few nodes each can be a tempting option, the ability to accurately recover properties of the networks that *are* measured usually makes measuring fewer networks with more nodes the best choice.

2 Measuring the Right Network

The first step in any networks study is identifying the network that is of interest. Although at first blush this may seem like a straightforward task, in practice, specifying a theory of which network matters and why can be challenging, warranting careful thought and substantive knowledge of the context under study. All design decisions hinge on having a carefully specified network theory.

2.1 Specifying a Network Theory

A network among a set of nodes N (here usually taken to be a set of people) is a record of the presence or absence of a link (also called a "tie" or an "edge") between every pair in N. For any set of nodes N, pairs of nodes could be linked in many ways.

Formulating a network theory requires three steps:

1. Define a set of nodes
2. Specify the type of links
3. Specify the function of the links

Table 1: Constructing a Network Theory

Attribute	Question to Ask	Examples
Nodes	What is the set of things that are connected?	Villagers; adult females; members of an organization; voters
Link Type	What is the relevant connection between two nodes?	Friendship; trust; shared geographic space; experience as colleagues; kinship
Link Function	What does a link between two nodes do? Why is the presence of such a link important?	Transmits information; spreads disease; exposes one to other's opinions; conveys peer pressure

In many studies, the nodes of interest will be people. When this is the case, a network theory requires answering the questions: which people are included in the network, how are they connected, and what do the connections do? The final question is easy to overlook, but is essential to the design of network elicitation. Networks are theorized to matter because the ties *do* something: transmit information, spread disease, indicate opportunities for learning, generate obligation, and so on. Being precise about their theorized function will simplify many of the design choices. Table 1 summarizes these steps.[2]

As an example, in a study of how information spreads through a village, a theory may hold that people receive new information from people they trust. This theory would regard the set of village residents as the nodes, trust as the link type of interest, and the transmission of credible information as the link function.

As another example, in a study of how people decide whether to vote, a theory may hold that a person can feel compelled to vote if her most intimate friends would judge her badly if she abstains. For this theory, a voting district's population could be the nodes, the link type is close friendship, and the ability to convey effective peer pressure is the link function.

An obvious but important point is that for any set of nodes, there are many different links that could connect them in theory. This is particularly true for studies theorizing that the "social network" is the network of interest. There is no single "social network" among a group of people, even in theory. Links connoting blood relatives, friends, business confidants, coworkers, conversation partners, and schoolmates are all sociable. Researchers must first carefully specify *why* sociable ties may matter, and then define and measure ties that correspond to this function. If ties matter in theory because they let people share trusted

[2]The link type may be the same as the link function, but it is important to specify the link function as precisely as possible. To study the spread of new technology through a village, specifying the link type as "friendship" and the link function as "connoting friendship" is not very useful. *Why* might friendship matter for the spread of technology? The answer to this question– perhaps that friends model each other's behavior– will be a more useful link function to specify.

information face-to-face, then the ties must capture the ability to share news face-to-face as directly as possible. If ties matter because they let people verify uncertain information, then the ties must capture the ability to verify information. While it may be tempting to remain agnostic and say the social network entails all of these link types, this imprecision in the theory can result in data that mask network effects.

For each of the design issues raised in the remainder of the article, we will discuss how to evaluate the tradeoffs with respect to the network theory underlying the project. The more precisely the nodes, link type, and link function can be specified, the easier it will be to evaluate the tradeoffs inherent in the study design.

2.2 Operationalizing a Network Theory

Once the network theory is specified by following steps (1)-(3), network measurement requires two additional steps:

4. Determine which nodes to survey

5. Operationalize link type

To complete step (4), identifying the boundary of a group of interest and then surveying everyone within this boundary is often ideal. Whether smaller samples can be informative depends on how important capturing structural features is. We consider this issue in section 4.

Step (5) entails pinning down a set of rules to establish when a link will count as present in the data. Doing so requires first selecting a procedure for measuring links. We introduce the options for measuring links in the next section, and then return to a discussion of how to operationalize links in ways that are tailored to these options in Section 3.2.

3 Eliciting Networks

Researchers have a large option set for how to elicit the identities of network ties. Different techniques are optimal in different contexts. How respondents are prompted to give names, which tie types are measured, the order of the questions, and how missing respondents are handled while in the field all affect the resulting picture of the network in the data; theory-driven choices will ensure that the data are informative for the research question.

3.1 How to Collect the Ties

The two main techniques for directly eliciting social networks are asking respondents to name names– the "name generator" approach– and asking respondents to select names from a list.

The list approach has obvious virtues. Respondents are less likely to forget ties if they are selecting them from a full census (Brewer, 2000). Consequently, list methods are likely to recover more weak ties or ties that are less salient to the respondent. Just how many more varies by context: different studies have shown that name generators recover only ten percent of ties (Hammer, 1984) or as much as eighty or ninety percent of ties that list methods recover (Sudman, 1988; Brewer, 2000).[3]

The list method is only possible for situations in which a census is available, and only appropriate when the census is also a theoretically meaningful restriction on ties. For instance, a village roster will let researchers use the list method if ties to other villagers exhausts the relevant ties. Measures of connectivity to other villagers, or paths throughout the village

[3]Because list methods draw out weak ties, they risk over-reporting to such an extent as to render the network meaningless. Pretesting is important to reveal whether respondents are inclined to answer "yes" to the presence of a tie to every single person on the list. If so, changing the question to capture a more restrictive relationship or restricting the number of names that may be chosen may be important.

through which messages could spread can be captured well. Overall popularity could not be, since this may depend on ties to individuals outside the village.[4]

Researchers do have the option of first assembling a roster in order to use the list method, but should consider whether this is appropriate given the study. For example, in Larson and Lewis (Forthcoming), the study's aim was to track the spread of brand new information carefully seeded. The act of assembling a roster would have risked contaminating the natural spread of information from the few sources.

List methods also risk exhausting respondents if the census is large. A list of hundreds of names (or photos, for contexts of low literacy) may introduce error as respondents become less willing to carefully consider options. Using a pretest to gauge the length of time required to consider all names in the census will inform the decision, and randomizing the list order is a useful safeguard.

When the boundary of network ties is straightforward, a census is readily available, the length of the census is not too long, and erring on over-inclusion is helpful to the study, list methods can be best.

Name generators are questions of the form "With whom do you discuss important matters?" or "Name up to 5 people with whom you have shared an office in the past." These questions rely on the respondent recalling and reporting the links of interest.

Name generators need not restrict the option set of ties (though they can by prompting respondents to only consider people within some set: "Name up to five ... *in this village*") and are fine in any population size. This method has been shown to collect relatively strong ties, and the ties offered in response to a network prompt tend to be socially related to one another as well (Fiske, 1995; Brewer, 1995; Marin, 2004), suggesting this method may detect

[4] For more on selecting the boundaries of a network, see Laumann, Marsden and Prensky (1989).

clustering well.[5]

When the boundary of network ties is not straightforward, when a census is not readily available or assembling one would damage the study, when the population is large, or when distinguishing strong ties is helpful to the study, name generators can be best.[6]

Most of these determinants are matters of feasibility. The issue of strong tie detection is a matter of theory. Given the link function, will the research question be better answered if the respondent gives every single person who meets a criterion, even if she needs to be reminded by seeing that name on a list, or will it be better answered if the respondent only gives the names most salient to the respondent that meet the criterion? If the former, over-including weak ties is best, which favors the list method; if the latter, prioritizing strong ties is best, which favors the name-generator method.[7]

The logistics of recording names offered by respondents can introduce data cleaning issues, especially spelling discrepancies, which we cover in Section 5.

---

[5]Shakya, Christakis and Fowler (2017) find that, in a comparison of twelve name-generators used in rural villages in India, those measuring domestic interactions detect the highest values of clustering.

[6]For more on the choice between specific kinds of name generators, especially with applications to sociology, see Klofstad, McClurg and Rolfe (2009); Bidart and Charbonneau (2011); Sokhey and Djupe (2014).

[7]For example, for a theory that holds that social ties are important because they spread awareness of a new product, over-including weak ties may be best. This kind of information probably spreads easily across lots of links. Conversely, for a theory positing that social ties let people confide their highly sensitive secrets about a regime with one another, aiming to capture the strongest ties may be most informative.

3.2 Operationalizing Link Types

Given a method for collecting the ties, there are three issues in operationalizing the link type: what to measure, how to prompt respondents to reveal the link, and how many different measurements of a link type to take.

The more precisely the link function is specified in the theory, the more straightforward it will be to identify what should be measured: the presence of opportunities for that function to occur. For example, in Barr, Ensminger and Johnson (2009), theory suggested that the relevant link type was social relationships, which function to make villagers behave in a trusting and trustworthy way with one another. The authors operationalized this link type by asking about an activity that likely connotes trust: "Who do you usually talk to about any kind of problem in this village" (p. 74). As another example, because theory suggests that social ties function as sources of political discussion (Klofstad, McClurg and Rolfe, 2009), the 1985, 1987, and 2004 versions of the General Social Survey operationalize the link type by inquiring about people with whom a respondent discusses important matters: "From time to time, most people discuss important matters with other people. Looking back over the last six months – who are the people with whom you discussed matters important to you?"

Eliciting networks through surveys relies on respondents recalling their network ties. In general, people are better at recalling salient, specific, concrete things. A useful starting point for operationalizing the link type as concretely as possible is to build elicitation questions that capture the link function. Sometimes the question can directly ask about the link function. If the function is still too abstract, the question can ask about a more concrete instance of the abstract concept.

For example, the elicitation question in Barr, Ensminger and Johnson (2009) did not ask "whom do you trust?" but asked about a specific activity that should indicate trust. Rather than ask "with whom do you have transactional relationships," Banerjee et al. (2013) asked about specific activities like borrowing and lending rice. In Paluck, Shepherd and Aronow (2016), to identify social referents, respondents were asked "whom did you choose to spend time with, face to face or online?" This elicitation question picks up susceptibility to influence without directly asking about susceptibility by asking about the specific event of spending time with others.

Asking about concrete activities not only helps with recall, but also avoids a related problem with prompts that are too vague. Respondents in different contexts may understand a vague concept to be different than the meaning intended by the researcher. The concept of "friendship" is particularly susceptible to this problem: a mismatch between researchers' and respondents' understanding of the concept has been documented in US samples (Fischer, 1982) and has been anecdotally noted by researchers conducting pretests in the developing world. Asking about activities that should indicate what the researcher means by friendship is a more promising approach than asking directly about friendship.[8] For instance, if friendship is theorized to matter because friends spend time together and expose one another to their political views, asking about those with whom the respondent spends a lot of time may be preferable to asking about those whom the respondent regards as a friend. In this case, appealing to the link *function* rather than type– why friendship is theorized to matter in a particular context– is more likely to yield meaningful network data.

If the link function can be captured reasonably well by a single question, then money,

[8]A clever approach to overcome this difficulty can be found in Rao, Mobius and Rosenblat (2007). The researchers were interested in friendship. Rather than ask respondents to list friends, respondents were asked to list others with whom they would like to play a game. In the game, one person would be asked a basic question about the other, and correct answers would earn money. This strategy incentivized respondents to name others who were most likely to know this information about themselves, focusing respondents' attention on the same aspect of friendship that the authors were interested in.

time, and attention are saved by asking just the one question to elicit ties. If the link function is facilitating the borrowing and lending of rice, then asking about the borrowing and lending of rice should suffice. If instead the theoretical tie of interest does not have a single obvious operationalization, or if the function is abstract and can only be made more concrete by asking about multiple aspects separately, then multiple measures of the tie type may be necessary. For instance, the link function relevant to Larson and Lewis (Forthcoming) is facilitating the spread of news by word-of-mouth. The authors operationalized this relatively abstract function by asking about seven concrete activities that should generate opportunities for spreading news, including sharing meals, talking on the phone, and visiting homesteads. Section 4.1.3 considers the related issue of measuring multiple tie types.

As a final note, some link types are easier to measure than others, which can generate temptation to operationalize the link type with an easier-to-measure proxy. For instance, the availability of GPS devices makes geographic networks easy to measure, so that two individuals share a link if their homes are geographically close. Shared membership in groups can be easier to learn via surveys than the individuals to whom a person is connected. Common origin can also be relatively easy to learn. While there are some research questions for which these networks will be meaningful, there are many others for which they will not.

The theorized link function helps sort the two. First ask: given the theorized link function, how likely is it that all links with this function will be captured by the proxy? Then ask: given the theorized link function, how likely is it that all links detected by the proxy serve this function? The greater the likelihood *for both questions*, the better the proxy. If the answer to the first question is "very likely" and the answer to the second question is "not very likely," then the correct links are likely to be a subset, possibly a small subset, of the links contained in the proxy. This effectively over-counts links. For some research questions, over-counting links results in poor inferences: see Section 4.1.2. For more on using proxies in place of direct elicitation survey questions, see Gross and Jansa (2016).

3.2.1 The Order

As with any survey, researchers should think carefully about the order in which networks questions appear. Since people recall things that are more salient to them, early parts of a survey can make those topics more salient to a respondent when answering later questions.

One implication is that if documenting the presence or absence of ties between individuals with certain attributes is important, if there is a chance that questions about the attributes will tip the respondent off about the research question, asking about the network ties first can be best. For instance, if a researcher is interested in documenting the extent of ties between long-time residents and new migrants in a village, beginning with a battery of questions about migration status may prime a respondent to over-report ties to the other group.

A second implication is that if multiple network elicitation questions appear and one network could in theory be a subnetwork of another, asking about the broader network first can maximize the difference detected between the two. For instance, if researchers are interested in people with whom respondents spend time in general, and also people with whom the respondent shares meals, the latter is in principle a subnetwork of the former. However, the respondents may also spend time with many people with whom they do not share meals. Asking about the subnetwork, in this case the shared meals network, first risks making these ties more salient during the survey, inducing the respondent to list the same people again for the time network. Asking about time first increases the chance of detecting additional distinct ties in the more general network.

3.2.2 Nested or Separate Questions

When multiple network elicitation questions appear, nesting the network questions can save time. However, this approach will only generate meaningful data for a narrow set of research questions.

Nested questions ask first about one kind of tie, and then further classify only those names offered. For example, a researcher would ask “with whom do you share meals?” and then ask *of those with whom the respondent shares meals* “do you exchange gifts with any of these? If yes, which?”

This method necessarily constrains the subsequent networks to be subnetworks of the first. The only gift exchange ties that will be recorded will be those that are also shared meals ties. If the research question is strictly concerned with the gift exchange network and knowing whether or not they also share meals, this method is fine. However, if the interest is in any comparison between a meals and a gift exchange network, this method is inappropriate. Nesting introduces distortions; for instance, the gift exchange network will mechanically appear weakly less dense than the meal network, and likewise will mechanically have weakly smaller out-degree.

The most common reason for eliciting multiple networks is that theory is weak or a goal is to adjudicate among different networks.[9] For these purposes, it is important that respondents have the opportunity to offer their ties separately for each.

4 Level of Precision

A precisely-measured network among a set of nodes N is a record of the presence or absence of a link for every pair of nodes in N. Precision is expensive: network portions of surveys take a substantial amount of time. The longer the survey, the fewer can be collected on a fixed budget and the greater the risk of exhausting respondents’ attention before the survey’s end.

The key determinant of how precisely a network must be measured is the extent to which the network *structure* is of interest. At one extreme are studies concerned exclusively with

[9]See Banerjee et al. (2013).

independent attributes of nodes. These are uninterested in how the nodes relate to one another and instead seek to estimate, say, the mean age or education status of a population. Small random samples of nodes will recover estimates of population values reasonably well. At the other extreme are studies interested in the way each person is interconnected with every other person. Studies interested in how central each person is in the network relative to each other person (eigenvector centrality), or on average how socially proximate a person is to every other person (closeness centrality) are closer to the second extreme. Sampling a number of nodes less than the full population, even randomly, can poorly represent these values; small samples perform particularly badly (Lee, Kim and Jeong, 2006; Chandrasekhar and Lewis, 2011).[10]

Figure 1 helps classify types of network attributes along the continuum from fully independent to fully relational data. Some network attributes can be recovered with small samples. Out-degree– the number of links from a person to others– can be measured without perfectly representing the full network (and we discuss simplified versions for doing so below). As the network feature becomes more intricate and depends on how others relate to more and more other people, samples perform worse and worse (and hence small samples become a bigger problem).

A good rule of thumb for studies interested in characterizing the network structure and capturing these relational features is to make sampling the largest possible number of nodes the top priority.[11]

[10]Researchers can calculate statistics like centrality for the measured networks, but the value and even the ranking of nodes' centrality scores may bear little resemblance to the scores in the true network.

[11]Relatedly, missing network data poses large problems as well. Research suggests that returning to individuals who were unavailable to be surveyed the first time, or even asking someone else to

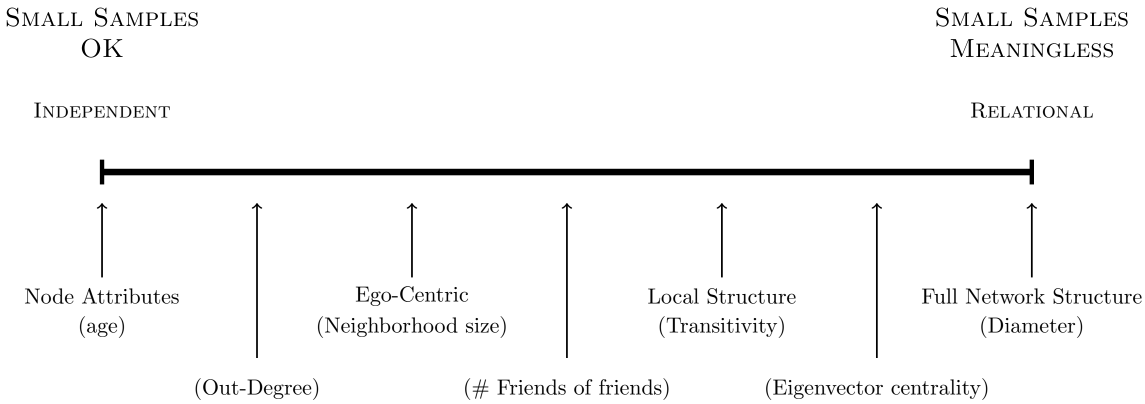

Figure 1: The more the network attributes of interest depend on interconnections throughout a network, the more precisely a network needs to be measured. Studies interested in network features like centrality, path lengths, transitivity, reciprocated ties, etc. should prioritize surveying as many nodes as possible.

4.1 Reductions in Precision and Their Costs

In a perfect world, researchers interested in network structure would measure the network with full precision– capture every node of interest, and measure the presence or absence of every possible link among them. There are options for reducing precision, all of which shorten the survey at the cost of introducing some error. Here we present the tradeoffs surrounding five of them. Researchers should take these tradeoffs into careful consideration before deciding to reduce the precision with which the network is measured.

Each will refer to the respondent offering her network information as an "ego" and the names she lists as her "alters."

answer on an unavailable respondent's behalf, may be preferable to trying to account for the missing node (or ignoring it) ex post (Marsden, 1990; Sudman et al., 1994; White and Watkins, 2000). In these cases it will be important to note which observations were recorded by proxy and assess the robustness of results to noise in these reports.

4.1.1 Reducing Precision by Restricting the Set of Candidate Alters

Suppose the population of interest is a village. In theory, individuals in this village may have ties to others in the village, as well as ties to individuals outside the village. One option is to restrict the set of individuals who may be named by respondents to others in the village, either by restricting the names on the roster, or instructing respondents to only offer these names (see Section 3).

Restricting the option set reduces information about a broader network. Whether this matters depends on theory. Take the borrowing network measured with the above elicitation question which measures borrowing behavior *within the village*. If the research question is strictly about borrowing within the village, or if people never borrow from anyone except members of their village, then this restriction is sensible. However, if the research question is about borrowing in general, then this restriction may be detrimental, especially for comparisons of node-level network attributes. For instance, people who show up as having low out-degree in the restricted network may in fact have high out-degree (many people from whom they have borrowed) in a multi-village network. Likewise, people who show up as peripheral in this network may in fact be highly central in a multi-village network.

In short, if network position with respect to people outside of a set is important to the theory, then ties to people outside of the set should be allowed; if it is not, these ties can be excluded.

4.1.2 Reducing Precision by Capping the Number of Links

Any one respondent in N can have as many as $n-1$ links to others in N, and possibly many more to others outside of N. Researchers have the option of letting a respondent list or select from a roster every single alter that meets a criterion.

The downside is that each extra name offered or considered adds time to the survey. If the network is dense, lots of respondents will have many names to name or select. Researchers have the option to cap the number of alters that each respondent can offer.

A cap on the number of alters tends to take the form "name *up to* x people with whom you have ..." or "select *up to* x people from the following list with whom you have ..." The larger the value of x, the longer the survey will take. Research has shown that eliciting three to five alters produces reliable information on network density (Marsden, 1993), and that eliciting one alter for multiple networks is worse than eliciting multiple alters for fewer networks (Marin and Hampton, 2007).

Two considerations should guide the choice of x. First, is variance in out-degree important to the research question? A cap on the number of alters right-censors out-degree. Individuals who offer x names all appear the same in the data even though some may truly have more than x out-links in the network. If a key analysis will entail distinguishing respondents based on their degree centrality or popularity, then x must be large enough to generate variance.[12] Smaller values of x compress variance.

A good rule of thumb is to begin with $x = 5$ and pretest the networks questions. If all pretested respondents provide 5 names, increase the number. If none do, keep 5 for the survey.

Note that degree is not the only network statistic that relies on variance in out-degree. Sometimes reciprocated ties– a names b and b names a– are used to proxy for strong ties. This measure works best when any tie is given sufficient opportunity to show up as reciprocated in the data. If everyone names one name, few ties that are genuinely reciprocated will show up in the data as reciprocated. The larger is x, the more reciprocated ties will be detected.

A second consideration is whether the absence of a link is important to answering the

[12]For example, Eveland Jr, Hutchens and Morey (2013) shows that recording too few ties in a political discussion network misses detecting hubs– the few people with many more ties than average– who play important roles in political discussion.

research question. Is the absence of a link from a to b important to document? Projects focused on the way that people are connected can afford to miss links; projects focused on the way that people are *not* connected need to more carefully document the absence of links. In a project interested in how social ties are arranged, setting $x = 1$ could be informative. In a project interested in whether two tribes are connected via social ties, setting $x = 1$ would be a poor design. The latter is interested in the absence of ties between tribes and so should offer a strong test for missing ties. The smaller the x, the weaker this test.

4.1.3 Reducing Precision by Limiting the Number of Network Types Measured

When a study's aim is to adjudicate between different network theories, multiple network types may need to be measured. Each network measured substantially lengthens the survey. To collect as few networks as possible, the networks should be as conceptually distinct as possible. Multiple distinct tie types help fill out a multidimensional network (Marin and Hampton, 2007), and will improve the efficiency of attempts to distinguish the consequences of each network. Starting by enumerating the categories of ties important to the research question – economic transactions, social interactions, information transmission, familial links, etc– and then selecting one concrete tie type from each can be a useful starting point.

An important consideration will be: to what extent are names offered in response to this question necessarily also going to be offered in response to another question? For instance, a question that asks "whom do you regularly visit" and another that asks "with whom do you regularly spend time" are not maximally distinct. The latter likely contains many of the former. Nested networks are redundant and will waste survey time. For the second scenario, the researcher must consider: how can I cover the full scope of this concept– say, trust– with the fewest questions? A question that asks whom one trusts with a secret and a question that asks whom one would trust with the news that she plans to run for office will likely contain more overlap, and hence more time-wasting redundancy, than a second question about trusting to pay back a loan.

4.1.4 Reducing Precision by Measuring Households Instead of Individuals

An option for shortening the total time devoted to surveying networks is to measure networks at the household instead of the individual level. In this case, each node would be a household, and a link would be present between households if any individual in the household lists an individual in the other household in response to a networks question.

The advantage is that this effectively reduces the sample size, often on the order of at least four or five times. The disadvantage is that this compresses potentially interesting information within household. If the goal is to measure the reach of training in fertilizer use, losing within-household variation may be fine. Having one person in the household trained means that even if the others have not yet learned, they have easy access to this information. Connections among households then could be meaningful channels along which this information spreads.

If the research question focuses on attributes of ties, however, then the household may be an inappropriate level of resolution. This will be true for any potential node or link attribute on which there is heterogeneity within household. Consider cross-ethnic ties. If every member of a household is the same ethnicity, then ties between households can be categorized as cross-ethnic or not in a way that captures well whether the individual ties are cross-ethnic or not. However, if households can be mixed ethnicity, then characterizing ties between households as cross-ethnic or not can at best give a partial cross-group tie score that may poorly capture the real relationships at the individual level.

Two issues should guide the choice of aggregating up to the household level versus keeping networks at the level of the individual. First, is the research question at the level of the household or the individual? If at the level of the household, then so long as cross-household links are defined and operationalized carefully, this can be the best network to measure.[13] Second, does the research question require information about nodes or links for which there is within-household variance? Answering the second question may be difficult ex ante, but can be estimated with pretesting or by contacting local experts in advance. If yes, then aggregating at the household level may be inappropriate.

4.1.5 Reducing Precision by Using Proxies for Network Position

One way to dramatically save survey time is to measure properties of the network without measuring the network itself. The most popular version of these measures aims to estimate the size of a person's social network. Rather than asking respondents to name individuals, the researcher asks for the number of ties with a series of questions of the form "how many X do you know" where X is a reasonably popular first name. Responses can be used immediately as a relative measure of popularity, or summed or scaled to estimate degree in a social network (Killworth et al., 1998; McCarty et al., 2001; McCormick, Salganik and Zheng, 2010; Calvo and Murillo, 2013).

The number of people with certain first names that a person knows gives a measure of popularity or general social network size. These simplifications can also be used to estimate degree in more specific networks by asking questions of the form "with how many people do you Y" where Y is an activity or specific relationship. Using a single question that asks respondents to estimate the number of people in their network has been popular in communication studies and american politics (Moy and Gastil, 2006; Eveland and Hively,

[13]See the study of enrollment in a microfinance program in Banerjee et al. (2013). Networks are measured at the household level because enrollment is at the household level. Likewise, in Cruz, Labonne and Querubin (2014), networks are measured at the family level, where political support is mobilized.

2009; de Zúñiga and Valenzuela, 2011).

These simplifications offer substantial time savings; if the only network feature that the researcher needs to collect is degree, these can be worthwhile substitutes for eliciting precise networks and tend to perform reasonably well (Bell, Belli-McQueen and Haider, 2007). Moreover, these methods better recover true zeros: listing no names in response to a network elicitation question is a weaker signal that that person in fact has no ties of that sort than responding "zero" to the question of how many ties of a sort a person has (Fischer, 2009).

The most obvious downside is that no additional details of the network are measured. These methods have two other limitations. First, measures of popularity that depend on counts of the number of names a person knows can be difficult to relate to the precise underlying network. A person who knows more names has a larger social network in some sense, but many research questions require knowing more than this. A person who knows more people does not necessarily have the most access to credit or the most sources of gossip or the largest number of trusted contacts.[14]

Second, the flip side of this method providing a more conservative test for zero ties is that it provides a weaker test for the presence of ties. Studies that hinge on the presence or absence of strong, deep connections may do better eliciting names to recover this variance.

[14]While proxies for degree (network size) are most common, research suggests it may also be possible to detect who holds the most central positions in a network without measuring the network. Banerjee et al. (2014) shows that people can fairly reliably identify people who would be the best injection points for information diffusion in a network; if research hinges on identifying them, it may be possible to ask some people to do so rather than elicit whole networks.

5 Cleaning the Data

Even the most carefully collected data need to be cleaned. Network data present special challenges for cleaning because responses are interconnected. Even if measurement error is random with respect to links, the error may not equally affect nodes. Making sure ties are reported correctly, and carefully deciding how to assemble the ties into a network for analysis, makes or breaks ultimate inferences. There are steps the researcher can take when designing the survey to make cleaning the data easier after collection.

5.1 Identity Matching

Making sure names that appear in multiple places in the data– a respondent's own name and a tie offered by a different respondent, say– refer to the same person is essential to properly representing the network and its structural features.

Three actions can help match identities. First, if it is feasible to assemble a set of photographs in advance, individuals can distinguish between common names by identifying the picture of the person being named (Kim et al., 2015). Second, researchers should collect the version of names that is most likely to be unique. This requires knowledge of local context in advance. In American samples, this would entail first and last names. In the Ugandan sample of Larson and Lewis (Forthcoming), for example, this entailed collecting respondents' and their alters' Ugandan and Christian name. In other settings, for instance in the Philippines, nicknames are more likely to uniquely identify individuals than full names, which may overlap due to familial naming conventions.

Third, researchers can collect an additional identifier along with each name. The best identifier will depend on the context. The simpler the identifier, the less time this will add to the survey. For instance, a question "about how old is he/she" could accompany the elicitation question. Then, so long as respondents are sufficiently good at estimating age, the researcher can not only match names offered as ties to surveyed egos (as long as researchers ask about age in the survey); the names offered can be matched to each other as well.[15]

5.2 Spelling Discrepancies

Spelling discrepancies introduce a related problem of identity: names that are different in the data could in actuality refer to the same person. If Robert lists Susan and Gary lists Suson, whether Susan and Suson are treated as two people or as one whose name was spelled two different ways can result in quite different networks. If Susan and Suson are in fact the same person but treated as two, Susan's in-degree will appear lower than it truly is. If Gary is highly connected, Susan's eigenvector centrality will appear lower by missing the link to Gary. And of course if Susan and Suson were two different people, aggregating them into a single node will inflate the degree of either of them, as well as the centrality of any neighbors.

If the goal is strictly to measure the number of names offered by those surveyed– their out-degree– then spelling discrepancies are irrelevant.[16] If the goal is to measure any other property of networks, including the related in-degree or overall degree, then spelling discrepancies can result in a significantly mis-measured network.

In addition to the identity matching techniques of the last section, three straightforward actions can reduce this problem: asking for the spelling during the survey, using electronic tablets rather than paper surveys, and/or performing an enumerator handwriting check. The

[15]This can also mitigate a related problem where different respondents can have the same name. An identifier like age can help distinguish which of a set of same-named people a respondent is referring to. Contacting a local expert to inquire about the extent of shared full names in advance will help determine whether adding this kind of identifier will be necessary.

[16]Unless the goal is to measure the number of distinct names offered across multiple network elicitation questions– then spelling discrepancies even affect the measure of out-degree.

first may maximize the chances that the same person's name is spelled the same way, and the latter two minimize spelling discrepancies arising from handwriting issues. Of course none of these measures eliminate the problem. The first requires literacy, and more strongly, that respondents know the correct spelling of each of their ties. The latter two do not preclude typos.

Once data are collected, there are options for cleaning that account for the presence of same names spelled differently. The first is a qualitative approach. While in the field, assemble a list of similarly-spelled names that appear in the data and take them to a local expert, for instance a village leader. He or she will have the best idea for whether there are people with the different names or whether they likely refer to the same person.

The cleaning can also be automated in a way that accounts for the unique structure of network data. The list of egos is a set of people known to be distinct. This can be used as a key to see if any of the alters are spelled suspiciously similarly to any of the egos. A straightforward approach is a string-matching metric to assess similarity without human judgment. For instance, Levenshtein Distance records the number of single-character edits that would make two strings (names) be the same (Levenshtein, 1966). The raw data can be converted into cleaned datasets in which names which differ by certain values of Levenshtein Distance are regarded as the same.

For instance, the Levenshtein-1 dataset would be the dataset cleaned in the following way: loop over every alter named in the data. Any alter's name that differs from an ego by a single character (for instance, ego Susan Olson v. alter Suson Olson) is replaced by that ego's name. Then all remaining alters that differ from each other by a single character are treated as the same. The Levenshtein-2 dataset would be the same for names that differ by at most two characters, and so on.

The question becomes what number of different characters to tolerate. Of course the Levenshtein-x dataset, where x is the maximum number of characters in any name in the data, would replace all alters with the first ego. This is clearly too much tolerance. It could be that the Levenshtein-1 dataset is too little tolerance, leaving many same names appearing to be different nodes. While this is ultimately a judgment call, there are two ways to proceed. One is to select a range of Levenshtein distances that produce sensible matches, and perform analyses on all of them to test robustness. The second is to use a known global measure to select the Levenshtein distance in a principled manner. For instance, if the sample was a village, each choice of Levenshtein distance results in a potentially different total number of names in the dataset (included as egos or alters). A statistic like the village population can lend some insight. If the ties are likely to others in the village and a choice of Levenshtein distance results in a number of names that is many times the population of the village, then a higher Levenshtein distance is probably needed to properly clean the data. Village population can help tune the cleaning.[17]

5.3 Ties to Unsampled Alters

Even if respondents are asked to name people within a defined set, and even if researchers *try* to survey everyone in that set, there could still be alters in the data who were not surveyed. In this case, researchers have three types of names in their data: those in the population who were surveyed ("respondents"), those in the population who were named by a respondent but who were not themselves surveyed ("unsurveyed alters"), and those in the population who were neither surveyed or named.

Pooling respondents with unsurveyed alters can substantially increase the proportion of

[17]If respondents are free to list ties to anyone, including outside the village, the survey could ask about each one "does this person live in your village?" Then the diagnostic would compare the number of distinct names for whom the answer was "yes" to the known village population to select the appropriate Levenshtein distance.

the population that are included in the data. For this reason, doing so is tempting. For questions that rely on comparing in-degree, this is helpful. However, for other network measures, pooling respondents and those named may not be advisable. The data contain information on respondents measured in two ways: network information for respondents comes from names they offered and from others naming them. Network information for unsampled alters comes only from others naming them; they were never eligible to offer names themselves. This difference can introduce distortions, especially in out-degree and centrality.[18]

A useful guiding principle is to check the robustness of analyses using structural features of the full network to instead using the closed network– the network that contains only respondents and links among them.

6 Conclusion

As theories of political phenomena increasingly account for interconnectedness between people, researchers must adjust our methods of data collection to capture these relationships. The above is intended as a guide, in order to help researchers think through design issues before eliciting networks in the field. Doing so is an increasingly common research practice in political science, and a necessary one since data on relevant networks are rarely available via observational data.

We have walked the reader through several important trade-offs to consider when designing a study that will measure networks in the field. Throughout, we have stressed the hazards of importing sampling techniques from settings where observations are independent, and the

[18]To see why, note that those named but not surveyed are more likely to have fewer of their links measured. This means that their degree, closeness centrality, and eigenvector centrality will be under-valued relative to respondents.

importance of considering design trade-off before arriving in the field. Careful thought ex ante maximizes the chance that scholars make the best use of finite time and resources in the field and return with data that can meaningfully answer research questions. In particular, the stronger the underlying network theory – which nodes matter, exactly which links among them matter and why – the easier the design will be. While research teams heading to the field can consider theory case by case, this guide also highlights the value of strong theory on its own.

Our hope is that this guide helps to make the collection of network data in the field easier and less error-prone, and encourages those who would have otherwise been deterred by such an onerous task to reconsider. The more political science is able to pin down the relationship between networks and political outcomes, and can be precise about which links in networks affect which outcomes, the more fruitful the empirical study of networks will be.

References

Alatas, Vivi, Abhijit Banerjee, Arun G Chandrasekhar, Rema Hanna and Benjamin A Olken. 2015. Network structure and the aggregation of information: Theory and evidence from Indonesia. Technical report National Bureau of Economic Research.

Baldassarri, Delia. 2015. "Cooperative Networks: Altruism, Group Solidarity, Reciprocity, and Sanctioning in Ugandan Producer Organizations." *American Journal of Sociology* 121(2):355–395.

Baldassarri, Delia and Guy Grossman. 2013. "The Effect of Group Attachment and Social Position on Prosocial Behavior: Evidence from Lab-in-the-Field Experiments." *PLoS ONE* 8(3):e58750.

Banerjee, Abhijit, Arun G Chandrasekhar, Esther Duflo and Matthew O Jackson. 2013. "The diffusion of microfinance." *Science* 341(6144):1236498.

Banerjee, Abhijit, Arun G Chandrasekhar, Esther Duflo and Matthew O Jackson. 2014. Gossip: Identifying central individuals in a social network. Technical report National Bureau of Economic Research.

Barr, Abigail, Jean Ensminger and Jeffrey C Johnson. 2009. Social networks and trust in cross-cultural economic experiments. In *Whom Can We Trust*, ed. Karen S Cook, Margaret Levi and Russell Hardin. Russel Sage Foundation pp. 65–90.

Bell, David C, Benedetta Belli-McQueen and Ali Haider. 2007. "Partner naming and forgetting: recall of network members." *Social networks* 29(2):279–299.

Bidart, Claire and Johanne Charbonneau. 2011. "How to generate personal networks: Issues and tools for a sociological perspective." *Field Methods* 23(3):266–283.

Brewer, Devon D. 1995. "The social structural basis of the organization of persons in memory." *Human Nature* 6(4):379–403.

Brewer, Devon D. 2000. "Forgetting in the recall-based elicitation of personal and social networks." *Social Networks* 22(1):29–43.

Breza, Emily, Arun G Chandrasekhar and Horacio Larreguy. 2015. "Network Centrality and Institutional Design: Evidence from a Lab Experiment in the Field." *NBER Working Paper* 20309.

Butts, Carter T. 2008. "Social Network Analysis: A Methodological Introduction." *Asian Journal of Social Psychology* 11:13–41.

Calvo, Ernesto and Maria Victoria Murillo. 2013. "When Parties Meet Voters Assessing Political Linkages Through Partisan Networks and Distributive Expectations in Argentina and Chile." *Comparative Political Studies* 46(7):851–882.

Chandrasekhar, Arun and Randall Lewis. 2011. "Econometrics of sampled networks." Unpublished manuscript, MIT.
URL: *goo.gl/Cp6T02*

Conley, Timothy G and Christopher R Udry. 2010. "Learning about a new technology: Pineapple in Ghana." *The American Economic Review* pp. 35–69.

Cruz, Cesi, Julien Labonne and Pablo Querubin. 2014. Politician family networks and electoral outcomes: Evidence from the Philippines. In *Annual Meeting of the American Political Science Association*.

de Zúñiga, Homero Gil and Sebastián Valenzuela. 2011. "The mediating path to a stronger citizenship: Online and offline networks, weak ties, and civic engagement." *Communication Research* 38(3):397–421.

Easley, David and Jon Kleinberg. 2010. *Networks, Crowds and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

Eveland Jr, William P, Myiah J Hutchens and Alyssa C Morey. 2013. "Political network size and its antecedents and consequences." *Political Communication* 30(3):371–394.

Eveland, William P and Myiah Hutchens Hively. 2009. "Political discussion frequency, network size, and "heterogeneity" of discussion as predictors of political knowledge and participation." *Journal of Communication* 59(2):205–224.

Fafchamps, Marcel and Flore Gubert. 2007. "Risk sharing and network formation." *The American Economic Review* 97(2):75–79.

Fischer, Claude S. 1982. "What do we mean by "friend"? An inductive study." *Social Networks* 3(4):287–306.

Fischer, Claude S. 2009. "The 2004 GSS finding of shrunken social networks: An artifact?" *American Sociological Review* 74(4):657–669.

Fiske, Alan P. 1995. "Social schemata for remembering people: Relationships and person attributes in free recall of acquaintances." *Journal of Quantitative Anthropology* 5(4):305–324.

Fowler, James H, Michael T Heaney, David W Nickerson, John F Padgett and Betsy Sinclair. 2011. "Causality in Political Networks." *American Politics Research* 39:437–480.

González-Bailón, Sandra and Ning Wang. 2016. "Networked discontent: The anatomy of protest campaigns in social media." *Social Networks* 44:95–104.

Gross, Justin and Joshua M Jansa. 2016. Relational Concepts, Measurement, and Data Collection. In *The Oxford Handbook of Political Networks*, ed. Jennifer Nicoll Victor, Alexander H Montgomery and Mark Lubell. Oxford.

Hammer, Muriel. 1984. "Explorations into the meaning of social network interview data." *Social Networks* 6(4):341–371.

Jackson, Matthew O. 2010. *Social and Economic Networks*. Princeton University Press.

Jackson, Matthew O, Tomas Rodriguez-Barraquer and Xu Tan. 2012. "Social capital and social quilts: Network patterns of favor exchange." *The American Economic Review* pp. 1857–1897.

Killworth, Peter D, Christopher McCarty, H Russell Bernard, Gene Ann Shelley and Eugene C Johnsen. 1998. "Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach." *Evaluation Review* 22(2):289–308.

Kim, David A, Alison R Hwong, Derek Stafford, D Alex Hughes, A James O'Malley, James H Fowler and Nicholas A Christakis. 2015. "Social network targeting to maximise population behaviour change: a cluster randomised controlled trial." *The Lancet* 386(9989):145–153.

Klofstad, Casey A, Scott D McClurg and Meredith Rolfe. 2009. "Measurement of Political Discussion Networks A Comparison of Two "Name Generator" Procedures." *Public Opinion Quarterly* pp. 462–483.

Kolaczyk, Eric D. 2009. *Statistical Analysis of Network Data*. Springer.

Larson, Jennifer M and Janet I Lewis. 2016. "Rumors, Kinship Networks, and Rebel Group Formation." Working Paper.
URL: *goo.gl/0dBMbt*

Larson, Jennifer M and Janet I Lewis. Forthcoming. "Ethnic Networks." *American Journal of Political Science* .

Laumann, Edward O, Peter V Marsden and David Prensky. 1989. "The boundary specification problem in network analysis." *Research Methods in Social Network Analysis* 61:87.

Lee, Sang Hoon, Pan-Jun Kim and Hawoong Jeong. 2006. "Statistical properties of sampled networks." *Physical Review E* 73(1):016102.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady*. Vol. 10 p. 707.

Marin, Alexandra. 2004. "Are respondents more likely to list alters with certain characteristics?: Implications for name generator data." *Social Networks* 26(4):289–307.

Marin, Alexandra and Keith N Hampton. 2007. "Simplifying the personal network name generator alternatives to traditional multiple and single name generators." *Field Methods* 19(2):163–193.

Marsden, Peter V. 1990. "Network data and measurement." *Annual Review of Sociology* pp. 435–463.

Marsden, Peter V. 1993. "The reliability of network density and composition measures." *Social Networks* 15(4):399–421.

Marsden, Peter V. 2005. "Recent developments in network measurement." *Models and Methods in Social Network Analysis* 8:30.

McCarty, Christopher, Peter D Killworth, H Russell Bernard, Eugene C Johnsen and Gene A Shelley. 2001. "Comparing two methods for estimating network size." *Human Organization* 60(1):28–39.

McClurg, Scott D. 2006. "The electoral relevance of political talk: Examining disagreement and expertise effects in social networks on political participation." *American Journal of Political Science* 50(3):737–754.

McCormick, Tyler H, Matthew J Salganik and Tian Zheng. 2010. "How many people do you know?: Efficiently estimating personal network size." *Journal of the American Statistical Association* 105(489):59–70.

Moy, Patricia and John Gastil. 2006. "Predicting deliberative conversation: The impact of discussion networks, media use, and political cognitions." *Political Communication* 23(4):443–460.

Paluck, Elizabeth Levy, Hana Shepherd and Peter M Aronow. 2016. "Changing climates of conflict: A social network experiment in 56 schools." *Proceedings of the National Academy of Sciences* 113(3):566–571.

Parkinson, Sarah Elizabeth. 2013. "Organizing rebellion: Rethinking high-risk mobilization and social networks in war." *American Political Science Review* 107(03):418–432.

Rao, Neel, Markus M Mobius and Tanya Rosenblat. 2007. "Social networks and vaccination decisions." Working Paper Series Federal Reserve Bank of Boston.
URL: *http://hdl.handle.net/10419/55601*

Scholz, John T, Ramiro Berardo and Brad Kile. 2008. "Do Networks Solve Collective Action Problems? Credibility, Search, and Collaboration." *Journal of Politics* 70:393–406.

Shakya, Holly B, Nicholas A Christakis and James H Fowler. 2017. "An exploratory comparison of name generator content: Data from rural India." *Social Networks* 48:157–168.

Siegel, David A. 2011. "Social Networks in Comparative Perspective." *PS: Political Science and Politics* 44.

Sinclair, Betsy. 2012. *The social citizen: Peer networks and political behavior*. University of Chicago Press.

Sokhey, Anand E and Paul A Djupe. 2014. "Name generation in interpersonal political network data: Results from a series of experiments." *Social Networks* 36:147–161.

Sudman, Seymour. 1988. "Experiments in measuring neighbor and relative social networks." *Social Networks* 10(1):93–108.

Sudman, Seymour, Barbara Bickart, Johnny Blair and Geeta Menon. 1994. The effect of participation level on reports of behavior and attitudes by proxy reporters. In *Autobiographical memory and the validity of retrospective reports*. Springer pp. 251–265.

Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.

White, Kevin and Susan Cotts Watkins. 2000. "Accuracy, stability and reciprocity in informal conversational networks in rural Kenya." *Social Networks* 22(4):337–355.