31

Collecting Social Network Data

Tobias H. Stark

Introduction

For decades, sociologists have been interested in the effects of social networks on people's behavior, attitudes, and economic success (Borgatti et al. 2009; Granovetter 1973). But also scholars in other fields such as medicine (Christakis and Fowler 2007), public health (Cornwell et al. 2014), and social psychology (Wölfer et al. 2015) have acknowledged the importance of social networks for phenomena in their discipline. Researchers in all of these fields use similar methods to assess the characteristics of people's network contacts and to get an understanding of the social structure that surrounds people. While social network data can be collected in many different ways, including archival records, tracking devices, or from mining the Internet, the vast majority of social network researchers still make use of surveys to gather information about the social connections of their subjects.

Comprehensive reviews of survey methods for social network data have appeared recently (Cornwell and Hoaglin 2015; Marsden 2011). Accordingly, the goal of this chapter is not to give an additional overview of these methods. Rather, the focus is on the challenges survey researchers

T.H. Stark (✉)
Utrecht University, Utrecht, The Netherlands
e-mail: t.h.stark@uu.nl

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_31

working with social networks are experiencing. Promising areas for future research are also identified that might help overcome some of these challenges.

In principle, it is possible to distinguish two types of network studies. In a *whole network study*, each member of a predefined social network completes an interview in which he or she indicates with which other persons in the network a relationship exists. A necessary prerequisite for this type of network study is that every member of a predefined social network can be identified in advance and that everyone is reachable for an interview. Accordingly, whole network studies are typically limited to one social context such as a school or a school class. The advantage of whole network studies is that researchers can get a very accurate picture of the social structure of the entire network (i.e., who is connected with whom). This contrasts with *ego-centered network studies*. Here, respondents (called egos) are asked to name their social contacts and these contacts do not need to be part of a predefined social network or belong to the same social context. Instead of interviewing all contacts, the survey respondents are asked to answer proxy questions about their network contacts. With this approach, only the direct network of respondents can be mapped and the larger, surrounding network that also includes the contacts of respondents' contacts remains invisible. The advantage of ego-centered network studies is that they can be included in regular surveys because the focus is not on a predefined social network of which all members need to be interviewed. Many large-scale national representative studies such as the American National Election Study, the General Social Survey (GSS), the Netherlands Life Course Survey, or the German Socio-Economic Panel have from time to time implemented ego-centered networks in their study design.

Whole Network Studies

Whole network studies typically focus on small social settings with clear boundaries to identify all members of the underlying social network. These can, for instance, be business leaders who interact in the same industry, employees within one company, or students in the same schools. All relationships that these people might have with people outside of the particular social setting are outside of the scope of a whole network study. This means every person of the sampling frame needs to be interviewed in a whole network study. As a consequence, whole network studies are typically case studies and cannot be implemented as part of a representative survey.

Survey researchers have to make a number of decisions when designing a whole network study. Typically, respondents are presented with a list of names of all members of the social network and are then asked to identify with whom they have a relationship. Two approaches have been used in the past. Some researchers print ID numbers next to the names of all network members and asked the respondent to write down the ID numbers of their contacts. Other researchers prefer to provide a name list with check boxes and ask the respondent to check the box next to the names of their contacts. No research has evaluated which method might yield more accurate representations of a network even though both methods face potential problems. The ID number method might encourage underreporting of network contacts because considerably more (mental) effort is needed when contacts have first to be found on one list and an associated ID number has then to be copied to another form. Moreover, this method adds a potential source of measurement error because copying ID numbers from one sheet to another may be prone to more error than simply checking a box. The check-box method, in contrast, might invite overreporting of network contacts because it is very simple to mark a large number of names on a list. For instance, in my own research, I have encounter students in school studies who report to be "best friends" with all of their 30 classmates. While possible, this seems highly unlikely. One possible solution to this problem is to only study network connections that are based on mutual nominations of both persons involved in a relationship. However, whether this is possible depends on the type of relationship that is studied (e.g., bullying relationships tend to be one-sided) and on the research questions (e.g., sometimes it is of interest under which circumstance a network nomination is reciprocated).

Researchers have also to decide whether they want to allow respondents to identify up to a certain number of network contacts or that respondents can identify as many contacts as they please. For instance, the National Longitudinal Study of Adolescent to Adult Health (AddHealth) allowed students to only identify their five best male friends and their five best female friends out of all students of their high school. European researchers often focus on school classes instead of entire schools and typically allow students to identify as many contacts among their classmates as they wish (e.g., Stark 2015). Research has found that both methods yield similar results but the unlimited nominations approach seems to be more valid when nominations are made that gauge social status (Gommans and Cillessen 2015).

Additional challenges for survey researchers

- Gaining access to name records in advance: To prepare the name lists, researchers need to have access to the names of all members of the social setting that is being studied. This is problematic if participant consent can only be obtained at the moment of data collection when the name lists have to already be prepared.
- Informed consent: Respondents who complete a social network questionnaire give information about "secondary subjects" (their network contacts). These secondary subjects have to be identifiable by the researcher in order to assess the structure of the social network. This means that data might be collected on people who have refused to participate in the study and have thus not given permission for any data being collected about them. Typically, network researchers have to make a strong argument with Institutional Review Boards (IRB) to justify their method.
- Cognitive burden: In large social settings (e.g., schools in AddHealth), respondents have to go through a long list of names to identify their contacts. Digital questionnaires (e.g., computer-assisted self-interviewing (CASI) or Internet surveys) can help reduce the cognitive burden if they are linked to a database with all names. Respondents start typing the names of their contacts and the computer can suggest matching names.

Ego-Centered Network Studies

Ego-centered network studies focus on the "core personal networks" (Marsden 2011) of respondents instead of the complete social network in a given social setting. In a first step, respondents are asked to identify their network contacts. This is done with "name generator questions." The most well-known name generator is used in the GSS and asks respondents, "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? Just tell me their first names or initials." Other name generator questions ask for names of people whom the respondents feel close to or from whom they could borrow money. The choice of name generator depends on the type of social relationship a researcher is interested in. Some researchers rely on one of these questions to assess the core personal network of their respondents but asking at least two different name generator questions seems to produce more accurate measures of network size (Marin and Hamilton 2007).

Name generators can be compromised if people forget network contacts (recall bias, see Bell et al. 2007; Brewer 2000), if people misinterpret the name generator question (e.g., Bearman and Parigi 2004; Brashears 2011; Small 2013), and due to random error (Almquist 2012). Unfortunately, there is no gold standard to achieve highest validity and reliability of name generator questions. Probes that call respondents' attention to different contexts and to people that may be close to already named contacts seem to reduce forgetting social contacts (Marin 2004). Recent research also suggests that asking respondents to go through their cell phone book to check for names they might be forgetting reduces the recall bias (Hsieh 2015).

To assess the dynamics of ego-centered social networks over time, such as with longitudinal surveys, Cornwell and colleagues (2014) developed a roster matching technique for name generator questions. After respondents have been interviewed for the second time in a longitudinal study, the names of the network contacts reported in the previous wave are matched to the new answers given. Respondents can then be asked to verify the matches and correct mistakes. This technique allows following up with questions about the reasons for changes in the network compositions that would otherwise only be detected in the data analysis phase of a research project.

Because the network contacts of survey respondents are typically not interviewed in an ego-centered network study, respondents have to report the characteristics of their contacts. These proxy reports are done in follow-up questions about each contact that are called "name interpreter questions." These questions either ask about characteristics of each of the contacts (e.g., "Is [NAME 1] a man or a woman?", "Is [NAME 2] a man or a woman?") or about relationships between the contacts (e.g., "Does [NAME 1] know [NAME 2]?"). Thus, information about people in the network and the structure of the social network relies entirely on the perception of the survey respondents.

There has been extensive research on the quality of answers given about network contacts in such follow-up questions. In this volume the chapter by Cobb (2017) focuses entirely on the quality of these proxy reports. In general, answers given by respondents about their contacts do often not correspond with the answers these contacts give themselves when they are also interviewed. This is particularly true for questions about non-observable characteristics, such as network contacts' attitudes. However, the accuracy of proxy reports about network members is often less of a concern in social network analysis than it is when proxy reports are used in a regular survey to replace a hard-to-reach target respondent. The reason for this is that, in

network analysis, researchers are typically interested in how the network influences their respondents. For this purpose, the perception of the network by the respondent might often be more important than the objective characteristics of the network (Cornwell and Hoaglin 2015). Yet, the ultimate test of this assumption is still lacking. A study that compares the impact of perceptions of a person's network on that person's attitudes or behaviors to the impact of objective measures of the network by also interviewing the network connections has, to the best of my knowledge, not been conducted yet.

Additional challenges for survey researchers:

- Cognitive burden: Repeatedly answering the same follow-up questions for each network connection or pair of network contacts may impose a substantial cognitive burden on respondents and reduce the data quality (Hsieh 2015; Matzat and Snijders 2010; Tubaro et al. 2014; Vehovar et al. 2008).
- Size of the network: Answering multiple follow-up questions about each network contact takes up valuable interview time. Moreover, the number of pairs of contacts that have to be evaluated to assess the structure of a person's social network (i.e., who knows whom) increases exponentially with the size of the network (McCarty et al. 2007). Accordingly, most researchers limit their respondents to a maximum of five network contacts and this number allows producing reliable estimates of many network characteristics such as network composition and network density (Marsden 1993).
- Mode effects: Research found that the number of connections between respondents' network contacts (e.g., "Does [NAME 1] know [NAME 2]?") was exaggerated in an online survey compared to a face-to-face survey (Matzat and Snijders 2010). More research is needed comparing different modes.
- Interviewer effects: Interviewers learn that more network contacts increase the length of an interview due to the follow-up question about each contact and have been found to shorten an interview by falsely reporting no or very few network contacts (Eagle and Proeschold-Bell 2015; Paik and Sanchagrin 2013). An interesting approach for future research might be to make use of mixed-mode designs. The names of the network contacts could be collected in CASI mode while an interviewer could collect the follow-up questions about the contacts in computer-assisted personal-interviewing (CAPI) mode.

Data Collection

Despite potential interviewer effects, face-to-face interviews or telephone interviews in which interviewers can motivate respondents to answer repetitive follow-up questions effortfully are still considered the best way to collect ego-centered network data (Marsden 2011). However, independent of the mode, a computer is necessary to handle the complexity of an ego-centered social network questionnaire because the names of the network contacts have to be pasted into the follow-up questions about the contacts (e.g., "Is [NAME 1] a man or a woman?"). Thus computer-assisted telephone-interviewing (CATI), CASI, CAPI, or Internet surveys can be used for ego-centered network questionnaires.

When a research question requires data on whole networks, paper-and-pencil questionnaires can be used in addition to the computer-assisted modes. The reason is that researchers using a whole network design typically only want to know who the network contacts are. Follow-up questions about these contacts are not necessary because these people are also part of the study and complete a questionnaire on their own.

Some design recommendations have been made for ego-centered network studies that use self-administration of the questionnaire (CASI, Internet). Most importantly, asking about one attribute of all network contacts in follow-up questions before asking about another attribute of all contacts leads to less item nonresponse, less drop out (Vehovar et al. 2008), and more reliable data (Coromina and Coenders 2006) than does asking all follow-up questions about one contact before asking all questions about the next contact. Also the number of name boxes displayed under the name generator question should be well considered because respondents tend to match the number of names they give to the number of name boxes they see (Vehovar et al. 2008).

Recently, graphical software tools have been developed that make use of these design recommendations and try to make the process of answering ego-centered network questionnaires less repetitive and thus more enjoyable. These tools make use of visual aids to reduce the cognitive burden for respondents. The survey tool PASN (Lackaff 2012) derives names of respondents' social networks by accessing their Facebook profiles whereas the tool TellUsWho (Ricken et al. 2010) mines respondents' email accounts for names. Subsequently, respondents can answer questions about their contacts by dragging and dropping the names or the Facebook profile pictures of their contact into answer boxes. The software ANAMIA EGOCENTER (Tubaro et al. 2014) lets respondents

draw a picture of their network in great detail, which give valuable information about connections and cliques but may make completing a network survey rather complex.

A new class of graphical data collection software lets respondents answer follow-up questions about their network contacts through interacting with a visual representation of the network. Such approaches have been implemented in the programs OpenEddi (Fagan and Eddens 2015), netCanvas (Hogan et al. 2016), and GENSI (Stark and Krosnick 2017). Questions about the network contacts can be answered by either clicking on the names of the contacts (for dichotomous questions) or by dragging and dropping the names into answer boxes (an example using GENSI is shown in Fig. 31.1). Connections between the network contacts (i.e., who knows whom) can be indicated by the traditional approach of asking separately for each pair of network contacts whether a relationship exists or by drawing lines between the names of two or more contacts in the figure of the network (Fig. 31.2). OpenEddi also allows indicating connections by sorting the names of network contacts into piles.

All of this new software has been developed to reduce respondent burden and increase data quality. However, very little research has been done to examine respondents' perception of these graphical tools and the quality of data collected. An evaluation study found that GENSI produces data of equal quality as a traditional ego-centered questionnaire (Stark and Krosnick 2017). However, respondents enjoyed completing the questionnaire more

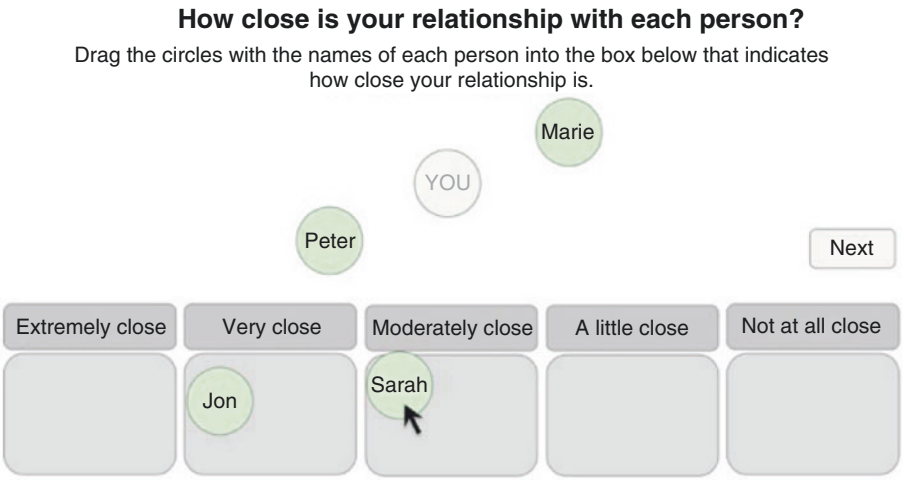

Fig. 31.1 Drag-and-drop question in GENSI. Answers are given by moving name circles into answer boxes

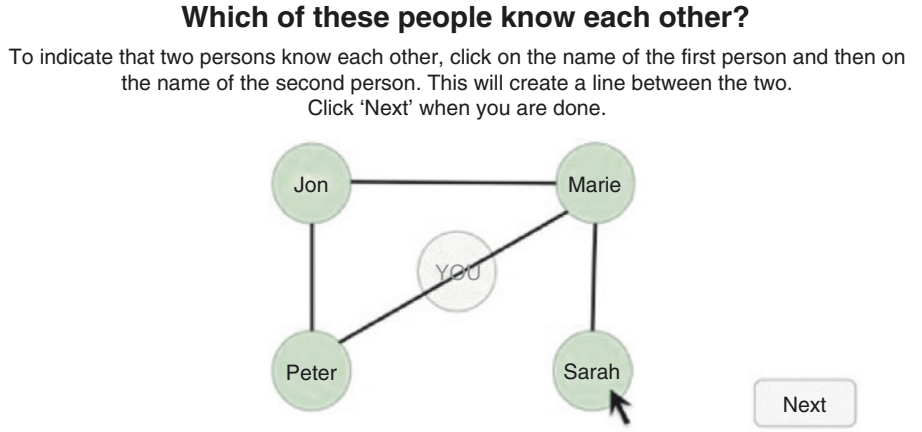

Fig. 31.2 Question for network relationships in GENSI. Relationships are indicated by drawing lines between related network contacts

with GENSI than with the traditional design. The new tool also seems to solve what past researcher have considered to be a problem with online administration; exaggerated numbers of connections between the contacts of a respondent (Matzat and Snijders 2010). Given these promising results, future research that compares the various programs with each other and with traditional ways to collect ego-centered network data seems highly valuable.

Social Media

With the abundant availability of social network data produced by social media, one might wonder why few network researchers make use of social media data but rather use surveys to collect their data. A central challenge for this research is that many social media websites that have a large amount of information on their users (e.g., Facebook, Instagram) do not allow access to these data. For instance, Facebook recently changed its application programming interface (API) to no longer support automatic downloads of user data. It is thus only possible to access and code public user profiles by hand but not in a less time-consuming automatized fashion. In contrast, social media data that are publicly available and can be automatically downloaded typically do not offer much information on a particular user (e.g., Twitter), which makes it difficult to link this data to survey data.

There are a number of additional limitations that make surveys still a preferred mode for collecting social network information. First, social

media networks are limited to relationships between people who are users of the social media website. This means that important parts of a population might be missed when a study relies completely on social media data. Whether this is a problem or not depends, of course, on the research question at hand. A study of, for instance, social influence through social media will not be affected by this limitation whereas a study that is interested in social influence between people in general might overlook influential actors that are not members of the social media network. Another limitation of social media data is that there is typically no way to understand the nature of the relationship between two people on the website. The pure existence of a connection says little about the actual relationship because on most websites everybody is linked in the same way (e.g., friends on Facebook, or contacts on LinkedIn). Researchers could enrich these data by accessing and coding the communication between connected people on the website. However, interpreting the meaning of the communication is typically difficult. Moreover, important communication may take place outside of social media. It is possible that people who interact on a daily basis, and are thus very relevant network contacts, make less use of communication through social media. A final limitation is that the available information on social media websites is restricted to users' behavior on the website whereas other information that is of interest for many researchers (e.g., age, sex, attitudes) is often missing. Recently, researchers have started to overcome this problem partially by combining social media information with survey data (Schober et al. 2016). This seems to be a valuable direction for future research.

Areas for future research:

- The existing best-practice recommendations for network questionnaire designs are based on small samples that are not representative for any population. Tests of these recommendations with data from random probability samples are needed.
- It is still unclear whether people's perceptions of their social network or objective measures of their networks have more impact on people's attitudes and behaviors.
- Tests of mode effects are needed, both for the collection of whole network data and ego-centered network data.
- Evaluation studies of the various existing graphical tools to collect ego-centered network data in comparison to traditional survey tools are needed. Do the graphical tools reduce cognitive burden and produce better measures of social networks?

- A combination of whole networks to gauge relationships within a social setting with ego-centered networks to assess relationships outside of this setting might overcome the weaknesses of both approaches. Such data are rarely collected because different research questions typically motivate the collection of whole or ego-centered networks. One noteworthy exception forms the CILS4EU study for which classroom network data were collected among more than 18,000 students from 958 classrooms in four European countries. Students were also asked to complete ego-centered network questions about up to five friends outside of their classroom.
- Research that links social media information with network survey data might give insights in how relevant these ties are compared to the network that is typically assessed with traditional approaches. The chapters in this volume by Pasek (2017), and Blaermire (2017) respectively give an overview of the opportunities and challenges associated with linking survey data with such external data.

References and Further Reading

Almquist, Z. W. (2012). Random Errors in Egocentric Networks. *Social Networks*, *34*(4), 493–505.

Bearman, P., & Parigi, P. (2004). Cloning Headless Frogs and Other Important Matters: Conversation Topics and Network Structure. *Social Forces*, *83*(2), 535–557.

Bell, D. C., Belli-McQueen, B., & Haider, A. (2007). Partner Naming and Forgetting: Recall of Network Members. *Social Networks*, *29*(2), 279–299.

Blaermire, B. (2017). Linking Survey Data with the Catalist Commercial Database. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research*. New York: Palgrave.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network Analysis in the Social Sciences. *Science*, *323*(5916), 892–895.

Brashears, M. E. (2011). Small Networks and High Isolation? A Reexamination of American Discussion Networks. *Social Networks*, *33*(4), 331–341.

Brewer, D. D. (2000). Forgetting in the Recall-Based Elicitation of Personal and Social Networks. *Social Networks*, *22*(1), 29–43.

Christakis, N. A., & Fowler, J. H. (2007). The Spread of Obesity in a Large Social Network over 32 Years. *The New England Journal of Medicine*, *357*, 370–379.

Cobb, C. (2017). Proxy Reporting. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research*. New York: Palgrave.

Cornwell, B., Schumm, L. P., Laumann, E. O., Kim, J., & Kim, Y. J. (2014). Assessment of Social Network Change in a National Longitudinal Survey. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences*, *69*, S75–S82.

Cornwell, B., & Hoaglin, E. (2015). Survey Methods for Social Network Research. In T. P. Johnson (Ed.), *Health survey methods* (pp. 275–314). Hoboken, NJ: Wiley.

Coromina, L., & Coenders, G. (2006). Reliability and Validity of Egocentered Network Data Collected Via Web – A Meta-Analysis of Multilevel Multitrait, Multimethod Studies. *Social Networks*, *28*(3), 209–231.

Eagle, D. E., & Proeschold-Bell, R. J. (2015). Methodological Considerations in the Use of Name Generators and Interpreters. *Social Networks*, *40*, 75–83.

Fagan, J., & Eddens, K. (2015). *OpenEddi*. Paper presented at the XXXV Sunbelt Conference. Brighton, UK.

Gommans, R., & Cillessen, A. H. N. (2015). Nominating Under Constraints. A Systematic Comparison of Unlimited and Limited Peer Nomination Methodologies in Elementary School. *International Journal of Behavioral Development*, *39*(1), 77–86.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*, 1360–1380.

Hogan, B., Melville, J. R., Ii, G. L. P., Janulis, P., Contractor, N., Mustanski, B. S., & Birkett, M. (2016). *Evaluating the Paper-to-Screen Translation of Participant-Aided Sociograms with High-Risk Participants*. Paper presented at the Human Factors in Computing, San Jose, CA.

Hsieh, Y. P. (2015). Check the Phone Book: Testing Information and Communication Technology (ICT) Recall Aids for Personal Network Surveys. *Social Networks*, *41*, 101–112.

Lackaff, D. (2012). New Opportunities in Personal Network Data Collection. In M. Zacarias & J. V. De Oliveira (Eds.), *Human-Computer Interaction*. Berlin: Springer.

Marin, A. (2004). Are Respondents More Likely to List Alters with Certain Characteristics?: Implications for Name Generator Data. *Social Networks*, *26*(4), 289–307.

Marin, A., & Hamilton, K. (2007). Simplifying the Personal Network Name Generator: Alternatives to Traditional Multiple and Single Name Generators. *Field Methods*, *19*, 163–193.

Marsden, P. V. (1993). The Reliability of Sociocentric Measures of Network Centrality. *Social Networks*, *15*, 399–422.

Marsden, P. V. (2011). Survey Methods for Network Data. In J. Scott & P. J. Carrington (Eds.), *The SAGE Handbook of Social Network Analysis* (pp. 370–388). London: Sage.

Matzat, U., & Snijders, C. (2010). Does the Online Collection of Ego-Centered Network Data Reduce Data Quality? An Experimental Comparison. *Social Networks*, *32*(2), 105–111.

McCarty, C., Killworth, P. D., & Rennell, J. (2007). Impact of Methods for Reducing Respondent Burden on Personal Network Structural Measures. *Social Networks*, *29*, 300–315.

Paik, A., & Sanchagrin, K. (2013). Social Isolation in America: An Artifact. *American Sociological Review*, *78*(3), 339–360.

Pasek, J. (2017). Linking Knowledge Networks Web Panel Data with External Data. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research*. New York: Palgrave.

Ricken, S. T., Schuler, R. P., Grandhi, S. A., & Jones, Q. (2010). TellUsWho: Guided Social Network Data Collection. *Proceedings of the 43rd Hawaii International Conference on System Sciences*.

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social Media Analyses for Social Measurement. *Public Opinion Quarterly*, *80*(1), 180–211.

Small, M. L. (2013). Weak Ties and the Core Discussion Network: Why People Regularly Discuss Important Matters with Unimportant Alters. *Social Networks*, *35*(3), 470–483.

Stark, T. H. (2015). Understanding the Selection Bias: Social Network Processes and the Effect of Prejudice on the Avoidance of Outgroup Friends. *Social Psychology Quarterly*, *78*(2), 127–150.

Stark, T. H., & Krosnick, J. A. (2017). GENSI: A New Graphical Tool to Collect Ego-Centered Network Data. *Social Networks*, *48*, 36–45.

Tubaro, P., Casilli, A. A., & Mounier, L. (2014). Eliciting Personal Network Data in Web Surveys Through Participant-Generated Sociograms. *Field Methods*, *26*(2), 107–125.

Vehovar, V., Manfreda, K. L., Koren, G., & Hlebec, V. (2008). Measuring Ego-Centered Social Networks on the Web: Questionnaire Design Issues. *Social Networks*, *30*(3), 213–222.

Wölfer, R., Faber, N. S., & Hewstone, M. (2015). Social Network Analysis in the Science of Groups: Cross-Sectional and Longitudinal Applications for Studying Intra- and Intergroup Behavior. *Group Dynamics-Theory Research and Practice*, *19*(1), 45–61.

Tobias H. Stark is an Assistant Professor at the European Research Centre on Migration and Ethnic Relations (ERCOMER), Utrecht University, The Netherlands. He studies how social networks affect the development and spread of prejudice, as well as how prejudice hinder the development of interethnic relationships. He uses insights of this research to develop anti-prejudice interventions in schools. His dissertation "Integration in Schools. A Process Perspective on Students' Interethnic Attitudes and Interpersonal Relationships" was awarded with a Research Prize of the Erasmus Prize Foundation and won the prize for best dissertation of the Dutch Sociological Association (NSV). Dr. Stark's research has been funded by the

European Commission (Marie-Curie International Outgoing Fellowship) and the Netherlands Organisation for Scientific Research (Veni grant). His work has appeared in sociological and social-psychological high-impact journals such as *Social Networks, Social Psychology Quarterly, Public Opinion Quarterly*, and *Personality and Social Psychology Bulletin*.