Lessons Learned from Generating, Consolidating, and Analyzing Large Scale, Longitudinal Social Network Data

Dr. Jack Elliott, Iron Range Engineering, Minnesota State University, Mankato

Jack Elliott is an assistant professor in Integrated Engineering at the Iron Range Engineering Program, a part of Minnesota State University Mankato. His research areas include student social support networks in engineering education, experimental fluid dynamics, and developing low-cost technology-based tools for improving fluid dynamics education.

Dr. Angela Minichiello, Utah State University

Angela (Angie) Minichiello is a military veteran, licensed mechanical engineer, and associate professor in the Department of Engineering Education at Utah State University. Her research examines issues of access, equity, and identity in the formation of engineers and a diverse, transdisciplinary 21st century engineering workforce. Angie received an NSF CAREER award in 2021 for her work with student veterans and service members in engineering.

Generating, Consolidating, and Analyzing Social Network Data: Lessons Learned from a Large-Scale, Longitudinal, Network Study

Introduction

This methods paper provides recommendations for engineering education researchers considering Social Network Analysis (SNA) to answer their Research Questions (RQs) in current or future studies, particularly for studies of large-scale networks. Over the last several decades, engineering educators have increasingly recognized the role interpersonal interactions play in shaping engineering student outcomes. These interactions span various modes, including online, face-to-face (f2f), student-to-student, student-to-instructor, and instructor-to-instructor interactions [1], [2], [3]. Concurrent with this growth, the development of engineering education as a discipline included an increasing number of engineering education researchers adopting sociological research methods [4], [5]. Among these sociological research methods, SNA applies network theoretic concepts to interpersonal networks [6], allowing researchers to explore how interpersonal connections form, evolve, and relate to outcomes of interest. Recognizing the intersection of interpersonal network importance, and the ability of SNA to study these networks, engineering education researchers have increasingly adopted SNA to identify and promote positive interpersonal networks in engineering education.

Throughout the relevant literature, engineering education research applying SNA to the undergraduate student context has demonstrated the importance of interpersonal connections for students' academic performance and affective outcomes [7]-[10]. These studies not only quantitatively assess the importance of connections between individuals, but also inform which interaction types and frequencies lead to positive or negative outcomes. For example, Ellis, Han, and Pardo [11] found that students who engaged in 'effective' collaboration worked closely, in infrequent amounts, with small groups of peers. Ineffective collaboration included large groups with frequent interactions. In another study, Elliott and colleagues [20] identified that student-to-student interactions increased in effectiveness as small groups became more close-knit over time. These and similar studies have demonstrated that cross-sectional networks applying simple SNA may overlook certain relationships between student interactions and outcomes.

Beyond these issues, our review of the relevant engineering education literature demonstrates a prevalence of studies regarding interactions in the online context. These studies have provided important observations of how increased interactions relate to performance for remote and/or hybrid instruction overall [12], [13], [14]. However, we believe that this emphasis on online interaction over f2f interaction may not reflect the scale of research need, but the ease of data collection for SNA regarding online interactions. Specifically, f2f interactions are a less studied, but major component of students' interactions.

To overcome these issues, our research group, familiar with SNA from small studies, conducted a large-scale (1000+ individuals) SNA study at a large, public university in the United States [15]. This study sought to extend the current understanding of student networks to a more holistic level by a) sampling student networks several times throughout individual semesters, b) sampling student networks for the first two years of students' undergraduate careers, c) asking participants to identify peers they studied and/or socialized with inside the academic context, and

d) asking participants to identify peers they studied and/or socialized with outside the academic context. Key results demonstrate how student networks extend beyond the bounds of single classrooms in enrollment and time, how students form and evaluate their peer relationships, and how interactions within and/or outside the academic context relate to positive and/or negative student outcomes.

Apart from the fundamental results of this work, we also identified several important considerations in the design, implementation, and evaluation of SNA studies, particularly in the large, f2f context. Primary issues include generating interaction data from a population of interest, consolidating interaction data to usable forms, and then analyzing the results in a manner that accurately represents the underlying network. These issues are further compounded by implications of the study goals, including how bounded the network should be, whether interactions are online or f2f, and the temporal resolution of the sampled network.

The purpose of this paper is to disseminate recommendations from our experience including implications of the desired study network for collecting and consolidating interaction data, how the bounds of the network inform open-ended vs. close-ended responses, and the benefits/drawbacks according to each type of survey selection. Beyond collecting network data, consolidating networks into usable forms presents a major hurdle for large-scale studies. This paper presents practices and freely available resources for consolidating networks according to data collection types. Finally, the analysis of social networks is an ever-growing field and includes sophisticated methods that require a threshold of interaction data resolution and confidence. For this reason, we present several accessible resources for conducting fundamental and advanced statistical SNA methods.

Social Network Analysis

SNA provides researchers with a method for quantitatively and visually describing and analyzing the interactions between individuals. Within SNA literature, the interactions between individuals are referred to as *ties* or *edges* and the individuals in the network are referred to as *actors*. When researchers consider a single individual in the network as the focal point, that actor is referred to as the *ego*. The individuals who interact with the ego are referred to as the ego's *alters*. The combination of actors and the ties between actors comprises a *network*.

To conduct SNA there are several primary steps (adapted for engineering education research from Borgatti and colleagues [16]). Steps one, two, and six are necessary for conducting SNA which only considers the networks (e.g., network changes over time, connectedness of students, etc.). Steps three, four, and five are also necessary for researchers hoping to compare network traits and actor traits (e.g., student connectedness vs. grades, the level of demographic mixing in student networks, etc.). Each of these steps are as follows:

1. <u>Identifying the Study Network.</u> Before conducting SNA, researchers must consider the *who, what,* and *when* aspects of the network that they want to study. This requires well-describing the actors (e.g., students enrolled in a specific engineering course, engineering club members, etc.), the interaction types (e.g., extra-curricular interactions, formal group interactions, etc.), and the time (e.g., a single course meeting, a full year, etc.) that

comprise the study network. The quality of these descriptions will determine the quality of the subsequent data collection and analysis.

2. Generating Network Data. After identifying the study network, data generation includes identifying and validating data collection methods, choosing a sampling frequency, and consolidating network data to an acceptable level of confidence.

3. Identifying Actor Traits. Parallel to identifying networks, identifying the actor traits such as demographics (e.g., age, race, ethnicity, etc.), academic outcomes (e.g., course grades, retention, etc.), and affective outcomes (e.g., motivation, feelings of belonging, etc.) is typically a key component of SNA studies in engineering education. Particularly in the engineering education context, identifying academic and/or affective outcome traits allows researchers to identify relationships between interactions and outcomes.

4. Generating Actor Trait Data. After identifying the actors and desired actor traits, researchers must identify and validate data collection methods, choose a sampling frequency, and consolidate alter trait data to an acceptable level of confidence.

5. Integrating the Network and Alter Trait Data. For studies that gather both network data and alter trait data, consolidating these into a single usable format is necessary before analysis.

6. Analyzing the Consolidated Data. After data is generated and consolidated into a usable manner, actual SNA begins. These steps include visual analysis of sociograms (network graphs of interpersonal networks), quantitative analysis of SNA measures (numeric descriptors of the network and/or alter), and statistical tests of SNA measures to alter traits.

Axiologically, the first author's pragmatic research paradigm suggests that effective research begins with a question, and then identifies methods to answer that question. To make this process easier for those considering SNA while aligning with this paradigm, this paper will focus on providing recommendations for researchers who begin with a research question and are evaluating SNA as a method for answering that research question. Further, our experience is particularly in conducting SNA in the f2f undergraduate engineering context and includes large, longitudinal networks. To make this experience available to the broader engineering education community, this paper generally presents considerations for researchers considering SNA but will trend toward considerations in large-scale, open-ended networks as these have proven the most difficult.

Network Types

SNA includes three broad types of networks: ego-, whole, and sub-networks [6], [16]. *Ego networks* isolate the interactions that are in and/or out of single individuals within the study network. An example of ego-network analysis includes researchers exploring the relationship between how many study partners each student in a class has and those students' grades [1], [17], [18]. *Whole networks* include all the individuals and interactions within the study network. An

example of whole network analysis includes considering the longitudinal evolution of connection density between all students in a specific cohort [19]. Between whole and ego-networks, *sub-networks* focus on a specified collection of individuals and interactions within the whole network. An example of sub-network analysis includes researchers studying how a single study group changes the number of reciprocal connections throughout a single semester [20]. Whole networks are typically built from a collection of ego-networks but require that a significant number of participant egos provide network data to be an accurate representation of the whole network. As a result, resource requirements are generally highest for generating whole networks and lowest for generating ego-networks. Further, whole network data often allow for subsequent sub- and ego-network analysis.

For whole and sub-network analysis, a minimum threshold of network completion is necessary to ensure useful results. Specifically, inaccurate whole and sub-network data may lead to differences in network traits being a result of errors in the data, rather than changes in the network. For example, a simple measure of *reciprocity* is the proportion of reciprocal connections (i.e., both individuals have reported the link) relative to the number of observed connections. If a larger number of participants within the network are sampled, the reciprocity may increase due to the number of sampled participants, rather than an actual increase in the social network's reciprocity. Further, several ego-network analysis methods rely on an estimate of the whole network. For example, ego-network measures such as *eigenvector centrality* (the connectedness of an individual considering how connected the surrounding network is) and in-degree centrality (the sum of incoming connections for an individual) rely on an estimate of the surrounding network, despite being ego-network traits.

For these reasons, researchers should carefully consider if their network is well-bounded enough for whole or sub-network analysis, and if they anticipate gathering sufficient network data to complete the network before analysis when determining which network type to study. While whole networks are desirable due to their broad usefulness, the resources required to gather an accurate whole network representation are high, scaling on the size of the study network. Researchers should carefully consider if their RQ requires whole, sub-, or ego-network data and select the minimum acceptable level.

Beyond network bounds in actors, networks may be either *static* (do not change with time) or *dynamic* (change with time). Like whole and ego-networks, researchers can build dynamic networks from a collection of static networks. Further, researchers can convert dynamic networks into static networks by aggregating interactions over a data collection period. Network data generation strategies also depend on and/or inform static vs. dynamic networks as explained in the network sampling section. Overall, building dynamic networks from multiple cross-sectional networks is most likely to require the greatest effort on the part of the researcher in both data collection and subsequent analysis.

Network Data Generation

Deciding on methods for network data generation involves a careful balance of the RQs to be answered and the resources available to the researcher. Overall, the network type and characteristics of interest according to RQs should be the primary determinants of the method.

However, researchers should also be aware of the resource costs to anticipate for each data collection method. Resource requirements for data collection scale on a) the size of the network, b) how automated the methods are (i.e., recorded through location, recorded through interaction medium, or not recorded automatically), and c) the characteristics of the interactions (e.g., perceived strength of ties, classifying friends or study partners, etc.). These factors are interrelated, and each case is likely to be unique. Recognizing these issues, we present a general strategy as a beginning guide for researchers interested in deploying SNA methods.

Network Data Collection

Several methods for gathering social network data are common in the relevant literature. General categories include automated data collection (i.e., recorded through the interaction medium), *name generator surveys* (asking participant egos to identify their alters), and interviews/protocols (i.e., asking participant egos about their interactions individually or observing their interactions) [21], [22], [23]. Within name-generator surveys, there are close- and open-ended name-generators. These categories are decided by whether (close-ended,) or not (open-ended,) the participant is provided a set list of names to choose from in the survey instrument. To elaborate on these points, we use this section to present each data collection method individually and summarize the key benefits and drawbacks of each.

Pull Data from LMS. Most LMS provide an Application Programming Interface (API), which helps users familiar with coding pull interaction data from the chatrooms, comments on assignments, etc. Further, strategies for pulling data without the use of an API are available for those who are adept at relevant coding. These strategies provide a wealth of information with a high temporal resolution and low uncertainty (i.e., interactions in online chat are known). However, these strategies require that the RQs may be answered when interactions are limited specifically to online interactions within the LMS platform. For instance, researchers studying connectedness between the instructor and students vs. course outcomes may miss email communication without additional methods. Overall, pulling network data from the course LMS provides high accuracy and resolution information at a fixed cost regardless of network size. These benefits come at the expense of strict actor and interaction-type bounds.

Pull Data from Social Media. Similar to LMS data, most social media platforms provide an API for pulling relevant data, and methods are also available for scraping this data without the API for adept coders. Unlike LMS data, social media can provide researchers with strategies for identifying friendship and similar networks. This difference in the bounds for alter types is a key benefit to social media and comes with a similar level of low uncertainty and high temporal resolution. However, similar limitations as in LMS data exist in social media data. These limitations include that the interaction types are not well documented, relationships between alters are limited to those the social media platform targets, and interactions are only recorded through platform-based interactions.

Close-Ended Name Generator Survey. Close-ended name-generator surveys ask potential participants to select whom they interact with for a specific purpose from an *a priori* list. Methods for providing this list include the use of a drop-down menu, a search function, or a written table. This network data collection method allows researchers to select the interaction

types (e.g., friends, study partners, etc.) and potential alters (e.g., classmates, club members, etc.) without accepting the bounds of LMS or social media data. In this regard, close-ended name generators provide a significant improvement in freedom for RQs which can be answered, maintain confidence in the alters identified, and reduce *recall error* (errors in participants forgetting who they interact with). However, this method requires that the researchers know and provide all potential alters in the survey instrument. Further, the temporal resolution of the network and network completion are limited by the survey sampling frequency and response rates.

Open-Ended Name Generator Survey. Open-ended name-generator surveys ask potential participants to identify alters whom they interact with for a specific purpose without providing a list of potential participants. This method removes bounds to potential alters and allows researchers to answer broader RQs with fewer prior assumptions than LMS, social media, and close-ended methods. This benefit comes at the expense of introducing increased uncertainty and resource costs. Specifically, the likelihood for recall error increases in open-ended name generator surveys, and at a scale larger than ~100 participants, the need for Entity Resolution (ER) methods increases. Like close-ended surveys, the temporal resolution of the network and network completion in open-ended data are limited by the survey sampling frequency and response rates. Generally, open-ended name-generator surveys allow researchers to generate exploratory network data at the expense of network completion and accuracy.

Interview/Protocol. Interview and/or protocol data collection methods provide researchers with a network data collection method that is not limited to interaction types, alters, or recall errors. These methods can also provide deeper insights regarding interaction and/or alter quality. For example, interviewers may ask about the strength of certain ties, in each context of interest, which would be limited by survey fatigue in name-generator surveys and is not possible in automated data collection methods. The cost of these benefits is the resource requirement for each interview/protocol. Generally, interviews are not scalable to the extent of automated data or name-generator data due to their individual time requirements.

Summary of Network Data Collection Methods

To summarize the primary trade-offs between each network data collection method, Table 1 presents a brief description of significant tradeoffs according to network data collection method.

Table 1. Summary of the study implications according to network data collection type.

Data Collection Method	Temporal Resolution	Scalabi-lity	Need for Entity Resolution	Interaction Uncertainty	Alter Bounding	Interaction Type Bounding
LMS	High	High	Med.*	Low	High	High
Social Media	High	High	Med.*	Low	Low	High
Close-Ended Name Generator	Low	Med.	Low	Med.	Med.	Med.

Open-Ended Name-Generator	Low	Med.	High	Med.	Low	Med.
Interview/ Protocol	Low	Low	Low	Low	Low	Low

*Entity resolution, discussed in "Data Consolidation," may be necessary for pairing alter trait data to network data.

As illustrated in Table 1, no single data collection method provides the optimal value for all study implications. In our experience, the resources required to collect and consolidate the data are a key limiter for large studies. To aid researchers in the decision process for whom resources are also a consideration, Fig. 1 provides a decision tree for minimizing resource cost while answering the RQs. Relative resource cost associated with each data collection method is indicated by color, where dark green indicates a low resource cost, and red indicates a high resource cost.

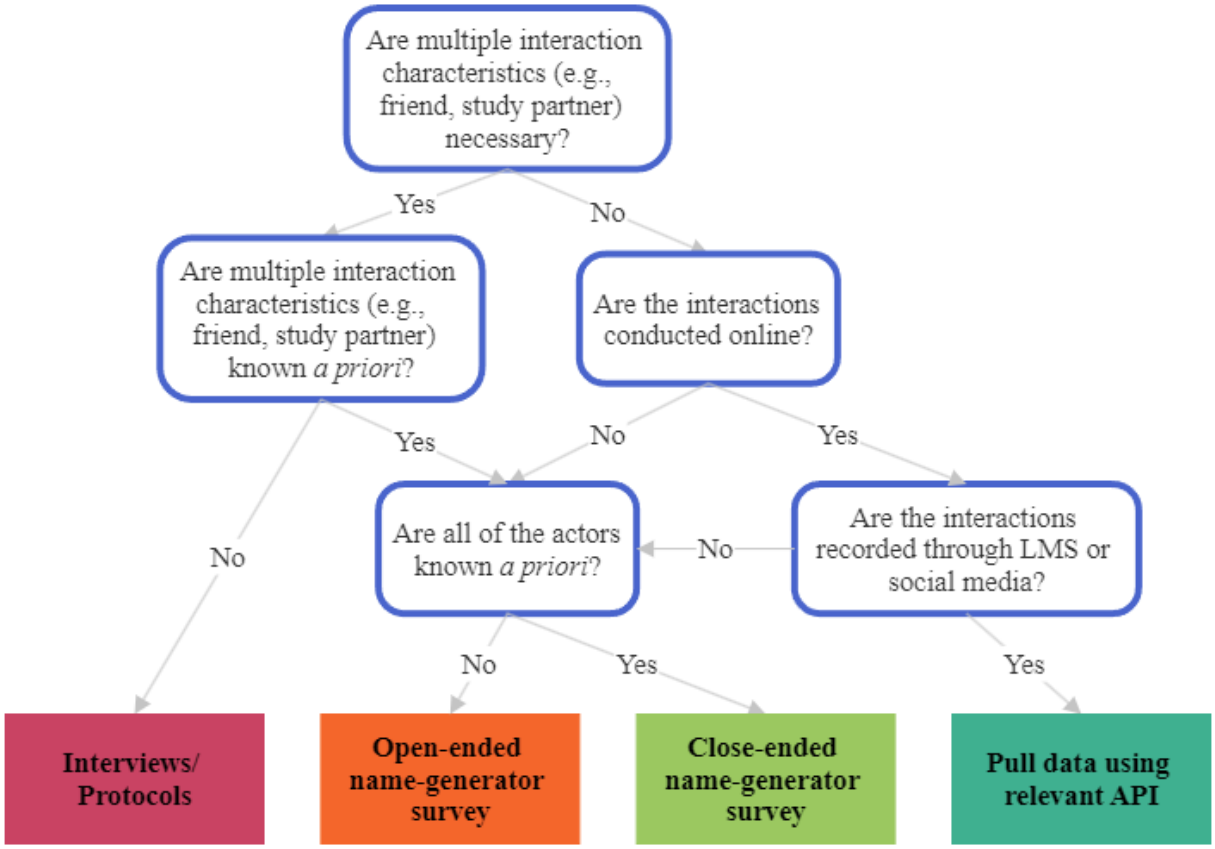

Figure 1. Decision tree for selecting network data collection methods according to desired network characteristics and minimizing resource cost. Relative resource cost associated with each data collection method is indicated by color, where dark green indicates a low resource cost, and red indicates a high resource cost.

Each data collection method in Figure 1 has specific trade-offs that balance the researchers' ability to measure interaction characteristics and extend network bounds against the resource cost. For example, a significant amount of SNA research has been conducted on Learning Management System (LMS) interactions vs. performance in courses. However, these studies are limited to the online context in both interaction types and participants. Similarly, interviews provide high-confidence data without limits on the interaction type, but at a very high resource cost. Researchers who cannot accept these limits to answer their RQs without the necessary resources may consider a hybrid approach (e.g., pulling automated course data and using interviews with several participants) for multi-methods or mixed-methods studies [24]-[27]. This option is a strong technique for capturing the positive aspects of multiple data collection types and overcoming the limitations of a single type at a balanced resource cost.

Network Data Sampling

Network data sampling frequency is first determined by the data collection method. Automated data collection methods such as pulling interaction platform data (i.e. Twitter, Canvas, Discord) or using location data (i.e., the Copenhagen network study [22]) have an effectively unlimited temporal resolution. However, for those studies that do not involve automated recordings of interactions (i.e., interviews and name-generator surveys), survey fatigue becomes a primary issue [28], [29]. In our experience, survey participation rates on name generator surveys follow a decreasing participation rate according to time in the study and the survey frequency. Figure 2 demonstrates the survey participation rates according to the number of survey iterations per semester in the pilot semester of the same study. Similarly, Figure 2 demonstrates the survey participation rates on a large, open-ended name generator survey-based study according to time in the semester. Note that the increase in survey participation at the end of the semester is likely due to extra credit incentives.

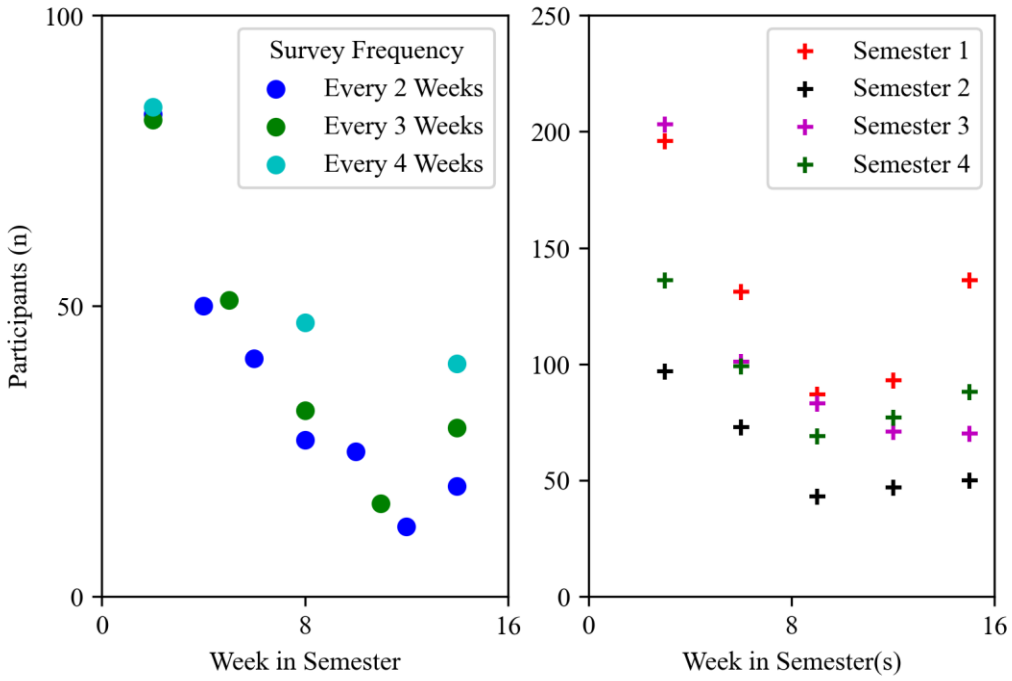

Figure 2. Name generator survey participation rates vs. time for changing survey frequency (left, data from [30]) and constant one survey per three weeks frequency (right).

Together, these results demonstrate that researchers deploying name-generator surveys or interviews should be aware of and sample at the minimum acceptable temporal resolution of the final network. In our experience and review of similar studies, within-semester changes in a network would be captured at a survey frequency of three times per semester without significant losses in network evolutions [19], [20].

Network Data Consolidation

After researchers collect the raw network data, the interactions must be consolidated for analysis. For LMS data, social media data, and close-ended name generators, this process includes the straightforward process of matching the recorded *identities* (accurate names) of egos and alters to their traits of interest (i.e., academic outcomes, demographics). However, a key issue in open-ended name generator surveys is the lack of recorded identities mixed with the number of possible alters. For example, if a “Jane Smith”, is identified in the survey data, who that *reference* represents in the real world is not a trivial question. For studies with a limited scope (i.e., less than 100 potential alters), this issue is solvable through careful consolidation by hand. However, when the study exceeds this scope, ER methods become necessary [31].

Entity resolution describes the process of matching ambiguous references in raw interaction data to the real-world identity they are meant to represent. Several libraries are available online for conducting ER on ambiguous references [32], [33], including a GitHub repository published for novice coders conducting SNA [31]. These libraries require the user to provide records of the

interaction data, any known identities available, and thresholds according to the ER algorithms in the library for consolidating reference-identity similarities. The documentation in these libraries is generally helpful for identifying thresholds and methods for unfamiliar users.

Overall, ER makes large, open-ended network studies possible, while introducing the potential for added error through incorrect consolidation. For this reason, we suggest researchers use conservative estimates of the ER thresholds, validated by hand on a sub-sample of the data. Further, conducting statistical analysis on only participant egos ensures that only actors with a high confidence of correct identity-reference consolidations are used for hypothesis testing. After the networks are consolidated with an acceptable level of confidence, researchers must prepare the network data for analysis by SNA software.

Standards importing formats for SNA software include *edge lists* (a list of interactions from an ego to an alter) or *adjacency matrices* (a matrix where each interaction is recorded as a value in the matrix from an ego [row index] to the alter [column index]). Social networks, which are typically sparse, may be stored as edge lists to save memory. Further, these interaction data are typically mixed with *attribute lists* (a list of alters and their relevant academic outcomes, demographics, etc.).

Network Data Analysis

Once researchers have prepared all necessary data representations, there are several options for conducting visual and statistical analysis on whole, sub-, and ego-networks. Useful tools for visualizing networks are the freely available Gephi [34] and SocnetV [35]. In our experience, SocnetV is simpler to use, and Gephi provides more features such as analyzing longitudinal networks. We recommend either resource for visually analyzing whole networks, visually analyzing sub-networks, and conducting basic statistical analysis. An example output from Gephi is in Fig. 3, demonstrating the results of a clustering algorithm available within Gephi applied to a large undergraduate engineering student network.

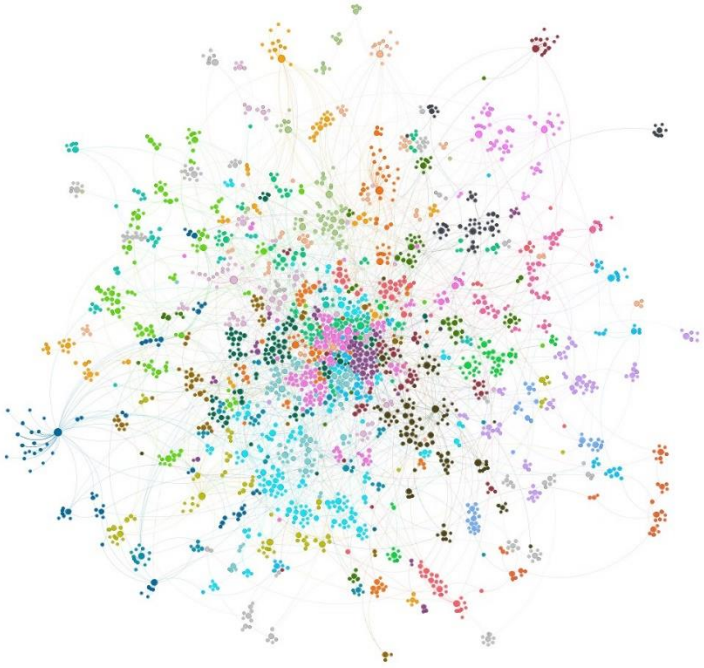

Figure 3. A Gephi-generated sociogram of a first- and second-year undergraduate engineering student network clustered through modularity clustering.

For thorough quantitative analysis, researchers should include a group member who has some familiarity with coding. The Python-based NetworkX library [36] and R-based Statnet library [37] are excellent resources with a wealth of methods and online resources for conducting statistical analysis of network data beyond those available in SocnetV and Gephi. Within each of these resources are strategies for analyzing whole networks, sub-networks, and ego-networks. Each of these libraries also provides methods for storing and analyzing networks with multiple types of connection between nodes called *multigraphs*. For novice coders and/or new researchers to SNA, each of these packages are likely to provide sufficient methods for analyzing cross-sectional networks through simple statistical methods.

However, a growing body of engineering education research applying SNA has recognized Exponential Random Graph Modelling (ERGM) and Stochastic Actor-Oriented Modelling (SAOM) for cross-sectional and longitudinal data respectively. These methods, to our knowledge, are not well developed and documented in openly available Python-based libraries. Further, Python-based libraries are not well prepared for longitudinal network analysis. Opposite this, R-based libraries such as Statnet have several readily available packages for running ERGM and analyzing longitudinal data. Researchers considering more advanced statistical analysis without manually coding the methods should consider doing so in R.

A final consideration for network analysis is applying clustering methods. For researchers who have identified a whole network or large sub-network, smaller sub-networks are readily

identifiable through clustering methods such as k-means for a prescribed number of clusters, or modularity clustering for an unknown number of clusters [38], [39]. A useful example of clustering in engineering education is identifying the size, density, and reciprocity of small friendship or study groups within a larger course network [20], [40], [41]. Clustering provides an efficient method for identifying sub-networks and is available in the recommended SNA tools.

Conclusion

This paper is meant to provide engineering education researchers with a brief overview of and recommendations for conducting SNA in the engineering education context. These recommendations were developed from our review of relevant literature and personal experience generating, consolidating, and analyzing large-scale, longitudinal social network data. This experience has guided our discussion to focus on the challenges of SNA for studying less understood, resource-intensive, large-scale networks (i.e., 1000+ actors, f2f, and longitudinal). Our primary difficulties in conducting this and similar studies were identifying the optimal network data collection method, identifying an optimal sampling frequency, developing and deploying entity resolution, and identifying freely available network analysis tools that met our needs.

Recognizing these issues, key recommendations for researchers undergoing the SNA process outlined in this paper are to a) capture the smallest network that will accomplish the goals of the study, understanding the implication of whole, sub- and ego-network data for subsequent analysis; b) use the network data collection method which maximizes network data accuracy and minimizes resources used according to the decision tree (Figure 1) and Table 1; c) sample at the lowest frequency possible to maximize network completion per sample, recognizing that large network, within-semester changes may be observed at a frequency of three iterations per semester, d) apply ER methods for efficient consolidation of open-ended network data, and e) consider SocnetV and Gephi for network visualization, and NetworkX and Statnet for statistical analysis. We hope that this and similar discussions will provide engineering education researchers considering using SNA to answer their research question a starting point and will demonstrate methods for making large SNA achievable at a reasonable resource cost.

Acknowledgments

This material is based upon work supported by the first author's National Science Foundation Graduate Research Fellowship Program under Grant DGE1745048. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the sponsors.

References

[1] G. Putnik, E. Costa, C. Alves, H. Castro, L. Varela, and V. Shah, "Analyzing the correlation between social network analysis measures and performance of students in social network-based engineering education," *Int J Technol Des Educ*, vol. 26, no. 3, pp. 413–437, 2016, doi: 10.1007/s10798-015-9318-z.

[2] C. N. Gunawardena, N. V Flor, D. Gomez, and D. Sanchez, “Analyzing social construction of knowledge online by employing interaction analysis, learning analytics, and social network analysis,” *Q Rev Distance Educ*, no. 3, p. 35, 2016.

[3] S. Hall, C. T. Amelink, and S. S. Conn, “A case study of a thermodynamics course: Informing online course design,” *Journal of Online Engineering Education*, vol. 1, no. 2, 2010.

[4] B. K. Jesiek, L. K. Newswander, and M. Borrego, “Engineering education research: Discipline, community, or field?,” *Journal of Engineering Education*, vol. 98, no. 1, pp. 39–52, Jan. 2009, doi: 10.1002/j.2168-9830.2009.tb01004.x.

[5] J. E. Froyd, P. C. Wankat, and K. A. Smith, “Five major shifts in 100 years of engineering education,” *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1344–1360, 2012.

[6] Alber-Laszlo Barabasi, *Network Science*, 1st ed. Cambridge University Press, 2016.

[7] J. B. Buckley, B. S. Robinson, T. R. Tretter, C. Biesecker, A. N. Hammond, and A. K. Thompson, “Belonging as a gateway for learning: First-year engineering students’ characterizations of factors that promote and detract from sense of belonging in a pandemic,” *Journal of Engineering Education*, vol. 112, no. 3, pp. 816–839, Jul. 2023, doi: 10.1002/jee.20529.

[8] J. Schnittka and C. Schnittka, “‘Can I drop it this time?” Gender and collaborative group dynamics in an engineering design-based afterschool program,” *Journal of Pre-College Engineering Education Research (J-PEER)*, vol. 6, no. 2, Dec. 2016, doi: 10.7771/2157-9288.1120.

[9] L. Xu *et al.*, “Understanding the role of peer pressure on engineering students’ learning behavior: A TPB perspective,” *Front Public Health*, vol. 10, Jan. 2023, doi: 10.3389/fpubh.2022.1069384.

[10] N. Pearson, J. Major, A. Godwin, and A. Kirn, “Using social network analysis to study the social structures of inclusion,” in *ASEE annual conference & exposition*, 2018.

[11] R. Ellis, F. Han, and A. Pardo, “When does collaboration lead to deeper learning? Renewed definitions of collaboration for engineering students,” *IEEE Transactions on Learning Technologies*, vol. 12, no. 1, pp. 123–132, 2019, [Online].

[12] D. Lee, R. Rothstein, A. Dunford, E. Berger, J. F. Rhoads, and J. DeBoer, “‘Connecting online’: The structure and content of students’ asynchronous online networks in a blended engineering class,” *Comput Educ*, vol. 163, p. 104082, 2021, doi: https://doi.org/10.1016/j.compedu.2020.104082.

[13] M. Li and Z. Liu, “The role of online social networks in students’ E-learning experiences.” p. 1, 2009. doi: 10.1109/CISE.2009.5364232.

[14] T. C. Russo and J. Koesten, “Prestige, centrality, and learning: A Social network analysis of an online class,” *Commun Educ*, vol. 54, no. 3, pp. 254–261, 2005, doi: 10.1080/03634520500356394.

[15] Elliott, J., "Understanding Peer Interactions in Undergraduate Engineering Education," Ph.D. Dissertation, Dept. Eng. Ed., Utah State University, Logan, UT., 2024.

[16] S. P. Borgatti, M. G. Everett, and J. C. Johnson, *Analyzing Social Networks*. SAGE Publications Ltd, 2013.

[17] M. de Laat, V. Lally, L. Lipponen, and R.-J. Simons, “Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for social network analysis,” *Int J Comput Support Collab Learn*, vol. 2, no. 1, pp. 87–103, 2007, [Online].

[18] C. Stadtfeld, A. Vörös, T. Elmer, Z. Boda, and I. J. Raabe, “Integration in emerging social networks explains academic failure and success,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 3, p. 792, 2019, doi: 10.1073/pnas.1811388115.

[19] Lin, “Evolution of civil engineering students’ friendship and learning networks,” *Journal of Professional Issues in Engineering Education and Practice*, vol. 144, no. 4, 2018, doi: 10.1061/(ASCE)EI.1943-5541.0000390.

[20] J. Elliott, A. Minichiello, and J. Ellsworth, “Examining relationships between student interactions with peers and resources and performance in a large engineering course using social network analysis,” *ASEE Annual Conference and Exposition*. Virtual, 2020.

[21] C. Bidart and J. Charbonneau, “How to generate personal networks: Issues and tools for a sociological perspective,” *Field methods*, vol. 23, no. 3, p. 21, 2011.

[22] P. Sapiezynski, A. Stopczynski, D. D. Lassen, and S. Lehmann, “Interaction data from the Copenhagen Networks Study,” *Sci Data*, vol. 6, no. 1, p. 315, 2019, doi: 10.1038/s41597-019-0325-x.

[23] J. P. Martin, K. Gipson, and M. K. Miller, “Developing a survey instrument to characterize social capital resources impacting undergraduates’ decisions to enter and persist in engineering,” in *2011 Frontiers in Education Conference (FIE)*, IEEE, 2011, pp. F2H-1.

[24] J. W. Creswell and V. L. P. Clark, *Designing and conducting mixed methods research*. Sage publications, 2017.

[25] J. Schoonenboom and R. B. Johnson, “How to construct a mixed methods research design.,” *Kolner Z Soz Sozpsychol*, vol. 69, no. Suppl 2, pp. 107–131, 2017, doi: 10.1007/s11577-017-0454-1.

[26] P. Shannon-Baker and C. Edwards, “The affordances and challenges to incorporating visual methods in mixed methods research,” *American Behavioral Scientist*, vol. 62, no. 7, pp. 935–955, Jun. 2018, doi: 10.1177/0002764218772671.

[27] E. G. Creamer and C. D. Edwards, “Editorial: Introduction to the special issue, innovative approaches to visual methods in mixed method research in psychological fields,” *Methods in Psychology*, vol. 6, p. 100090, 203, doi: https://doi.org/10.1016/j.metip.2022.100090.

[28] S. R. Porter, M. E. Whitcomb, and W. H. Weitzer, “Multiple surveys of students and survey fatigue,” *New Directions for Institutional Research*, vol. 2004, no. 121, pp. 63–73, Jan. 2004, doi: https://doi.org/10.1002/ir.101.

[29] B. Fass-Holmes, “Survey Fatigue— What is its role in undergraduates’ survey participation and response rates?,” *Journal of Interdisciplinary Studies in Education*, vol. 11, no. 1, pp. 56–73, 2022.

[30] J. Elliott and A. Minichiello, “Work in progress: An investigation of the influences of peer networks on engineering undergraduate performance outcomes,” *ASEE Annual Conference and Exposition*, Virtual, 2021.

[31] A. Weaver and J. Elliott, “Uncovering students’ social networks: Entity resolution methods for ambiguous interaction data,” in *ASEE Annual Conference and Exposition*, Baltimore, MD, 2023.

[32] F. Gregg and D. Eder, “Dedupe,” https://github.com/dedupeio/dedupe.

[33] J. De Bruin, “Python Record Linkage Toolkit: A toolkit for record linkage and duplicate detection in Python,” https://github.com/J535D165/recordlinkage.

[34] M. Bastian, S. Heymann, and M. Jacomy, “Gephi: an open source software for exploring and manipulating networks,” *International AAAI Conference on Weblogs and Social Media*. 2009.

[35] D. Kalamaras, “SocNetV.” p. Social Network Analysis and Visualization Software, 2018. [Online]. Available: http://socnetv.org/.

[36] G. Varoquaux, T. Vaught, and J. Millman, “Exploring network structure, dynamics, and function using NetworkX,” in *7th Python in Science Conference*, 2008.

[37] P. Krivitsky *et al.*, “Statnet Development Team,” *statnet: Software tools for the Statistical Modeling of Network Data*. http://statnet.org.

[38] U. Von Luxburg, “A tutorial on spectral clustering,” *Stat Comput*, vol. 17, no. 4, pp. 395–416, 2007.

[39] A. Ng, M. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” *Adv Neural Inf Process Syst*, vol. 14, 2001.

[40] A. Harding and J. Engelbrecht, “Personal learning network clusters: A Comparison between mathematics and computer science students,” *J Educ Techno Soc*, vol. 18, no. 3, pp. 173–184, 2015.

[41] S. E. Carrell, B. I. Sacerdote, and J. E. West, “From natural variation to optimal policy? The importance of endogenous peer group formation,” *Econometrica*, p. 855, 2013, doi: 10.3982/ECTA10168.