ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #35917

Work in Progress: Developing Disambiguation Methods for Large-Scale Educational Network Data

Mr. Adam Steven Weaver, Utah State University

Adam Weaver is a B.S. Mechanical Engineering student at Utah State University. His research is focused on developing explicit disambiguation methods for large-scale social network studies. In addition, he works with applications of Particle Image Velocimetry (PIV), and wrote curriculum using PIV to teach energy conservation to high school students.

Mr. Jack Elliott, Utah State University

Jack Elliott is a concurrent M.S. in Engineering (mechanical) and Ph.D. in Engineering Education student at Utah State University. His M.S. research is in fluid dynamics including the application of PIV, and his Ph.D. work examines student collaboration in engineering education.

Disambiguation of Large-Scale Educational Network Data

Introduction

Social Learning Theory (SLT) suggests that social interactions have the potential to positively impact learning [1]. Research has produced evidence of these positive effects in multiple aspects of engineering undergraduate students' learning, including their retention rates [2], critical thinking skills [3], and grades [4]. Capturing the relationships between students' academic performance and interactions requires methods for identifying and analyzing various interaction traits, like sub-network *homophily* (the grouping of individuals along a similar trait) or an individual's *centrality* (the connectedness of a certain individual in a network). Social Network Analysis (SNA) enables such numerical scrutiny; providing educators the ability to examine students' interactions statistically and visually.

In educational contexts, researchers apply SNA studies to strictly online environments or single courses. While these efforts are important steps for understanding the role of students' sociality in academia, these studies' inherent network limitations may yield incomplete conclusions about social networks. To wholly capture student support networks, studies should include face-to-face interaction types and extend beyond single classrooms. However, efforts to capture these wholistic networks introduce a complexity associated with collecting and processing large network data: *entity resolution*, or the process of assigning ambiguous reported references to real world individuals.

Current work involving entity resolution for educational network analysis is sparse, limiting educational research on student support networks. To combat this issue, this paper presents our work-in-progress developing and implementing methods for disambiguating large scale (1000+ students) network data. The final disambiguation framework consists of four iterative stages to create a "best-guess" of a completely resolved network data set and provides a general structure for future algorithmic methods. Results of this work will better enable researchers to study larger, more holistic educational networks.

Background

Students benefit from social interactions in a variety of ways. For example, Kalaian et al.'s [5] meta-analysis identified that across 18 studies, formal small group settings enhance student's abilities to succeed academically—especially among first-year engineering students ($d = 0.84$). Studying students' online social media interactions, Su and Huang [6] found that students who frequently use social media for academic purposes reported an enhanced learning experience at a Chinese public university. Further, Hurst, Wallace, and Nixon [3] surveyed undergraduate students and concluded that students believe social interactions increased their enjoyment of learning, interest in topics, and level of responsiveness. Beyond academic success, engineers of the 21st century must effectively engage with peers [7], and as noted by Passow [7], teamwork and communication are among the most valuable skills in the engineering industry. By researching these benefits, educators can capture and incentivize the positive attributes of students' interactions. Further, such research often requires methods, like SNA, for characterizing interactions of the research participants.

Social Network Development

Social Network Analysis uses quantitative representations of individuals' network qualities to allow researchers to study connections between individuals' interactions and traits of interest. Exploring students' social interactions through SNA requires researchers to represent each individual student of interest as a *node* and connections between students as *ties*. For a particular *network* (a given collection of nodes and respective ties), researchers assign each individual in the network a row and column in a matrix. By filling this matrix with interaction *weights* (interaction strength or distance from row individual to column individual), researchers construct an adjacency matrix. With adjacency matrices and modern software such as Gephi [9] or SocNetV [10], researchers can conduct statistical analysis of network traits-of-interest and map networks in plots called *sociograms* for visual analysis (e.g., comparing students' study network ties to GPA as demonstrated by Figure 1).

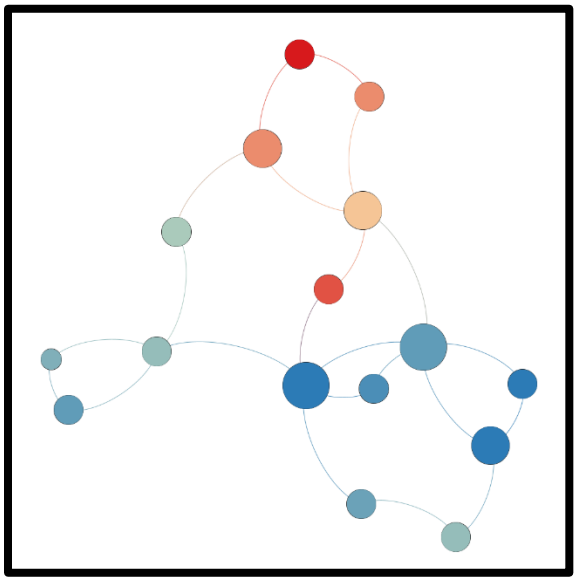

Figure 1. An example sociogram visualizing students (nodes), connections between students (curved lines), and students' grades (4.0 to 0.0: blue to red).

Within SNA, ego-centric networks represent an *ego* (person-of-interest) at the center of their network, and then surround the ego with *alters* (the ego's peers). To identify interactions between individuals in ego networks, researchers frequently use *name generator* surveys. For example, a study investigating friendship networks in classrooms may include a name generator survey question such as: "Considering all of your classmates, who are your closest friends?" Name generators guide participants to identify their perceived interactions within the network framed by the name generator. Name generators may be either close-ended or open-ended.

Researchers employing a close-ended name generator ask participants to identify their ties to people found on a list of names who belong to a network-of-interest. For example, Hansell [11] gave a class enrollment list to 5th and 6th grade students and asked them to rank their classmates

according to their fondness of each, identifying that cross-race and cross-sex ties were less likely to develop in the researchers' classrooms than same-sex and same-race ties. The close-ended name generator technique reduces the impact of respondents not remembering ties [12]. However, this technique strictly limits the development of social networks to a known population, and potentially ignores unexpected interactions. To circumvent this issue, researchers use open-ended name generators.

In contrast to close-ended name generators, which prompt individuals to identify ties *within* a network of interest, open-ended name generators prompt participants to freely identify ties to *create* a network of interest. For instance, Wellman [13] asked adult participants living in East York, Toronto, to provide detailed information about the participants' six closest friends residing outside their own home. This study determined that most residents of East York possess non-localized social networks and proved that kin networks are more densely clustered than other types of networks. As demonstrated by the Wellman [13] study, open-ended name generators produce valuable information which may not be observed in close-ended studies. However, because open-network name generator responses are ambiguous, such studies exhibit complications associated with entity resolution.

Simplification of Social Environments in Engineering Education

When applying SNA in engineering education, researchers balance accurate entity resolution inherent in close-ended network studies against authentic interaction data produced by open-ended network studies. In SNA research, properly identifying a population sample and interaction type to answer a given research question is a difficult but important task [12]. Further, SNA researchers must carefully consider the bounds of their study network in a trade-off between identifying a large enough sample to adequately gather all relevant interactions and maintaining a manageable study scope. Consequently, SNA research in engineering education has typically explored social learning in narrow social environments: online interactions and/or single courses.

Learning Management Systems (LMS) automatically record interactions between students. Therefore, researchers favor using LMS as a platform for SNA. For example, Llantos & Estuar's [14] study on LMS collaboration data found that administrators drive information distribution in student networks. In addition, Gupta [15] monitored students' interactions within an LMS to establish that students with higher influence and popularity levels earn better grades than other students. However, both studies were limited to LMS interactions, neglecting aspects of face-to-face collaboration and other aspects of online collaboration.

Outside of online learning environments, a common strategy for bounding networks in engineering education is to limit studies to a single course. This limit allows researchers to implement a close-ended network approach, reducing recall error. Grunspan et al. [16] had students select interactions from a drop-down menu containing the names of students belonging to a single biology class, pointing out that such menus reduce errors and simplify data processing. However, Grunspan et al. [16] noted that the single-class network study approach limits potentially important relationships students possess outside of their classmates.

Current Efforts to Represent more Holistic Student Networks

To date, well-bounded (limited to online interactions or a single course) studies have revealed important implications for practice, including encouraging dense, small networks [5], and a need for more structured conceptual support from teachers [17]. Yet, students' networks reach far beyond single classrooms—and are often tied together by a blend of online and face-to-face interactions [18]. Therefore, a comprehensive analysis of authentic relational data is founded in more holistic networks resulting from reduced bounding. Indeed, the most current literature shows researchers are making efforts to consider open networks beyond the bounds of single classes or online interaction types. We identified three of these most recent studies.

The first of these more holistic network studies, the Copenhagen Network Study [19] measured students' use of text messaging, social media interactions, and proximity to estimate social interaction. The results of this study showed that sub-networks are time activated, and that students' extraversion was strongly correlated to the number of their Facebook friends. To arrive at these conclusions, Stopczynski et al. [19] distributed phones among students and monitored the use and location of the phones. This study represented a very large network with various interaction types but relies on assumptions regarding cell phone usage and location data, including using location proximity to determine study interactions.

The second of the more holistic network studies was conducted by Lin [20], who monitored students' social networks by means of a self-report name generator survey during a three-year Social Network Analysis. This study found that students who room with each other during their first year of school are likely to develop and retain friendships. Accordingly, Lin [20] suggests schools can make more methodical decisions about dormitory rooming assignments to foster positive social interactions among students. Although this study was not limited to a single classroom, the sample population was small (approximately 50 students) and bounded to students studying civil engineering at a geographically isolated Taiwanese university.

The third and final large network study, conducted by Stadtfeld et al. [21], measured multiple types of interactions (i.e., students' friends and study partners) between 226 students at a Swiss university over a one-year period. This study identified that students with more incoming social ties were likely to perform better than students with fewer incoming ties, and friendship ties were 16 times more likely to lead to study ties than the opposite. While these findings show promise for engineering educators, this study was conducted at a prestigious, small, private university and results may have limited generalizability.

The Overarching Study

To further the current understanding of larger, less bounded networks, our research group is performing a SNA study to draw links between students' friendship interactions, formal interactions, informal interactions, and academic outcomes through the first and second years of an undergraduate degree. To gather network data, we sent all first- and second-year engineering students at the study institution (~1200 students) an invitation to identify peers they interacted with in an open-response name generator survey [22].

Due to the loosely bounded, open response nature of the survey, names provided by the participants were ambiguous. For example, many names were misspelled, found in different formatting order, or lacked information (e.g., known nicknames or missing last names). The COVID-19 pandemic also caused many students to rely more heavily on online interactions. Accordingly, students often knew their peers by only virtual tags instead of traditional names. Notwithstanding each ambiguity circumstance, it is still important to resolve each name to the correct student entity so educators can form conclusions around an accurate social network. Therefore, the purpose of this study is to answer the following research question (RQ): How can large-scale educational social network data be efficiently disambiguated?

Methods

To resolve the open-response network data by disambiguating each participants' response(s), we developed our disambiguation strategies concurrent with data collection of a large scale (1000+ node), loosely bounded educational network. After completing this concurrent network resolution, we delineated our disambiguation strategy. By qualitatively reflecting on our entire disambiguation process and then improving the framework, we developed a refined and complete disambiguation procedure for future studies.

Data Collection

During the Spring 2021 semester, we conducted a network study to understand the formation and effects of undergraduate peer interactions and outcomes. In accordance with an approved IRB protocol, we identified the first- and second-year engineering students at a public land-grant university through the university registrar. We added the 1,259 qualifying students to an online LMS course, where we informed the participants of their eligibility for the study and briefly described the study purpose and methods. We invited these students to participate in the study name generator surveys every three weeks through LMS announcements for a total of five iterations. Participation consisted of a Qualtrics name generator survey, where we asked each participant to list: their first and last name, any nicknames by which they may be known, and up to 20 peers they had interacted with in the past three weeks to either study, conduct group assignments, or for purely social purposes. Our recruitment efforts generated 1088 complete survey responses resulting in approximately 3,997 *ambiguous names* (individual names provided in the name generator responses) as shown in Figure 2.

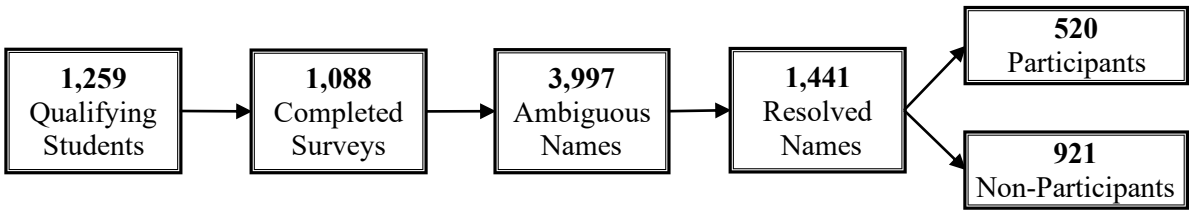

Figure 2. Spring 2021 network name and participant scope.

After participants completed each name generator survey, we uploaded their responses to a main spreadsheet. As shown in Table 1, the main spreadsheet contained the participant's name,

a provided nickname (if any), and peers they interacted with for studying, socialization, or group work.

Table 1. Example of Survey Responses. The green cell highlights the participant's own name, blue cells demonstrate ambiguous full name variances, and red cells demonstrate ambiguous partial names.

Name	Nick-name	Peer 1	Peer 2	Peer 3	Peer 4	Peer 5
John Chase Deer	N/A	Alex Sociogram	Bob Survey	Gerry Network	Earl Excel	
Bob Survey	Bobby	John Deer	Gerry Network	Hannah Nodal	Rick Social	Lindsey Analysis
Earl Excel	N/A	J.C. Deer	Matt Response	Rick Social		
Gerry Network	Jerry	Hannah Nodal	John, Deer	Bob		
Rick Social	N/A	Matthew Response	Lindsey Analysis	John, D	Rick Social	
Lindsey Analysis	N/A	Gerry Network	John	Alex Sociogram		
Hannah Nodal	N/A	Jon Deer	Earl Excel	Lindsey Analysis	Jared Interaction	
Jared Interaction	N/A	Lindsey Analysis	J-Dawg	Gerry Network	Hannah Nodal	Rick Social
Alex Sociogram	N/A	Deer, John	Bob Survey			
Matthew Response	Matt	John, D	Alex Sociogram	Hannah Nodal	Lindsey Analysis	

Resolving Full Names

To analyze the interaction data, we first needed to disambiguate the names we found in the survey responses. We started by resolving participants' own names, which appeared in full-name form. To keep track of each newly resolved full name, we dedicated a spreadsheet as a key, and then secured the key in a cloud folder accessible only by the research team for de-identification purposes. In the key, we assigned each resolved name a number and recorded known nicknames (whether provided by the participant or found in the interaction data) as shown in Table 2.

Table 2. Example of the Key demonstrating resolved number, full names associated with each number, and nicknames/other common spellings used to resolve the full name.

Number	Name	Nickname(s)
1	John Chase Deer	Jon, J.C.
2	Bob Survey	Bobby
3	Earl Excel	
4	Gerry Network	Jerry
5	Rick Social	
6	Lindsey Analysis	
7	Hannah Nodal	
8	Jared Interaction	
9	Alex Sociogram	
10	Matthew Response	

Beyond participants' provided names, the remaining ambiguous full names came in varied formatting orders and spellings. After thinking of several possible misspellings (i.e., common misspellings—like John Deer and Jon Deer), we used Excel's find function to search for these *minor* variances (e.g., vary by only one character/common misspellings and/or had the order of names reversed). After we found the full-name variances, we used Excel's replace function to consolidate the full name variances with their respective resolved name (e.g., Jon Deer to John Deer).

Developing the key also required resolving ambiguous names to students outside the inclusion criteria (i.e., other students not enrolled in the engineering program or beyond the first and second year). Accordingly, we repeated a similar process of identifying ambiguous full names (this time, for those full names within the interaction data), finding variances to these ambiguous full names, and consolidating the variances to their respective ambiguous full name.

Resolving Partial Names

After resolving full name minor variances, we realized many of the participants knew their peers strictly by first names, gamer-tags, or nicknames (e.g., John Deer known as J-Dawg). Some participants also listed their peers by their first and middle initials only (e.g., John Chase Deer listed as J.C.). We also recognized that as we identified connections between resolved names, we could make educated guesses about who these ambiguous partial names belonged to through the additional resolved network information.

One way to identify connections between resolved participants' names and their ambiguous alters (ambiguous partial names) was to compare all iterations of a single participant's survey responses. Because participants often learned their peers' last names over the course of the semester, we examined each single participant's later responses in attempt to find more information about the remaining partial ambiguous names (typically, the last name was identified later in the study). If this process did not yield the information needed for entity resolution, we took advantage of known network information to connect potentially related ego-networks.

Specifically, we took note of *potentially matching names* (resolved names whose first names matched the partial ambiguous names), and then compared the sub-networks surrounding these potentially matching names with the sub-network surrounding the ambiguous partial name. As shown by Figure 3B, we checked the alters belonging to the participants who listed the ambiguous partial name and the potentially matching name (in the case of Figure 3B, both Jared Interaction and Bob Survey had five alters). Further, if four alters belonging to these participants were exact matches, we consolidated the potentially matching name with the ambiguous partial name.

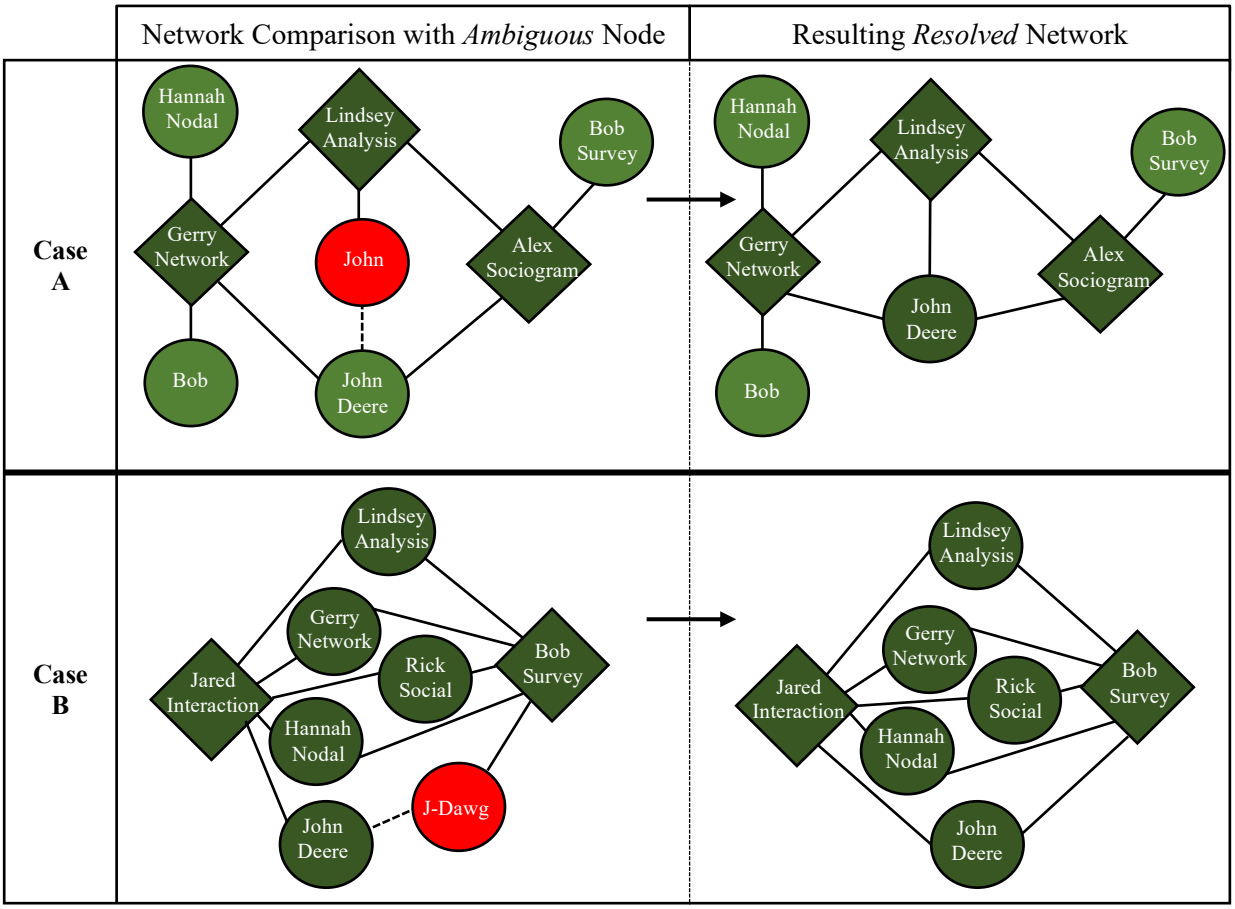

Figure 3. Sociograms of Sub-Network Comparisons. Participants' resolved sub-networks are represented by dark green nodes, with parent nodes rendered as diamonds. Light green nodes represent resolved names belonging to an expanded sub-network. The red nodes represent an ambiguous partial name. We expand an original participant's sub-network (A), or contrast two potentially related sub-networks (B) to prove that the ambiguous partial name is the potentially matching name (depicted by a dashed line)

We also took note of *sure ties*: each parent node that a potentially matching name was associated with, or, if the potentially matching name was a participant—all their alters. As shown in Figure 3A, if two or more sure ties shared a common parent node with the partial ambiguous name, we consolidated the potentially matching name and the partial ambiguous name. Subsequently, for an ambiguous partial name, either Case A or Case B shown in Figure 3 had to

be fulfilled by the sub-network comparisons for entity resolution. If a partial ambiguous name was not resolved at the conclusion of these comparison processes, we did not have enough information to confidently resolve that entity. Therefore, we resolved these remaining partial ambiguous names as a new node.

Reflective Process Evaluation

At the conclusion of the Spring 2021 semester network study, we reflectively considered our manual disambiguation strategy. To complete this evaluation, we began by writing down each step we took to disambiguate the student network data and considered the name variances associated with each disambiguation strategy. To visualize this process, we graphically represented each type of disambiguation in a flow chart. With all the disambiguation strategies represented graphically; we placed the steps in a procedural order. To iteratively evaluate this flow-chart process, we then considered various types of name variances and imagined what step (if any) the ambiguous names would be resolved at. If an ambiguous name passed through these steps without being resolved, we refined the process and tested the flow chart again. After completing this process for each name variance type, we arrived at a complete strategy for disambiguating interaction data. Organizing this strategy allowed our efforts to become more repeatable and became a framework for our ongoing automation efforts.

Results

To resolve the open-response network data, we pass each ambiguous name through four successive stages: each stage delineating processes for resolving ambiguous interaction data. Splitting the overall disambiguation task into consecutive stages also allows the researcher to be more methodical about automating each ambiguity circumstance as demonstrated by Figure 4. Further, the likelihood of error in the entity resolution increases with each stage, with the highest confidence of accurate entity resolution for names disambiguated at stage one, and the lowest confidence if a name is disambiguated at stage four.

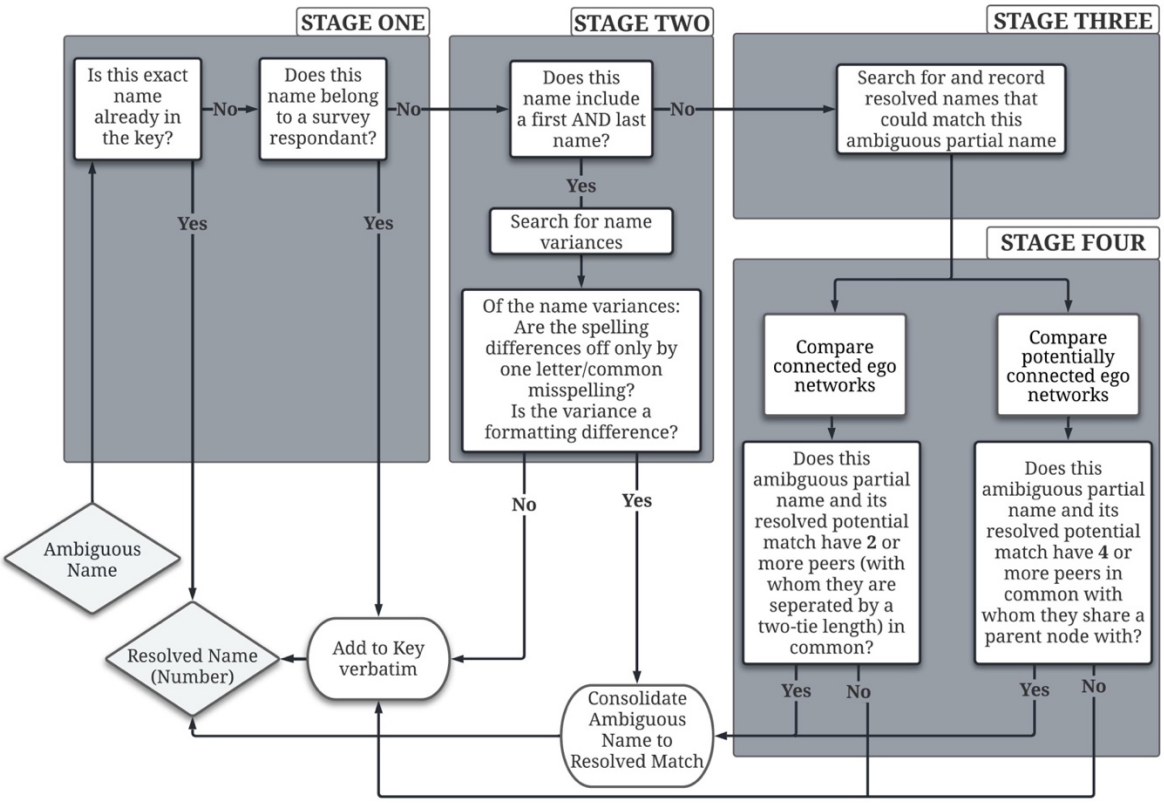

Figure 4. Overarching Disambiguation Strategy

The Four Stages of Disambiguation

Stage 1 identifies ambiguous names which could be resolved in high-confidence and resolves them. High-confidence ambiguous names includes participants' names obtained by the registrar and user-provided names which are in full-name form. In the resolution process, we find ambiguous full names (or their exact matches), add these matches to the key, and replace instances of these ambiguous names with correct number(s) from the key.

In Stage 2, we identify resolved full names who have name-variances by searching potential variances using Excel's find function. We then consolidate these ambiguous full names to their resolved variances (replacing the name with the key number) if the variants deviate only by minor spelling differences (i.e., one letter off or common variant spellings) or formatting order.

In Stage 3, we identify ambiguous partial names, and find resolved names that could be matched with these names. To identify ambiguous partial names, we manually highlight every partial name (i.e., first-name only, initials, nicknames) in the interaction data and record them. For every highlighted name, we then use Excel's find function to identify resolved names whose first names or initials match with the ambiguous partial names and record these potentially matching names for use in Stage 4.

In Stage 4, we perform sub-network comparisons to identify how many peers an ambiguous partial name has in common with a resolved potentially matching name. Specifically, we compare the surrounding network(s) of the ambiguous partial name and the related potentially matching name. If these sub-network comparisons find either two matching peers in connected sub-networks, or four matching peers in potentially related sub-networks (Figure 4), we consolidate the ambiguous partial name to the respective potentially matching name. We provide names without sufficient sub-network similarity a unique key number and resolve them as-is.

Automation Efforts

All disambiguation stages exhibit potential for automation. To date, we have automated Stage 1: resolving exact matches of survey respondents' own names and/or registry data. For resolving exact names, we wrote a Python script that adds registry names to a key then finds and adds participant-provided names contained in the survey responses to the key. The script then concatenates the interaction data and replaces all instances of an exact match from the key appearing in the interaction data. Resultantly, the script produces an automatically disambiguated version of the interaction data for the entirety of Stage 1. The autonomous process we wrote to complete Stage 1 has one remaining limitation: two instances of exact names belonging to different students would be automatically consolidated into one entity. Therefore, we remove these names from the running key and manually resolve them.

Future Work

Future work consists of developing a primarily algorithm-based disambiguation process. To date, we have automated Stage One through a python script. Further, we are actively incorporating string matching methods for automating stage two and stage three. To begin stages two and three, Snae [23] cites that *Levenshtein* distance (text-based similarity measurement) [24] in addition to the *Metaphone* algorithm (language-based similarity measurement) [25] are used for string matching methods. Horng-Jyh et al. [26] describes using *fuzzy-name* matching via language-based and text-based similarity (i.e., generating a 0-1 scale value of the estimated likelihood an ambiguous name matches a given resolved name), indicating that such methods are valuable tools for public directories. Menger et al. [25] used a Levenshtein distance of 1 to develop an automatic de-identification process for medical data that achieved a 94% accuracy. For the large-scale educational data in our study, we will develop and implement both algorithms: compare phonetic representations of an ambiguous name to resolved names, then compare an ambiguous name's Levenshtein distance to resolved names.

Combining these measures in a fuzzy name matching algorithm will allow us to threshold name similarities determined by the manually disambiguated data. Comparing the names added to the key after stage one with each remaining name in the interaction data, we will compute and save each Levenshtein distance and Metaphone value. After saving these string-matching values, we intend to cluster the social network(s) according to a hierarchical clustering method. In studies where researchers know some network information, hierarchical clustering identifies the network proximity of two seemingly different names. For example, Malin [27] successfully applied agglomerative hierarchical clustering to disambiguate a network of 180,000 Hollywood

actors. Ferreira et al. [28] also outlined similar disambiguation strategies in bibliometric analysis by noting a difference between author name grouping (where authors are grouped according to similarity) and author assignment (assigning each entity an author and then determining if that author is correct using a similarity function). We will use these clustering methods to consolidate the overall network data, allowing path lengths to represent sub-network comparison proximity.

Combining these sub-network comparisons with the fuzzy name matching values will provide us a numeric estimation of name similarity and sub-network proximity for each ambiguous name compared to any given resolved entity, allowing us to resolve names in stages two through four. Determining the appropriate weights/thresholds is readily achievable through our already manually resolved network. By determining these parameters, we will automate the stage four comparisons and thereby, the entire network resolution process. To assess the robustness of our automation methods, we will further generate synthetic interaction data set(s), run the automated disambiguation strategies, and compare resulting resolved networks. The result of this automation will provide educators employing SNA a means to resolve large scale interaction data outside of traditionally limited contexts.

Conclusion

Our current disambiguation method yielded a best guess of a complete and large student network for the purpose of conducting longitudinal SNA of undergraduate engineering students' informal, formal, and friendship connections. Our strategy also provided a framework for future work to include more effective and efficient disambiguation methods. Researchers hoping to disambiguate large-scale education data can complete the disambiguation task through four iterative stages. These stages filter out each commonly encountered group of ambiguity circumstances. The stages include:

Stage 1. Resolving high confidence names: locate and resolve exact matches of registry or participants' own names.

Stage 2. Resolving name variances: locate and resolve name variances belonging to resolved full names.

Stage 3. Matching partial names: locate and note ambiguous *partial* names (e.g., initials, first name only, nickname) and their resolved potentially matching names.

Stage 4. Resolving partial names: perform sub-network comparisons to see if ambiguous partial names should be consolidated to their resolved potentially matching name.

The ambiguity circumstances embedded in the interaction data yielded from a loosely bounded network study are addressed in each iterative stage. An important note, that as names progress through these stages, the likelihood of accurate resolution decreases. For instance, resolving an exact match of a unique name that was registry identified (stage one) is more likely to be a correct entity resolution than matching a first name only with resolved name based on similar sub-networks (stage four). Beyond the manual framework, every stage exhibits possibility for automation. To date, we successfully automated in Stage 1. Our efforts are now guided toward accurately automating Stages 2 through 4, which is made possible by metaphone, Levenshtein distance, and hierarchical clustering.

Continuing to refine and develop disambiguation methods will be vital for allowing Social Network Analysis to extend to more wholistic student networks, identified through more holistic interaction data and open-ended name generators. Conclusions drawn from such studies will enable educators to obtain accurate data regarding networking of students and learn how to incentivize positive and lasting ties in engineering education.

Acknowledgements

This material is based upon work supported by the second author's National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745048. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

Works Cited

[1] A. Bandura, "Albert Bandura- Social Learning Theory," *Simply Psychology*, 1977.

[2] N. A. Bowman, L. Jarratt, L. A. Polgreen, T. Kruckeberg, and A. M. Segre, "Early identification of students' social networks: Predicting college retention and graduation via campus dining," *Journal of College Student Development*, vol. 60, no. 5, 2019, doi: 10.1353/csd.2019.0052.

[3] B. Hurst, R. Wallace, and S. B. Nixon, "The impact of social interaction on student learning," *Reading Horizons*, vol. 52, no. 4, 2013.

[4] D. J. Zimmerman, "Peer effects in academic outcomes: Evidence from a natural experiment," *Review of Economics and Statistics*, vol. 85, no. 1. 2003. doi: 10.1162/003465303762687677.

[5] S. A. Kalaian, R. M. Kasim, and J. K. Nims, "Effectiveness of small-group learning pedagogies in engineering and technology education: A meta-analysis," *Journal of Technology Education*, vol. 29, no. 2, 2018, doi: 10.21061/jte.v29i2.a.2.

[6] X. Su and J. Huang, "Social media use and college students' academic performance: Student engagement as a mediator," *Social Behavior and Personality: an international journal*, vol. 49, no. 10, 2021, doi: 10.2224/sbp.10797.

[7] ABET, "Criteria for accrediting engineering programs, 2020-2021." 2021.

[8] H. J. Passow, "Which ABET competencies do engineering graduates find most important in their work?," *Journal of Engineering Education*, vol. 101, no. 1, 2012, doi: 10.1002/j.2168-9830.2012.tb00043.x.

[9] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," May 2009.

[10] D. Kalamaras, "Social Network Visualizer (SocNetV)." 2015.

[11] S. Hansell, "Cooperative Groups, Weak Ties, and the Integration of Peer Friendships," *Social Psychology Quarterly*, vol. 47, no. 4, 1984, doi: 10.2307/3033634.

[12] S. P. Borgatti, M. G. Everett, and J. C. Johnson, *Analyzing Social Networks*. SAGE Publications, 2013. [Online]. Available: https://books.google.com/books?id=dHhpBAAAQBAJ
[13] B. Wellman, "The Community Question: The Intimate Networks of East Yorkers," *American Journal of Sociology*, vol. 84, no. 5, 1979, doi: 10.1086/226906.
[14] O. E. Llantos and M. R. J. E. Estuar, "Characterizing instructional leader interactions in a social learning management system using social network analysis," in *Procedia Computer Science*, 2019, vol. 160. doi: 10.1016/j.procs.2019.09.455.
[15] A. Gupta, "Application of Human Factors Engineering Principles to the Development Of Social Network Analysis (SNA) Assessment Tools for Use By Teachers Within A Collaborative Educational Environment," Medford, MA, 2014.
[16] D. Z. Grunspan, B. L. Wiggins, and S. M. Goodreau, "Understanding classrooms through social network analysis: A primer for social network analysis in education research," *CBE Life Sciences Education*, vol. 13, no. 2, 2014, doi: 10.1187/cbe.13-08-0162.
[17] C. I. Damşa and M. Nerland, "Student Learning Through Participation in Inquiry Activities: Two Case Studies in Teacher and Computer Engineering Education," *Vocations and Learning*, vol. 9, no. 3, 2016, doi: 10.1007/s12186-016-9152-9.
[18] L. Barkhuus and J. Tashiro, "Student socialization in the age of facebook," in *Conference on Human Factors in Computing Systems - Proceedings*, 2010, vol. 1. doi: 10.1145/1753326.1753347.
[19] A. Stopczynski *et al.*, "Measuring large-scale social networks with high resolution," *PLoS ONE*, vol. 9, no. 4, 2014, doi: 10.1371/journal.pone.0095978.
[20] S. C. 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.
[21] C. Stadtfeld, A. Vörös, T. Elmer, Z. Boda, and I. J. Raabe, "Integration in emerging social networks explains academic failure and success," *Proc Natl Acad Sci U S A*, vol. 116, no. 3, 2019, doi: 10.1073/pnas.1811388115.
[22] J. Elliott, A. Minichiello, and J. D. Marquit, "Work in Progress: An Investigation of the Influences of Peer Networks on Engineering Undergraduate Performance Outcomes." Jul. 2021.
[23] C. Snae, "A Comparison and Analysis of Name Matching Algorithms," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 1, no. 1, 2007.
[24] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10. pp. 707–707, Feb. 1966.
[25] L. Philips, "The Double Metaphone Search Algorithm," *C/C++ Users Journal*, vol. 18, pp. 38–43, Jun. 2000.
[26] P. W. Horng-Jyh, N. Jin-Cheon, and C. K. Soo-Guan, "A hybrid approach to fuzzy name search incorporating language-based and text-based principles," *Journal of Information Science*, vol. 33, no. 1, 2007, doi: 10.1177/0165551506068146.
[27] V. Menger, F. Scheepers, L. M. van Wijk, and M. Spruit, "DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text," *Telematics and Informatics*, vol. 35, no. 4, 2018, doi: 10.1016/j.tele.2017.08.002.
[28] B. Malin, "Unsupervised Name Disambiguation via Social Network Similarity," *SIAM SDM Workshop on Link Analysis, Counterterrorism and Security*, vol. 1401, 2005.

[29] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender, “A brief survey of automatic methods for author name disambiguation,” *SIGMOD Record*, vol. 41, no. 2. 2012. doi: 10.1145/2350036.2350040.