12-9-2021

Disambiguation of Large-Scale Educational Network Data for Social Network Analysis

Adam Weaver
Utah State University, adamweaver2000@gmail.com

Disambiguation of Large-Scale Educational Network Data for Social Network Analysis

Introduction

Research shows that under certain conditions, social interactions relate to student performance and retention. As a result, researchers frequently deploy Social Network Analysis (SNA) methods for identifying and incentivizing positive social conditions. SNA is a research method that quantifies connections between individuals that form a network according to traits of interest. Researchers mathematically represent connections using an *adjacency matrix* (Figure 1), and then analyze this data using contrast methods, or by visualizing them through *sociograms* (Figure 1). Unfortunately, genuine educational networks often exhibit ambiguity (i.e., names with not-obvious connections), and this steers many researchers away from studying these types of networks, even though they are most authentic to the educational context. To address this issue, this presentation describes our current work to disambiguate large scale social network data.

A

Name	Nickname	Peer1	Peer2	Peer3
John Deer	None	Alex Sociogram	Bob Survey	
Bob Survey	Bobby	John Deere	Gerry Network	Hannah Nodal
Earl Excel	None	J.D	Matt Response	Rick Social
Gerry Network	Jerry	Lindsey Analysis	J-Dawg	Bob
Rick Social	None	Matthew Response	Lindsey Analysis	John, D
Lindsey Analysis	None	Gerry Network	John	Alex Sociogram
Hannah Nodal	None	Jon Deer	Earl Excel	
Jared Interaction	None	Lindsey Analysis	John, Deer	
Alex Sociogram	None	Deer, John	Bob Survey	
Matthew Response	Matt	John, D		

B

Node/ Node	1	2	3	4	5	6	7	8	9	
1	0	1	0	0	0	0	0	0	1	0
2	1	0	0	1	0	0	1	0	0	0
3	1	0	0	0	1	0	0	0	0	1
4	1	1	0	0	0	1	0	0	0	0
5	1	0	0	1	0	1	0	0	0	1
6	1	0	0	1	0	0	0	0	1	0
7	1	0	1	0	0	0	0	0	0	0
8	1	0	0	0	0	1	0	0	0	0
9	1	1	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0

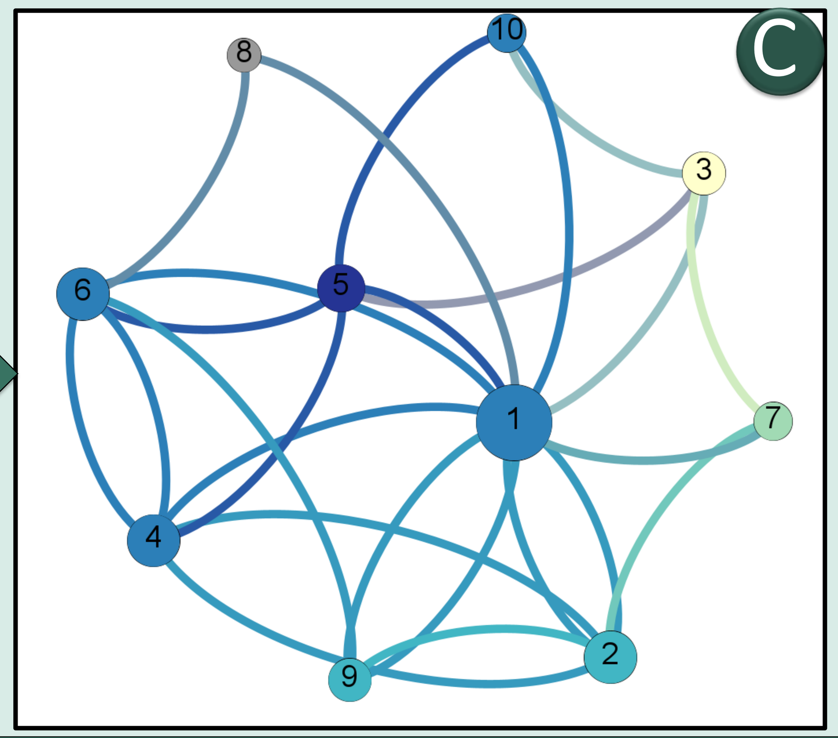
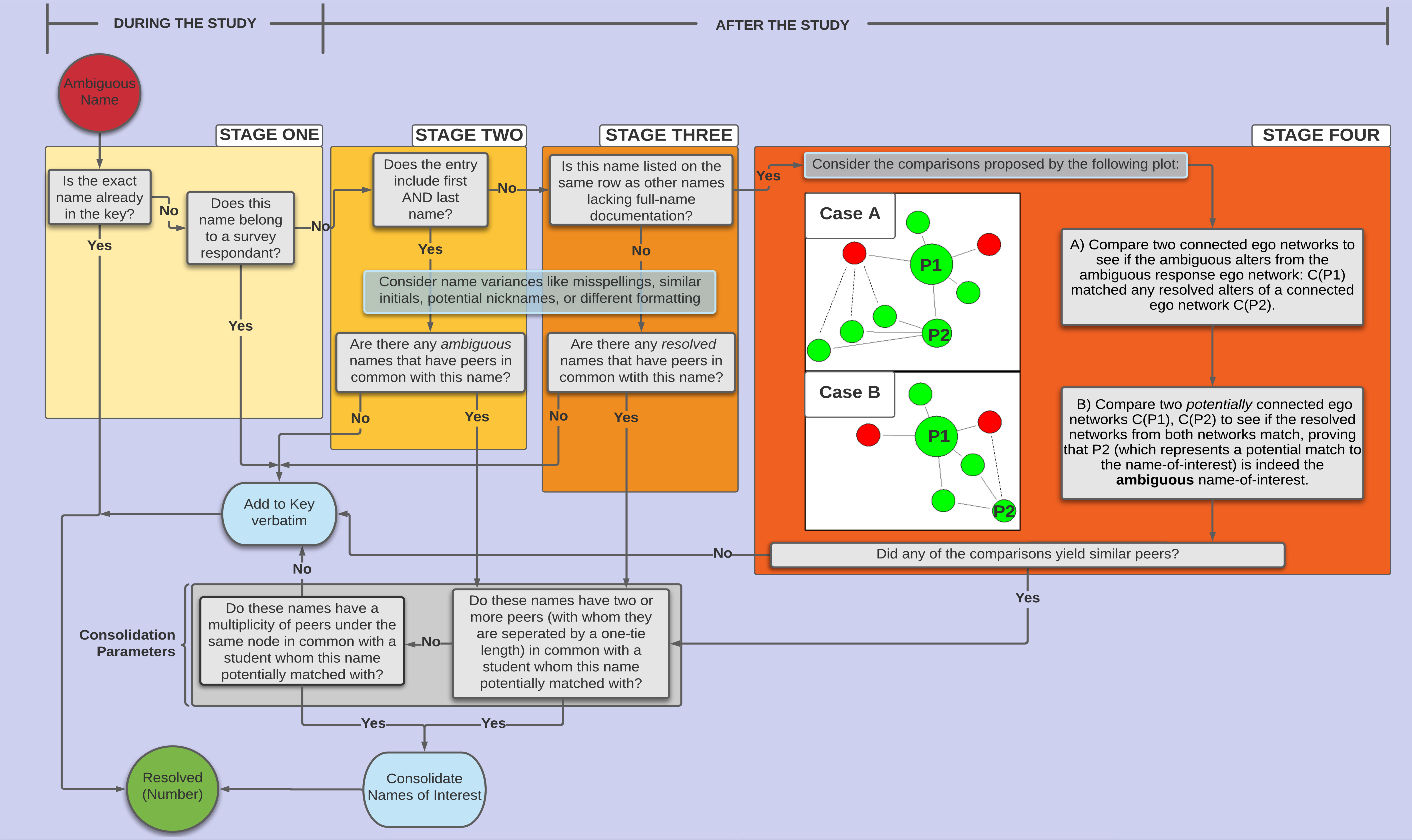


Figure 1. Survey responses are downloaded into an **excel spreadsheet (A).** Each person in the network is assigned a "node", and ties between nodes are consolidated into an **adjacency matrix (B).** The adjacency matrix is analyzed statistically and used to create **sociograms (C)** for visual analysis.

Methods

We organized the overarching network development task into discrete stages to filter responses according to unique name-ambiguity circumstances. To complete these stages, we relied on a hybrid blend of automation rooted in Excel and python, with following manual substitution.



Stage 1: Resolve Exact Names
Concurrent with data collection, Stage 1 identifies a "key" of high confidence names (i.e., user-provided or full names) and resolves them.

Stage 2: Match Resolved Names
Stage 2 consolidates ambiguous full names to resolved names if they varied only by formatting or spelling.

Stage 3: Match Ambiguous Names
Stage 3 finds resolved names that could be matched with ambiguous partial names.

Stage 4: Matches Double-Ambiguous Names
Stage 4 performs sub-network comparisons on ambiguous names.

Figure 2. The overarching disambiguation task stages, with a survey response beginning ambiguous (red) and filtering to a resolved node (green).

Results

The methodology outlined by Figure 2 was effective in consolidating the ambiguous network data; creating a best estimate of the complete network for further study. This work developed procedures for future researchers to efficiently consolidate network data.

Conclusions

This manual disambiguation process provided two key results outside of the primary study:

1. A framework for coding hybrid and disambiguation methods
2. A "best estimate" of a disambiguated network for testing and validation of automated methods

We will use these results for preparing an algorithmic equivalent, deploying agglomerative hierarchical clustering, to provide efficient means for large scale network development.

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

UtahStateUniversity

Adam Weaver
Utah State University
Department of Engineering Education
adamweaver2000@gmail.com