Visualizing Systematic Literature Reviews to Identify New Areas of Research

Allison Godwin
School of Engineering Education
Purdue University
West Lafayette, IN
godwina@purdue.edu

***Abstract*—This research paper describes the application of linguistic analysis through scaled co-occurrence networks to create visual representations of systematic literature reviews. This approach uses currently available and free tools in this new application. Co-occurrence networks are a method for visualizing relationships between concepts within written material. The results of a co-occurrence analysis are a network that reflects the relationships between words based on meaning similarities among words and text segments. To illustrate this approach, I applied the methodology of a systematic literature review search that resulted in 391 unique journal articles, books, and reports on identity published in the period from 1995 to 2015. The abstracts, titles, and keywords of these journals were analyzed via linguistic analysis to create a word co-occurrence network of these articles related to identity in science and engineering. Clusters within this network were identified based on word frequency. The results of this research illustrate a hole in the current identity literature in understanding diversity beyond traditional definitions of race, class, and gender. This method has the potential to powerfully convey and embed information about large amounts of data in a single image and offers a new way to report findings from a systematic literature review.**

Keywords—linguistic analysis; co-occurrence networks; network analysis; systematic literature review

I. Introduction

Since the publication of Borrego, Foster, and Froyd's *Journal of Engineering Education* paper describing the methodology for a systematic literature review [1], the frequency and popularity of these reviews in engineering education to critically appraise and summarize research has grown dramatically. The last and essential step in a systematic literature is a synthesis of the articles identified. This process involves mapping, critiquing within studies, and critique across studies. Often in the mapping step of these reviews, the numerous journal articles that are identified for inclusion via database searches are reported tables that are long and complex to understand. As an alternative, I present a process for creating a single visual network created with linguistic analysis. This method is not new but has not been used in engineering education in conjunction with systematic literature reviews. This process is illustrated with a specific example of a systematic review of identity and diversity research in science and engineering education.

Linguistic analysis allows for visually identifying connections and holes within the current body of literature. While this technique does not replace the full synthesis in a systematic literature review, it does offer a new way to examine how papers are connected to one another in the literature and to visually represent these connections. This method has been successfully used in information retrieval, educational technology and other pattern recognition problems with complex data [2], [3]. In a systematic review that can include hundreds of papers, this method can provide ways to map large bodies of complex linguistic data. In the following sections, I present summaries of the method of linguistic analysis and the methodology of systematic literature reviews as well as a step-by-step example of how these methods were used in combination to understand the STEM education literature on student identity and diversity.

II. Linguistic Analysis

This study used techniques from linguistic analysis to understand and visually represent the body of knowledge generated in STEM education on identity and diversity over the last two decades. The underlying assumptions for this work are that language through written text is the mechanism for the exchange and transmission of knowledge. Linguistic analysis extracts the knowledge presented in published journal articles, books, and reports to analyze it and organize it in a systematic way. This approach is based on the science of science, or the study of science communication, which has been used to understand author networks of citations, maps of scientific concepts, or visualizations of conference topics [4].

Word co-occurrence networks model the relationships between sources of text based on which words are contained in contexts (e.g., documents, paragraphs, sentences, etc.) and between words within and across contexts. The underlying concept is that these contexts in which a particular word appears gives information about the meaning of the word as well as the context and relationships among contexts. This approach takes raw data and develops a co-occurrence network of words with the nodes (or vertices) as the keywords and the edges as the weighted connections based on the frequency of words among documents. In other terms, nodes in a co-occurrence network represent important terms and location of these terms in a network reveal close "neighbors" (connected

by edges) as well as "neighborhood" or words that commonly appear in text together (i.e., clusters). The structure of these networks can provide information on how different publications in the literature base are using words around topics and visually identify "holes" in the literature that offer potential new areas of study.

III. Systematic Literature Review

Systematic literature reviews are a collection of methodologies that provide ways to critically appraise and make sense of large bodies of literature to inform policy and practice. Systematic reviews have been used to highlight gaps in the literature or highlight concepts that are accepted as true with little evidence [5]. Borrego, Foster, and Froyd highlight four steps in conducting a systematic literature review [1]: (1) defining inclusion and exclusion criteria; (2) finding and cataloging sources; (3) critique and appraisal and; (4) synthesizing. Systematic reviews are appropriate "when a general overall picture of the evidence in a topic area is needed to direct future research efforts" [5].

Inclusion and exclusion criteria set the bounds for what artifacts (e.g., archival journal papers, books, reports, conference proceedings, etc.) will be included in the review. Inclusion criteria are aspects of artifacts that are essential for incorporation into the study. Exclusion criteria are aspects of artifacts that require removal from the study. These criteria should be defined before gather sources to reduce potential bias in results. These criteria should be liberally applied, only excluding studies that clearly fit the exclusion criteria [6].

The next step of a systematic literature review is searching multiple databases to locate every study that potentially could be used to understand the research questions. This step involves using keywords to search multiple databases and extract all relevant sources. At this point, a time window can be imposed on what source are selected to include only current literature. This date restriction should be consistent with the research questions. Cataloging sources requires that the results of the search be exported into a usable format and the data cleaned for duplicate entries.

Once the data are collected, the sources must be critiqued using the inclusion and exclusion criteria along with descriptions of the studies (e.g., number of authors, topic, level of education, methods, etc.) and measures of quality of the study. This step is often undertaken by a group of researchers rather than an individual for interrater reliability.

The final step in a systematic literature review is the synthesis of the included artifacts. The purpose of this step is to pool the findings from all of the artifacts to provide an underlying outcome that shows what the current body of literature on the topic has found as well as what gaps remain in our current understanding. Often this step is a synthesis of long tables of information included from the coding step of data gathering. It can be difficult for researchers as well as reader to interpret and find meaning in numerous artifacts, especially in understanding gaps in the literature.

IV. Identity and Diversity in Science and Engineering Education

To illustrate how I have combined a systematic literature review with a linguistic analysis to create a visual network, I provide an example of current literature on identity (especially work on diversity) in science and engineering education. This example is only one application of this technique; however, it provides a clear example of how this approach might be used in engineering education. In this work, I define engineering identity as the ways in which students describe themselves in the role of being an engineer. The authoring of their stories as engineers is the central process of envisioning themselves as engineers. Engineering identity is more complex and nuanced that just feeling like an engineer or seeing oneself as an engineer. There are many different facets and contexts that play into the development of a central engineering identity for students. Engineering identity is an important factor for engineering career choice, retention, and personal development [7]–[16]. This topic of research has grown in popularity in engineering education in recent years. However, it has historically been more widely researched across STEM education, especially in science education [17].

Engineering identity is not formed in isolation. Other social identities like race, ethnicity, class, and gender impact the formation of an engineering role identity. Some studies have documented the impact of being a woman of color in science and engineering fields which highlight different concerns than women as a whole who are majority white women [18]–[22]. Understanding how work on identity and diversity have been framed in the current science and engineering education literature can provide ways to understand the history of this research topic as well as new areas for research.

Additionally, I was interested in how identity has been understood in the context of unobservable aspects of diversity like sexual orientation, disability status, mindsets, and personality. These invisible aspects of an individual's identity can be covered or expressed as the individual chooses. The findings of this analysis are inclusive of these forms of diversity as well as more traditionally researched aspects of gender identity/expression, race, ethnicity, and class.

V. Applying Linguistic Analysis to a Systematic Literature Review

To understand the current field of research on identity and diversity in engineering, I conducted a systematic literature search for work on identity as well as identity in the context of diverse populations within science and engineering education. I followed the steps for a systematic literature review to define the inclusion and exclusion criteria, gather the data, and critique sources. After compiling a complete database of current literature (i.e., literature from 1995 to present) on identity and diversity in science and engineering education, I used linguistic analysis to understand the relationships between words used in this work by creating scaled word co-occurrence networks from the titles, keywords, and abstracts of these artifacts. These networks were analyzed for clusters within the structure of the data and these clusters were coded by theme to understand connections and gaps in the literature. This work is

consistent with other research in engineering education that uses network analysis to find trends in research and holes for innovation [23]–[25]. Below, I describe each step of the process in combining systematic literature reviews with co-occurrence network visualization including which tools I used and how they work to create the final visual network.

A. Gathering Data

I searched Engineering Village, Scopus, ERIC, Education Full Text, and Web of Science databases. This search resulted in 391 unique journal articles, books, and reports. The keywords used in this search were *identity* in conjunction with *science*, *engineering*, *mathematics*, *technology*, *STEM*, and *education*. These words were chosen to capture as many artifacts on how identity is address across STEM education. Papers related to identity in STEM with a focus on science and engineering published in the period from 1995 to present (2015) were included in this analysis. Only archival journal papers, books, and reports were included in this analysis to report the research that has gone through the most rigorous peer review process. Since the focus of this work is on students' engineering identities, I excluded papers related to teacher identity development.

B. Cleaning Data

Many of the databases allow for comma-separated values files to be exported with detailed information about each paper including authors, title, year of publication, journal, volume, issue, page numbers, keywords, and abstracts or book and report summaries. Other databases allow for a limited number or incomplete export of records as text files. Records from these databases were downloaded in batches and filled in by hand for complete records.

The text file data were cleaned by first being sorted by individual record (i.e., single journal paper, book, or report) and then imported into SAS statistical programming software [26]. Because the data for each field (i.e., author, title, etc.) were contained in multiple cells from the download, the data were divided into separate datasets for the authors, titles, journals, etc. as well as times of download so that the full information could be concatenated into one observation for each journal paper, book, or report. After concatenating all data fields for each variable, the datasets were merged back into one large dataset by using a common identification variable. These records were combined with the complete records into a single comma-separated values dataset. The data were cleaned of redundant entries for a total of 391 unique archival journal articles, books, and reports.

C. Using Sci2 Tool to Create Co-occurrence Network

Once the data were cleaned, Sci2 Tool was used to conduct the linguistic analysis [27]. In an efficient and effective analysis, documents and words typically undergo a series of pre-processing steps to allow common meanings to be extracted from words with slightly different grammatical forms or usage within the text. The text was converted to lowercase letters to homogenize the text example. Then, the text was tokenized into individual words using a common delimiter for spacing between words. In Sci2 Tool the default delimiter is a vertical bar character (|). Then, the words were converted to stems which removed common or low-content prefixes and suffixes to identify the core concept. Finally, a list of stopwords was applied to remove common low-token words like "of" or "and" from the analysis. The stopwords dictionary is built into the Sci2 Tool, but can be edited to add additional words. For this analysis, the default stopword dictionary was used. To illustrate the preprocessing process, the phrase, "Identity development of professional engineers," would become, "ident|develop|profession|engin" after converting to lowercase letters, tokenizing, stemming, and removing stopwords. This allows common words with similar meaning like engineer, engineering, engineers, engineered, etc. to all be treated as the same base word.

Once the data were preprocessed a co-occurrence word network was extracted. The similarity of topics in single words, as well as aggregate units of text, can be calculated through co-occurrence. Documents that share more words in common are assumed to have higher topical overlap. The co-occurrence network analysis in Sci2 Tool creates a weighted, undirected network where each node is a word and edges connect words to each other, where the strength of an edge represents how often two words occur in the same body of text together. The resulting network was cleaned by removing any unconnected nodes and scaled to determine the underlying structure.

Network scaling reveals the underlying structure and organization of the data using similarities, correlations, or distances to trim a network and prioritize similarities between nodes. Many networks, including topical networks, have interesting and unusual topological properties that are often valuable when visualized graphically. However, it is difficult to visualize raw networks that have not been scaled, especially when the size of the network grows proportionally with the data included in the analysis. Specific algorithms have been developed to deal with this visualization issue and prune raw networks to the most important structures for visualization. This approach is commonly used to visualize complex semantic structures. One of the most well-known algorithms is the Pathfinder algorithm [28]. The Pathfinder algorithm is frequently used because it is able to conserve the triangle inequalities among paths with any number of links, model asymmetric relationships, and represent the most salient relationships in the data [28], [29]. This means that the algorithm prioritizes links with the greatest weights and paths with lower weights are trimmed from the network. However, this algorithm has a long run-time. The MST-Pathfinder algorithm uses a Minimum Spanning Tree approach (a greedy algorithm) [30] coupled with the traditional Pathfinder algorithm to address the limitation. A more detailed description of this algorithm and its function can be found in Quirin and colleagues' paper [31]. I used an MST-Pathfinder scaling algorithm to trim the network to the essential structure.

Finally, the network was analyzed to determine the graph descriptive statistics. Network statistics were run on the resulting model using the Network Analysis Toolkit in Sci2 Tool. This function performs basic analysis on a network. The toolkit calculates the number of weak component clusters, strong component clusters, self-loops, parallel edges, whether the network appears to be directed or undirected, the attributes

present on both nodes and edges, number of nodes, number of edges, and the density of the network. All of the descriptive statistics gives information about the composition of the co-occurrence network. The resulting undirected network had 1266 nodes and 1257 unique edges with no self-loops.

D. Using NodeXL to Visualize Co-occurrence Network

The resulting scaled network was exported from Sci2 Tool as a comma-separated values file and imported into NodeXL for analysis and visualization. NodeXL is a free add-on package for Microsoft Excel and is useful for simple network visualizations. In order to understand the thematic clusters within the network [32], a Clauset, Newman, and Moore (CNM) clustering algorithm was applied to the network. This method is a bottom-up agglomerative clustering which continuously finds and merges pairs of clusters trying to maximize modularity of the community structure in a greedy manner [33]. Modularity is a property of a network and a specific proposed division of that network into communities or clusters [33]. It measures when the division between clusters is good by examining when there are few edges between clusters but many edges within clusters. The number ranges from -0.5 to 1, and a positive number indicates that the number of edges within a group exceeds the number based on chance. For this network, the modularity was 0.93 which indicates that the clusters are densely connected within clusters but sparsely connected between clusters. This algorithm reveals large-scale patterns present in the co-occurrence network that can be used to interpret the network. This clustering technique allowed closely connected words (i.e., close "neighbors") to be grouped into "neighborhoods" to simplify common themes and understand how general topics in the literature are connected to one another.

After clusters were identified using the CNM algorithm, the clusters were coded based on the words associated with each node in the cluster. This process was similar to thematic coding in qualitative research and was used to create meaning about each cluster within the network [34]. First, each node was examined within a cluster to understand which words were associated with one another. Next, each cluster was assigned a word or phrase that identified what the unit of analysis, each cluster was about and/or what it meant. This part of the analysis was based on researcher interpretation of each cluster.

The network was laid out using a Harel-Koren Fast Multiscale format. This format is appropriate for drawing undirected graphs with straight-line edges [35]. This is one of NodeXL's two force graphs that makes the edges to appear about the same length and to minimize line crossings, which makes for a more readable graph. This approach was chosen to most easily show a two-dimensional representation of the resulting network, create an image with interpretable clusters, and highlight the hole in the literature found in this analysis (see Fig. 1).

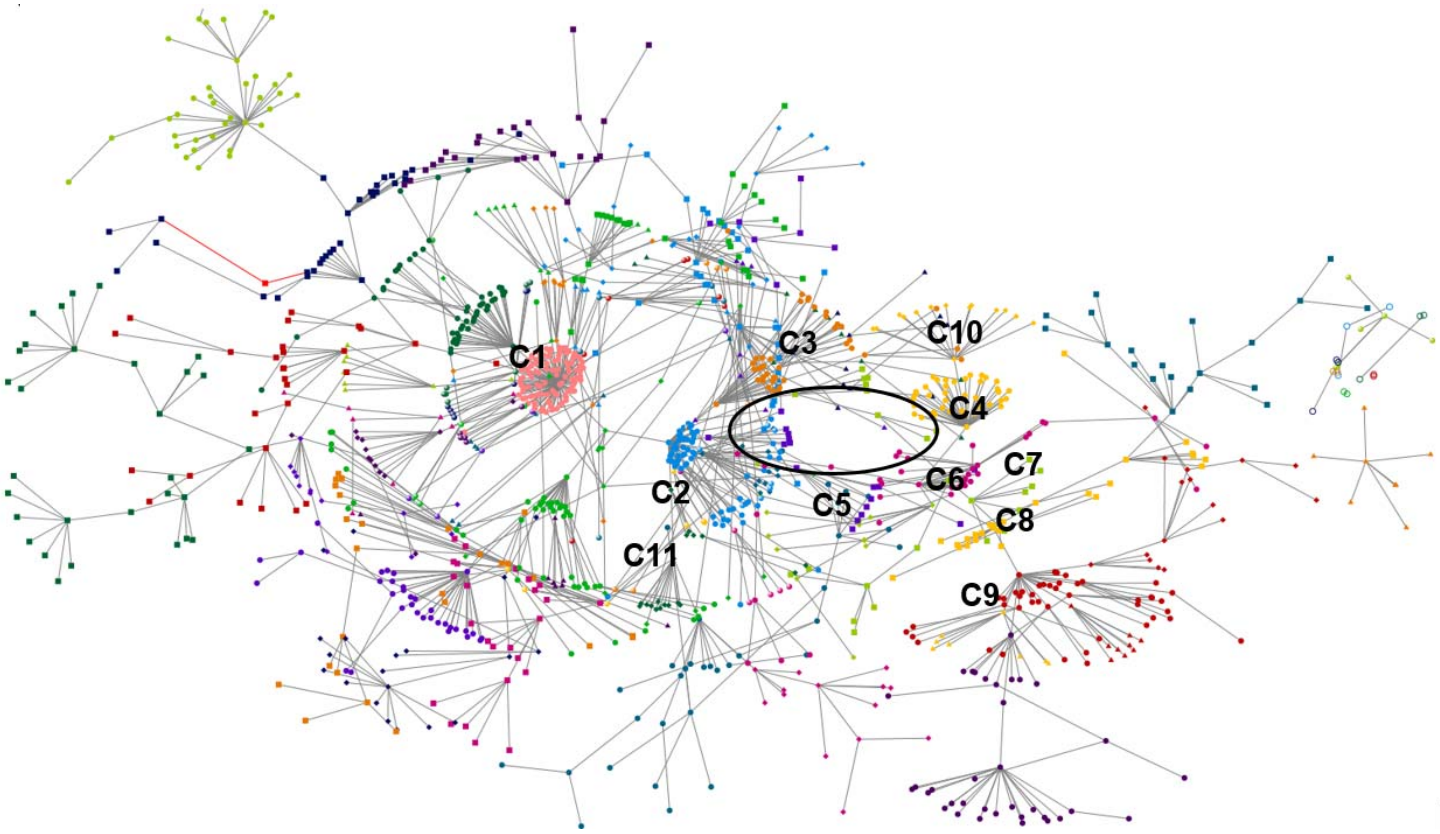

Fig. 1. Network map of word co-occurrence in published identity literature from 1995-2015. Important clusters are labeled in the figure: C1 = K-12; C2 = Identity; C3 = Diversity; C4 = Invisible; C5 = Science and Engineering; C6 = Engineering; C7 = Gender; C8 = College Instruction; C9 = Engineering Education; C10 = Work Environment; C11 = Race/Ethnicity.

VI. Results and Discussion

Fig. 1 shows the resulting topical network with each cluster as a different color and node shape. Each node represents common words in titles, keywords, abstracts (or summaries) of the journal papers, books, and reports. The edges show how each of the words is connected to one another throughout the literature. Important clusters (identified through the CNM algorithm and coded for themes) are labeled in Fig. 1 to illustrate the connections between the concepts of diversity, gender, identity, engineering, and engineering education in the last two decades. The graph shows that while there has been some research on ideas surrounding less visible diversity categories (1.5% of papers found), this specific topic has not been well-researched. This hole in the current literature is marked in Fig. 1 with an oval.

This network shows a divide at the center with the left side of the Fig. 1 stemming from connections with K-12 literature (C1). On the other side, much of the work stems from the central node of identity (C2) to topics in higher education. A large cluster on diversity topics is closely tied to identity (C3). This connection is not surprising as identity has been used as a theoretical framework to understand how students see themselves as engineers and highlight ways to support diverse students in developing these identities. The cluster coded as Race/Ethnicity (C11) is also tied closely with identity. Gender and gender identity appeared as topics throughout multiple sections of the network with a larger cluster labeled as Gender (C7) that encompasses studies solely focused on gender as a topic. The presence of gender across multiple clusters is not surprising considering that many studies in engineering education include gender differences as an additional research question within larger studies. Other similarly themed clusters branch off from these connections including the fields of study of Science and Engineering (C5), Engineering (C6), College Instruction (C8), and Engineering Education (C9). One interesting structural aspect is the connection of the Engineering Education (C9) cluster to the general cluster of Engineering (C6) through the cluster of College Instruction (C8).

Other smaller and less connected clusters were found for the theme Invisible (C4) and Work Environment (C10). The Invisible cluster included ideas of diversity that are not readily visible including disability status, sexual orientation, and gender identity rather than gender expression. It is interesting that this cluster was closely connected to the cluster of Work Environment. This connection may tie the culture of the engineering workplace with concepts of "covering" or "passing" to fit in. These terms describe behaviors of stigmatized groups to manage how they present their identities to others. In fact, three nodes within the cluster directly address the intersections of social identities and invisible diversity for LGBTQ students passing or covering in engineering [36].

This network also highlights a gap in the current identity literature on research related to less visible diversity as shown by the oval in Fig. 1. The concept of "invisible diversity" has been applied to invisible *social* identities of sexual orientation and disability status rather than mindsets and attitudes [36]. However, little to no studies examine the students' affective and cognitive profiles as aspects of diversity. These aspects are one reason why we desire to have students with a variety of lived experiences in engineering. Paul Eremenko, founding CEO of Airbus Group Silicon Valley technology and business innovation center and former director at Google, captures this idea well,

> It strikes me that there are two families or reasons why [we need diversity] …One is social justice. That there should be representation commensurate with the representation of everyone in our society at all levels, including engineering. A different one is that we believe that diversity improves the quality of innovation. But it is possible that those aren't congruent - that there are groups that we want represented for social justice reasons that have nothing to do with the quality of innovation or those outcomes. It's entirely possible, right? But, I don't have data either way…I think you might do different things depending on which of those are the end game or purpose. So particularly if you care about ideas, the focus, it strikes me, should be on output-centric measures…The way that we segment for diversity, gender, race, etc., the traditional ways of segmenting, are not the right ways to do it [understand diversity in engineering] [37].

Innovation is the key to economic growth and prosperity, and engineering is a critical driver in industrial innovation [38]. Many companies are discovering that increased and more diverse approaches to problem solutions contribute to product innovation, global competence, and other successful corporate outcomes [39]–[42]. However, engineering persistently lacks the diverse mindsets and ways of thinking needed to solve complex problems facing our world [43], [44].

Much of the research based on innovation has operated under a key assumption that external markers of diversity (e.g., age, race, gender expression, etc.) will automatically increase the diversity of engineering solutions. The literature shows inconsistent and mixed findings for this assumption. Some research shows that teams including more variety in diversity indicators like age, race, and gender do not show improved innovation [45], [46] while other research has found that minority dissent that actively challenges the basis for engineering decisions can improve innovation [47], [48]. These findings suggest that diversity in approaches, problem-solving, and ways of thinking improves innovation in engineering design more reliably than does diversity along the lines of age, race, gender, etc. However, the process of enculturating students into engineering through engineering curriculum often creates homogeneity in students' approaches to problems [49], [50], ways of thinking [51], and attitudes [52]. This homogenization reduces variability in students' innovation [53] and can create a mismatch between how students perceive engineering as a field and how they perceive themselves as engineers, often resulting lack of belonging and ultimately, attrition [53]–[55]. As a result, a gap of understanding how to develop students with diverse and innovative mindsets in engineering education remains.

The visual representation of a systematic literature review presented in this paper was the basis for a recently funded grant

titled, "CAREER: Actualizing Latent Diversity: Building Innovation through Engineering Students' Identity Development," to explore students' diverse attitudes and mindsets in engineering, termed *latent diversity*. These latent attributes are present, but are not visible or actualized, and have the capacity to become or develop into opportunities for innovation in the future.

VII. Limitations and Future Work

One limitation of this method is the work of the researcher to make meaning out of the numerous resulting clusters. Other more advanced linguistic analysis techniques could be used to automate this process. For example, Latent Semantic Analysis is a method that uses high-dimensional linear models to analyze large bodies of text and create a representation that captures the similarities of word and text passages [56]. This approach models the relationships among documents based on word frequency and usage within and across text contexts (e.g., documents, paragraphs, sentences, and phrases) [57]. The underlying premise is similar to the analysis conducted in this paper but more rigorously represents how individual units of analysis (i.e., words or phrases) have meaning within the texts. Latent semantic analysis does not use researcher-constructed dictionaries, knowledge bases, semantic networks, grammar, syntactic parsers, or morphologies. Rather, it takes only the raw input text parsed into individuals words or strings and separated into meaningful passages like paragraphs, sentences, etc. [58]. This approach approximates human language processing is a powerful way to understand relationships among text. However, latent semantic analysis does have some drawbacks. This method does not use word order, syntactic relations or logic, to understand the meaning within text. It only uses differences in word usage and frequency to understand how passages of text are related to one another. This approach could offer another way to understand large bodies of text from a systematic literature review using big data techniques.

Future work includes examining different approaches to synthesizing systematic literature reviews through linguistic analysis to understand the current body of literature in a new, highly visual way. This approach does not replace the current synthesis of a systematic literature review which gleans meaning of the current literature and the findings of that work from an expert understanding. However, this method can be used for gap analysis and a preliminary understanding of how the topics in the body of literature are related to one another.

VIII. Implications

This paper offers a step-by-step process for implementing linguistic analysis methods commonly employed in other areas of research into engineering education. This paper presents the application of linguistic analysis to create visualizations of systematic literature reviews both methodologically as well as in an applied example of identity and diversity in science and engineering education. This approach uses current tools for exploring the science of science, Sci2 Tool, and visualizing networks, NodeXL, that are readily available to the community. Use of widely available tools provides accessible and replicable methods for use by the engineering education community. This paper offers a new way to visualize systematic literature reviews as interconnected networks of word use and commonalities among published work rather than tabular lists of papers commonly seen with systematic reviews. This approach is not a replacement but rather an augmentation of current systematic literature review practices.

IX. Conclusions

This paper presents a first step in using networks to visualize connections between data gathered in a rigorous and systematic method and can be used in addition to traditional approaches in systematic literature reviews. The process for this approach is presented in and example understanding how identity and diversity are framed in the current science and engineering education literature. A large gap in understanding nonvisible aspects of diversity in identity development was highlighted and a new concept, *latent diversity*, was presented to fill this identified gap. Use of a visual network to understand gaps in current literature can provide new and quick ways to promote research in under-researched areas as well as highlight areas with significant published work.

Acknowledgment

The author thanks Krishna Madhavan for his expertise and advice in developing this approach to understanding the current literature. She also thanks Thomas Godwin for his help in cleaning data.

References

[1] M. Borrego, M. J. Foster, and J. E. Froyd, "Systematic Literature Reviews in Engineering Education and Other Developing Interdisciplinary Fields," *J. Eng. Educ.*, vol. 103, no. 1, pp. 45–76, 2014.

[2] U. Connor and A. Mauranen, "Linguistic analysis of grant proposals: European Union research grants," *English Specif. Purp.*, vol. 18, no. 1, pp. 47–62, 1999.

[3] P. Kenis, V. Schneider, and others, "Policy networks and policy analysis: scrutinizing a new analytical toolbox," *Policy networks Empir. Evid. Theor. considerations*, pp. 25–59, 1991.

[4] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "Science mapping software tools: Review, analysis, and cooperative study among tools," *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 7, pp. 1382–1402, 2011.

[5] M. Petticrew and H. Roberts, *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons, 2008.

[6] T. Meline, "Selecting studies for systematic review: Inclusion and exclusion criteria," *Contemp. Issues Commun. Sci. Disord.*, vol. 33, no. 21–27, 2006.

[7] A. Jocuns, R. Stevens, L. Garrison, and D. Amos, "Students' changing images of engineering and engineers," in *American Society for Engineering Education Annual Conference & Exposition*, 2008, p. 28.

[8] R. Stevens, K. O'Connor, L. Garrison, A. Jocuns, and D. M. Amos, "Becoming an engineer: Toward a three dimensional view of engineering learning," *J. Eng. Educ.*, vol. 97, no. 3, pp. 355–368, 2008.

[9] R. Stevens, K. O'Connor, and L. Garrison, "Engineering student identities in the navigation of the undergraduate curriculum," in *Association of the Society of Engineering Education Annual Conference*, 2005, p. 8.

[10] H. H. M. Matusovich, R. R. A. Streveler, and R. L. R. Miller, "Why do students choose engineering? A qualitative, longitudinal investigation of students' motivational values," *J. Eng. Educ.*, vol. 99, no. 4, pp. 289–304, 2010.

[11] C. E. Foor, S. E. Walden, and D. A. Trytten, "'I Wish that I Belonged

More in this Whole Engineering Group:' Achieving Individual Diversity," *J. Eng. Educ.*, vol. 96, no. 2, pp. 103–115, Apr. 2007.

[12] A. Godwin and G. Potvin, "Fostering female belongingness in engineering through the lens of critical engineering agency," *Int. J. Eng. Educ.*, vol. 31, no. 4, pp. 938–952, 2015.

[13] A. Godwin, G. Potvin, Z. Hazari, and R. Lock, "Identity, Critical Agency, and Engineering: An Affective Model for Predicting Engineering as a Career Choice," *J. Eng. Educ.*, p. In Press, 2016.

[14] G. Downey and J. Lucena, "When students resist: Ethnography of a senior design experience in engineering education," *Int. J. Eng. Educ.*, vol. 19, no. 1, pp. 168–176, 2003.

[15] K. L. Tonso, "Student Engineers and Engineer Identity: Campus Engineer Identities as Figured World," *Cult. Stud. Sci. Educ.*, vol. 1, no. 2, pp. 273–307, Jun. 2006.

[16] K. L. Tonso, "Teams that work: Campus culture, engineer identity, and social interactions," *J. Eng. Educ.*, vol. 95, no. 1, pp. 25–37, 2006.

[17] *Identity Construction and Science Education Research: Learning, Teaching, and Being in Multiple Contexts*. M. Varelas, Ed. Rotterdam: Sense Publishers, 2012.

[18] A. Johnson, J. Brown, H. Carlone, and A. K. Cuevas, "Authoring identity amidst the treacherous terrain of science: A multiracial feminist examination of the journeys of three women of color in science," *J. Res. Sci. Teach.*, vol. 48, no. 4, pp. 339–366, Apr. 2011.

[19] E. D. Tate and M. C. Linn, "How Does Identity Shape the Experiences of Women of Color Engineering Students ?," *J. Sci. Educ. Technol.*, vol. 14, no. 5, pp. 483–493, 2005.

[20] A. Calabrese Barton and E. Tan, "We Be Burnin'! Agency, Identity, and Science Learning," *J. Learn. Sci.*, vol. 19, no. 2, pp. 187–229, Apr. 2010.

[21] D. Verdín, A. Godwin, and J. L. Morazes, "Qualitative Study of First-Generation Latinas: Understanding their Motivation for Seeking an Engineering Degree," in *American Society for Engineering Education Annual Conference & Exposition*, 2015, p. 19.

[22] M. Ross, "Stories of Black Women in Industry – Why they Leave," in *Frontiers in Education Conference (FIE)*, 2015, p. 5.

[23] K. Madhavan, L. Zentner, V. Farnsworth, S. Shivarajapura, M. G. Zentner, N. Denny, and G. Klimeck, "NanoHUB. Org: Cloud-based Services For Nanoscale Modeling, Simulation, And Education," *Nanotechnology Reviews*, vol. 2, no. 1. pp. 107–117, 2013.

[24] X. Chen, N. Sambamurthy, C. M. Schimpf, H. Xian, and K. Madhavan, "Weighted Social Tagging as a Research Methodology for Determining Systemic Trends in Engineering Education Research," in *American Society for Engineering Education Annual Conference & Exposition*, 2011, p. 20.

[25] A. Johri, G. A. Wang, X. Liu, and K. Madhavan, "Utilizing Topic Modeling Techniques to Identify the Emergence and Growth of Research Topics in Engineering Education," in *Frontiers in Education Conference (FIE)*, 2011, p. T2F–1.

[26] SAS Institute Inc., "Base SAS® 9.4 Procedures Guide." Cary, NC, 2011.

[27] Sci2 Team, "Science of Science (Sci2) Tool." Indiana University and SciTech Strategies, 2009.

[28] D. W. Dearholt and R. W. Schvaneveldt, "Properties of Pathfinder networks," in *Pathfinder associative networks: Studies in knowledge organization*. R. W. Schvaneveldt, Ed. Norwood NJ: Ablex Publishing Corp., 1990, pp. 1-30.

[29] D. D. Reese, "PFNET translation: A tool for concept map quantification," *2003 Annu. Proceedings-Anaheim Vol. 2*, p. 314, 2003.

[30] J. B. Kruskal, "On the shortest spanning subtree of a graph and the travelling salesman problem," *Proc. of the Amer. Math. Soc*, 1956, vol. 2, pp. 48–50.

[31] A. Quirin, O. Cordón, V. P. Guerrero-Bote, B. Vargas-Quesada, and F. Moya-Anegón, "A quick MST-based algorithm to obtain Pathfinder networks (8, n- 1)," *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, no. 12, pp. 1912–1924, 2008.

[32] Smith, M., N. Milic-Frayling, B. Shneiderman, E. Mendes Rodrigues, J. Leskovec, and C. Dunne, "NodeXL: A Free and Open Network Overview, Discovery and Exploration Add-in for Excel 2007/2010." Social Media Research Foundation, 2010.

[33] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, no. 6, p. 66111, 2004.

[34] J. Saldana, *The coding manual for qualitative researchers*, 2nd ed. Thousand Oaks, CA: Sage Publications, 2013.

[35] D. Harel and Y. Koren, "A fast multi-scale method for drawing large graphs," *J. Graph Algorithms Appl.*, vol. 6, no. 3, pp. 179–202, 2002.

[36] E. A. Cech and T. J. Waidzunas, "Navigating the heteronormativity of engineering: The experiences of lesbian, gay, and bisexual students," *Eng. Stud.*, vol. 3, no. 1, pp. 1–24, 2011.

[37] P. Eremenko. Informal Meeting, Topic: "Lifelong Engineering Access & Diversity for Broadening Participation." Purdue University, West Lafayette, IN. Oct. 7, 2014.

[38] M. C. Thursby, "The Importance of Engineering: Education, Employment, and Innovation," *Bridge*, vol. 44, no. 3, pp. 5–10.

[39] National Science Board, *The Science and Engineering Workforce: Realizing America's Potential*. 2003.

[40] D. E. Chubin and E. L. Babco, "Diversifying the engineering workforce," *J. Eng. Educ.*, vol. 94, no. 1, pp. 73–86, 2005.

[41] J. L. Keith, D. B. Ayer, E. Rees, D. V Freda, J. K. Lowe, and J. Day, "Brief of Amici Curiae Massachusetts Institute of Technology, Leland Stanford Junior University, EI Du Pont de Nemours and Company, International Business Machines Corp., National Academy of Sciences, National Academy of Engineering, National Action Council for Minorities in Engineering, Inc., in Support of Respondents in the Supreme Court of the United States (Grutter v. Bollinger and Gratz v. Bollinger)," *Grutter v. Bollinger, al*, no. 02–241, 2003.

[42] W. A. Wulf, "Diversity in Engineering," *Bridg.*, vol. 24, no. 4, 1998.

[43] Committee on Prospering in the Global Economy of the 21st Century:An Agenda for American Science and Technology, *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future*. National Academies Press, 2007.

[44] NAE, *Educating the Engineer of 2020: Adapting Engineering Education to the New Century*. Washington, D.C.: The National Academies Press, 2005.

[45] S. H. Cady and J. Valentine, "Team Innovation and Perceptions of Consideration What Difference Does Diversity Make?," *Small Gr. Res.*, vol. 30, no. 6, pp. 730–750, 1999.

[46] N. D. Fila, R. E. H. Wertz, and S. Purzer, "Does diversity in novice teams lead to greater innovation?," in *Frontiers in Education Conference (FIE), 2011*, 2011, p. S3H–1.

[47] C. K. W. De Dreu and M. A. West, "Minority dissent and team innovation: the importance of participation in decision making," *J. Appl. Psychol.*, vol. 86, no. 6, pp. 1191–1201, 2001.

[48] P. L. McLeod, R. S. Baron, M. W. Marti, and K. Yoon, "The eyes have it: Minority influence in face-to-face and computer-mediated group discussion," *J. Appl. Psychol.*, vol. 82, no. 5, pp. 706–718, 1997.

[49] L. L. Bucciarelli and S. Kuhn, "Engineering Education and Engineering Practice: Improving the Fit," in *Between Craft and Science: Technical Work in U.S. Settings*, S. R. Barley and J. E. Orr, Eds. Cornell University Press, 1997, pp. 210–229.

[50] P. M. Leonardi, M. H. Jackson, and A. Diwan, "The Enactment-Externalization Dialectic: Rationalization and the Persistence of Counterproductive Technology Design Practices in Student Engineering," *Acad. Manag. J.*, vol. 52, no. 2, pp. 400–420, 2009.

[51] T. Becher and P. Trowler, *Academic tribes and territories: Intellectual enquiry and the culture of disciplines*. McGraw-Hill Education (UK), 2001.

[52] K. J. B. Anderson, S. S. Courter, T. McGlamery, T. M. Nathans-Kelly, and C. G. Nicometo, "Understanding engineering work and identity: a cross-case analysis of engineers within six firms," *Eng. Stud.*, vol. 2, no. 3, pp. 153–174, 2010.

[53] M. Lumsdaine and E. Lumsdaine, "Thinking preferences of engineering students: Implications for curriculum restructuring," *J. Eng. Educ.*, vol. 84, no. 2, pp. 193–204, 1995.

[54] E. Seymour and N. Hewitt, *Talking about Leaving: Why*

Undergraduates Leave the Sciences. Boulder, CO: Westview Press, 1997.

[55] B. N. Geisinger and D. Raman, "Why they leave: Understanding student attrition from engineering majors," *Int. J. Eng. Educ.*, vol. 29, no. 4, pp. 914–925, 2013.

[56] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman, "Using latent semantic analysis to improve access to textual information," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1988, pp. 281–285.

[57] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Process.*, vol. 25, no. 2–3, pp. 259–284, 1998.

[58] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.