



SUMMER 2014

## Tools for large-scale data analytic examination of relational and epistemic networks in engineering education

KRISHNA MADHAVAN  
Purdue University  
West Lafayette, Indiana

ADITYA JOHRI  
George Mason University  
Fairfax, VA

HANJUN XIAN  
Microsoft Corporation  
Bellevue, WA

G. ALAN WANG  
Virginia Tech  
Blacksburg, VA

AND

XIAOMO LIU  
Thomson Reuters R&D  
New York, NY

### ABSTRACT

The proliferation of digital information technologies and related infrastructure has given rise to novel ways of capturing, storing and analyzing data. In this paper, we describe the research and development of an information system called Interactive Knowledge Networks for Engineering Education Research (iKNEER). This system utilizes a framework that combines large-scale data mining techniques, social network mapping algorithms, and time-series analysis, to provide a mechanism for analyzing and understanding data about the engineering education community. We provide a detailed description of the algorithms, workflows, and the technical architecture we use to make sense of publications, conference proceedings, funding information, and a range of products derived from research in EER (also known as knowledge products). Finally, we demonstrate one possible application of iKNEER by applying topic modeling techniques to a subset of the data to identify the emergence and growth of research topics within the community thereby illustrating the unique epistemic value of this knowledge platform. The system can be found at <http://www.ikneer.org>.



**Key Words:** Interactive knowledge platform, academic collaboration, research trends, big data

## INTRODUCTION

In today's globally competitive economy, success is increasingly driven by the ability to act on knowledge and intellectual capital. Academic communities that have developed a corpus of knowledge artifacts over decades or sometimes centuries of research are uniquely positioned to capitalize on their expansive knowledge bases. Yet, this process is fraught with difficulties. To be innovative, an organization [or community] has to be adept at exploiting existing knowledge as well as exploring new ways of producing knowledge [1]. To do so, a community must have a holistic, deep, and accessible understanding of what it knows.

Engineering education has recently undergone a resurgence and reorientation that mirrors growing recognition of the challenges faced by both engineering educators and learners in the 21<sup>st</sup> century. A new field of Engineering Education Research (EER) has emerged, in part coalescing around theories of how people learn in the domain of engineering. Yet as the EER community expands, it is becoming increasingly difficult to develop and sustain *community memory*. This has the potential to significantly hinder progress as the inability of a field, discipline, or more generally – a problem space – to recognize what it knows increases the risk that isolated researchers and groups will tackle similar problems using relatively similar approaches. Lack of easy access to community knowledge also hinders the acculturation of newcomers into the community. The expansion of engineering education over the past decade has led the field to a critical juncture that demands new tools and methods to enable the community to expand and build on prior work. In this paper, we address this challenge by describing the development and deployment of an interactive knowledge platform – entitled Interactive Knowledge Networks for Engineering Education Research (iKNEER). This system is designed with the goal to help members of this growing community explore the current state of knowledge within EER, identify future directions for research, and find collaborative partners.

The engineering education community has a vision of improving and innovating how engineers are trained and prepared to make them more competitive in the global economy. To pursue this goal the community has coalesced around several initiatives such as those that have produced *The Engineer of 2020* [2] and the draft report on *Engineering Education for the Global Economy* [3]. The National Science Board report entitled *Moving Forward to Improve Engineering Education* [4] explicitly points to the need for “expanding research and data collection related to engineering education”. Inherent in this call is the need to expand the rate and volume of data collected and analyzed in EER. Yet the question remains: once such large scales of data are collected, what sorts



of analytics and informatics can be applied to them to derive *actionable knowledge*? This actionable knowledge could take the form of new policies, organizational structures, student learning practices and others that may have a direct impact on not only how we conduct research in EER, but also how it impacts student learning. Prior efforts and reports provide us with a blueprint of where the community needs to head, especially in terms of supporting desired outcomes for engineers who are prepared to practice effectively in the 21<sup>st</sup> century. Yet we do not have specific pathways and broad knowledge of how engineering education research can help transform engineering teaching and learning to cultivate engineers of the future. This challenge is further compounded when one considers the international and highly interdisciplinary state of this problem space, with researchers in many different countries and regions perhaps undertaking similar research on engineering education and professional practice.

Extant literature in engineering education [5] and numerous other disciplines including learning sciences [6] and cyberinfrastructure [7, 8] have called for radically rethinking education research to include large scale data and collaborations. Throughout this paper, we use the term large-scale data to indicate data from a variety of contexts - research, learning, policy, and funding - that is growing in volume at high velocity. For example, there are new research papers appearing every day; each time a student interacts with learning materials within learning environments it produces archival data; each time a funding decision occurs at any funding agency (such as the US National Science Foundation) this produces data. By large data or big data - we mean an accumulation of all of this data that happens daily, growing in volume, and most importantly capable of expanding at very high speeds.

One important question for growing fields such as engineering education is: How do we know when large-scale research collaborations are happening? Also, how do we know that research utilizing large datasets attracts a large number of researchers to utilize these datasets? Can we take a data-driven approach to clearly point out trends in research productivity and collaboration? Information retrieval research (e.g., search engines) often helps address such problems by improving the aggregation of data and focusing on what any given document is about (i.e., word-level content analysis). However, for scientific communication, it is equally important to know who writes the document and how the document is positioned in the process of knowledge emergence as this provides credibility and grounding for the research outcomes. Improving access to such information demands different types of analytic tools.

Traditionally, analyzing large-scale academic data has been the domain of a few computer scientists and engineers. It requires computational techniques to acquire and manage data, analyze large-scale networks, and identify trends and patterns. To allow a broader range of researchers, educators, and other stakeholders in the EER community to drive the exploration of the problem



space, the data gateway must not only handle the underlying computational components, but also provide insightful representations and a user-friendly interface. In this paper, we describe an infrastructure and associated problems of infrastructure construction that can eventually allow us to characterize and provide the type of insights required by the community. To achieve this goal we utilize large-scale data from a variety of knowledge products. By knowledge products we mean publications in journals and conferences in engineering education, National Science Foundation (NSF) award information, articles published by international organizations, reports on engineering education, and a range of other such academic artifacts.

The purpose of this paper is to describe a system called Interactive Knowledge Networks for Engineering Education Research (iKNEER) – a prototype of a knowledge mining and visualization gateway designed primarily for the engineering education research community. As of April 15, 2013, iKNEER includes a total of 164,213 documents from 22 different publications including Journal of Engineering Education, International Journal of Engineering Education, Frontiers in Education, ASEE conference proceedings, IEEE Transactions on Education, and the US National Science Foundation. Overall, our system covers almost 50 years of data with the oldest document dating back to 1963. The document repository is continuously expanding to reach broader inclusion of publication sources and longer coverage. Even as we increase coverage, we are also particularly careful to avoid losing focus with respect to the data elements we provide. There are multiple perspectives on which academic outlets count within the engineering education space. Our strategy allows members of the community to recommend specific knowledge products we need to archive. We track these requests and if there are multiple requests for archiving a specific data source – our system then crawls the metadata related to that source. Human experts monitor this process, thereby keeping it focused on engineering education. Out of the whole dataset, we have developed the capability to understand the scientific profile of 137,039 authors and 262,670 keywords (including 197,216 keyphrases extracted from document full text automatically).

#### **CURRENT APPROACHES TO CHARACTERIZING A RESEARCH DOMAIN USING DATA AND VISUALIZATION**

Analyzing academic publications to characterize a research domain, usually involves a manual, in-depth analysis of relevant papers. The aim is to synthesize literature related to popular research topics, recognize patterns and trends, and chart a literature roadmap for future development. In engineering education, researchers have produced review papers that cover a variety of research topics. The fundamental problem with traditional review papers is that they are extremely time and



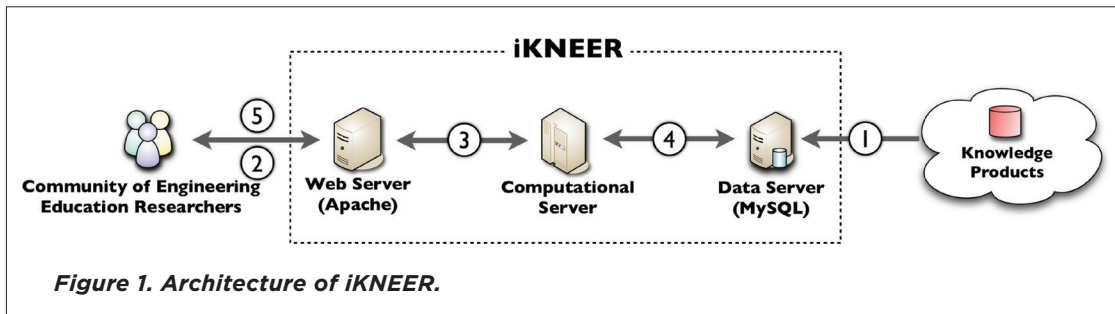
resource intensive to produce. Therefore, it is difficult to conduct these studies on an on-going basis. For example, some researchers attempt to identify main research areas in engineering education based on a review of a relative broad range of journal and conference proceedings papers. By conducting a keyword analysis of Journal of Engineering Education (JEE) papers from 1993 to 2002, Wankat [9, 10] identified that major research topics such as *teaching*, *design*, *computer*, *ABET*, and *assessment* became increasingly popular during the second half of the last decade. Chen et al. [11] also reviewed JEE papers to recognize core topics in engineering education wherein the results were based on a social tagging activity. Jesiek et al. [12, 13] performed a review of international conference papers in engineering education and revealed different patterns in topics and collaboration within and across nations. Instead of studying engineering education as a whole to recognize topical trends, some projects have focused on specific research areas. Madhavan et al. [14] provided a synthesis of *cyberlearning environments* in engineering education based on a qualitative analysis of JEE articles from years 2000 to 2009. Other similar studies examining specific topics in engineering education include Prince [15] who evaluated the effectiveness of *active learning* by synthesizing relevant literature and Dutson et al. [16] who investigated the topic of *teaching engineering design*. Review papers help other researchers quickly develop comprehensive and insightful understanding of a discipline or a research topic. However, the effort behind writing a review paper is extremely high. While such studies are valuable and provide rich insights, they do not provide mechanisms to capture the zeitgeist of the research as it occurs – primarily because it takes significant time and effort to manually capture topical and other trends. Just the number and frequency of publication of significant review papers that include the type of analyses discussed above is evidence that the process of collecting enough papers and then subsequently performing in-depth analyses is a difficult and resource intensive process. Further, it is infeasible to review every topic on a regular basis based on all the relevant literature. Instead, authors of review papers usually selectively cover a small set of top publications [17].

As performing trending and high-level topical analyses manually is too costly, many scholars seek a more automatic solution to analyzing academic articles. Frameworks and tools have been developed to identify significant trends and patterns based on publication metadata such as titles, authors, abstracts, keywords, affiliations, citations, and other attributes. For instance, some studies identified prominent scholars in a specific area [18] and characterized main research foci and trends [19, 20] by performing co-citation analyses. Based on statistical analysis of keywords/keyphrases of each paper, researchers revealed trends and patterns that chart the emergence and development of a field. Some scholars [21] working on topic modeling studied author-topic models for academic publications to link a research topic to its contributing authors and produce author profiles based on authors' academic production. Other researchers have tracked the longitudinal evolution of topics



in a specific domain [22] and the development of scholarly collaboration networks among authors [23]. All these studies uncovered trends and patterns based on a statistical analysis of large-scale bibliographic data. With iKNEER, our goal is to provide the EER community with an infrastructure that makes such studies easy to undertake repeatedly and on-demand.

To provide more interactions and meaningful representations to any research community, visual analytics experts have implemented tools to help understand individual academic articles as well as the entire field. Uren et al. [24] developed a visual tool named *ClaiMapper* to allow users to sketch a map of arguments to summarize main ideas in individual papers. They defined a taxonomy of rhetorical link types, which were denoted by edges on the argument map. Strobel et al. [25] presented a technique called Document Cards to display a brief summary of any article in a conference proceedings. The document summary was automatically generated with key terms and figures and could direct readers to the original context in the full text upon readers' request. Besides efforts on visualizing individual articles, researchers have proposed information visualizations to demonstrate paradigm shifts, topical trends, and author-topic relationships. McCain [18] performed author co-citation analysis on publications in the ISI databases and represented domains and top authors as clusters in a network. Similarly, White et al. [26] analyzed the co-citation relationships based on journal papers in information science and clustered authors by their research specialties. He et al. [27] also visualized clusters of authors based on co-citation relationships but developed an additional web-based search engine for querying the citations database. Chen et al. [19] proposed a 3-D visualization tool for exploring the semantic relationships between articles in the ACM Hypertext conference. The software SCIMap [28] also aimed to generate co-citation maps and ontology based on scientific publications in the natural sciences [29]. Another popular visual analytics tool for research domain analysis called VxInsight [30] used a terrain view to demonstrate the popularity of topics and the commonality between them. A recent study by Bergström et al. [31] combined existing visualization techniques such as tree map view, force-directed network, and circular network to develop an intelligence system, PaperCube, to facilitate researchers' interaction with a digital library and exploration of different dimensions of bibliographic metadata. Börner et al. [32] summarized present knowledge domain visualizations and proposed guidelines for appropriate use of these techniques to analyze bibliographic data in different contexts. These visualization tools address various aspects of bibliographic data and offer users interactive interfaces to navigate any problem space. However, none of the above approaches has been applied to provide an insightful and comprehensive overview of engineering education research. More importantly, the methods, tools, and algorithms identified above require significant expertise in data mining to use on a regular basis. They represent a set of approaches that need to be integrated into a single unifying analytical framework and workflow. iKNEER is designed for non-experts in data mining



and visual analytics. It brings together a variety of algorithmic techniques into a single problem space – namely, EER.

## METHODOLOGY

Figure 1 illustrates the architecture and workflow of iKNEER. As a data-intensive gateway, iKNEER first (1) collects knowledge products such as academic articles and grant proposals from a variety of sources periodically using focused crawling. As a cyber-tool for researchers to explore the field, the web-based interface of iKNEER (2) processes user operations on the website, which then (3) triggers the underlying computational components to (4) compute the output. The result is then (5) represented in a visual form. In this section, we present our design and implementation of iKNEER by elaborating on the three major components: data management, computation, and representation.

### Data acquisition and management

iKNEER's ultimate goal is to archive every knowledge product in engineering education. To achieve this goal, the data server acquires metadata and full texts (when feasible) of academic articles relevant to engineering education from online publication data sources such as IEEE Xplore, Web of Science, and EBSCO. A full listing of all archived data is provided in Table 1. Our data acquisition process focuses on data sources that have open metadata access. The list of relevant publications is derived from the feedback from a large number of users in the community and is constantly expanding to include a broader literature base. As explained earlier, even though the number of products we archive is increasing, the active involvement and curation of data sources by humans allows the system to stay focused on engineering education.

To keep our database constantly updated, we automated the acquisition process by detecting updates from monitored sites periodically. Once new issues and volumes are published, the detectors will inform iKNEER's crawlers to download the relevant data. Occasionally, we import



Knowledge Product	Number of Documents	Available Years
<i>Advances in Engineering Education</i>	62	2007–2012
<i>American Society for Engineering Education (annual conference)</i>	17,938	1996–2012
<i>Australasian Journal of Engineering Education</i>	50	2000–2009
<i>Computer Applications in Engineering Education</i>	505	1997–2011
<i>Engineering Education</i>	2,292	1969–1991
<i>European Journal of Engineering Education</i>	1,953	1975–2011
<i>Frontiers in Education (conference)</i>	7,082	1982–2010
<i>Global Journal of Engineering Education</i>	322	2000–2007
<i>IEEE Transactions on Education</i>	3,331	1963–2011
<i>International Journal of Continuing Engineering Education and Life-Long Learning</i>	185	2004–2009
<i>International Journal of Electrical Engineering Education</i>	1,491	1969–2010
<i>International Journal of Engineering Education</i>	1,563	1997–2011
<i>Journal of Chemical Education</i>	1,823	2005–2008
<i>Journal of Engineering Education</i>	541	2000–2011
<i>Journal of Professional Issues in Engineering Education and Practice</i>	842	1994–2011
National Science Foundation (Awards information)	123,054	2001–2011

**Table 1. Partial set of knowledge products currently accessible via iKNEER.**

data manually from optical media when target data are not available on the Internet. To overcome issues with different data formats owned by different publishers, we have developed a robust set of data adapters to transform publication metadata into a unified format before including them in our data archives. The need for developing new adapters for new data sources sometimes leads to a minor gap between data collected and data accessible by the public. This is however unavoidable as we need to ensure data integrity and also broad inclusion representative of engineering education research. Table 1 shows a partial list of the knowledge products currently accessible via iKNEER. (Note that publications may be missing in certain years when no articles were published.) Other resources that have been collected by iKNEER but yet to be published are: Australasian Association for Engineering Education, Education for Chemical Engineers, International Conference on Engineering Education, International Conference on Engineering Education Research, SEFI, and World Conference on Continuing Engineering Education. We are currently working on making these new resources accessible on iKNEER.

iKNEER collects a significant amount of funding (awards) information from the National Science Foundation public database. In the long run, one of the potential applications of iKNEER is to understand the impact of federal funding on engineering education. In order to gain a full picture,





we archive and provide secondary mining capabilities for all of the NSF awards. While iKNEER focuses primarily on engineering education, we utilize the entire body of awards information to create scientific profiles for researchers. We provide a fuller description of how individual researcher profiles are constructed below. Additionally, funding information is also used as one of the markers for helping disambiguate researcher names and institutions.

Once data are systematically collected in our system, there are significant issues of optimization and assuring data quality that need to be addressed. The former aims to reduce query-processing time, whereas the latter ensures that publication information is accurately represented. Author name and product title ambiguity is the biggest barrier to establishing data quality at this stage of the process. Meaning, publishers follow their own naming conventions for author names such as abbreviating first names and ignoring middle initials. As a result, it is common that one author has published multiple papers under two or more versions of their names. Author name ambiguity produces erroneous results when computing how many authors are working on a given topic, collaboration models, and other metrics such as publication counts. To overcome this issue, we designed a recommendation-based system to allow users to disambiguate duplicate items used primarily when the automatic processes are not able to resolve the ambiguity.

Figure 2 demonstrates how iKNEER creates multiple groups of author names that are detected as potential duplicates. Based on the recommendation list and the corresponding authors' publication

**1 To Be Processed**  
Submit All Refresh

- ADEEB RAHMAN
  - A. RAHMAN
- M. RAHMAN
  - M. M. RAHMAN
  - MUHAMMED ABEDUR RAHMAN
- P. K. RAJU
  - P K RAJU
- BALA RAM
  - B. RAM
- BALA RAMACHANDRAN
  - B. R. RAMACHANDRAN
- J. RAMOS
  - J. C. RAMOS
  - JAIME RAMOS
  - JUAN CARLOS RAMOS
- MIGUEL RAMOS
  - M. J. RAMOS

**2 Confirmed Groups of Author Duplicates**  
Clear All

- MARY B. RABER
  - MARY RABER
- DAVID F. RADCLIFFE
  - DAVID RADCLIFFE [Remove](#) [Set Primary](#) [Compare](#)
- WIM RAVESTEIJN
  - W. RAVESTEIJN
- LUISA M. REGUERAS
  - L. REGUERAS

Authors Comparison

Author	Title	Publication	Year	Coauthor with
DAVID F. RADCLIFFE	The Competence Dilemma in Engineering Education: Moving Beyond Simple Graduate Attribute Mapping	Australasian Journal of Engineering Education	2007	JOACHIM WALTHER;
	Shaping the discipline of engineering education	Journal of Engineering Education	2006	JOACHIM WALTHER;
	Innovation as a meta-attribute for graduate engineers	International Journal of Engineering Education	2005	JOACHIM WALTHER;
DAVID RADCLIFFE	Analysis of the use of an accidental competency discourse as a reflective tool for professional placement students	Frontiers in Education Conference	2007	JOACHIM WALTHER; LLEWELLYN MANN;

**Figure 2. The author name disambiguation system that aids in grouping and removing duplicates.**



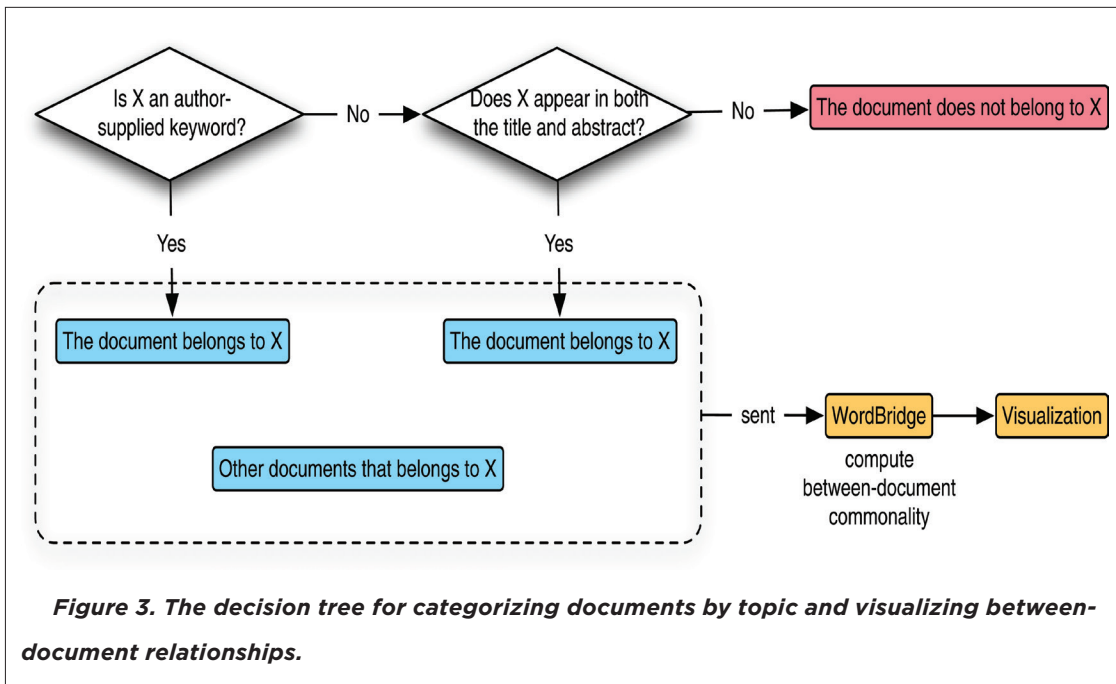
activities, members of the iKNEER design team determine whether to group the seemingly similar author names together or not. In Figure 2, the author names compared within the group share the same co-author, which is a clear indicator that these two names refer to the same author. Therefore, these two names are marked as the same. We are adding grants information from NSF to further strengthen this disambiguation process. Our name disambiguation system supports rollback operations so that mistakenly grouped name duplicates will not result in permanent changes in the database. This process is completely transparent to the end users, who never have to perform this process or interact with the disambiguation system directly. However, iKNEER provides mechanisms for end users to identify issues with the automatic disambiguation system and report them.

### **Data-centered computational components**

Based on the data collected, the computational server supports the presentation layer (or user interface for end users) by running appropriate algorithms. To support composite searches with multiple constraints such as author and publication time, we developed a sophisticated search and compilation methodology that is tuned to provide short response times in the order of microseconds (whereas traditionally these results took days to compile). To produce co-author networks, iKNEER computes and caches results from social network analysis based on the co-authorship information found in the entire dataset. To provide insights on how a topic evolves, we implemented a computational component that aggregates relevant knowledge products and groups them by time of publication.

To produce concise views related to any individual knowledge product we provide end users with a collection of weighted keyphrases (which go beyond just singular keywords to retain context). We utilized existing author-supplied keywords and also designed an automatic smart tagging system. Describing an academic article with a list of keywords has been commonly used and often been mandated by publishers to give readers a general sense of what an article is about. However, some publishers do not impose this requirement and leave many articles without keywords. In a system like iKNEER, manually assigning keywords to documents is infeasible because of the time cost and volume of data. Therefore, we created a smart tagging system that generates keywords based on the frequency of word occurrence in the full text of a given document. We maintain a stop word list to filter common words such as *the*, *of*, *is*, and *a* so that they will not be identified as keywords. For example, the top four keywords generated from a JEE paper [33] are *<mentor, 74>*, *<experience, 63>*, *<gender, 57>*, and *<cooperative, 42>*, where values indicate the number of word occurrence in the document. We continue to investigate other methods to identify appropriate tags for documents and other knowledge products.

To enable better navigation by topics, we utilize a rule to determine the likelihood of a document belonging to a certain category. Depending on word occurrences in the title, abstract, and keywords,



a document is characterized as belonging to one or more topics. We define a simple taxonomy and ontology of engineering education research so that inter-topic relationships can emerge automatically. For example, *workplace diversity* should be contained in *workplace* and if a user searches for *workplace*, articles on all sub-topics will be returned. Relationships between documents are passed to an algorithm called WordBridge for producing a visualization showing the commonality between the two, as outlined in Kim et al. [34]. Figure 3 provides an overview of the process. We discuss the types of visuals that result later in the paper.

To open our database to other researchers in EER who may want direct access to the data unfiltered by our interface, iKNEER provides a set of remote procedure application interfaces formatted in JavaScript Object Notation - Remote Procedure Call (JSON-RPC). For example, a developer can pass the JSON packet in Table 2 to request information about the first ten papers with the keyword

```
{
  "params": {
    "tag": ["assessment"],
    "publication": "Frontiers in Education Conference",
    "publicationYear": {
      "beginYear": 2000,
      "endYear": 2009
    },
    "output": "PaperInfo",
    "range": {
      "beginIndex": 0,
      "endIndex": 9
    }
  },
  "method": "advancedQuery",
  "id": 8818
}
```

**Table 2. A JSON request for getting the first ten papers related to assessment published in FIE.**



assessment published in FIE. Other procedure calls include computing co-author networks, keyword trends, and papers written by a given author.

### Visualizations and user interface

We discussed the design and implementations of data management and computational components above, which involve computational techniques such as data mining, social network analysis, and time-series analysis. iKNEER, however, is designed for a non-expert in data mining and visual analytics. To this end, we created a web-based user interface for users to explore the data in a visual and intuitive manner. No application or plug-in installation is required to use iKNEER. The user interface primarily provides the following capabilities:

### Full text phrase search (example in Figure 4)

One of the features that we have developed and implemented in iKNEER is the capability to perform full text search across our entire archive. Users can search any word, phrase, author name, institution, NSF program name, and specify other constraints such as time range. Search is intrinsically tied to all other aspects of the iKNEER site. We treat every data element as searchable and assume that every search result eventually maps to other larger more powerful data points. iKNEER then shows the most relevant publications along with an academic profile of the total number of academic papers, grant proposals, authors, PIs, and funding amount associated with the user input. These elements on the search result page are interactive to allow further exploration. Full text search requires us to process and return a significant amount of data at any given time. To perform this effectively,

The screenshot displays the search results for the query "active learning". At the top, it indicates "You are searching for active learning (0.35 seconds)" and "Displaying 282 documents". Below this, there are three search results:

- Article:** "The Key Factors Of An Active Learning Method In A Microprocessors Course". Authors: A Carpeño, J Arriaga, J Corredor, J Hernández. Publication: Education, IEEE Transactions On. Year: 2011. Volume: 54. Issue: 2. Keyword: teaching, cooperative learning, electronic engineering education, microprocessor chips, microprocessors course, active learning method, teacher activity, active, student, european, method, factors, key, these, microprocessors coursethe. [Show abstract](#)
- Grant Proposal:** "Collaborative Research: Development of Adaptable Web Modules to Stimulate Active Learning in Hydrology using Data and Model Simulations". PI: Upmanu Lall. Category: EHR -->DUE -->Tues-type 2 Project. Program Officer: Susan Finger. Award ID: 1123039. Active from: 2011 Oct.1 to 2013 Sep.30. Award amount: \$ 98,324. Current status: Active. [Show abstract](#)
- Grant Proposal:** "Collaborative Research: Development of Adaptable Web Modules to Stimulate Active Learning in Hydrology using Data and Model Simulations".

On the right side, there is a "Search" box and a "Statistical Overview" section:

**Search:** Find authors, documents, keywords, ...

**Statistical Overview:**

Num. of academic papers:	200
Num. of grant proposals:	82
Num. of authors:	166
Num. of distinct PIs:	80
Total amount:	\$24.70m
Avg. amount per award:	\$301k

**Figure 4. Full text search of "active learning" in iKNEER. Documents that mention "active learning" in title or full text are returned, along with a statistical overview of total number of publications, authors, and funding amount associated with this term.**



we have researched and developed a testbed using an open-source indexing system called SOLR (Apache SOLR. Available online at <http://lucene.apache.org/solr/>) - which is a web extension to a high performance, full-text search engine library called Lucene (Apache Lucene. Available online at <http://lucene.apache.org/java/docs/index.html>). Our SOLR-based implementation indexes all archived documents, takes any text as input, and returns a list of documents that mention the given search text. Figure 4 shows the search result within iKNEER using the phrase “active learning”. The left window lists all academic papers and grant proposals related to active learning with the most relevant ones on top. The statistical overview shows the academic and financial profile of active learning. All these results are computed and returned in about 0.35 seconds.

By clicking on the search results page, users can navigate to see the full profile of a scholar, a topic, or an NSF program. Figure 5 shows a scholar’s academic profile, which includes his overall academic performance (measured by number of publications and funds), major research interests, collaborators, NSF program officers who approved the awards, and the researcher’s publications. The profile page takes into consideration name disambiguation, incorporates keyphrases extracted from full text, and presents the large dataset in a concise way. Again, all elements on the page are interactive such that clicking on a research topic, a node in the collaboration network, or a program officer’s name in the table will open a new profile page for the selected entity.

### Visualizing relationships within iKNEER

One of the more powerful aspects of iKNEER is its ability to visualize relationships between authors and co-authors of papers and conference proceedings. We generate these relationship maps interactively and in real-time based on user interactions. Users generally get the visuals in a matter of seconds – a process that used to take several hours. Figure 6 provides a simple collaboration network for one scholar. Here the brightness of lines indicates the number of papers produced between the scholar and his collaborators.

While iKNEER can easily handle relationships between people (authors), we also apply the same approaches to visualize thematic areas, keywords, journals, and other entities. For example, we could pose the question “who are the top 50 researchers working in the problem space ‘assessment’ and what is the collaboration network between those researchers?” Figure 7(a) provides a simple answer to the question quickly and accurately, whereas Figure 7(b) includes all scholars working in ‘assessment’.

While we have made good progress on the algorithmic aspect of iKNEER, we still need to ensure that we have complete data coverage. We are examining techniques to scale our work to include a larger set of data. These types of maps can also be generated based on timescales. iKNEER already has this feature built into it. For example, any user can generate a time-scaled version of the collaboration networks that allows us to understand how a person’s collaboration network evolves over time.



## Tools for large-scale data analytic examination of relational and epistemic networks in engineering education

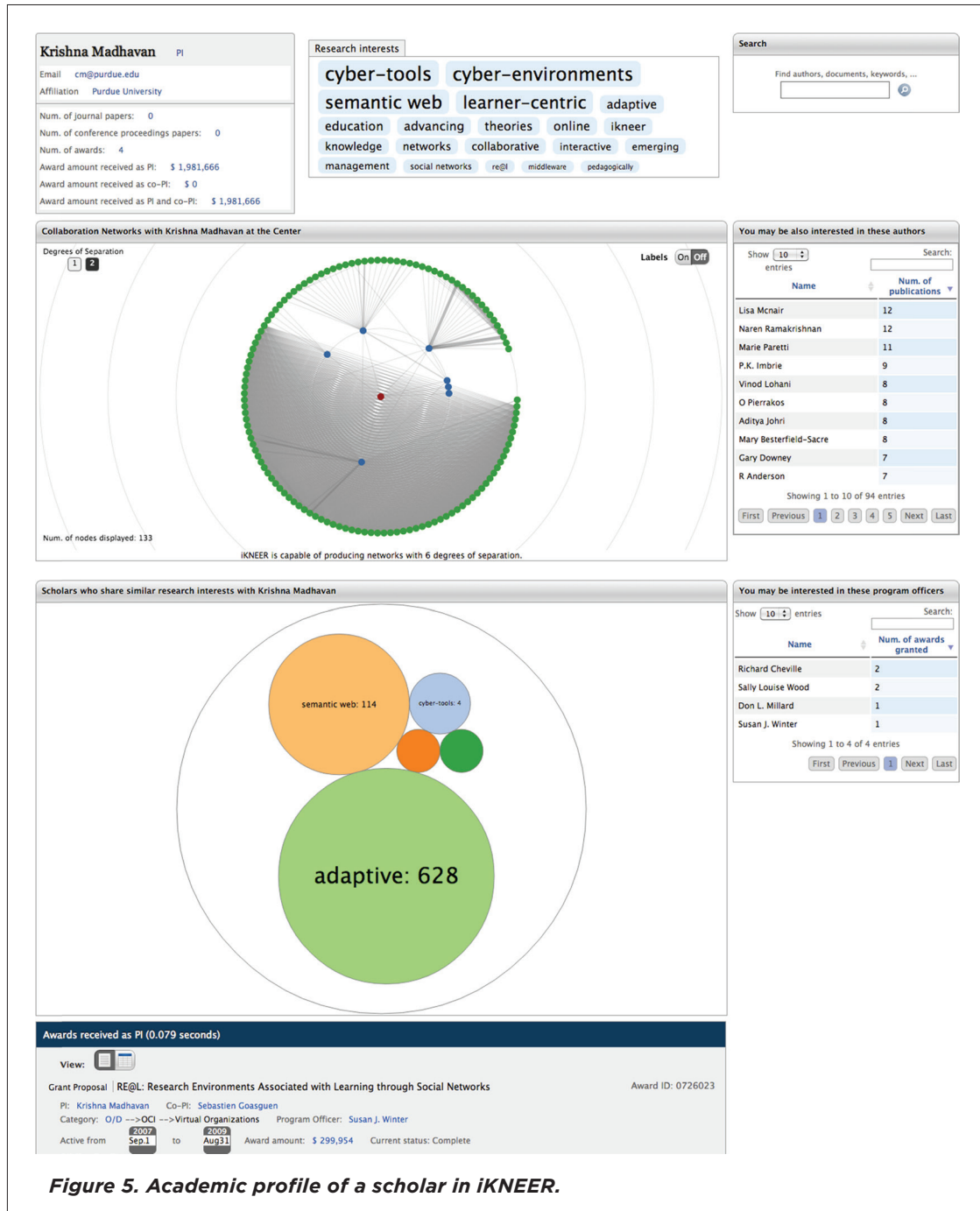
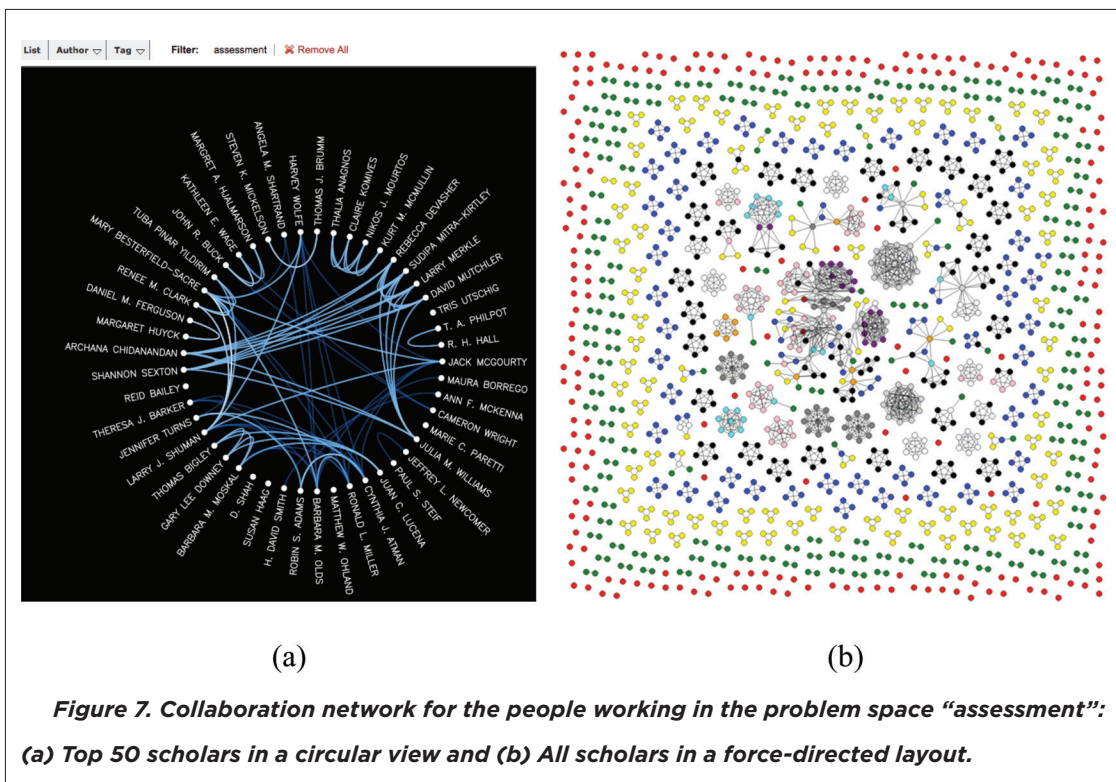
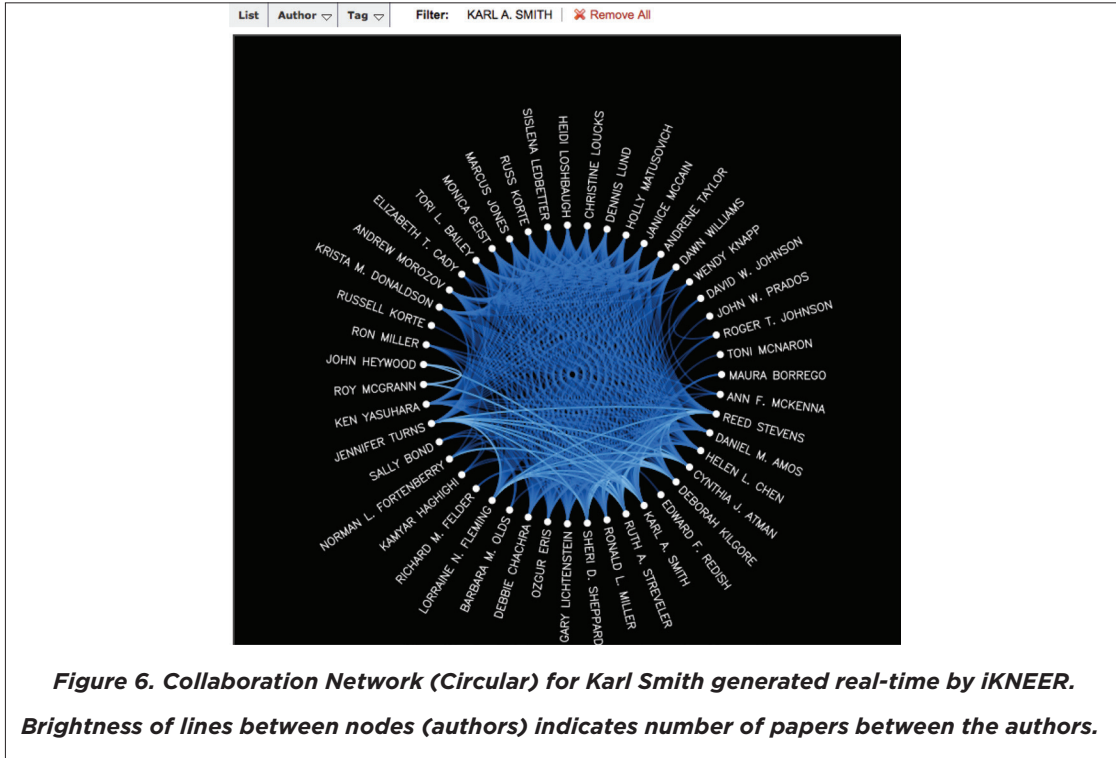
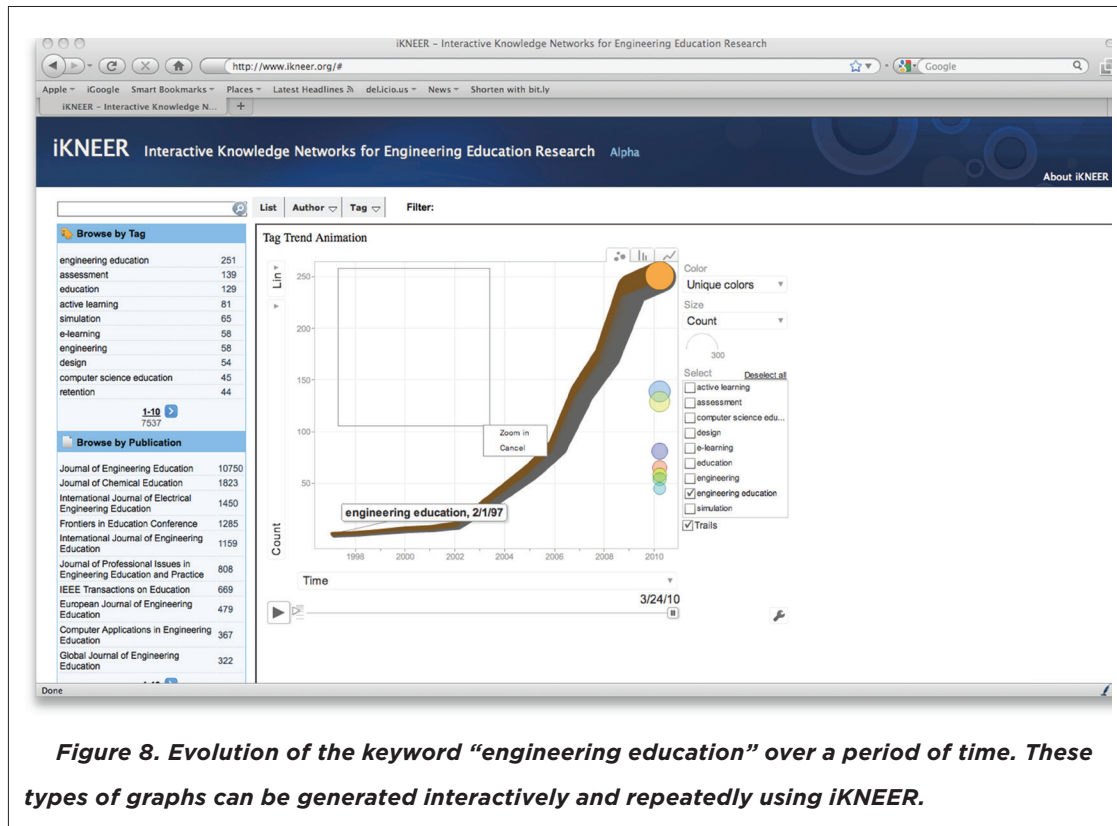


Figure 5. Academic profile of a scholar in iKNEER.

### iKNEER's advanced trend tracking capability

In the previous sections, we highlighted some of iKNEER's search and network mapping capabilities. In this section, we examine some of iKNEER's advanced trend mapping capabilities. iKNEER includes tools that allow users to understand how various concepts, keywords, and tags evolve over



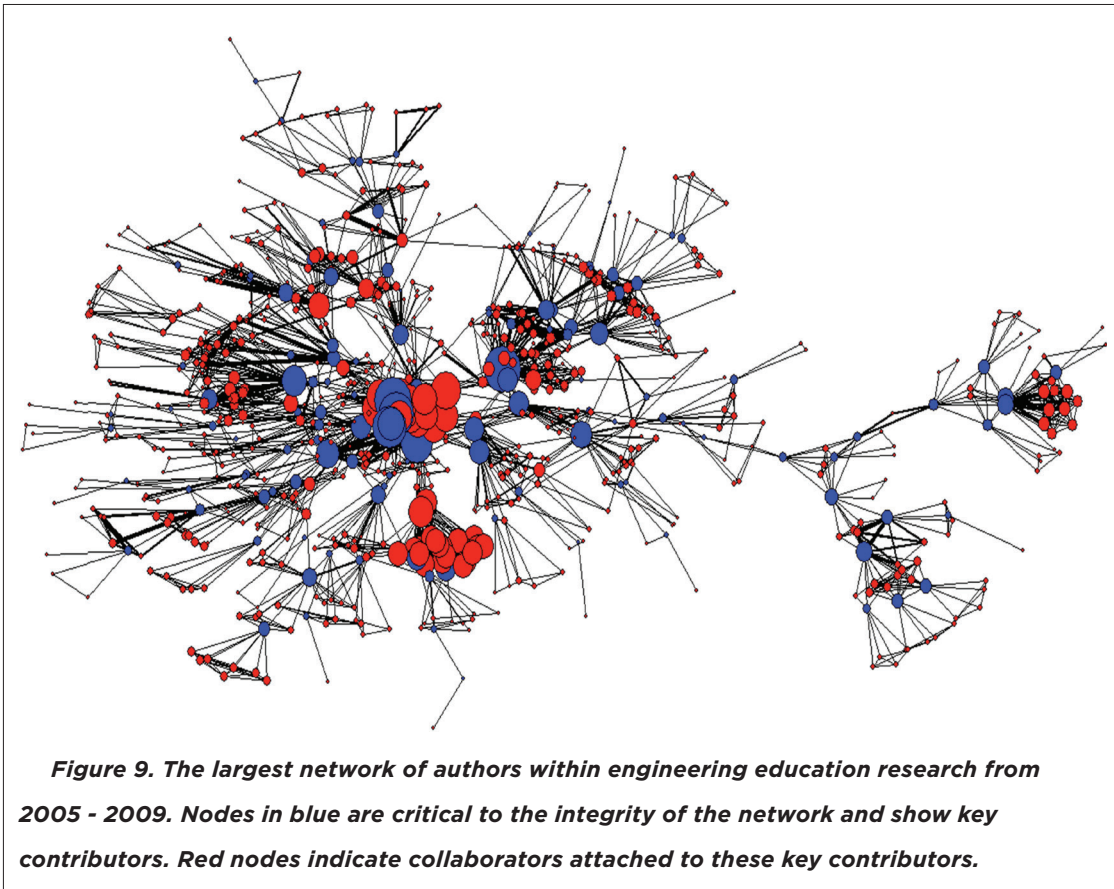


time. Figure 8 shows how the use of the keyword “engineering education” has evolved over a period of time. We allow users to plot these either on a direct frequency scale or a logarithmic scale. We use time-series analysis as an intrinsic part of these analyses. Any user can create these trending graphs based on a simple search for a topic of interest.

### ENABLING UNIQUE INSIGHTS

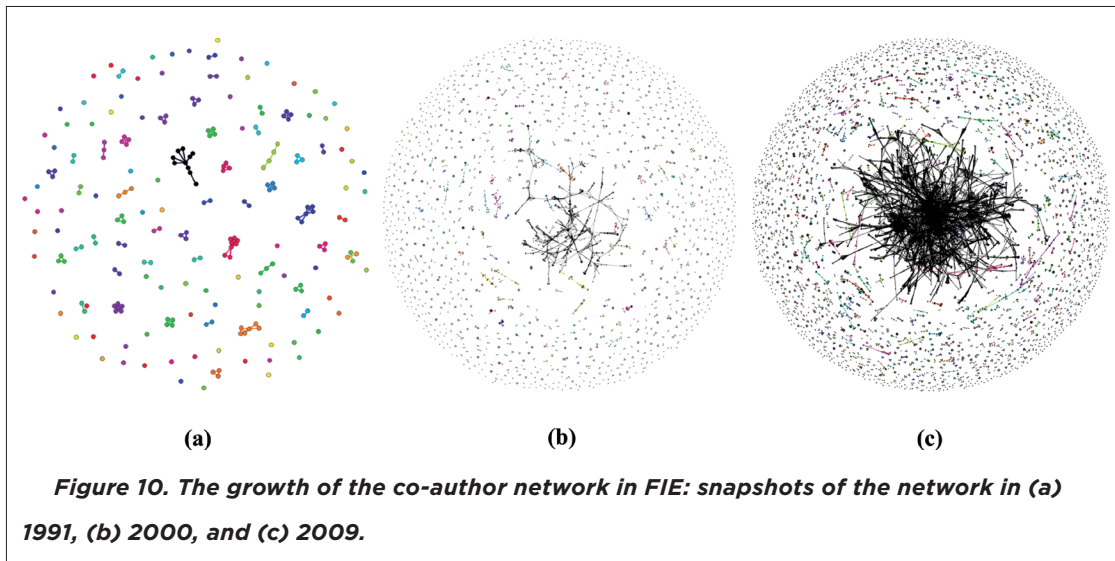
While iKNEER acts as a unique knowledge mining platform for the engineering education community, it also is being used to generate very unique insights about the field of engineering education research. This is indeed one of the main goals of iKNEER. Figure 9 provides a visualization of the largest network within the field of engineering education research based only on published work (journals and conference proceedings). The largest network in the field of engineering education research between the years 2005 – 2009 has 814 nodes (restricted by size of dataset). Future analyses with larger datasets may show a larger network emerging. This image was generated using





a software environment called UCINET (which is fairly complex for most users to utilize). However, the core dataset needed for this work was generated by iKNEER in a few seconds (a process that previously took months). We are building towards being able to generate these types of insights automatically and on-demand.

One of the key questions for any new discipline such as engineering education research is the question of capacity building and capability to propagate innovations. When looking at a network graph as shown in Figure 10, we ask the question - how is this network related to the larger community of engineering educators? Furthermore, is this network capable of propagating innovations? What does it mean that 814 researchers are connected in a single network? To showcase iKNEER's capability to answer such questions - we undertook a network visualization of the community fostered by the Frontiers in Education (FIE) conference. This effort used data from the proceedings of the FIE conference from 1991 to 2009 - a significant amount of data to analyze manually. The resulting visualization [35] showed that through the papers presented at the FIE conference, a larger community of researchers was being united into a powerful network. This network showed not only



the characteristics that demonstrate significant capacity – but also the size of the largest network showed tremendous potential to propagate pedagogical and theoretical innovations. Key points in the growth of the network fostered by the FIE conference are shown in Figure 10.

#### **BUILDING EPISTEMIC UNDERSTANDING – TOPIC MODELING TO IDENTIFY THE EMERGENCE AND GROWTH OF RESEARCH TOPICS IN ENGINEERING EDUCATION**

To test the value and scope of the infrastructure beyond the visualizations and analysis tools we had already incorporated in the system, we developed a case study where we undertook a detailed epistemic analysis of the field. We use the term epistemic consistent with its normal usage – of relating to or involving knowledge [36]. Different scientific fields and disciplines develop their own epistemic cultures which shape what is considered knowledge and the nature of knowledge that is valued within the field, with implications for the needed digital infrastructure [37]. As we noted earlier, engineering education research has undergone significant changes over the past decade. There has been an increase in the number of scholars and practitioners involved in the field, particularly those that are applying rigorous research principles to advance understanding of engineering education. In this context, it is important to understand the topics, approaches, and ideas that have caught the imagination of people in the community. Therefore, one of the analyses undertaken in this paper is to apply topic modeling and associated techniques to chart the emergence and growth of research topics in engineering education research over 9 years, from 2000-2008. Since this nature of work



has not been done in relation to EER, a significant part of the effort described here is innovative and exploratory in nature where different techniques were tested with the goal to identify diversity of topics that are of interest to the community.

As discussed earlier, although the roots of engineering education go back over a century, when *Journal of Engineering Education* published its first issue, in recent years there has been increased focus on improving the empirical foundations of the field. Numerous initiatives to develop the field have been created and implemented. Any maturing research field can reap significant advantages from a holistic understanding of its past and current efforts, particularly what topics found favor with researchers earlier, how they have changed, and what are some novel and recurring problems that need to be addressed. Yet, empirical efforts to do so at a smaller scale, such as through interviews and surveys, suffer from problems of bias, validity, and reliability. Recognizing the limitations of other approaches, one of the starting points for this research project was the question: How can we identify and study a research field over time, noting periods of gradual development, major ruptures, and most importantly the major topics that have been of interest to members of the field?

Faced with this question, we decided to leverage emerging advances in the data mining and analytics techniques. In particular, our investigation of observing such insights is operated on the unsupervised topic modeling method, Latent Dirichlet Allocation (LDA) [38] that has also been implemented in iKNEER. As a comparison, we also extract the most meaningful noun phrases and keywords from documents for topic detection and topic trend analysis. These approaches have been applied to various scientific corpora such as Proceedings of the National Academy of Sciences (PNAS), CiteSeer (a computer and information science paper collection), Proceedings of Neural Information Processing Systems (NIPS), and others. This approach has shown great capabilities for capturing the dynamics of research. To analyze topics in engineering education we developed a corpus of more than 2,500 articles from two journals and one conference on engineering education: *Journal of Engineering Education* (JEE), *International Journal of Engineering Education* (IJEE), and *Proceedings of Frontiers in Education* (FIE). These publications cover most major research topics across engineering education. We are in the process of adding the *Proceedings of Annual Conference of ASEE* to the corpus as well but our preliminary analysis suggests that the topics remain the same with or without that data.

## RELATED WORK

There have been many studies of the dynamics of scientific research. Using LDA models to capture the trends of topics has become popular in recent years. Griffiths and Steyvers [39] analyzed



the hot and cold topics of PNAS articles between 1991 and 2001 as a way of gaining insights into the dynamics of science. They presented a basic analysis based on the post-hoc examination of the estimated probability of a topic to a document produced by the LDA model. Hall et al. [40] applied a similar method to understand historical trends in the major conferences within Computational Linguistics from 1978 to 2006. They also introduced a model of the diversity of ideas, topic entropy, which is able to show the topic diversities of difference conferences. Wang and McCallum [22] extended the original LDA model by directly incorporating topic changes over time. Their model parameterizes a continuous distribution over time associated with each topic. Their experiments on several real-world data sets showed the discovery of more salient topics that are clearly localized in time than the plain LDA model. Despite the popularity of using LDA family models for trend analysis, other methods based on noun phrases and keywords were proposed and proven to be effective. For example, Jo et al. [41] addressed the problem of detecting topic trends using the correlation between the distribution of n-gram noun phrases that represent topics and the link distribution in the citation graph where the nodes are documents containing the phrases. Their approach is based on the intuition that if a phrase is relevant to a topic, the documents containing the phrase have denser connectivity than a random selection of documents. In another example, Mane and Börner [42] denoted topics as highly frequent words with a sudden increase in usage, a phenomenon called “burst”. Their major sources of these words came from keywords indexed by Institute for Scientific Information (ISI) and MEDLINE’s controlled vocabulary, also called MeSH terms. In order to determine the trends of keywords, top 10 most meaningful words were selected by domain experts. The frequency changes of these words over time were used to indicate the trends of each domain.

## METHODOLOGY

In this section, we describe the topic modeling technique that we used to analyze the research trends in engineering education. To perform the analysis we approached trends from different perspectives – topics, noun phrases and keywords. The system provides great flexibility in terms of selecting which data to analyze, including its context and time range. The data controller enables the selection of input corpus and allows a combination of any journals and conferences. The context controller enables us to choose the context for topic analysis. It can either be the title, the abstract or keywords in a paper. The model controller allows us to choose from a variety of extraction approaches such as LDA, noun phrase extraction or keyword extraction. The time controller enables us to choose a time range to calculate topic trends. It can be either individual



years or individual months. The findings indicate that some topics have remained constant over the years but some topics, such as global issues and assessment, have seen significant interest in the past five years.

### Topic modeling

Topic modeling techniques such as the Latent Dirichlet Allocation model (LDA) [38], aim to identify semantic topics given a text corpus. LDA is a generative probabilistic model of a corpus. It assumes that documents in a corpus are generated as random mixtures over latent topics. Let us assume that there is a corpus with  $D$  documents that contain a mixture of multiple topics  $\{z_1, \dots, z_T\}$ . LDA specifies the following distribution over words within a document:

$$p(w_i) = \sum_{j=1}^T p(w_i | z_j) p(z_j)$$

where  $T$  is the number of topics. Let  $p(w | z_j) = \phi^{(j)}$  refer to the multinomial distribution over words for topic  $z_j$  and  $p(z) = \theta^{(d)}$  be the multinomial distribution over topics for document  $d$ . The two sets of parameters,  $\phi^{(j)}$  and  $\theta^{(d)}$ , indicate which words are important for which topic and which topics are important for a particular document, respectively. Two symmetric Dirichlet distributions with hyper-parameters  $\alpha$  and  $\beta$  are introduced to the estimation of  $\theta^{(d)}$  and  $\phi^{(j)}$ , respectively, in order to achieve smoothed topic and word distributions. Smoothing is necessary to avoid zero probability assigned to words that do not appear in a training corpus. Those parameters are posterior probabilities that cannot be assessed directly. The values of the hyper-parameters depend on number of topics  $T$  and vocabulary size. Steyvers suggests that  $\alpha = 50/T$  and  $\beta = 0.01$  should work well with many different text collections. However, we still need to determine the number of topics  $T$  in the corpus. Perplexity is commonly used in language modeling to test the fitness of a text model given a training corpus. A lower perplexity score indicates better generalization performance in new documents. Therefore, we can obtain the best approximation of the topic numbers for the data by minimizing the perplexity as:  $T \arg \min_T \{\text{perplexity}(D_{\text{test}}/T)\}$ . Therefore, we can learn the topic models using a subset of documents as training data and evaluate the perplexity using the rest of the documents as testing data:

$$\text{perplexity}(D_{\text{test}} | T) = \exp\left(-\frac{\sum_{d=1}^{|D_{\text{test}}|} \log p(w_d | T)}{\sum_{d=1}^{|D_{\text{test}}|} N_d}\right)$$

### Noun phrase extraction

Frequently occurred noun phrases can also capture the major semantic concepts from a corpus. A noun phrase normally consists of a head noun and optionally a set of modifiers. It is an



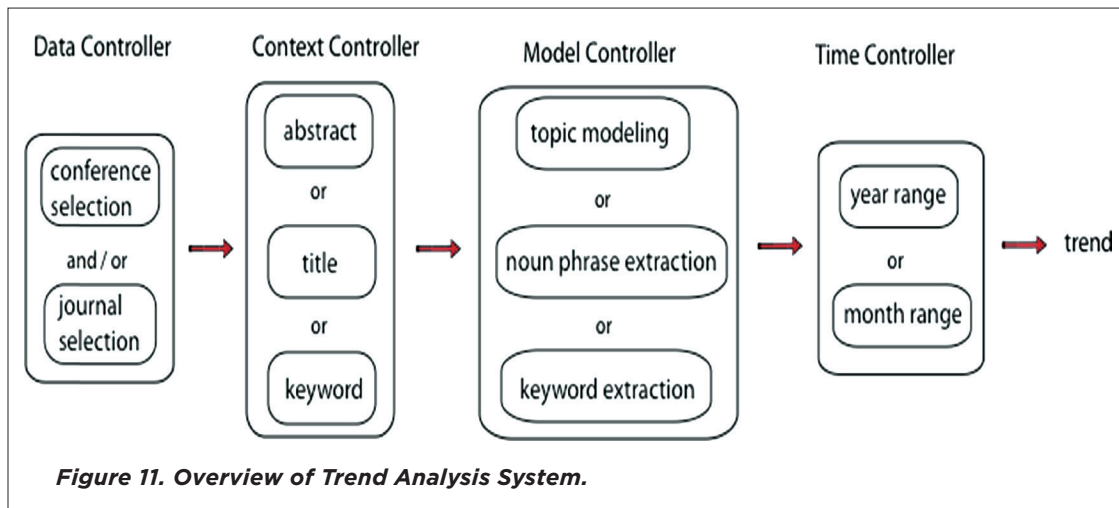
important grammatical unit of texts in many languages such as English, French, and Chinese. In natural language processing (NLP), there are two major noun phrase extraction methods, namely static parsing and machine learning. The static parsing method relies on a set of rules pre-defined by linguists. These rules are often described using finite state automation (FSA). However, the effectiveness of this method is strongly dependent on the accuracy and comprehensiveness of the rule set. On the other hand, machine learning methods aim to overcome the drawbacks of static parsing. They rely on various statistical learning techniques to identify important noun phrases by analyzing the part-of-speech (POS) tags of texts. Existing methods such as transformation-based method, memory-based method, maximum entropy, hidden Markov model, conditional random field, and support vector machine have been reported to be effective in noun phrase extraction. Existing studies show that there is not a single method that always performs better than others in different applications [43–45]. The NLP approach we applied uses conditional random field to extract noun phrases.

#### **Keyword extraction**

Keyword extraction is straightforward. It simply tokenizes the text to individual words. After removing common stop words (i.e., “a”, “the”), corpus-specific stop words such as engineering and education (in this particular study) are removed. Finally, words are stemmed to their roots (e.g., “studied” to “studi”) so as to obtain an accurate vocabulary of the corpus. For example, a sentence “engineering education has recently undergone a resurgence and reorientation” would be represented by the following keyword tokens {*recent, undergon, resurg, reorient*}.

### **DESIGN AND IMPLEMENTATION**

Based on the LDA topic modeling technique, we propose a topic trend analysis system that consists of 4 modules (see Figure 11). The data controller allows the user to specify the scope of the input corpus by selecting a combination of journals and/or conferences. The context controller asks the user to specify information (title, keyword, or abstract) to be included in the corpus for each publication. The model controller enables to choose the models of extracting topics or concepts in the corpus. It can be either of topic modeling using LDA, noun phrase extraction or keyword extraction. The time controller enables to choose a time range to calculate topic trends. It can be either individual years or individual months. Through different selections, a mix of inputs can be obtained giving a view across time and based on different data corpuses. This mechanism ensures that user can apply different lenses on the data.



**EXPERIMENT AND DATA ANALYSIS**

**Data preparation**

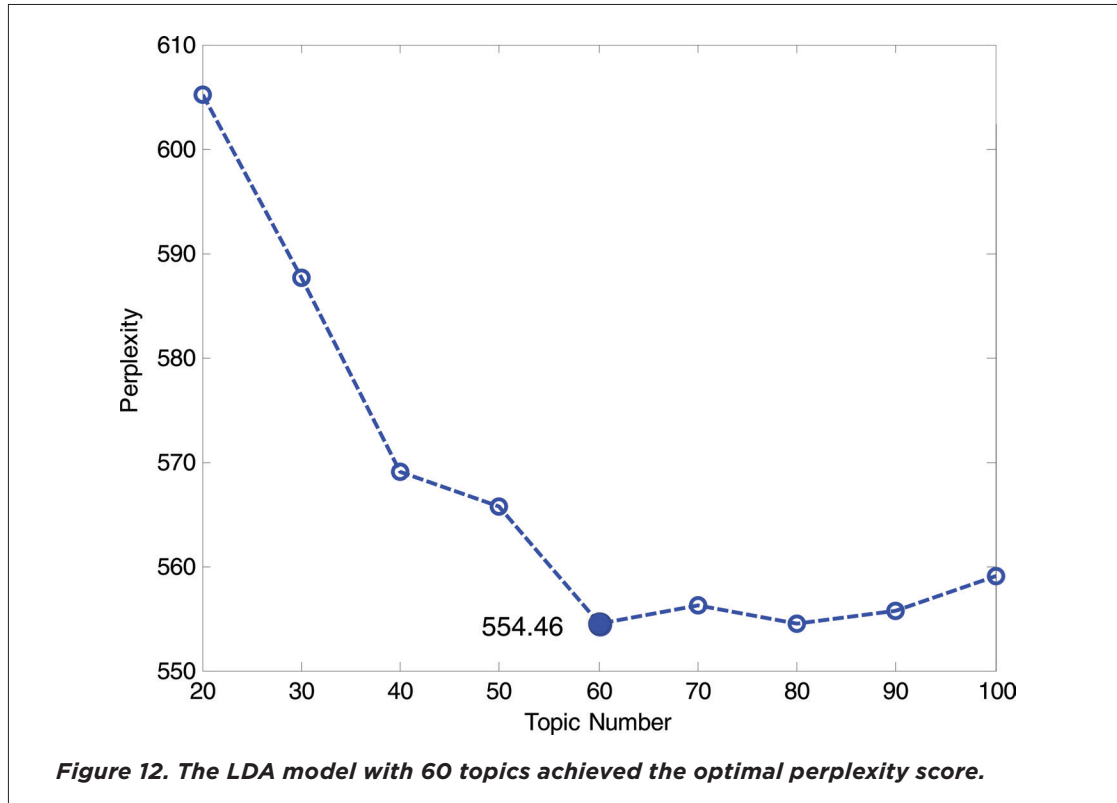
We analyzed the topic trends using a corpus which consisted of articles published in Journal of Engineering Education (JEE) and International Journal of Engineering Education as well as the conference proceedings of Frontiers in Education (FIE). The corpus should cover major research topics in Engineering Education. Table 3 summarizes the scope of our corpus.

**LDA model estimation**

We used an open source LDA package, namely GibbsLDA++ (see <http://gibbslda.sourceforge.net/>), for our LDA model estimation. The package is a C++ implementation of LDA using Gibbs sampling technique for parameter estimation and inference. Gibbs sampling is a form of Markov Chain Monte Carlo, which is easy to implement and efficient when extracting a set of topics from a large corpus. It is a standard parameter estimation approach in LDA [46, 47]. For LDA parameter settings and Gibbs sampling experimental settings, we followed those suggested by [37]. We split the original corpus into 90% for training and 10% for testing. The LDA parameters were set as

Data	D	V	W	Range
JEE, IJEE, FIE	2,645	7,768	203,453	2000-2008

**Table 3: Data Corpus (D is the number of documents, V is the size of vocabulary, W is the total number of words).**



$\alpha = 50/T$  and  $\beta = 0.01$ . For Gibbs sampling, we chose to run 1,000 iterations for estimation and 50 iterations for inference. As shown in Figure 12, the LDA model with approximately 60 topics achieved the optimal perplexity score. Therefore, we consider that our corpus mainly contains 60 topics.

### Topic trends

For illustration, we randomly selected 15 topics out of all 60 topics with top 10 words associated with each topic (see Table 4). We analyzed the trends of the 15 topics between 2000 and 2008. These trends are shown in Figure 13 & 14. The mean  $\theta$  values in the two figures measure the topic strength in each year. They are calculated based on the observed topic probabilities of each document belonging to each topic within a certain time frame such as a year. Following [38], we define topic strength as  $\hat{p}(z | y)$ , the empirical probability that an arbitrary paper  $d$  written in year  $y$  is about topic  $z$ :

$$\begin{aligned} \hat{p}(z | y) &= \sum_{d \in D(y)} \hat{p}(z | d) p(d | y) \\ &= \frac{1}{C} \sum_{d \in D(y)} \hat{p}(z | d), \end{aligned}$$

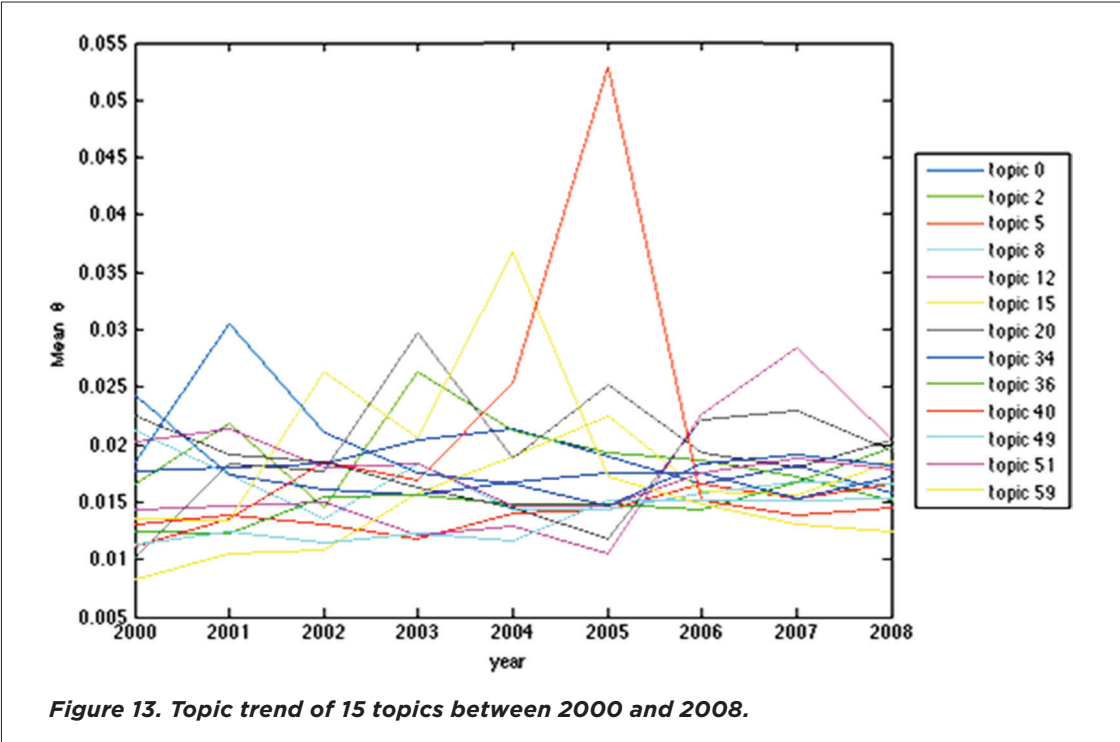
where  $C$  is the number of papers published in year  $y$ .

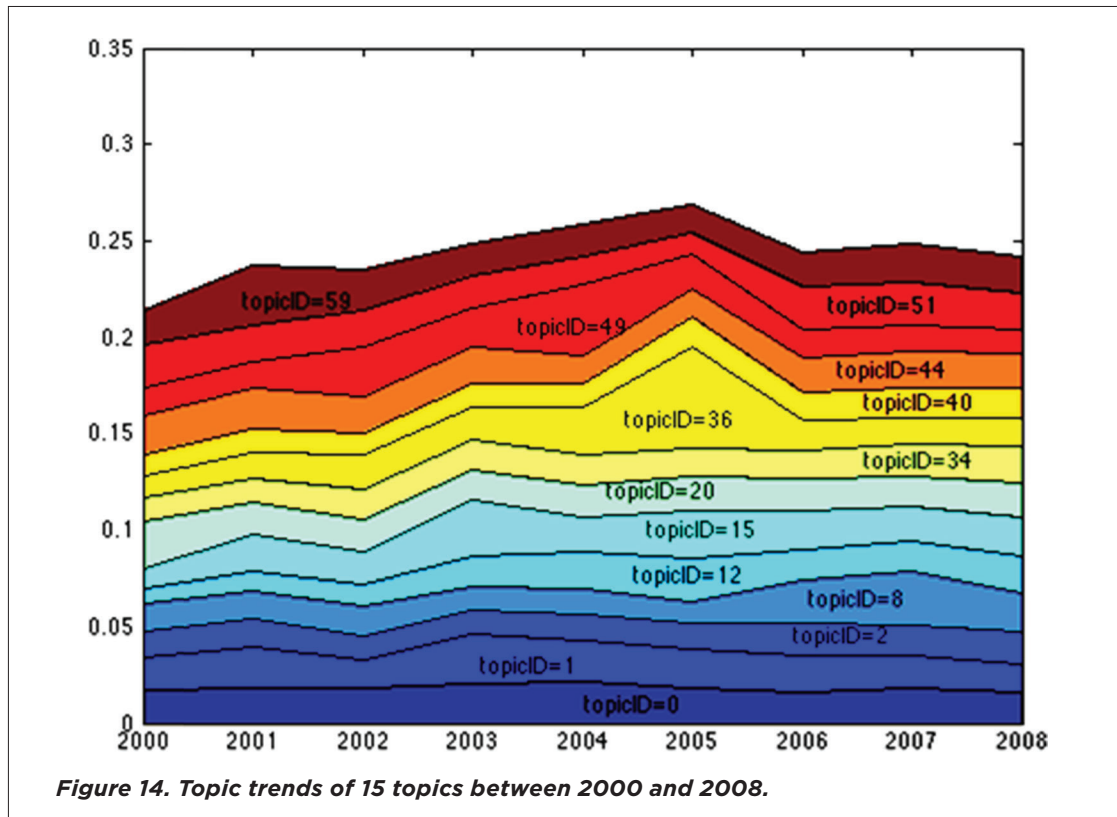




Topic ID (numbers do not represent ranking)	Top 10 Words in Each Topic
0	Student perform academic study factor significant higher level examination success
1	Design process engineering build idea open support pattern incorporate hand
2	Global intern competition culture university country state unit institution paper
5	Learn instruct base effect strategy cognition evaluation think tradition understand
8	School science teacher high student active stem middle career math
12	Device digit application mobile system embed base present logic implement
15	Laboratory lab experiment robot virtual remote control equipment simulation hardware
20	Survey study response result percept relate question rate complete determine
34	Data analysis collect inform analyze quality quantity method generate develop
36	Control simulation electron matlab power circuit paper present operate require
40	Software develop platform paper source potential open provide formal tool
44	Skill community develop technic student profession compete leadership knowledge integrate
49	Method chemic transfer energy numer spreadsheet flow calculate heat fluid
51	Student retent college mentor program success academy freshman increase university
59	Project student design capston require involve senior experiment final manage

**Table 4. Top 15 topics.**





### Keyword trends

We extracted the top 20 keywords of the entire corpus in Table 5 and analyzed their frequency trends over the time. The keywords were calculated based on paper titles and abstracts. The trends of two representative keywords, “laboratori” and “undergradu”, are shown in Figures 15 and 16, respectively.

### Noun phrase trends

We extracted the top 20 noun phrases of the entire corpus in Table 6 and analyzed their frequency trends over the time. The trends of two representative noun phrases, “engineering education” and “a case study”, are shown in Figures 17 and 18, respectively.

## FINDINGS AND DISCUSSION

### Topic trends

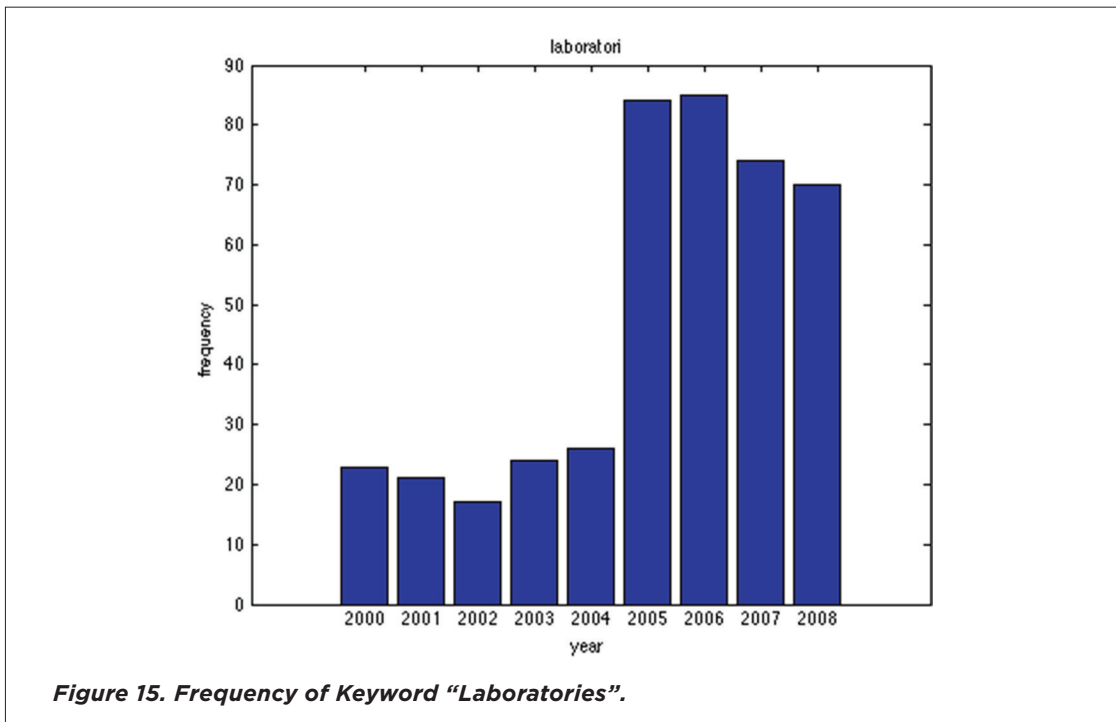
Overall, the findings from these analyses show that some topics remain constant over time whereas other topics become more popular - meaning, significantly more work is done on them

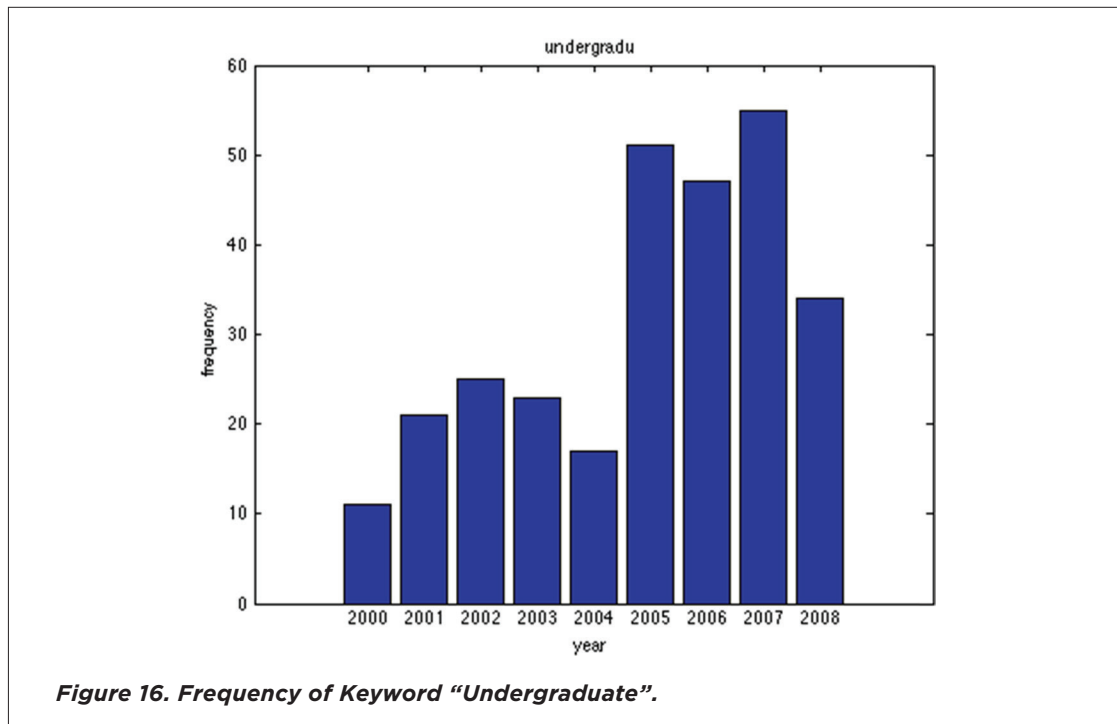


Keyword	Frequency	Keyword	Frequency
Learn	633	Project	252
Student	542	Assess	235
Teach	486	Model	234
Base	475	Approach	214
Design	469	Analysis	211
Laboratory	424	Study	211
Chemistry	384	Control	204
Experiment	336	Program	176
Develop	301	Simulate	158
Undergraduate	284	System	158

**Table 5. Major Keywords and Their Frequency.**

as compared to previous years – at certain time periods. For instance, since 2005 the topics *global and international aspects of engineering education* have seen a significant spike. This interest can partially be attributed to the discussion of international aspects of educating engineers in the NAE publications (*Engineer of 2020 & Educating the Engineering of 2020*) as well as the publication of *The World is Flat* by Thomas Friedman, which had a significant influence on science and engineering

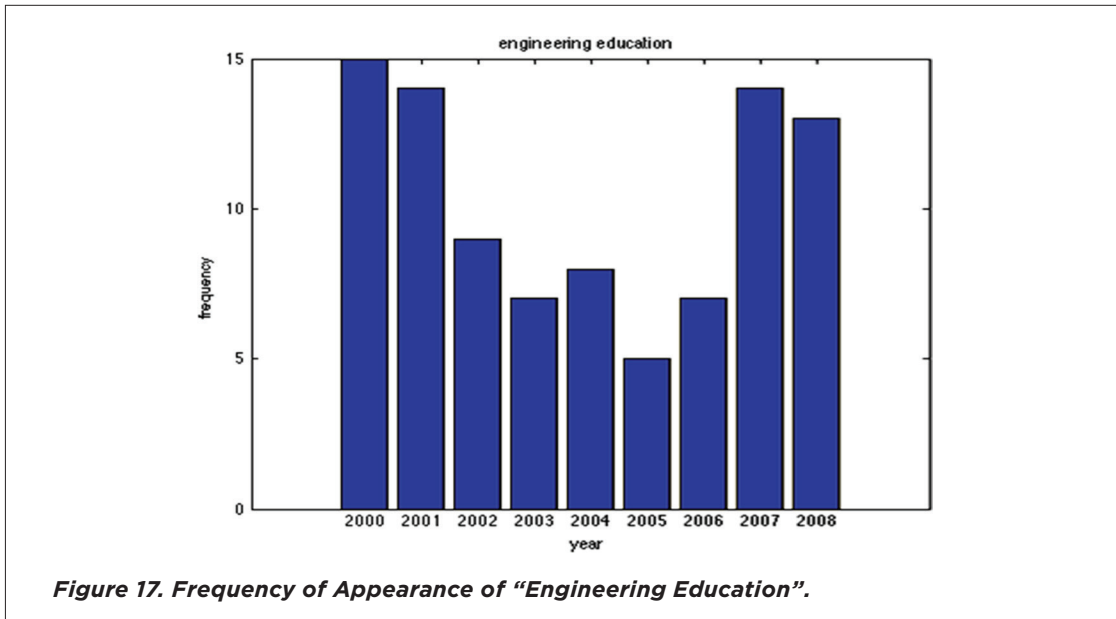




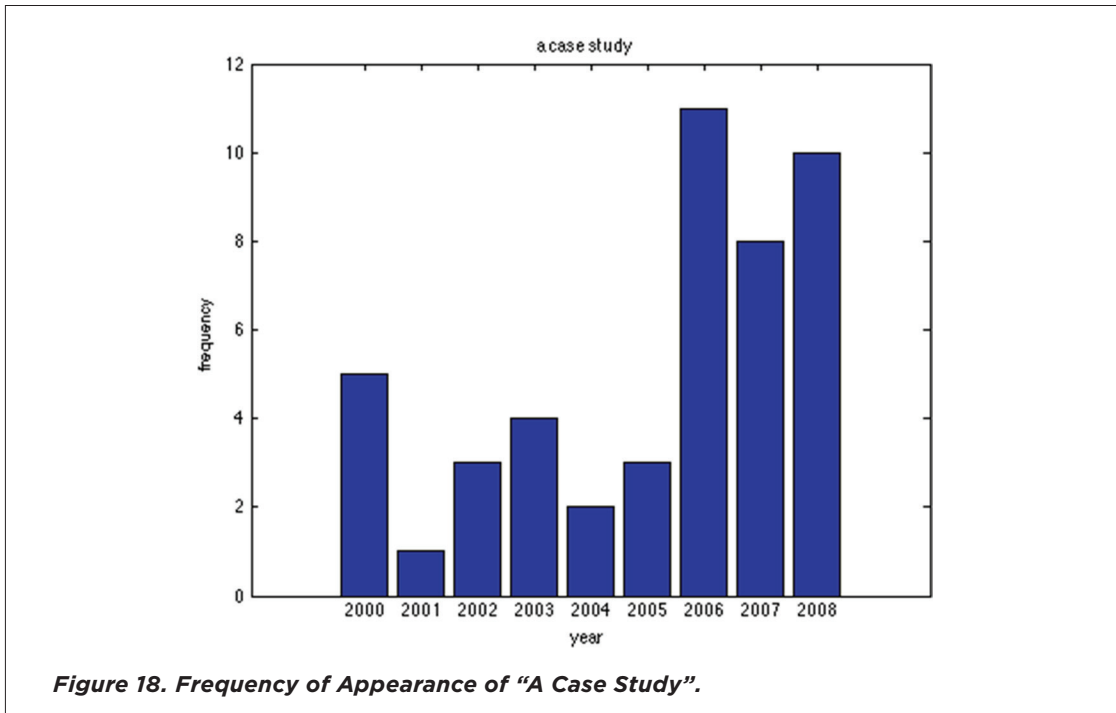
public policy in the United States. Of course, there are other factors at play that make a topic popular such as funding or uptake of a technology by the industry. The findings from the topic analysis also shed light on several methodological issues that emerged as the primary methods of interest to the community – *experiments, case studies* and *survey-based studies*. The results from the analysis also show that certain engineering related software and data analysis tools, such as *MATLAB*, are popular

Noun Phrase	Frequency	Noun Phrase	Frequency
students	161	synthesis	40
design	96	evaluation	39
development	93	technology	39
engineering education	92	the role	36
chemistry	64	the impact	32
assessment	56	the use	32
analysis	53	an experiment	31
engineers	48	course	31
a case study	47	research	31
matlab	41	science	31

**Table 6. Noun Phrases and Their Frequency.**



topics given their use in engineering education. In terms of disciplinary areas, *electronic and communications engineering* and *chemical engineering* were found to be common areas addressed by scholars. Efforts such as *mentoring* and *community development* were also frequently present in the list of topics. The use of technology in learning was another dominant area of research and several





topics (across the analyses) related to technology were identified, such as, *robotics* and *mobiles*. Not surprisingly, another major topic was *design*, given the central role of design in engineering practice and engineering learning and cognition. Results from topic modeling also suggested that *capstone projects* and *freshmen projects* are an area of interest across the community. Professional skills such as *leadership*, *communication*, and *teamwork* were also part of list of topics that were of interest to a significant number of scholars. Finally, another topic common across all results was *assessment*.

### **Potential concerns**

One area of potential concern, from the perspective of knowledge generation, which emerges from the analysis of topics is the paucity of theoretical or analytical keywords in the data. For a growing and maturing field it is essential to develop mechanisms to accumulate a body of knowledge in a meaningful manner [48-53]. This body of knowledge can then serve as the basis for productive future research, which avoids the pitfalls of earlier efforts. For any academic discipline, particular a social science or interdisciplinary discipline such as engineering education research, a strong core of theoretical or analytical ideas can coalesce scholars to make meaningful contributions [49]. For instance, no psychological, sociological, or learning sciences theory was present as a keyword. Issues of concern such as student motivation or student identity were also absent from the list of topics. This finding is of significance as it alerts us to a gap between practice and theory and the still greater effort needed to develop a more cohesive scholarly agenda in the field. Given the applied nature of engineering education, a lack of theoretical focus is not entirely surprising.

Another area of concern that emerged from the analyses was the disproportionate attention to undergraduate education and a lack of attention to graduate education within the community. Graduate students, in addition to being students of engineering, are also highly involved in both engineering teaching and research. Furthermore, the number of graduate students and their involvement in the engineering and engineering education community is steadily increasing. Therefore, more attention is needed to issues that focus on graduate engineering education. In a related issue, there was no mention of K-12 experiences either, which is also a growing area of interest within engineering education. As the field continues to grow it has to look beyond undergraduate students and steps have to be taken to include graduates and also K-12 students in engineering education. These are potential growth areas. As we further develop our data corpus to make it more inclusive and diverse, we are likely to uncover other areas of interest and of concern to engineering educators.

We do not aim to provide our analysis as prescriptions for the field but as outcomes or guidance that can be generated through a system such as iKNEER. Overall, by using different approaches to topic modeling we were able to provide a more comprehensive representation of the field than that achievable by other approaches. We combined LDA, noun phrase extraction,



and keyword extraction, and all three approaches provided a different lens on the data. We argue that for future work such a combined approach might be the ideal way to understand disciplinary communities and their interests and ideas. We highlight some of the key areas of interest for the community over the past years and identify emerging patterns as well as highlight an area of concern – the lack of theoretical or analytical topics with which the community engages. We also found that interpreting the results occurs best when someone from the disciplinary field looks at the findings.

One major limitation of our work is the exclusion of non-U.S. publication outlets. Although the journal and conferences in the sample publish international work, their representation is quite limited, therefore skewing the results towards issues that more pertinent to the U.S. In future work, we are trying to balance the data by including data from *European Journal of Engineering Education* as well as proceedings from *SEFI* and *REESE*. This finding essentially led to us expanding the offering of data sources within iKNEER to capture a more international flavor in the data. The goal is to make the dataset as comprehensive and diverse as possible. A secondary concern with the analysis methods adopted here is the frequent occurrence and identification of generic topics such as “students” or “learning.” We are cognizant of this issue but also believe that including such topics in the analysis and findings captures a more honest characterization of the field and present a diffuse but real representation of the ideas present in the field.

## CONCLUSION

In this paper, we describe the design and implementations of a data-intensive knowledge platform and show how the system- iKNEER – and data contained within the system can be used to develop a better understanding of engineering education research. iKNEER aims to document and present the evolution of engineering education research. We collect, index, and allow sense making of a large collection of data through intuitive and user-friendly interfaces. We apply topic modeling techniques to the data to understand the emergence and growth of research topics within the community. Researchers, educators, and other stakeholders in the engineering education research community can visually identify potential collaborators, research patterns, topic trends, and highly related articles. iKNEER is also starting to provide unique insights about the topology of the networks within engineering education research. It shows that the content and knowledge that rests within the networks formed by researchers are the fundamental mechanisms through which practices and methods unique to the field of EER can propagate. We acknowledge that the insights derived from iKNEER are highly linked to the amount and quality of data we index and process. See <http://youtu.be/etWIJ5ESKhU> for a full illustration of iKNEER.



### ACKNOWLEDGEMENTS

This paper is based on work conducted as part of a project supported through National Science Foundation Grant EEC-0957015, EEC-0935109, and EEC-0935124. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

### REFERENCES

1. March, J.G., *Exploration and exploitation in organizational learning*. Organizations Science, 1991. **2**(1): p. 71-87.
2. National Academy of Engineering, *The Engineer of 2020: Visions of Engineering in the New Century*. 2004: The National Academies Press.
3. American Society of Engineering Education (ASEE), *Engineering Education for the Global Economy: Research, Innovation, and Practice*, 2008.
4. National Science Board, *Moving forward to improve engineering education*. 2007.
5. Fortenberry, N.L., *Guest editorial: An extensive agenda for engineering education research*. Journal of Engineering Education, 2006. **95**(1): p. 1-5.
6. NSF Task Force on Cyberlearning, *Fostering learning in the networked world: The cyberlearning opportunity and challenge*, 2008.
7. Atkins, D.E., et al., *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*, 2003.
8. Cummings, J., et al., *Beyond being there: A blueprint for advancing the design, development, and evaluation of virtual organizations*. 2008.
9. Wankat, P.C., *Analysis of the first ten years of the Journal of Engineering Education*. Journal of Engineering Education, 2004. **93**(1): p. 13-22.
10. Wankat, P.C., *An analysis of the articles in the Journal of Engineering Education*. Journal of Engineering Education, 1999. **88**(1): p. 37-42.
11. Chen, X.C., et al. *Weighted social tagging as a research methodology for determining systemic trends in engineering education researcher*. in *Proceedings of the ASEE Annual Conference and Exposition*. 2011.
12. Jesiek, B.K., M.J. Borrego, and K. Beddoes. *Expanding global engineering education research collaboration*. in *Proceedings of the 2008 SEFI Annual Conferences*. 2008. Aalborg, Denmark.
13. Jesiek, B.K., L.K. Newswander, and M. Borrego, *Engineering education research: Discipline, community, or field?* Journal of Engineering Education, 2009. **98**(1): p. 39-52.
14. Madhavan, K.P.C., J.D. Schroeder, and H. Xian. *Evaluating the effectiveness and use of cyber-learning environments in engineering education: A qualitative analysis*. 2009. Austin, TX.
15. Prince, M., *Does active learning work? A review of the research*. Journal of Engineering Education, 2004. **93**(3): p. 223-232.
16. Dutson, A.J., et al., *A review of literature on teaching engineering design through project-oriented capstone courses*. Journal of Engineering Education, 1997. **86**(1): p. 17-28.





17. Webster, J. and R.T. Watson, *Analyzing the past to prepare for the future: Writing a literature review*. MIS Quarterly, 2002. **26**(2): p. R13-R13.
18. McCain, K.W., *Mapping authors in intellectual space: A technical overview*. Journal of the American Society for Information Science, 1990. **41**(6): p. 433-443.
19. Chen, C. and L. Carr. *Trailblazing the literature of hypertext: author co-citation analysis (1989-1998)*. in *Proceedings of the tenth ACM conference on hypertext and hypermedia: Returning to our diverse roots*. 1999. ACM.
20. Acedo, F.J. and J.C. Casillas, *Current paradigms in the international management field: An author co-citation analysis*. International Business Review, 2005. **14**(5): p. 619-639.
21. Rosen-Zvi, M., et al. *The author-topic model for authors and documents*. 2004. Banff, Canada: AUAI Press.
22. Wang, X. and A. McCallum. *Topics over time: a non-Markov continuous-time model of topical trends*. in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006. ACM.
23. Mccallum, A., A. Corrada-Emmanuel, and X. Wang, *The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email*, 2004: University of Massachusetts, Department of Computer Science.
24. Uren, V.S., et al., *Sensemaking tools for understanding research literatures: Design, implementation and user evaluation*. International Journal of Human-Computer Studies, 2006. **64**(5): p. 420-445.
25. Strobel, H., et al., *Document Cards: A top trumps visualization for documents*. IEEE Transactions on Visualization and Computer Graphics, 2009. **15**(6): p. 1145-1152.
26. White, H.D. and K.W. McCain, *Visualizing a discipline: An author co-citation analysis of information science, 1972-1995*. Journal of the American Society for Information Science, 1998. **49**(4): p. 327-355.
27. He, Y. and S.C. Hui, *Mining a web citation database for author co-citation analysis*. Information Processing & Management, 2002. **38**(4): p. 491-508.
28. Small, H.G., *A SCI-Map case study: Building a map of AIDS research*. Scientometrics, 1994. **30**(1): p. 229-241.
29. Small, H.G., *Visualizing science by citation mapping*. Journal of the American Society for Information Science, 1999. **50**(9): p. 799-813.
30. Boyack, K.W., B.N. Wylie, and G.S. Davidson, *Domain visualization using VxInsight® for science and technology management*. Journal of the American Society for Information Science and Technology, 2002. **53**(9): p. 764-774.
31. Bergstrom, P. and D.C. Atkinson. *Augmenting the exploration of digital libraries with web-based visualizations*. 2009.
32. Borner, K., C. Chen, and K.W. Boyack, *Visualizing knowledge domains*. Annual Review of Information Science and Technology, 2003. **37**(1): p. 179-255.
33. Ingram, S., S. Bruning, and I. Mikawoz, *Career and mentor satisfaction among Canadian engineers: Are there differences based on gender and company-specific undergraduate work experiences?* Journal of Engineering Education, 2009. **98**(2): p. 131-144.
34. Kim, K., et al. *WordBridge: Using composite tag clouds in node-link diagrams for visualizing content and relations in text corpora*. 2011.
35. Madhavan, K.P.C., et al., *The FIE story - 1991 to 2009*, 2010.
36. Cetina, K.K., *Epistemic cultures: How the sciences make knowledge*. 1999: Harvard University Press.
37. Merz, M., *Embedding digital infrastructure in epistemic culture*. Hine, CM New Infrastructure for Knowledge Production: Understanding E-Science. Hershey, PA: Information Science Publishing. p, 2006: p. 99-119.
38. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation*. Journal of Machine Learning Research, 2003. **3**: p. 993-1022.



39. Griffiths, T.L. and M. Steyvers, *Finding scientific topics*. Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(Suppl 1): p. 5228-5235.
40. Hall, D., D. Jurafsky, and C.D. Manning. *Studying the history of ideas using topic models*. in *Proceedings of the conference on empirical methods in natural language processing*. 2008. Association for Computational Linguistics.
41. Jo, Y., C. Lagoze, and C.L. Giles. *Detecting research topics via the correlation between graphs and texts*. in *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*. 2007. ACM.
42. Mane, K.K. and K. Bärner. *Mapping topics and topic bursts in PNAS*. in *Proceedings of the National Academy of Sciences of the United States of America*. 2004.
43. Jiang, M., et al., *A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries*. Journal of the American Medical Informatics Association, 2011. **18**(5): p. 601-606.
44. Tang, B., et al., *Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features*. BMC Medical Informatics and Decision Making, 2013. **13**(1): p. 1-10.
45. Zhu, F. and B. Shen, *Combined SVM-CRFs for biological named entity recognition with maximal bidirectional squeezing*. PLoS One, 2012. **7**(6): p. e39230.
46. Abbas, Z., et al., *A Semantic Grid-based E-Learning Framework (SELF)*. IEEE International Symposium on Cluster Computing and the Grid, 2005. CCGrid 2005., 2005. **1**: p. 11-18.
47. Xu, Z., et al. *Fast Inference in Infinite Hidden Relational Models*. in *5th International Workshop on Mining and Learning with Graphs*. 2007.
48. Jamieson, L. and J. Lohmann. *Creating a culture for scholarly and systematic innovation in engineering education. Phase 1 report*. 2009.
49. Johri, A., *Creating theoretical insights in engineering education*. Journal of Engineering Education, 2010. **99**(3): p. 183-184.
50. Kemnitzer, S., *The need for theory-based research in engineering education*, 2008: Video interview recorded at ASEE.
51. Shulman, L.S., *If not now, when? The timeliness of scholarship of the education of engineers*. Journal of Engineering Education, 2005. **94**: p. 11-12.
52. Streveler, R.A. and K. Smith, *Conducting rigorous research in engineering education*. Journal of Engineering Education, 2006. **95**: p. 103-105.
53. Watson, K., *Change in engineering education: where does research fit?* Journal of Engineering Education, 2009. **98**(1): p. 3-4.



### AUTHORS



**Krishna Madhavan** is an Assistant Professor in the School of Engineering Education at Purdue University. He is also the Education Director and co-PI of the NSF-funded Network for Computational Nanotechnology (nanoHUB.org). He specializes in the development and deployment of large-scale data and visualization based platforms for enabling personalized learning. His work also focuses on understanding the impact and diffusion of learning innovations. Dr. Madhavan was the Chair of the IEEE/ACM Supercomputing Education Program 2006 and was the curriculum director for the Supercomputing Education Program 2005.

In January 2008, he was awarded the NSF CAREER award for work on transforming engineering education through learner-centric, adaptive cyber-tools and cyber-environments. He was one of 49 faculty members selected as the nation's top engineering educators and researchers by the US National Academy of Engineering to the Frontiers in Engineering Education symposium.



**Aditya Johri** is an Associate Professor in the Department of Applied Information Technology in the Volgenau School of Engineering, George Mason University, Fairfax, VA, USA. He studies the use of information and communication technologies (ICT) for learning and knowledge sharing, with a focus on cognition in informal environments. He also examines the role of ICT in supporting distributed work among globally dispersed workers and in furthering social development in emerging economies. He is a co-editor of the Cambridge Handbook of Engineering Education Research (CHEER), Cambridge University Press (2014). He can be reached

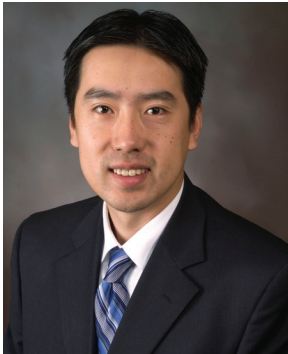
at [ajohri3@gmu.edu](mailto:ajohri3@gmu.edu). More information about him is available at: <http://mason.gmu.edu/~ajohri3>



**Hanjun Xian** is a Research Software Design Engineer in the Bing Data Mining team at Microsoft. He earned his PhD degree in Engineering at Purdue University in 2013. He holds Masters and Bachelors degrees in Computer Science. Dr. Xian's research interests include diffusion of innovations, big data validation and visualization, data mining, social network analysis, and user study.



## Tools for large-scale data analytic examination of relational and epistemic networks in engineering education



**G. Alan Wang** is an Associate Professor in the Department of Business Information Technology, Pamplin College of Business, at Virginia Tech. He received a Ph.D. in Management Information Systems from the University of Arizona, an M.S. in Industrial Engineering from Louisiana State University, and a B.E. in Industrial Management & Engineering from Tianjin University. His research interests include data mining and knowledge discovery, decision support systems, and social media analytics. He has published in *Decision Support Systems*, *Communications of the ACM*, *IEEE Transactions of Systems, Man and Cybernetics (Part A)*, *IEEE Computer*, *Group Decision and Negotiation*, *Journal of the American Society for Information Science and Technology*, and *Journal of the Association for Information Systems*.



**Xiaomo Liu** is a Research Scientist at Thomson Reuters R&D. He received a Ph.D. in Computer Science from Virginia Tech in 2013, an M.S. in Complex Adaptive Systems from Chalmers University of Technology in 2005, and a B.E. in Computer Science and Technology from Nanjing University of Science & Technology in 2003. His research interests include data mining, information retrieval and social computing. He has published more than 10 refereed journal and conference papers. His research appeared in journals such as *Information Systems Frontiers* and conferences such as *International Conference on Information Systems (ICIS)*, *Frontiers in Education Conference (FIE)*, and *China Summer Workshop on Information Management (CSWIM)*.