

Knowledge Creation Analytics for Online Engineering Learning

Hon Jie Teo

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State  
University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy  
in  
Engineering Education

Aditya Johri (chair)  
Vinod K Lohani (co-chair)  
Christopher B. Williams  
Michael A. Evans

June 2, 2014  
Blacksburg, Virginia

Keywords: knowledge creation, k-means clustering, non-parametric analysis, online  
engineering community

# Knowledge Creation Analytics for Online Engineering Learning

Hon Jie Teo

## ABSTRACT

The ubiquitous use of computers and greater accessibility of the Internet have triggered widespread use of educational innovations such as online discussion forums, Wikis, Open Educational Resources, MOOCs, to name a few. These advances have led to the creation of a wide range of instructional videos, written documents and discussion archives by engineering learners seeking to expand their learning and advance their knowledge beyond the engineering classroom. However, it remains a challenging task to assess the quality of knowledge advancement on these learning platforms particularly due to the informal nature of engagement as a whole and the massive amount of learner-generated data. This research addresses this broad challenge through a research approach based on the examination of the state of knowledge advancement, analysis of relationships between variables indicative of knowledge creation and participation in knowledge creation, and identification of groups of learners. The study site is an online engineering community, All About Circuits, that serves 31,219 electrical and electronics engineering learners who contributed 503,908 messages in 65,209 topics. The knowledge creation metaphor provides the guiding theoretical framework for this research. This metaphor is based on a set of related theories that conceptualizes learning as a collaborative process of developing shared knowledge artifacts for the collective benefit of a community of learners. In a knowledge-creating community, the quality of learning and participation can be evaluated by examining the degree of collaboration and the advancement of knowledge artifacts over an extended period of time. Software routines were written in

Python programming language to collect and process more than half a million messages, and to extract user-produced data from 87,263 web pages to examine the use of engineering terms, social networks and engineering artifacts.

Descriptive analysis found that state of knowledge advancement varies across discussion topics and the level of engagement in knowledge creating activities varies across individuals. Non-parametric correlation analysis uncovered strong associations between topic length and knowledge creating activities, and between the total interactions experienced by individuals and individual engagement in knowledge creating activities. On the other hand, the variable of individual total membership period has weak associations with individual engagement in knowledge creating activities. K-means clustering analysis identified the presence of eight clusters of individuals with varying lengths of participation and membership, and Kruskal-Wallis tests confirmed that significant differences between the clusters. Based on a comparative analysis of Kruskal-Wallis Score Means and the examination of descriptive statistics for each cluster, three groups of learners were identified: Disengaged (88% of all individuals), Transient (10%) and Engaged (2%). A comparison of Spearman Correlations between pairs of variables suggests that variable of individual active membership period exhibits stronger association with knowledge creation activities for the group of Disengaged, whereas the variable of individual total interactions exhibits stronger association with knowledge creation activities for the group of Engaged. Limitations of the study are discussed and recommendations for future work are made.

## Table of Contents

Abstract .....	ii
List of Figures .....	vii
List of Tables .....	ix
List of Definition.....	xi
Chapter 1 Introduction and Motivation for Study.....	1
1.1 Wide Range of Opportunities in Online Learning.....	1
1.2 Growth of Online Engineering Learning and Challenges of Assessment .....	3
1.3 Purpose of Study .....	6
1.3.1 Motivation for Study.....	8
1.3.2 Research Questions.....	8
1.4 Significance of Study.....	12
Chapter 2 Literature Review .....	16
2.1 Online Discussion Forums.....	16
2.1.1 Structure and Features of Online Discussion Forums.....	16
2.1.2 Benefits and Limitations of Online Discussion Forums .....	18
2.2 Theoretical Basis.....	21
2.2.1 The Knowledge Creation Metaphor.....	21
2.2.2 Knowledge Artifacts in Knowledge Creating Communities .....	26
2.2.3 Collaboration and Social Interaction in Innovative Knowledge Communities	29
Chapter 3 Methods .....	34
3.1 Data Collection .....	36
3.1.1 Data Collection From Individual Discussion Topics.....	37
3.2 Data Processing.....	42
3.2.1 Processing of Social Network Variables.....	42
3.2.2 Processing of EngineeringTerms .....	46
3.2.3 Processing of Online Resources and Digital Files.....	49
3.3 Data Organization .....	50

3.3.1 Benefits of Approach .....	51
3.4 Research Question 1: Descriptive Statistics .....	53
3.5 Method for Research Question 2 and 3.....	55
3.5.1 Research of Knowledge Creation Environments .....	55
3.5.2 Correlation Analysis .....	57
3.6 Method for Research Question 4 .....	59
3.6.1 Examining Groups in Research of Online Communities.....	59
3.6.2 K-means Clustering Analysis .....	60
3.7 Synthesis of Research Question, Theory and Methods .....	62
3.7.1 Variables Indicative of Knowledge Creation.....	65
Chapter 4 Findings.....	70
4.1. Participation Demographics.....	71
4.1.1 Community Participation.....	71
4.1.2 Topic Participation.....	73
4.1.3 Individual Participation.....	77
4.2 Association between Topic-Level Variables .....	81
4.2.1 Nonparametric Correlation Analysis .....	82
4.3 Association between Individual-Level Variables .....	85
4.3.1 Nonparametric Correlation Analysis .....	86
4.4 K-means Cluster Analysis.....	90
4.4.1 Nonparametric Correlation Analysis Based on Cluster Classification .....	95
Chapter 5 Discussion .....	99
5.1 State of Knowledge Creation .....	99
5.2 Strengths of Associations Between Topic-level Variables with Topic Length, Topic Duration and Topic Views.....	103
5.2.1 Strengths of Associations Between Topic Length and Topic-Level Variables .....	105
5.2.2 Strengths of Associations Between Topic Duration and Individual Variables .....	107
5.2.3 Strengths of Associations Between Topic Views and Topic-Level Variables .....	108

5.3 Strengths of Associations Between Individual-level Variables with Individual Total Interactions, Individual Total Membership Period and Individual Active Membership Period .....	109
5.3.1 Strengths of Associations Between Active Membership Period and Individual Variables .....	110
5.3.2 Strengths of Associations Between Individual Total Interactions and Individual Variables .....	112
5.3.3 Strengths of Associations Between Individual Total Membership Period and Individual Variables.....	113
5.4 Clustering Analysis and Correlation Analysis Based on Classification by Clusters .....	114
Chapter 6 Conclusions .....	118
6.1 Implications of Research.....	118
6.2 Contributions of Research.....	122
6.3 Limitations and Assumptions .....	124
6.4 Future Research Directions.....	125
References .....	128
Appendices.....	148
Appendix A: File Types of Digital Files.....	148
Appendix B: Sample Engineering Terms from IEEE Dictionary of Electrical and Electronics Terms .....	149
Appendix C: Scatterplots for Correlation Analysis .....	175
Appendix D: IRB Approval Letter and Application Documentation .....	181
Appendix E: Python Software Routine.....	187

## List of Figures

Figure 1.1 Main Forum Page for All About Circuits Online Community	... 7
Figure 2.1 The Three Metaphors of Learning	... 22
Figure 3.1 Overview of Data Collection and Analysis	... 34
Figure 3.2 Listing of Discussion Threads in a Sub-Forum (Homework Help Sub-forum)	... 38
Figure 3.3 Discussion Messages in a Discussion Thread	... 39
Figure 3.4 Extraction of Data from Discussion Message	... 40
Figure 3.5 MySQL Database of Processed Online Community Data	... 41
Figure 3.6 Social Network Analysis Process	... 45
Figure 3.7 Use of Gephi Software for Social Network Analysis	... 46
Figure 3.8 Extraction of Web Links from Discussion Message	... 49
Figure 3.9 Potential Variables for Predictive Relationships Between Topic-Level Variables	... 68
Figure 3.10 Potential Variables for Predictive Relationships Between Individual-Level Variables	... 69
Figure 4.1 Cumulative Message Count in Online Engineering Community	... 73
Figure 4.2 Heat Map of Topic Duration in the Online Community	... 76
Figure 4.3 Packed Bubbles Visualization of Individual Post Count	... 79
Figure 4.4 Histogram of Select Topic Level Variables	... 81
Figure 4.5 Histogram of Select Individual Level Variables	... 86
Figure 4.6 Scatter Plots of Individual Total Interactions with Individual Variables	... 88

Suggestive of Knowledge Creation	
Figure 4.7 Identification of Groups of Learners Through K-Means Clustering Analysis	... 91
Figure C-1: Scatter Plots for Individual Total Interactions and Variables Suggestive of Engagement with Knowledge Creation	... 175
Figure C-2: Scatter Plots for Individual Active Membership Period and Variables Suggestive of Engagement with Knowledge Creation	... 176
Figure C-3: Scatter Plots for Individual Total Membership and Variables Suggestive of Engagement with Knowledge Creation	... 177
Figure C-4: Scatter Plots for Topic Length and Variables Suggestive of Engagement with Knowledge Creation	... 178
Figure C-5: Scatter Plots for Topic Duration and Variables Suggestive of Engagement with Knowledge Creation	... 179
Figure C-6: Scatter Plots for Topic Views and Variables Suggestive of Engagement with Knowledge Creation	... 180



## List of Tables

Table 1.1 Online Communities for Electrical and Electronics Engineering Learners	... 4
Table 3.1 Forum Organization of All About Circuit Online Community	... 36
Table 3.2 Sample Interaction Log for Discussion Thread With Topic ID of 70000	... 51
Table 3.3 Research Questions, Theoretical Basis and Methods	... 64
Table 3.4 Variables Suggestive of Knowledge Creation at Topic-Level and Individual-Level	... 66
Table 3.5 Variables That Predict Knowledge Creation at Topic-Level and Individual-Level	... 67
Table 4.1 Participation Parameters at the Community Level	... 72
Table 4.2 Community Message Count from 9/27/2003 to 11/5/2012	... 72
Table 4.3 Descriptive Statistics of Topic-Level Variables	... 74
Table 4.4 Descriptive Statistics of Individual Participation in Community	... 78
Table 4.5 Correlation Between Topic-Level Variables with Topic Length	... 83
Table 4.6 Correlation Between Topic-Level Variables with Topic Duration	... 84
Table 4.7 Correlation Between Topic-Level Variables with Topic Views	... 85
Table 4.8 Correlation Between Individual-Level Variables with Individual Total Interactions	... 87
Table 4.9 Correlation Between Individual-Level Variables with Individual Active Membership Period	... 89

Table 4.10 Correlation Between Individual-Level Variables with Individual Total Membership Period	... 90
Table 4.11 Iterative K-Means Cluster Comparison	... 91
Table 4.12 Descriptive Statistics of Active Participation and Total Membership Period Classified by Cluster	... 93
Table 4.13 Kruskal-Wallis Score Means for Individual-Level Variables Classified by Cluster	... 95
Table 4.14 Non-parametric Correlation Analysis For Individual Variables with Individual Total Interactions, Classified by Cluster	... 96
Table 4.15 Non-parametric Correlation Analysis For Individual Variables with Individual Active Period, Classified by Cluster	... 98
Table A.1 File Types and Corresponding Extensions	... 148

## **List of Definition**

Discussion thread: Hierarchically organized collection of messages whereby the first post initiates the discussion and other messages are replies to earlier messages.

Knowledge creation metaphor: Collection of learning theories that propose that knowledge advancement in an online community is facilitated by the transformation of conceptual or material artifacts through collaboration over a sustained period of time.

Online community: Network of individuals who gather together voluntarily as a social collective to pursue shared interests through a computer-mediated online communication platform.

Online discussion forum: Computer-mediated communication platform to enable individuals to interact asynchronously, with one another without the constraint of time and physical place, through the use of text, image and video content.

Social network analysis: Methodology that is used to examine network links between connected groups of individuals to understand structural patterns and social relations.

## **Chapter 1**

### **Introduction and Motivation for Study**

#### **1.1 Wide Range of Opportunities in Online Learning**

With the advances in information communication technologies (ICT) in recent decades, researchers and policymakers have coalesced around the use of technology to advance learning opportunities beyond the formal environment (Atkins, 2010; Bell et al. 2009; NSF, 2007). In engineering education, researchers and practitioners have embraced computer technology and the Internet to drive innovation in engineering teaching and to enhance access to engineering learning resources (Bourne, Harris & Mayadas, 2005). Spurred by the increased affordability of computers and greater accessibility to the Internet, numerous educational innovations such as online discussion forums, Wiki, Open Educational Resources and MOOCs have been developed with unbounded potential to reach a large number of learners. These advances have in turn attracted the contributions of a wide range of instructional videos, written documents and discussion archives that provide opportunities for learners to expand their learning beyond the classroom. In the formal education system, higher education institutions have begun to offer online courses and distance learning programs to cater to the learning needs of individuals who do not establish a physical presence at educational institutions. A Sloan Consortium report “Online Nation” (Allen & Seaman, 2007) found that approximately 3.5 million students were taking at least one online course in the fall semester of 2006. This figure is double that of a preceding study conducted four years ago in the fall semester of 2002 in which it was reported that 1.6 million students took at least one online course. It was also reported

that public institutions on average offered courses to more than 1,400 online learners each semester. These trends suggest that the use of the Internet and technology-mediated learning platforms, such as online discussion forums, to teach and learn is gaining popularity among students in higher education institutions and K-12 schools.

The growth of online learning is expected to continue as the current generation of students possess learning habits and tendencies that rely on the Internet and information technology (Chubin, Donaldson, Olds & Fleming, 2008; Rideout, Foehr & Roberts, 2010). The current generation of engineering students, referred by many as Net Generation and Millennials, are highly proficient with computer devices and often use the Internet to support their learning outside the classroom (Johri et al., 2014). ‘Net Generation’ or ‘Millennials’ students, born between 1980 and the present (Howe & Strauss, 2000; Oblinger, 2003, Oblinger & Oblinger, 2005), grew up with readily available computer-based technology and almost unrestricted access to the Internet (Tapscott & Williams, 2008). This generation of youths is known to be proficient with digital technology and proficient at multitasking when using electronic and digital devices (Foehr, 2006). They perceive digital technology as omnipresent in their lives and often leverage technology to assist with their learning needs (Kvavik, 2005). Kvavik et al. (2004) who found that almost all the surveyed college students in their study had access to the Internet. They found that over 99% of survey participants used email for academic purposes and over 98% of survey participants used the Internet to access materials for their coursework. It has also been noted that students in STEM majors such as chemistry, computer science, engineering, math, and physics spent considerably more time online than other students (Anderson, 2001). These research studies provide evidence to support the notion that students are

increasingly capable of using the Internet and computer tools for learning. With an increasing amount of learning activities carried out through the Internet on online platforms (Bourne et al., 2005), the studies and reports outline the growing importance of understanding and evaluating engineering student learning on computer-supported learning platforms.

## **1.2 Growth of Online Engineering Learning and Challenges of Assessment**

The use of online communities to foster learning beyond the classroom is well-documented (Bruckmann, 2006; Stahl, 2006; van de Sande, 2012). In the academic area of engineering, there is an increasing presence of online communities across a variety of interest and content areas (Teo et al., 2013). In the area of Electrical and Electronics Engineering (EEE), online communities hosted on discussion forum platforms have attracted a larger number of learners and host an extensive number of contributions (see Table 1.1). The online communities enjoy significant reach which translate to vibrant discussions and massive membership bases, despite not receiving official support or endorsement by formal institutions in any manner. As described by Table 1.1, hundreds of thousands of learners have contributed and posted more than 2 million messages to five online engineering discussion forums. In addition to the forums listed below, newsgroups such as alt.binaries.electronics host thousands of discussion topics over a period of close to twenty years.

Table 1.1

*Online Communities for Electrical and Electronics Engineering Learners*

<b>Site</b>	<b>Message Count</b>	<b>Membership</b>	<b>Year of Establishment</b>
AllAboutCircuits.com	570,000	202,000	2003
Arduino.cc	1,176,000	128,000	2010
EdaBoard.com	1,174,000	483,000	2001
Electro-Tech-Online.com	796,000	191,000	2002
ElectronicsPoint.com	83,000	23,000	2006

*Note.* Descriptive data is rounded off to nearest thousand and accurate as of April 2013

Bourne and colleagues (2005) from the Sloan Consortium examined the status of online learning in American engineering colleges and highlighted opportunities for research and practice. They proposed an assessment framework for online learning based on the “five pillars of learning” and used three performance parameters to assess the viability of online platform in promoting learning (breadth, scale, quality). They urged colleges to explore opportunities in implementing online learning platforms to contribute to improve the quality of online courses, the ability to scale to more learners, and the breadth of coverage of engineering domains/courses. Another assessment consideration of online platforms was floated and focused on the need to provide feedback to learners and teachers. This is captured in National Education Technology Plan (United States Department of Education, 2010) which calls for assessments that provide students, instructors and education stakeholders with timely and actionable feedback about student

learning and participation to improve instruction and student achievement. According to the latest Horizon Report (Johnson, Smith, Willis, Levine, & Haywood, 2011), an annual survey conducted by the New Media Consortium and Educause, a total of six emerging technologies – electronic books, mobiles, augmented reality, game-based learning, gesture-based computing, and learning analytics were projected to gain widespread adoption and exposure in education. The authors of the report argued that these emerging technologies have the potential to disrupt learning and education significantly in the near future. Overall, the report raises new assessment and evaluation challenges for educators as online learning gains prominence and reach. The report has highlighted the challenges of evaluating educational data on online learning platforms such as online discussion forms and learning management systems. With increased student engagement with online learning spaces and use of digital media devices, it is expected that more educational data is expected to be located in online environments and represented in digital form.

There is growing awareness of the need to develop and incorporate ways to assess online learning and the quality of student contributions on online platforms (Bruckmann, 2006; Stahl, 2006; van de Sande, 2012). Various researchers have developed methods to assess learning in informal online communities. van De Sande (2011) examined an online help forum for mathematics and found that learners receive general forms of help that orientate the learners towards resolving homework challenges rather than detailed step that will directly lead to the solution. Her work has focused on analyzing the structure of online communities (van de Sande, 2009), users' experiences of learning on discussion forums and cognitive factors associated with online homework help forums (van de Sande & Greeno, 2012). Her research draws attention to the viability of online communities in



facilitating the learning of mathematics outside the classroom and outlined opportunities to conduct deeper inquiries of online communities for informal learners seeking common interests and purposes. Stahl (2006) has examined the use of online collaborative environments, such as chat platforms and shared digital whiteboards, to support mathematical learning beyond the classroom. In his research, he invited small groups of students to collaborate, discuss and solve ill-structured mathematical problems. He introduced the concept of Group Cognition to describe the group process in which individuals engage in discourse to accomplish a cognitive act. These studies are however carried out in the K-12 mathematics and a review of educational research literature suggests that assessment of learning on online engineering communities have been given scarce research attention.

### **1.3 Purpose of Study**

This research introduces an evaluation approach based on the description of the state of knowledge advancement, examination of relationships between variables indicative and supportive of knowledge creation, and identification of groups of learners. The site of study for this research is an online engineering community named “All About Circuits.” The online community can be accessed using a web browser through the global URL (Uniform Resource Locator) address of <http://www.allaboutcircuits.com>. The vBulletin® online discussion software (Internet Brands, 2013) was installed on web servers to provide a online asynchronous discussion platform to facilitate communication between members of this online community. As of November 1, 2012, a total of 182,783 registered members contributed a total of 503,908 messages over 5 discussion sections

(see Section 4.1). According to the titles of the main discussion sections, the main areas of discussion include electronics, circuits, software, communications and embedded systems.



Figure 1.1: Main Forum Page for All About Circuits Online Community.

This research draws from the theoretical basis surrounding the educational metaphor of knowledge creation to assess and evaluate the specified online engineering forums. The overarching focus is on deepening the understanding of learning and participation in online communities by evaluating the strengths of association between discussion characteristics and activities that suggest engagement with knowledge creation on online environments, and by identifying groups of learners based on their participation tendencies. This research takes into consideration that there is limited presence of active pedagogical support and instructor presence in an online learning environment. The motivation to address this broad research area is pertinent as we seek to further our understanding of how to build effective, sustainable and self-renewing online communities

that foster innovation and advancement of knowledge in engineering. These projections are consistent with the needs in the digital age as demonstrated by the insatiable appetite for online learning and engagement with online communities by new generations of learners.

### **1.3.1 Motivation for Study**

The author is motivated to undertake this research due to his experiences utilizing online discussion forums in the course of his undergraduate and graduate studies in Electrical Engineering. As an undergraduate in University of Minnesota, he majored in electrical engineering and took a variety of elective courses from the sophomore to senior level. In various engineering courses, he faced several challenges in understanding the material. This is at this point when he accessed a number of forums to clarify some of his understanding of electrical engineering concepts. He has learned much from this experience and felt that online communities provide a valuable source of information and rich silos of resources. Through further engagements with the online communities, the author has gained opportunities to make contributions and advance the knowledge of the community, which in turn has boosted his understanding of engineering content. Prior to conducting the research, the author believes that online forums are rich sites of learning that offer highly accessible opportunities to active participants to advance and create knowledge in engineering.

### **1.3.2 Research Questions**

The focus of the research is on understanding and assessing the state of knowledge creation of an online engineering community “All About Circuits” through the examination

of linguistic features, online artifacts and social interactions. The theoretical basis is based on the knowledge creation metaphor, which is considered as a collection of learning theories that propose that knowledge creation (in an community of learners) is facilitated by the transformation of conceptual or material artifacts through collaboration over a sustained period of time (see Section 2.2). The evaluation process is focused on uncovering patterns of use of engineering terms, the use of digital forms of engineering artifacts and the formation of social networks in the online discussions carried out by Electrical and Electronics learners. The strengths of association between characteristics and activities suggestive of engagement are also examined at both the individual and topic levels. This research will identify groups of learners based on their participation tendencies and verify if the strength of associations between individual characteristics and knowledge creation activities vary across the identified groups. It is hypothesized that numerous online community learners will draw on a wide range of linguistic features, participate in extensive use of engineering artifacts and form extensive social networks and that strong associations will be found between the identified variables identified in the literature and variables suggestive of engagement with knowledge creation.

The over-arching research question for this research is:

**What is the state of knowledge creation in online engineering communities?**

The over-arching research question will be examined by considering a case study of "All About Circuits" online community. Further, the following sub-research questions will be investigated to answer the above research question:

**1. What is the state of engineering knowledge creation at the topic and individual levels?**

This descriptive question is focused on examining the state of engineering knowledge creation through the use of engineering terms, formation of social networks and advancement of engineering artifacts in the forum. The first part of this descriptive question addresses the frequency counts of engineering term indicative of engineering devices and concepts based on the IEEE Standard Dictionary of Electrical and Electronic Terms. The second part of this descriptive question will examine the trends of use of engineering artifacts to facilitate discussion on engineering related topics. There will also be a focus on uncovering the frequency counts of web links, links to digital files hosted online and the contribution attachments by individual users. The answers to this question should include usage statistics of various types of engineering artifacts and order in which they are used. The third part of this descriptive question is focused on evaluating the degree of interaction among learners through social network analysis. The research outcomes to this question will include in/out degree of the participating learners and the network size of the learners including information about the number of established individuals (with extensive contribution and long period of continued contributions) in each discussion topic.

**2. What is the relationship between topic length, duration and views associated with participation in knowledge creation activities at the topic level?**

This research question aims to examine the strength of association between pairs of variables based on three variables representative of activities that promote knowledge

creation and six variables suggestive of engagement with knowledge creation at the topic level. The thread variables include calculated values of topic length (the total number of messages in a topic), topic duration (the time period between first post and the last post) and topic views (the total number of views received by a topic). The variables that are suggestive of knowledge creation in topics include Network Size, Quoted References, Engineering Terms, Messages, Resources and Files. In order to examine the strength of association between the chosen variables, two non-parametric correlation techniques (Spearman's Rho and Kendall's Tau) are employed.

### **3. What is the relationship between individual total interactions, active period and total membership period with individual participation in knowledge creation activities?**

This research question aims to investigate the strength of association pairs of variables from three variables representative of activities that promote knowledge creation and five variables suggestive of engagement with knowledge creating activities at the individual level. The three individual characteristics include calculated values of total active period (the time period between an individual first post and last post), total membership period (the time period between an individual registration date and end date of data collection) and individual total interactions (the total number of interactions with other learners in unique discussion topics). The variables that suggest engagement with knowledge creating activities include Quoted References, Engineering Terms, Messages, Resources and Files.

In order to examine the strength of association between the chosen variables, two non-parametric correlation techniques (Spearman's Rho and Kendall's Tau) are employed.

**4. How can learners be grouped based on their individual total interactions and active membership period? How do the correlation statistics vary across groups?**

This research question aims to uncover distinct groups of individuals in the online community based on similarity in their active membership period and total membership period in the online community. Based on the clustering analysis, individuals are classified by their clusters to examine the grouped individuals' correlations between Individual Total Interactions, Individual Active Period or Individual Total Membership Period and Individual Variables such as Quoted References, Engineering Terms, Messages, Resources and Files. The research outcomes to this research question includes the optimal number of clusters based on k-means clustering, descriptive statistics of individual characteristics in each cluster and the score means from Kruskal-Wallis Tests. In order to examine the strength of association between the chosen variables, two non-parametric correlation techniques (Spearman's Rho and Kendall's Tau) are employed.

#### **1.4 Significance of Study**

In recent years, there is increased focus on informal spaces of learning beyond the classrooms such as afterschool clubs (Barton & Tan, 2010), sports clubs (Nasir & Hand, 2008), online discussion forums (van de Sande, 2010) and online chat rooms (Stahl, 2006). With supporting evidence from their studies, researchers are quick to argue that learners pick up critical learning attitudes and knowledge through these avenues. A common

characteristic of these informal spaces of learning is that there is an absence of formal educational institutional influence: there is limited involvement by trained educational practitioners and a lack of alignment with institutional curricula. The focus of this research, All About Circuits online engineering community, is distinct from formal educational settings as expertise is voluntarily available and learners are not supported by education practitioners, but by a small number of established individuals with extensive tenure or records of contributions. Furthermore, the All About Circuits online engineering community adopts an open registration policy in which anyone with an email account can set up an forum account. This open registration policy therefore feature a large number of volunteers with different lengths of tenure and expertise levels. A review of educational research literature unveiled a number of studies such as online programming community (Bruckmann, 2006; Resnick et al., 2009) and online mathematics help forums (van der Sande & Leinhardt, 2007) and Virtual Maths Chat (Stahl, 2006). However, none of these studies have focused on online communities catered to engineering learners. With no prior studies conducted on online engineering communities supported by computer-mediated platforms, this research will significantly increase our understanding of whether online engineering communities are essential sites of learning and participation beyond the classroom.

Online communities are supported by online discussion forum software which allows for a large number of participants and accumulated archives of learning activities. The presence of a large number of learner-generated data and information draws attention to the difficulty of using manual assessment methods for the evaluation of online learning environments. This broadly stated challenge is of importance as it allows engineering



educators to uncover how educational institutions contrast against informal online settings and to understand how to develop successful online communities in support of formal education. While it has been argued that online discussion fosters social knowledge construction (Fischer & Weinberger, 2006), these distinct features also a gap of knowledge in the sense that it is unclear if informal online discussion are productive spaces for learning. This research aims to make an contribution to this knowledge pool by presenting an assessment approach that leverage descriptive statistics, social network analysis and text analytical processing techniques to characterize knowledge creation in an online community powered by an discussion forum platform. The discussed approach has immediate practical implications for online education and distance learning. This is particularly so as online discussion forums feature heavily in online distance learning courses (Garrison, 2007), in course management software such as Blackboard and Moodle (Unal & Unal, 2011) and MOOC (Mak et al., 2010).

This research outlines a plan to answer calls from engineering education stakeholders and leaders to leverage information technology in support of innovations in educational assessment (Bourne, Harris & Mayadas, 2005; Jamieson & Lohmann, 2009). Findings from this study can potentially inform the development of an analytic framework that enhances both the breadth and depth of the assessment of learning through computer-mediated discussion forums. The described analytical approaches in this research serve to assist educators and practitioners to derive a more holistic approach for the evaluation of knowledge creation in computer-mediated learning environments, by considering the use of engineering terms, quality of social interactions and advancement of engineering artifacts.

Furthermore, student use of discussion forums is expected to grow as an increasing number of educational innovations (such as MOOCs) are leveraging discussion forum technology to support a large number of learners. Based on a nexus of social network analysis and text analytical techniques, this research approach can contribute to the expansion of the existing repertoire of knowledge creation analytical approaches in online learning environments. To accomplish the goals of this research, a software routine was developed with Python programming language and applied to process multiple forms of learner trace data and discussion parameters from HTML web pages. It is potentially feasible to expand this software routine into an executable software package that is capable of processing webpages and outputting assessment reports. This envisioned software package is expected to be capable of extracting discussion and participation data from online communities supported by off-the-shelf commercial software such as the vBulletin software. As the vBulletin discussion platform is a popular choice for many online communities, this software routine can prove to be useful to researchers seeking to understand and assess online communities supported by this discussion platform.

## **Chapter 2**

### **Literature Review**

#### **2.1 Online Discussion Forums**

The “All About Circuits” online community is supported by an online discussion forum software (vBulletin) and this text-based platform facilitates asynchronous communication among registered members. Due to this setting, it will be relevant to review literature that is centered on the use of discussion forums to facilitate learning or education. Doing so will allow for a better understanding of the state of research regarding discussion forums and its role in learning. The review of research studies, conducted in the setting of online discussion forums, is presented in two main sections. Section 2.1.1 elaborates on the structure and features of online discussion forums with a focus on the functionality and how discussion forums are used by online learners. Section 2.1.2 describes the benefits and limitations of online discussion forums. This section discusses research that focus on effects of online discussion on student learning and participation.

##### **2.1.1 Structure and Features of Online Discussion Forums**

An online discussion forum refers to a predominantly text-based computer-mediated communication platform that enables individuals to interact asynchronously with one another without the constraint of time and physical place (Hew, Cheung & Ng, 2010; Naidu & Järvelä, 2006). Online discussion forums are platforms that rely primarily on asynchronous text-based communication between participants through the exchange of written contributions that foster collaboration on a group task. They are typically installed

on a computer server and accessed through a web browser on the Internet. This process commonly involves learners who write their opinions, analyze others' comments, respond to others' comments and reflect on their learning process. Discussion is carried on topic threads created by the learner or instructor to foster interaction and exchange of knowledge about a specific topic, which is summatively described by the title of the discussion thread. In this online platform, a discussion thread refers to a hierarchically organized collection of messages whereby the first post starts the discussion and all other subsequent messages are written as either replies to earlier (Hewitt, 2005). The messages are then sequentially arranged by time of post and organized into various discussion threads for reading (Hewitt, 2005). Online discussion is initiated when learners post a message on a discussion thread for others to read and respond to. A sustained dialogue or discussion is formed among learners when messages are exchange to and fro as responses and counter-responses.

There are various features that set online discussion forums apart from face-to-face classroom discussion. Online discussion forums are an environment that provides pervasive access to learners from any physical location and are generally available 24 hours a day which reduces the time restriction on learning and increases the accessibility to learning resources (Harasim, 1990). This allows for interaction between students and instructors on coursework to be extended beyond the class hours (Garrison et al. 2000). Learning via the use of online discussion forums is also pervasive in the sense that students' contributions are recorded, and can be visited many times and at any time (Hew et al., 2010). Another compelling feature of online discussion forums is "interactivity" (Harasim, 1990) which can be understood as a reciprocal process whereby learners are engaged in interaction through the process of posting original messages, reading and replying to the

messages of other learners. The exchanges are asynchronous and contribute to a culture of reflection as they allow time for learners to reflect on others' contributions and on their own writing before making preparations for subsequent contributions (Poole, 2000). The exchanges are not only text-based but can be complemented by rich multimedia content such as video (Snelson, 2008), and linked to external repositories of educational content (Cantor, 2009).

Since its introduction as a form of educational technology, online discussion forums have played an increasingly central role in online, blended and traditional courses to support teaching and enhance the quality of learning (Naidu & Järvelä, 2006). Formal education institutions have used online discussion forums as part of courseware (such as Moodle and Blackboard) and the online discussion forums typically serve as means to facilitate communication among learners and teachers. While discussion forums feature in formal coursework, there is increasing educational interest in understanding how to develop and utilize discussion forums to enhance learning beyond the classroom. For instance, researchers have developed discussion forums customized for the “knowledge building” pedagogy (Scardamalia & Bereiter, 2006) and for computer-mediated discussion and chatting on K-12 mathematical topics (Stahl, 2006). Overall, a review of literature found increased use of online discussion forums to foster and facilitate learning in and beyond the classroom.

### **2.1.2 Benefits and Limitations of Online Discussion Forums**

Online discussion forums emphasize on learner-centered dialogue in collaboration with others and facilitate the sharing of multiple perspectives. This focus on social

interaction and meaning making is consistent with the social constructivist paradigm towards learning (Harvard, 1997). While practitioners have used online discussion forums to facilitate collaboration activities such as problem solving activities, researchers found that online discussion enhances knowledge construction through focused task-oriented discussion (Schellens & Valcke, 2006). Online discussion forums are also noted to facilitate individual meaning making in which learners bring along their own unique experiences to interact with one another to construct shared understandings and engage in sharing of their own unique experiences (Pena-Shaff & Nicholls, 2004), Online discussion forums have been leveraged to facilitate inquiry and knowledge building where learners build knowledge and engage in idea improvement for benefit of the community (Scardamalia & Bereiter, 2006). The potential of online discussion forums are influenced by the role and tasks assigned by instructors: researchers have noted that learners are more likely to participate to engage in knowledge construction in an environment suited for role-based discussion and argumentation (Weinberger & Fischer, 2006; Wise & Chiu, 2011).

The asynchronous nature of exchange allows learners to deliberate, compose a reply, and to reflect on their own and others' comments at their own pace, all of which allow opportunities to thoroughly think through their writings and not just to write answers to questions but to critically reflect upon them (Vonderwell, 2003). With no time limit posed on the composition process, participants tend to compose messages that are concise and relevant with attention paid to punctuation, grammar and spelling (Garrison, 2003). In comparison to face-to-face learning, online discussion fosters more engaged and on-task learning (Garrison, 2007). Jonassen and Kwon (2001) studied group problem solving activities and found that learners engaged in online discussion contributed lesser messages

on average than face-to-face communication but comparatively, more messages were on-task. Furthermore, the process of expressing their opinions in writing and proactive revision of their ideas facilitate high level learning including analysis, synthesis and evaluation (Newman, Johnson, Webb & Cochrane, 1997), which allow learners to undergo a personal and cognitively complex learning process (Tam, 2000).

On the other hand, researchers have raised various concerns about the inability of discussion forums to facilitate high order thinking and engaged collaboration without the presence and intervention of an instructor. For instance, Cheung and Hew (2004) found that the majority of students on online discussion forums achieve low levels of participation marked by surface-level comments (Cheung & Hew, 2004). Similarly, Thomas (2002) suggest that students pay limited attention to the contents of other students' posts and often write replies that are not connected to other students' writings. Garrison and colleagues (2001) suggest that online discussion forums may not be able to facilitate high order thinking without the active involvement of the instructor in the design of discussion activities and in supporting discussion as they evolve (Garrison, Anderson, & Archer, 2001). This concern is echoed by Dennen (2005) who found that students learn most when instructors have high online presence and provide frequent feedback (Dennen, 2005).

Overall, a review of education literature and research performed in the setting of online discussion forums highlight the potential of online discussion forums for facilitating positive learning experiences through active involvement of the course instructors. It is also noted that most educational researchers examined online discussion forums that are tied to coursework or curriculum in a setting where efforts are either mandated by coursework or given academic credit. A review of literature on EBSCO database targeted

studies focused on online communities catered to informal engineering learners from the educational research stand point and found no studies with similar research focus. It is however noted that in other domains such as K-12 education, various researchers from the computer-supported collaborative learning (CSCL) domain have studied collaborative learning in informal online communities. The researchers are namely van De Sande (2010) who studied online homework help forums for Mathematics and Stahl (2006) who studied math virtual chat. Their research highlights the effectiveness of discussion forums and chat rooms in supporting learning.

## **2.2 Theoretical Basis**

The theoretical basis for this research is based on a set of learning theories with specific focus on knowledge creation in online communities. Section 2.2.1 highlights the origins of the knowledge creation metaphor together with the key principles and considerations of this theoretical basis. This section includes the discussion of theories that the metaphor is based on. Section 2.2.2 describes in detail what knowledge artifact means and prior work that has examined the development knowledge artifacts in collaborative learning environments. Section 2.2.3 describes the pertinence of social interactions and collaboration in knowledge creating communities.

### **2.2.1 The Knowledge Creation Metaphor**

The knowledge creation metaphor refers to a synthesis of a set of learning theories such as *Knowledge Building* (Scardamalia & Bereiter, 2002), *Expansive Learning* (Engeström, 1999) and *Organizational Knowledge Creation* (Nonaka & Takeuchi, 1995). These theories are commonly targeted at understanding collaboration



and learning through a focus on the development of new knowledge in communities of learners in school or organizational settings (Paavola et al., 2004; Paavola et al., 2009). The origins of the conceptualization of the knowledge creation metaphor can be traced to the introduction of the third metaphor of learning by Paavola and colleagues (2004) in response to the two metaphors of learning (Sfard, 1998). The acquisition metaphor suggests that learning can be understood through an individual gain in knowledge within his/her mind (Piaget 1970; Vgotsky, 1978) whereas the participation metaphor argues that learning is facilitated through social interactions between individuals in a community of practice (Brown & Duguid, 1991; Lave & Wegner, 1991).

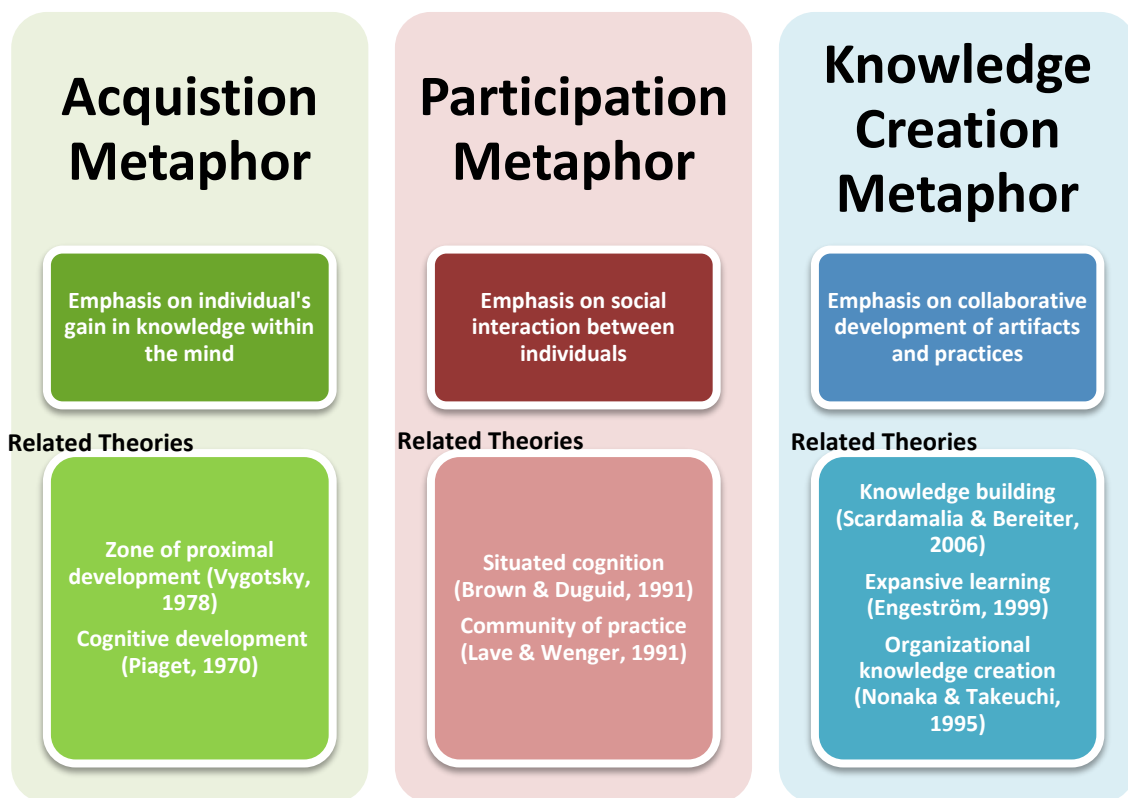


Figure 2.1: The Three Metaphors of Learning

To understand the standing of the third metaphor of learning – *knowledge creation*, it will be useful to first revisit the first two metaphors of learning (*acquisition*

and *participation*). Sfard (1998) introduced the acquisition and participation metaphors of learning to discuss how learning theories have been conceptualized by two dominant camps of researchers. Sfard's (1998) research was focused on the classification and synthesis of learning theories into two metaphors, and she argued that educators will not gain a complete picture of learning if they are to prefer a metaphor over the other. The distinction between the acquisition and participation metaphors can be best understood through a contrast of what learning means and a view of knowledge according to each metaphor. In the acquisition metaphor, the human mind is suggested to be a 'container' for storing knowledge and learning occurs when a teacher fill the 'container' with content and knowledge. On the other hand, the participation metaphor suggests that people learn when they engage in dialogue with other learners and by doing. In this view, learning is situated in individual lives in a certain social and cultural setting, and centers on the process of becoming a member of a community. The participation metaphor suggests that learning involves acquisition of skills to communicate and to act within the socially negotiated norms. According to the acquisition metaphor, the knowledge exist in the world on its own or in the minds of individuals whereas the participation metaphor suggests that knowledge is an aspect of participation and engagement in cultural practice, and that knowledge is accessible as a result of enculturation and interaction with other community members.

Paavola and colleagues (2004) introduced a third metaphor termed knowledge creation to supplement the distinction between the acquisition and participation metaphors. The knowledge creation metaphor combines the acquisition and participation metaphors: it posits that individuals participate in collaborative activities in a community

which allow them to acquire individual knowledge and create new knowledge that is usable for the community at large. These outcomes are evident through the engagement in the co-development of artifacts and practices. The knowledge creation metaphor of learning suggests that learning is understood as participation in processes of inquiry, typically focused on inquiry about something that is either new or substantially enhanced through a discovery and transformative process.

Educational researchers who introduced the knowledge building theory have proposed numerous principles and ideas that can be used to conceptualize the knowledge creation metaphor (Bereiter, 2002). In a knowledge building community, all participants have access and the rights to contribute which means that the knowledge is therefore commonly owned and forged by the participants (Scardamalia & Bereiter, 2006). The creation of knowledge can be understood as a collaborative process of building knowledge in which participants engage to co-construct knowledge through social interactions and engagement. In this process, the participants acquire knowledge and create new knowledge that is accessible and usable by a broader number of participants. Participants of the knowledge building community focus on authentic problems to identify gaps in their understanding of the problem at hand, improve the understanding of the problem to work towards solving it and to provide a diversity of ideas to advance the goals of collaborative work (Bereiter, 2002). The providence of ideas can be facilitated by sharing deepened explanations and developments of ideas with the entire community and by considering the overview of critiques and alternatives to better the ideas. Gaps of understanding may surface when an idea develops. Learning activities can then be undertaken to facilitate the refinement and collective advancement of the understanding

of the idea. This is manifested as learners work together collectively under the support of the instructor to advance and elaborate on knowledge artifacts over series of inquiry cycles conducted over a sustained period of time (Paavola et al., 2009; Scardamalia & Bereiter, 2006).

Paavola et al. (2004) also drew on the expansive learning theory (Engeström, 1999). Expansive learning theory was developed by Engestrom (1999) based on studies of learning cycles in work teams through the use of Cultural Historical Activity Theory. Expansive learning is understood as a process which undergoes transformation of social practices with an emphasis on the practices and activities that are transformed in collective processes involved in knowledge advancement (Engeström, 1999). The cycles of expansive learning are beyond the discussion in this research, but can be understood as a process in which learners question existing practices with the goal of modeling, proposing and testing new solutions for implementation. The theory views learning as an object-oriented activity structured by division of labor within communities supported by share values and mediated by tools (Engeström, 1999). Learners interact to collaboratively and jointly constructing knowledge objects which can be either abstract or concrete entities (Engeström, 1987), in pursuit of the advancement of shared knowledge (Paavola & Hakkarainen, 2005).

Overall, the knowledge creation metaphor represents a collection of theories with common principles that frame and deepen the understanding of knowledge advancement in online communities. In this research, the creation of discussion topics is viewed as an initiation of the process of knowledge creation whereas subsequent discussion based on the topic is understood as process of advancing existing shared understanding and

knowledge. One, it suggests that social interaction is key to knowledge creation. Participation in an online community through discussion in the online forums can be understood as a collaborative process of individuals working together to create knowledge for the collective benefit of the community. Thus, the frequency and the degree of collaboration will describe the extent of social interactions in support of knowledge creation processes. Second, the overarching theoretical framework suggests that knowledge artifacts are produced as a result of knowledge creation. These knowledge artifacts are thought to be either conceptual or material and are commonly shared, used and manipulated (Lipponen et al., 2004). The engagement of learners in collaborative activities leads to the development of new ideas and transformation of knowledge artifacts, which are indications that learning is taking place among the learners.

### **2.2.2 Knowledge Artifacts in Knowledge Creating Communities**

In the view of knowledge creation, learners embody or objectify knowledge by “putting” knowledge onto artifacts. Artifacts can be understood as either material or conceptual. The description of a conceptual artifact can be traced back to Bereiter’s description that a conceptual artifact can exist in the form of ideas, theories or model which are expressed and mediated by language or computers. Bereiter’s description was based on Popper’s World 3 (Popper, 1972) where it is argued that this realm include conceptual things such as theories and models. According to Popper (1972), it is important to consider this realm because human individuals develop understandings in the conceptual space in addition to physical or mental states. In other words, individuals engage with a world of cultural knowledge to produce cultural or conceptual artifacts that

are distinct from what is contained the minds of the participants, such as mental processes within the human brain. The produced conceptual artifact can be understood as a working object that is shared amongst a group of individuals and is constantly tested as learners engage in open-ended inquiry. On the other hand, a material artifact refers to physical artifacts that are evolved as a result of linguistic and social practices. Some examples of physical artifacts are drawings, prototypes and molds.

Conceptual artifacts are described as products of objects of thinking and reasoning that learners can access to foster an understanding and to improve upon (Bereiter, 2002). The shared artifacts can refer to vocabularies, taxonomies and ontologies (Locoro, Mascardi & Scapolla, 2010; Sun et al., 2010), and as theories, ideas, questions and models (Lipponen et al., 2004). A common characteristic is that both conceptual and material artifacts are susceptible to changes and subjected to continuous transformation in knowledge creating communities. In other words, it is not fixed and typically unstable. This is in line with the opportunities for engagement with intentional efforts to advance knowledge and suggests that shared artifacts are shared for manipulation and transformation by participants in the community. This also means that learners in the community will have the potential to surpass their existing achievement and advance their prior understandings.

Seitamaa-Hakkarainen and colleagues (2010) studied conceptual and material artifacts created by teachers and elementary school students in a science project based on the topic of light that lasted 13 months of collaborative inquiry across three phases. In total, the students created 1906 notes in examination of the past, present and future phases. They found that students worked in decreasingly smaller number of views as they

gained focus on their projects. For instance, they would be working in 14 team views as they evaluated the history of the lamp and converged towards 3 views towards the design of a lamp in the future phase of the project. In this research, the researchers were also interested in the knowledge practices in the classroom, with particular attention on those initiated by the teacher. To understand the enacted practices in the classroom, the researchers analyzed the contributions on the knowledge forum as well as the project diary which was updated by the teacher. Their findings suggest that the teacher played the role of an organizer who was focused on structuring and directing collaborative activities that foster knowledge creation rather than controlling multiple aspects of students' learning.

Researchers have found evidence to support the notion that learners constantly transform and modify conceptual artifacts in an innovative community (Chen et al., 2012; Zhang et al., 2007). Chen et al. (2012) developed a tool to explore concepts that demonstrate "promisingness" (or big ideas) and facilitate the selection of ideas by students. Based on the ideas that students have selected, the authors found evidence to support that the students work around conceptual artifacts through questions and facts. For instance, they observed that 40% of the "big ideas" are made up of facts and questions central to the discussion theme. Zhang and colleagues (2007) studied four months of online discourse data based on knowledge building discussion by 22 elementary grade students. They found that students have actively generated conceptual artifacts such as theories in their online discourse and designed experiments to obtain empirical data in examination of the theories that they have developed.

Overall, a review of literature highlight that shared objects of activities such as conceptual and material artifacts are subjected to advancement and transformation in innovative and knowledge creating communities. Conceptual artifacts such as ideas are continually improved by individuals who come together to collectively improve on and progressively make inquiries about the overarching topic over an extended period of time. As conceptual artifacts are typically mediated by language and computers, this also means that a researcher will be able to examine knowledge creation by analyzing linguistic features in learners' dialogue and by examining traces of learning on online learning environments. It is with this deepened understanding of theoretical underpinnings and empirical illustrations of knowledge creation that the research is positioned to focus to examine conceptual artifacts through engineering terms and material artifacts through engineering artifacts.

### **2.2.3 Collaboration and Social Interaction in Innovative Knowledge Communities**

Lipponen and colleagues (2004) have stressed on the importance of collaborative activities in the process of knowledge creation. Essentially, knowledge creation is fundamentally a social process, and the emergence of new ideas and understanding is evident among people rather than within individuals (Paavola et al., 2002). While individual activities are essential activities in knowledge creating environments, collaborative activities in contrast represent a social stream of activities and interactions where participants engage with one another. In other words, engagement with collaborative activities represents a collective series of tries and efforts to understand the matter at hand. Individual participants in a collaborative activity may initially possess incomplete or differing knowledge to accomplish the learning task but as they work with



each other, they collectively improve their understanding in this social interaction process while developing shared understandings of the topic at hand.

In a body of work that studies knowledge creation in Japanese organizations, Nonaka and Takeuchi (1995) argue that the knowledge creation begins with socialization, which is described as a phase that begins with individuals sharing their knowledge and experiences with other individuals in their group. This phase suggests that not only individuals are expected to interact with each other but they have to interact and collaborate closely with other individuals in order to develop a common understanding at the group level. This phase can be iterative and may involve extensive dialogues that feature disagreement and potential conflict, which eventually lead to individual challenging themselves to view their experiences and knowledge from another perspective. It is common that questions and problems of understanding will drive the knowledge creation process (Bereiter, 2002). Questioning and criticism of accepted practices forms the basis for the expansive learning cycle (Engeström, 1999) and often feature as part of social interaction and discussion between individuals (Nonaka & Takeuchi, 1995).

The quality of social interaction can be examined through social network variables derived through social network analysis (see Chapter 3.2.1 for a detailed discussion the method used to process and derive social network variables in this research). Social network analysis has been leveraged in studies of collaboration and social interaction in knowledge building communities. Sha and van Aalst (2003) used social network analysis on server log data of two classes of students that used the knowledge building software – the Knowledge Forum. Their research is targeted at

assessing participation and interactivity by evaluating individual measures that resulted from the use of forum features such as writing notes or posting comments with systemic network measures. Social network analysis aided the researchers with the identification of favorable online affordances and they found that students who used the forum features more extensively had higher measures of network reciprocity. They also found that there was a tendency for students to read and build on the notes of other students, rather than posting isolated messages.

Social network analysis is used extensively to uncover social interactions in learning environments that support knowledge creation. Palonen and Hakkarainen (2000) examined a total of 493 written communicative comments on the knowledge building software, Computer-Supported Intentional Learning Environments (CSILE), produced by 28 elementary grade students in a public school. With the use of social network analysis, they viewed students as network nodes and comments as vertices. Their research reveals that high density of interactions was accompanied with large individual differences in participation. The researchers also found that while the density of interactions was high and all students participated in the forums, students tend to interact with students of their own gender groups and that the achievement levels of students seem to influence the intensity of participation. They concluded that social network analysis is a technique that is capable of providing information about interaction patterns and structures from the interaction culture of students supported by knowledge building software and in this research, social network analysis allowed them to uncover gender differences in online participation.

Philip (2010) employed social network analysis to examine the frequency of face-to-face interactions in Grade 5 class and corresponding online interactions on the Knowledge Forums. Through a visualization example, the author suggested that social network analysis may be a useful approach for identifying students who is central and participating frequently in the forums, and for identifying students who are not participating in the knowledge building discussion. There was little variation in the in-degree centralities of individuals in social network obtained from student interactions from the activities of reading online notes. As the researcher examined a classroom with active use of the knowledge building pedagogy for the communication of ideas, it was also reported that face-to-face interactions appear to be sparse compared to interactions on the knowledge forums.

Software tools have been developed to support the visualization of social networks in knowledge building discourse data. Oshima et al. (2012) developed the Knowledge Building Discourse Explorer (KBDEX) aimed at visualizing network structures in discourse data. Through the analysis of two sets of data, the KBDEX was able to reveal the pivotal points in discourse that facilitated knowledge advancement and identify participants that contribute to the social advancement of knowledge. The findings from social network analysis mirror that of discourse analysis, where it is observed that both methods were able to arrive at the same conclusion about the quality of social knowledge advancement by comparing two sets of discourse data. They commented that while their tool was capable of identifying interactions that lead to knowledge creation, future research needs to consider the comparison of the learning processes of learners

with a set of benchmark data set by experts, in order to make valid interpretations of the social advancement progress.

In sum, this section has provided a discussion of literature focused on examining social interactions among students in knowledge creating communities. The reviewed research activities suggest that the frequency of social interaction and the directedness of the interaction have been widely characterized as variables that are indicative of individual engagement in the process of knowledge advancement. Social network analysis appears to be a popular method to examine social structures and social interactions in knowledge creating communities. Studies have leveraged social network analysis to pinpoint social practices in the classroom such as the identification of active contributors in the classroom and gender differences in student group participation. The theoretical basis and the empirical studies of social interactions in online communities directly inform the research design of this research. The focus of the research targets the examination of the quality of in-thread interactions and degree of collaboration between learners of differing participation levels. The data processing and organization approaches (see Section 3.2 and Section 3.3) will contribute to the derivation of network variables that will be used to examine the state of social interactions among learners in the All About Circuits online community.

## Chapter 3

### Methods

The research activities described in this research can be presented as four main stages in the following order: data collection, data processing, data organization and data analysis (see Figure 3.1). The research study began with data collection where the goal was to download all accessible discussion data in the form of web pages from the discussion forum of the All About Circuits online community (Section 3.1). The data processing phase begins after data collection by processing web pages using techniques such as (i.e. social network analysis and text analysis) to process social network data, discussion message content and user trace data to ready them for further organization (see Section 3.2).

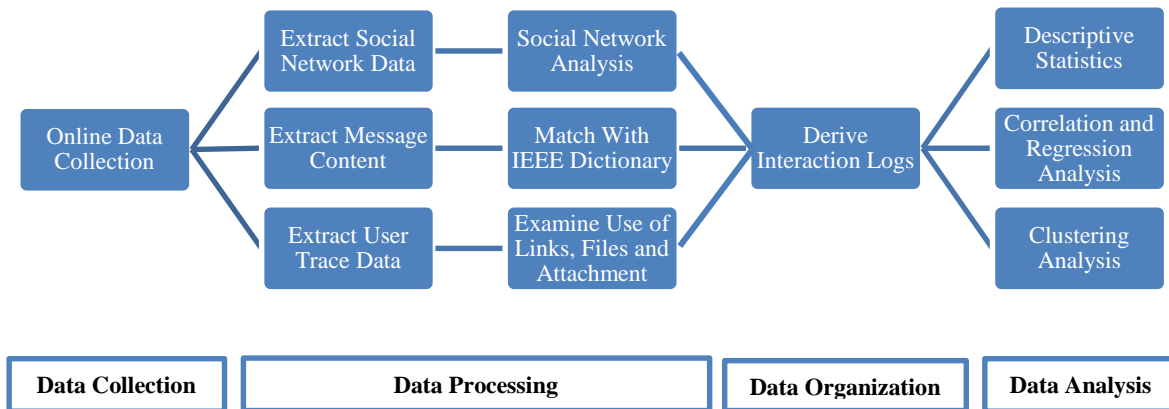


Figure 3.1: Overview of Data Collection and Analysis.

The third stage of the research – data organization – involves the derivation of interaction logs that describe participation data, social networks, use of engineering terms and artifacts at both the individual level and topic level to be captured and stored in a MySQL database (see Section 3.3). This is accomplished by means of a Python data processing software routine written to computationally organize interaction logs for every discussion topic ( $n = 65,209$ ) and to allow for the compilation of individual parameters for each individual ( $n = 31,219$ )

The last stage of this research involves quantitative data analysis (see Section 3.4). Descriptive statistic is first computed for all variables that are suggestive of engagement of knowledge creation and variables that are deemed to be supportive of knowledge creation, at both the topic and individual levels. Then non-parametric Spearman and Kendall correlation analysis is performed to examine the relationship between variables that are supportive of knowledge creation and variables that are suggestive of knowledge creation, at both the topic and the individual levels. K-means clustering analysis is then performed to detect clusters of individuals. Then, clusters with similar participation characteristics are classified into groups with the goal of examining the variation of correlation statistics across these groups. Overall, the data analysis phase employs the approaches of descriptive statistics, non-parametric correlation analysis and k-means clustering analysis. Section 3.5 and Section 3.6 delineates the rationale and justification for the choice of these statistical techniques. The last section of this chapter – Section 3.7 – synthesizes the research question, employed approaches and variables used to characterize engagement with knowledge creation in this online community.

### 3.1 Data Collection

Raw data for this research comes in the form of HTML pages. Data is collected by downloading a total of 87,264 HTML pages through accessing the forum software for the online community. A discussion of the forum organization will benefit the understanding of the forum structure and how the data was collected. As shown in Figure 3.2, the online community is supported by the Vbulletin online discussion forum software and organized into various academic areas of interest. There are four main sections of forums in the site of study (see Table 3.1). For example, there are four sub-forums in the Electronics discussion section and they are namely General Electronics Chat, The Project Forums, Homework Help and Electronics Resources. Data collection has focused on all the online sub-forums through the use of an automated downloader where the focus will be on ensuring all the web pages for each discussion topic are downloaded to facilitate the derivation and examination of the use of engineering terms, web links to external resources, social networks for all active and dormant discussion topics.

Table 3.1:

*Forum Organization of All About Circuit Online Community*

Discussion Section	# of Sub-Forums	Titles of Sub-Forums
Electronics	4	General Electronics Chat, The Project Forums, Homework Help, Electronics Resources
Software, Micro-computing, and Communications	4	Programmer's Corner, Embedded Systems and Microcontrollers, Computing and Networks, Radio and Communications

Discussion Section	# of Sub-Forums	Titles of Sub-Forums
Circuits and Projects	1	The Completed Projects Collection
Abstract	3	Maths, Physics, General Science
Community	3	Off-Topic, The Flea Market, Feedback and Suggestions

### 3.1.1 Data Collection From Individual Discussion Topics

Each online discussion thread is created by any individual with a registered forum account and represents a contained discussion described by a distinct title in the discussion forums and placed within a discussion section (see Table 3.1). Figure 3.2 displays the first page of the forum and lists the most active questions which have most recently received a reply or comment at the time of capture. The subsequent pages list the questions according to the order in which they last received a response. In order to download the discussion contents and the responses in the community forums, a Python Program based on the software library “Beautiful Soup” was written to look for topic IDs within these listing pages. The purpose of using this program is to compile a comprehensive list of web pages to download from the web site rather than to download the web pages based on an iterative algorithm that may cause unnecessary network stress on the web hosts. This list will include web links of the web pages that contain all the discussion topics and corresponding pages of discussion messages.



VOL. I - DC VOL. II - AC VOL. III - SEMICONDUCTORS VOL. IV - DIGITAL VOL. V - REFERENCE VOL. VI - EXPERIMENTS WORK SHEETS VIDEO 8

**All About Circuits**

All About Circuits Forum > Electronics Forums  
 Homework Help

User Name:  User Name ☐ Remember Me?  
 Password:

[Register](#) [Blogs](#) [FAQ](#) [Community](#) [Today's Posts](#) [Search](#)

**Homework Help** Stuck on a textbook question or coursework? Cramming for a test and need help understanding something? Post your questions and attempts here and let others help.

[New Thread](#) Page 1 of 437 1 2 3 4 5 6 7 8 9 10 11 51 101 > Last »

Threads in Forum : Homework Help Forum Tools Search this Forum

	Thread / Thread Starter	Rating	Last Post	Replies	Views
	Sticky: <a href="#">LaTeX Tutorial - The AAC Mathematical Formula Editor</a> Georacer		01-02-2012 10:51 PM by Georacer	2	4,348
	Sticky: <a href="#">PLEASE READ - Posting Questions in the Homework Help Forum</a> Dave		04-30-2011 03:58 PM by berfus	1	19,515
	<a href="#">Linearity questions</a> sukalpa mishra		Today 06:56 PM by wayneh	2	106
	<a href="#">Binary search trees from node</a> (1 2 3 ... Last Page) zulf100		Today 06:52 PM by WBahn	35	815
	<a href="#">Measure fuel level in closed tank ?</a> (1 2) ArFa		Today 06:27 PM by wayneh	10	167
	<a href="#">Expression for Current in a Small Signal Differential Amplifier</a> HunterDX77M		Today 05:25 PM by HunterDX77M	4	124
	<a href="#">9-0 down counter using d flip flop</a> annbarbie01		Today 04:41 PM by MrChips	2	52
	<a href="#">differentially compounded motor/inrush current</a> skiddman		Today 03:51 PM by MaxHeadRoom	1	50
	<a href="#">Transfer function for a system</a> mo2015mo		Today 02:42 PM by mo2015mo	0	106

Figure 3.2: Listing of Discussion Threads in a Sub-Forum (Homework Help)

The downloaded web pages represent ongoing or archived discussion topic; an example of a web page is presented in Figure 3.3. In Figure 3.3, the web page shows two messages; the first message of the discussion topic refers to a question initiated by the user “anhnhha” and the second message is a reply to the question by user “Bill\_Marsden”. Every web page has a maximum of 10 messages on one web page. There are 12 messages in this discussion thread titled “Monostable Multivibrator help” which means that 2 web pages will be downloaded and the second web page will contain 2 messages.



Figure 3.3: Discussion Messages in a Discussion Thread

With the successful collection of all targeted web pages, a Python program was used to parse each individual message on each downloaded web page. As stated earlier, each web page contains a maximum of a total of 10 messages and therefore, additional precaution was taken to identify discussion threads that feature more than 10 messages and to facilitate extraction of data in multiple web pages when the discussion topic comprises of more than 10 messages.

The Python program utilized the Python library “BeautifulSoup” for the purpose of processing and extracting HTML tags that contain information about the learning traces for each discussion message. The functional purpose of this program was to process each message on the web pages and capture replying user information, receiving user information, message content, order of occurrence, and time and date of post of message (see Figure 3.4). The process was then repeated for extraction process for all discussion topics in AllAboutCircuits.com and organized on a SQL database – a common output format for Python (see Figure 3.5). The end result of the data collection was an organized database of relevant data that had been extracted from each downloaded webpage.

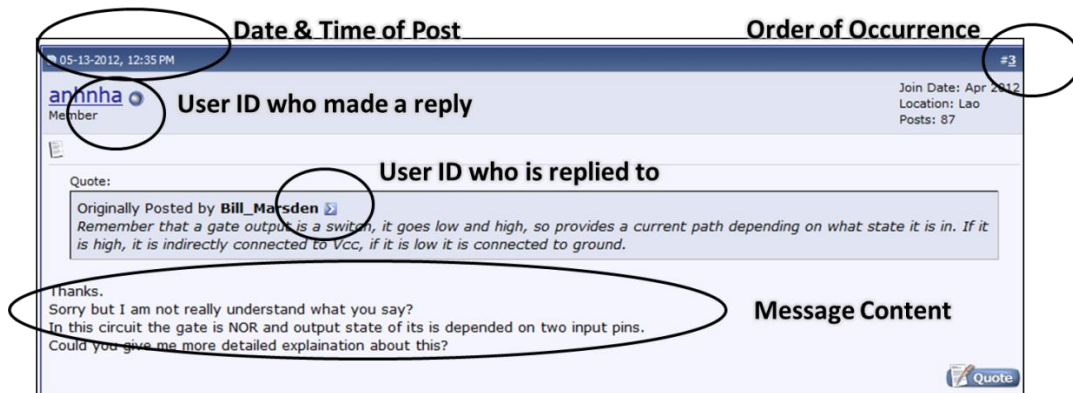


Figure 3.4: Extraction of Data from Discussion Message

No.	User_ID	Post_Date	Post_Timing	Post_ID	Post_Order	Topic_ID	Poster_ID	Role_of_ID	Topic_Total_Count	Message_content
99972	12669	2009-01-05	07:36 PM	109304	5	17683	37519	Helper	13	I have a bench-mounted (clamp-on) lighted magnifier. It ...
99973	28424	2009-01-06	04:15 AM	109396	6	17683	37519	Helper	13	Hey Greg [If you have 20/20 vision, you do not need any...
99974	10771	2009-01-06	12:55 PM	109443	7	17683	37519	Helper	13	I have a Luxo 17113 that I got at a clearance sale. I bel...
99975	3092	2009-01-06	02:25 PM	109449	8	17683	37519	Helper	13	Depends entirely on one's eyes. I used to be able to vi...
99976	37519	2009-01-06	05:08 PM	109468	9	17683	37519	Seeker	13	Thanks for all of the help, y'all. I may split the difference ...
99977	6856	2009-01-06	10:29 PM	109547	10	17683	37519	Helper	13	Hi [I have both one of the really expensive bench mount...
99978	28424	2009-01-07	03:45 AM	109599	11	17683	37519	Helper	13	Greg [If it turns on your wife, go for it! (Dan ...
99979	37519	2009-01-07	03:56 AM	109602	12	17683	37519	Seeker	13	Dan [It doesn't exactly turn her on, she would simply roll ...
99980	37519	2009-01-08	08:53 PM	110015	13	17683	37519	Seeker	13	Resolution:]I purchased a pair of 3X reading glasses (for...
99981	36969	2009-01-05	06:19 PM	109281	1	17694	36969	Seeker	1	Hello friends [I am doing my project on vhd_ams... [I hav...
99982	15094	2009-01-05	08:28 PM	109314	1	17696	15094	Seeker	16	Since we are debating enhancing the AAC back on trans...
99983	19834	2009-01-05	09:14 PM	109322	2	17696	15094	Helper	16	Actually Diamond is a semiconductor, since carbon falls in ...
99984	3092	2009-01-06	03:46 AM	109383	3	17696	15094	Helper	16	Do carbon, silicon, and germanium readily form cations? D...
99985	21582	2009-03-30	06:09 PM	130608	4	17696	15094	Helper	16	Gold is used in computers.
99986	19834	2009-03-30	11:54 PM	130709	5	17696	15094	Helper	16	Graphene transistors are well under development, graphen...
99987	19834	2009-04-03	03:43 AM	131783	6	17696	15094	Helper	16	Another story on Graphene...[http://www.physorg.com/...
99988	55308	2009-07-14	05:30 AM	156513	7	17696	15094	Helper	16	I saw the title Gang of four and thought Sheesh, maybe I...
99989	15094	2009-07-14	02:03 PM	156558	8	17696	15094	Seeker	16	Quote:]I saw the title Gang of four and thought Sheesh ...
99990	51987	2009-07-14	05:01 PM	156602	9	17696	15094	Helper	16	Good luck with the diamond transistor! Man, will those b...
99991	507	2009-07-14	06:43 PM	156631	10	17696	15094	Helper	16	Graphene has quite possibly made diamond obsolete - ex...
99992	19834	2009-07-16	01:39 PM	157037	11	17696	15094	Helper	16	Actually the current cost of diamonds is pretty artificial. I...
99993	507	2009-07-17	08:08 PM	157328	12	17696	15094	Helper	16	I can't give a link, but there is at least one outfit that will ...
99994	51987	2009-07-17	10:20 PM	157367	13	17696	15094	Helper	16	This is pretty interesting...so how well is graphene and di...
99995	19834	2009-07-18	12:52 AM	157376	14	17696	15094	Helper	16	Using dodec in either polarity? That sounds like a zener....
99996	507	2009-07-18	08:45 PM	157469	15	17696	15094	Helper	16	The structure of diamond does not conduct electricity, bu...
99997	8620	2009-07-19	06:38 AM	157604	16	17696	15094	Helper	16	Diamond's structure can support conduction. However, it ...
99998	37670	2009-01-05	09:16 PM	109323	1	17690	37670	Seeker	3	I have two audio effects units that have foot switch/ope...
99999	20420	2009-01-05	09:24 PM	109324	2	17690	37670	Helper	3	Hello [Do you have the schematics of those effect boxes...
100000	4483	2009-01-06	03:17 AM	109377	3	17690	37670	Helper	3	I don't see any problems if you can find out what voltage...

Figure 3.5: MySQL Database of Processed Online Community Data

In sum, data collection was been completed and a total of 4.08 Gigabytes of files was downloaded from AllAboutCircuits.com – corresponding to a total of 87,263 files HTML web pages representative of 65,209 discussion topics. These raw HTML pages however required additional data processing with focus on extracting rudimentary user trace data. The next phase, data processing, specifically focused on parsing and storing parameters of interest include a range of information including date of message posted, message ID, posting user ID, replying user ID and message contents. The third phase – data organization – derived 65,209 interaction logs from the entire dataset, and they consist of social network, linguistic and engineering artifacts from a database consisting of processed user and message parameters.

## **3.2 Data Processing**

A software routine based on the Python programming language was developed to process the three described categories of variables from the collected data. The design of the Python program was based on two main Python libraries (BeautifulSoup and NLTK) and built with the functionality of extracting the engineering terms, use of engineering artifacts from messages and to extract social network data from interaction trace data. The data extraction approaches have been utilized in a similar fashion in prior research on online communities (Teo & Johri, 2014; Teo et al., 2013). In line with the identified categories of variables in the review of knowledge creation literature, the data processing focused on three main sets of variables in social network (see Section 3.2.1), use of engineering terms (see Section 3.2.2) and advancement of online engineering artifacts (see Section 3.2.3).

### **3.2.1 Processing of Social Network Variables**

A social network is defined as a set of links between connected groups of individuals and is described by the set of individuals (nodes) and the relationship (ties) between the individuals (Wasserman et al., 1994). The characteristics of the linkages between individuals can be used to understand and interpret the social behavior of individuals in the networked community (Wasserman et al., 1994). It is understood in social network theory that high levels of reciprocal interactions represent strong ties between individuals where on the other hand, weak ties are characterized by non-reciprocal and low levels of interactions (Granovetter, 1973). Nodes represent individuals in a social network and an edge represents a social connection between two nodes or individuals (Wasserman et al., 1994).

Social network analysis (SNA) is a quantitative method used to derive person to person relations from communication traces in a networked community (Haythornthwaite, 2011; Wasserman et al., 1994). SNA allows the researcher to analyze the structural patterns of social relationships in a social network (Haythornthwaite, 2011; Wassermann et al., 1994). In educational research, social network analysis has emerged as a major research perspective in online learning research (Siemens, 2012; Wellman, 2012). SNA has found credibility as a research approach that is capable of illuminating interaction processes in of networks of learners based on relational properties and structures (Haythornthwaite, 2011; Suther et al., 2012). In this perspective, learning can be viewed as a relation that connects learners and as a network outcome of relations supported by interactions (Haythornthwaite, 2011, Ferguson et al., 2012). Social network analytics allows one to gain insights into the practices and interests of a social group (Haythornthwaite & de Laat, 2010), and to examine interactions between individuals based on sharing common knowledge and practices (Haythornthwaite 2006; 2008). The potential of social network analysis can be realized in an online learning setting where the quality of interactions between individuals can be examined by analyzing relational activities and events.

Within a classroom context, both learners and teachers can leverage data visualization and recommendation systems to guide classroom learning experiences (Duval, 2011; Verbert et al., 2011). These advances present insightful examples of how practitioners and researchers can evaluate the nature of interactions between learners, to understand the impact of learning activities and make evidence-backed pedagogical decisions. Primary uses of SNA include clarifying relationships and interactions between learners in a MOOC coursework (Fournier et al., 2011). For instance, Cambridge and

Perez-Lopez (2012) studied an online community of professional teachers and their interactions with content objects using bimodal social network analysis. They identified highly influential individuals through egocentric usage maps and found that these users are persistently engaged with the communities over a sustained period of time. Suthers and Chu (2012) studied a professional network of educators using social network analysis techniques such as community detection algorithms and illuminated prominent actors and described six major communities within the social network.

Advances in software development suggest that social network analysis can feed into a subsequent form of analysis to deepen one's understanding of learning, which supports the idea to complement text analytical techniques with SNA in this research. Software tools have been developed in the field of learning analytics to support social network analysis. Social Networks Adapting Pedagogical Practice or SNAPP (Bakhari & Dawson, 2011) is one tool which is capable of extracting data from existing Learning Management Systems (LMS) such as Blackboard and Moodle to visualize the discussion activities, mainly retroactive analysis of student interaction. In a demonstration of this tool, Macfadyen and Dawson (2010) utilized SNAPP to extract user data from Blackboard Vista LMS to model if students are "at risk of failure". Zhuhadar and Yang (2012) devised the design of a recommender system which is capable of proposing learning resources to the learners facilitated by data mining techniques and social network analysis on user log files.

#### **3.2.1.1 Use of Social Network Analysis**

The social network analysis process for this research is to be carried out on the Gephi™ software platform (Bastian et al., 2009). Gephi is an open source software that is capable of visualizing network graphs and calculating the individual network

characteristics in a social network. This software was written in Java and is free-of charge. To perform social network analysis on Gephi, the Gephi software has to be installed on a Windows PC provided that additional steps are taken to allocate more memory to the software (the default memory allocated to Java-based programs is 96mb and is insufficient for this research). By importing network data into Gephi, this research will leverage the program for three main calculations that will determine the individual social network characteristics and provide information about the frequencies of interaction in the forums (see Figure 3.6). They are the out-degree distribution for each user, the in-degree distribution for each user and the network size of each individual user.

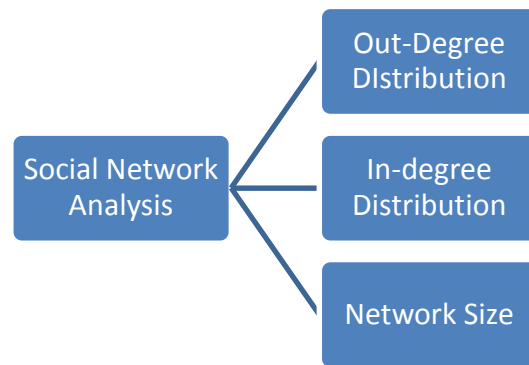


Figure 3.6: Social Network Analysis Process

In-degree and out-degree measures the direction and degree of interactions to and from a node (or individual) in the social network. The out-degree distribution refers to the number of connections made by a node to other nodes and it will be used to examine a user's frequency of contact with other users. The in-degree distribution refers to the number of connections received by a node from other nodes and it will be used to examine an individual's frequency of received interactions from other users. Network size is defined as the number of direct and unique individuals that have interacted as a



sender or receiver with an individual. It is also understood as the total number of direct ties that connect an individual to other individuals in the social network. The three measures (in degree, out degree and network size) were directly read off Gephi once the network data has been derived and imported into Gephi. In Figure 3.7, the use of Gephi is illustrated whereby there are two nodes (Node 1 and Node 2) are connected by an edge (interaction).

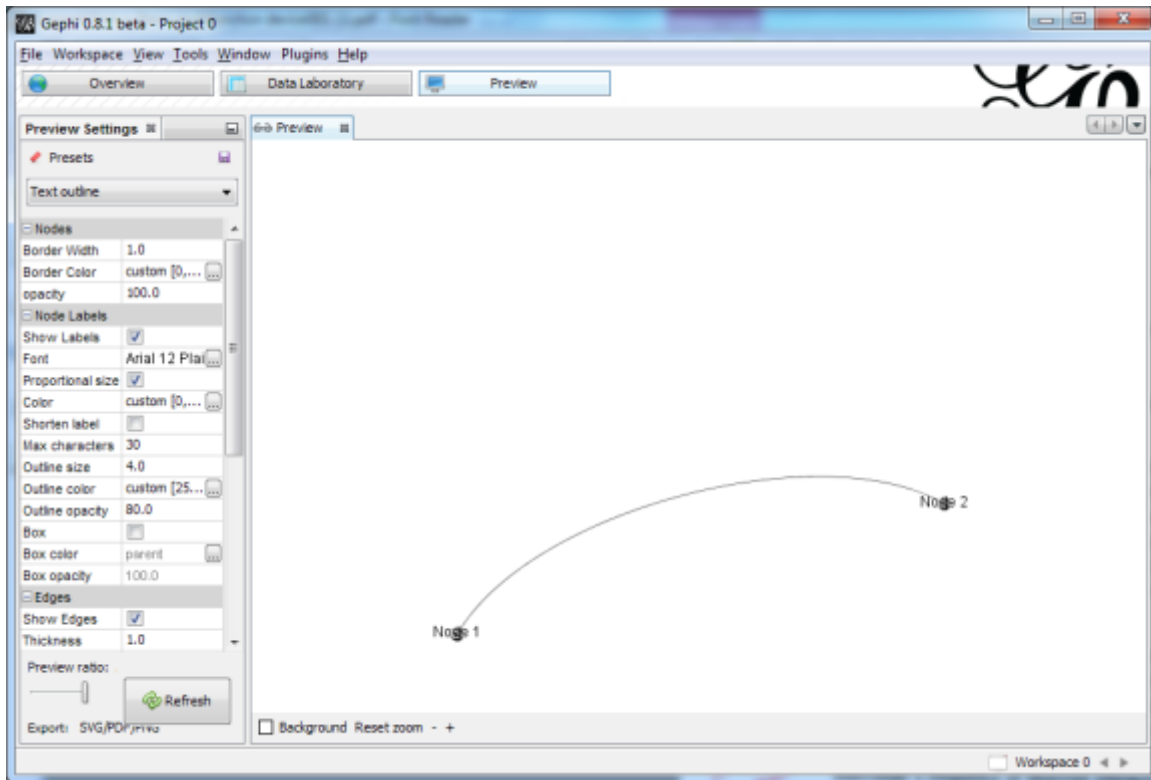


Figure 3.7: Use of Gephi Software for Social Network Analysis

### 3.2.2 Processing of EngineeringTerms

Engineering terms may refer to engineering concepts, devices, abbreviations and standard units that are parts of a sentence that provides information to understand textual content within an engineering context. One can grasp the contexts in which a noun is

used by referring to the corresponding segment of one word (unigrams) and segment of two words (bigrams). The purpose of this research activity is therefore to identify learners' use of concepts and devices as part of their online discussion pertaining to topics related to the Electrical and Electronics Engineering domain. In prior research, the author has adopted a similar approach to conduct text analysis or word count analysis to identify all linguistic features including nouns, verbs or pronouns (Teo et al., 2013). This prior study uncovered words that learners frequently use which includes a range of engineering terms and words associated with socialization, and demonstrated that text analysis is capable of identifying key engineering concepts and devices in student conversations.

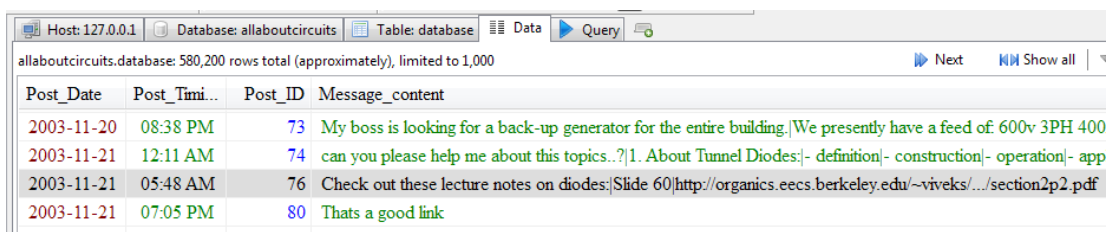
The extraction of engineering terms from the IEEE Dictionary of Electrical and Electronic Terms (Radatz, 1997) will be performed on the Python programming platform and this process is aided extensively by a software package named Natural Language Toolkit (NLTK). One popular and efficient Python library used for natural language processing and computational linguistic analysis is the Natural Language Toolkit (NLTK). The NLTK is an established open source Python toolkit with very proficient libraries for text processing and analysis (Bird et al., 2009). NLTK has been well-utilized in educational research to understand learning and interaction among students. Haythornthwaite and Gruzd (2007) utilized NLTK to explore noun phrases in an online bulletin board for a graduate class and based on the prevalence of words such as “thanks” and “agree”, they argued that the bulletin board was a supportive avenue for learning. On the other hand, Worsley and Blikstein (2010) used NLTK to explore students' speech about electronics and mechanical devices and found that speech markers (such as adverbial modifiers) are

indicative of domain-specific expertise. While these research works are exploratory in nature, they highlight the potential of NLTK in deepening our understanding of students' discourse and provide ways of assessing written contributions in online environments. These studies also suggest that linguistic markers are capable predictors of the quality of learning.

A list of engineering terms was extracted from the IEEE Standard Dictionary of Electrical and Electronics Terms (Radatz, 1997). The PDF files for the dictionary were first downloaded and converted to Microsoft Word format. Then, the bolded words that represent unique engineering terms were extracted from each Word file. In this research, a total of 12,739 engineering terms was extracted from the terms listed in the IEEE Standard Dictionary of Electrical and Electronics Terms (see Appendix B for an alphabetically ordered sample). The goal of extracting a list of engineering terms from the IEEE Standard Dictionary of Electrical and Electronics Terms is to allow for the comparison against the database of learner-produced written contributions in the All About Circuits online community (Radatz, 1997). A software routine, based on the Python programming language and the NLTK software library, will facilitate the process of counting the frequency of engineering terms used in the dataset against the derived database of engineering terms and derivation of the frequency counts for each discussion message in a separate database. The expected outcomes from this language processing technique are a list of engineering terms from the IEEE Standard Dictionary of Electrical and Electronics Terms and a database that shows the frequency counts of the use of engineering terms.

### 3.2.3 Processing of Online Resources and Digital Files

The focus of the third research sub-question is on the examination of the use of engineering artifacts (such as the web links to external resources, schematics, sketches and CAD files) in online discussion carried on discussion topics in the engineering online community. In order to examine this research question, data processing is carried out on a total of 503,908 messages contained in the database of 65,209 discussion topics. This process is carried out by scanning each message for the presence of digital file attachments and URL web links. An example of this process is illustrated in Figure 3.8, where the highlighted row represents a message (with the message identification number of 76) of a discussion topic (with the topic identification number of 26) created on 11/21/2003. In this illustration, the described process of scanning the text contents of this discussion message results in the capture of a web link to an external resource hosted a computer server hosted by the University of California, Berkeley. The URL is: <http://organics.eecs.berkeley.edu/~viveks/ee130/lectures/section2p2.pdf>. It is expected that the process is repeated for every message in the data set to derive frequency counts of use of web links and attachments.



Post_Date	Post_Time	Post_ID	Message_content
2003-11-20	08:38 PM	73	My boss is looking for a back-up generator for the entire building. We presently have a feed of 600v 3PH 400
2003-11-21	12:11 AM	74	can you please help me about this topics..?1. About Tunnel Diodes:- definition- construction- operation- app
2003-11-21	05:48 AM	76	Check out these lecture notes on diodes; Slide 60  <a href="http://organics.eecs.berkeley.edu/~viveks/.../section2p2.pdf">http://organics.eecs.berkeley.edu/~viveks/.../section2p2.pdf</a>
2003-11-21	07:05 PM	80	Thats a good link

Figure 3.8 Extraction of Web Links from Discussion Message

### 3.3 Data Organization

To verify the viability of the research approach, a Python prototype program was created to extract the parameters of interest (as suggested by the three categories of engineering artifacts, social network and engineering terms) from the comprehensive MySQL database of processed community data. The goal of the approach is to perform arithmetic calculations to derive the parameters from the 1<sup>st</sup> column to 7<sup>th</sup> column of Table 3.2. The computational derivation of the engineering terms (in the 8<sup>th</sup> and 9<sup>th</sup> row of Table 3.2) is functional. A list of engineering terms from the alphabet A to Z will be derived from the IEEE dictionary of Electrical and Electronics (See Appendix A).

For the purpose of presenting a sample process of data collection and analysis, the engineering terms (described in the 8<sup>th</sup> column of Table 3.2) are derived through computational means. Table 3.2 describes an interaction log for discussion topic with identification number of 70,000 and each row contains parameters that will inform the examination of three research sub-questions. For instance, the 2<sup>nd</sup> row of Table 4 suggests that the 2<sup>nd</sup> message of the discussion topic was sent from user 19834 to user 163782 in which an hour has elapsed since the discussion topic has created. This discussion message has resulted in an out degree of 1 and network size of 1 for user 19834, and has resulted in seven engineering terms (*gate, state, circuits, capacitor, voltage, input and output*). The use of interaction logs will facilitate the descriptive examination of parameters across the three knowledge creation dimensions in each discussion topic. Overall, a total of 65,209 interaction logs corresponding to a similar number of discussion topics are derived and stored in a MySQL database.

Table 3.2

*Sample Interaction Log for Discussion Thread With Topic ID of 70000*

Order	From User	To User	In/Out Degree	Network Size	Count of Eng. Artifacts	Time Elapse	Engineering Terms	Term Count
1	163782	-	0/0	0	1	0	Gate, State, Circuits, Capacitor, Voltage, Input, Output,	7
2	19834	163782	0/1	1	0	1	Gate, output, switch, current, ground	5
3	163782	19834	1/1	1	0	1	Circuit, output, gate, state, input	5
4	20420	163782	0/1	1	0	1	Gate, TTL	2
5	163782	20420	2/2	2		2	Gate, logic, capacitor, circuit	4
6	41645	163782	0/1	1	3	2	Pulse, Input, Gate, Voltage, Capacitor, Resistor	6
7	163782	19834, 20420, 41645	3/5	3	1	2	Gate, Output, Capacitor	3
8	41645	163782	1/2	1	1	3	Gate, CMOS, input, diode, circuit, current, resistor	7
9	163782	19834, 20420, 41645	4/8	3	0	3	Pulse, plate, capacitor, logic	
10	163782	41645	4/9	3	0	4	-	0
11	41645	163782	2/4	1	0	5	Capacitor, Circuit	2
12	2867	163782	0/1	1	0	10	Circuit	1

### 3.3.1 Benefits of Approach

This research leverages the Python programming platform and accompanying software libraries to derive structured data sets, which in turn facilitated the examination of linguistic features, social networks and engineering artifacts contained within each discussion topic. The benefits of this approach include step-by-step organization of data

that accounts for temporal data distribution, relatively quick process and accuracy of data organization. The benefits are described in more details in the following paragraphs.

The contributions made by each participant are saliently represented according to the order in which they occur. Each row of the interaction log consists of information that represents the most current information at the point of capture. The advantage of capturing accurate and up-to-date data through the interaction logs is evident in Table 3.2. In Table 3.2, a sample interaction log is derived from the discussion topic with the identification number of 70,000. It will be useful to examine, for instance, the 5<sup>th</sup> row of the interaction log. The 5<sup>th</sup> row of the interaction log is representative of the 5<sup>th</sup> message in the discussion topic and was contributed by the learner with the identification number of 163782. The contributions took place about 2 hours after the creation of the discussion topic. The learner has experienced a total of 4 interactions with other participants which can be further described as 1 incoming and 1 outgoing interaction with each of the users with the identification numbers of 19834 and 204020, as described 4<sup>th</sup> in the 7<sup>th</sup> column of the 5<sup>th</sup> row. The learner has used four engineering terms (*logic, gate, capacitor and circuit*) and three engineering artifacts in the form of three attachments in the discussion topic. Overall, the interaction log is capable of capturing up-to-date parameters central to research focus on engineering terms, engineering artifacts and social networks.

The social network characteristics of individual users can be swiftly calculated from the interaction log. From Table 3.2, the social interaction data from the 2<sup>nd</sup> and 4<sup>th</sup> columns in each row of the interaction log will allow the researcher to conduct social network analysis at the individual level as well as to examine social network data associated with each discussion topic. These parameters from the interaction logs will be

critical for the examination of social network characteristic and patterns throughout at the individual and community level.

This data organization approach allows for the organization of parameters across the three dimensions of interest (social interactions, conceptual artifacts and material artifacts) from each discussion topic into one interaction log. The organization of data into interaction logs is of importance as each discussion topic is distinct from the others in terms of its focus and the user who created it. This manner of organization will allow the analysis process to be framed by the discussion topic in which learning took place. Furthermore, the approach is accurate and error-free. A manual check was conducted to verify if the derived parameters matches that of the actual web pages. This manual check was performed by going through line by line to verify the accuracy of extracted parameters from 1<sup>st</sup> to 7<sup>th</sup> column for the total of 12 messages in the discussion topic (with identification number of 70000). Based on the comparison, it is verified that the prototype software routine is accurate and precise in calculation and computation.

### **3.4 Research Question 1: Descriptive Statistics**

The descriptive research approach is focused on describing existing conditions by examining individuals, groups and institutions with the goal of describing what, how or why a particular event of interest occurs (Fraenkel, Wallen & Hyun, 2011). This allows the researcher to interpret extant information and provide details on a given state of affairs by examining practices, the ongoing processes and the trends that are developing (Cohen, Manion and Morrison, 2007). This is attained by focusing on the calculation and tabulation of descriptive statistics which include the statistical values of mean, mode, variance and range of the examined variables (social network parameters, engineering



words and web links in this research) for the purposes of analysis, comparison and interpretation of the educational entities and events of interest (Creswell, 2009).

The selection of a descriptive approach in this study requires a more detailed discussion. As established earlier, a review of literature suggested that limited studies have examined knowledge advancement in an online engineering community. The research addresses problems that are novel in its examined context with no directly related prior work to build on. Thus, the use of a descriptive approach will yield information from the examined states to provide for a more insightful theoretical account of knowledge creation in online communities oriented towards engineering content. The purpose of selecting a descriptive approach will allow for a more informative description of the events leading to knowledge creation and understanding of the processes in online discussion as they unfold in the context of online communities. A descriptive approach will help focus attention on the characteristics of what parameters are the most important for the assessment of knowledge advancement and which events will likely lead to online discussion featuring rich knowledge creation processes. These research projections are expected to deepen the discussion of the question – are online communities rich sites of knowledge creation?

With the employment of a descriptive statistical approach, the purpose of the first research question is to address the broad research question of the state and level of knowledge creation on All About Circuits. The description research is focused on examining two variables that describe the characteristics and quality of created knowledge and three network variables that describe social interactions in the forums. The two variables describing the quality of knowledge creation are the use of engineering

terms and the use of web links to external resources. The three network variables that describe social interactions are individual network measures in degree, out degree and network size. This descriptive set up is used as research question is targeted at examining the state of knowledge creation and it matches the need to address the broad research approach of characterizing and measuring the advancement of knowledge.

### **3.5 Method for Research Question 2 and 3**

To address Research Question 2 and 3, this research will require a statistical approach that is capable of examining the relationship between the variables representative of knowledge creation activities in an online learning environments. In order to select the most appropriate approach, Section 3.5.1 describes a literature review of the research focused on the examination of relationships between distinct activities carried out by individual learners within online environments. Based on the literature review, two non-parametric statistical approaches (Spearman's Rho and Kendall's Tau) were selected and the statistical background of the two approaches is described in Section 3.5.2.

#### **3.5.1 Research of Knowledge Creation Environments**

A review of literature in the domain of online learning suggests that non-parametric correlation analysis approaches such as Spearman's Rho and Kendall's Tau has been frequently used to examine relationships between variables in knowledge building and creation environments. Chuy et al. (2013) studied science dialogue in an elementary school classroom setting where the teacher committed to the knowledge building pedagogy. Their findings revealed that the majority of students' contributions were dedicated to theorizing and working with evidence, based on a list of six major

contribution types. Based on non-parametric Spearman correlation analysis, the researchers found that theory improvement in scientific dialogue is moderately associated with an ability to use evidence or references to support an idea whereas learner formulation of explanatory questions is moderately associated with the explanation proposal and synthesis of ideas. Hong et al. (2010) studied knowledge building in a university course in teacher education where the knowledge building theory is used to guide the course with the Knowledge Forum as a central point for discussion. They found that learners tended to perceive the climate of the knowledge building environment they were engaged as highly supportive for knowledge creation. Non-parametric Spearman correlation analysis found that students who contributed more written notes to the community also tended to be consistently more active in other contributive activities in a knowledge building environment.

Gan (2008) studied drawings or sketches generated by students in the process of production of ideas and writing in an elementary school and found that children who drew more while writing produced more ideas and used more words in longer written sentences. Non-parametric Spearman correlation analysis found a positive correlation between drawing scores and the quality of produced ideas. Zhang et al. (2010) studied knowledge building activities in Wikipedia based on a comparison of entity creation, article revision, and link evolution made by contributors in a six year period from 2001 to 2007. Based on non-parametric Kendall correlation analysis, they found that the ranks of entities have positive correlations with time of creation, the number of revisions, and number of unique contributors. This finding suggest that the highly ranked entities are

correlated with higher number of revisions made by large numbers of unique contributors, suggesting that highly ranked entities were of superior quality.

A review of various studies suggests that non-parametric correlation analytical approaches such as Spearman's Rho and Kendall's Tau are appropriate for analyzing relationships between variables in online learning environments, particularly when the distribution of quantitative data does not meet the assumptions of normality in cases where data is either sparsely distributed or widely varied.

### **3.5.2 Correlation Analysis**

Correlation refers to the measure of the strength of the linear relationship between two random variables. Correlation analysis assumes that individual observations are statistically independent and does not imply causal relationships between the examined variables. An often used correlation measure includes Pearson product-moment correlation, which refers to the strength of association between sets of bivariate observations (Chen & Popovich, 2002). Assumptions of parametric correlation include normality, linearity and homoscedasticity (Creswell, 2009). Normality considers if the probability distribution of the data conforms to normal whereas linearity refers to the assumption that there is a linear relationship between the variables examined. Homoscedasticity on the other hand refer to the expectation that the dependent variable will demonstrate equal or similar levels of variance across the range of values for an independent variable. If the above conditions are not met, non-parametric correlation may be used as it does not require that the analyzed data is normally distributed.

Non-parametric correlation coefficients measures the strength of the monotone relationship between sets of x and y data. Commonly used statistical approaches include Spearman's Rho and Kendall's Tau. Spearman's Rho is the Pearson product moment correlation of the ranks of data points applied to the scores of data after the data points have undergone ranking from the smallest to the largest based on two variables (Chen & Popovich, 2002; Spearman, 1904). Spearman's Rho is determined using Equation 1:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (1)$$

where  $d$  refers to differences in the ranks of the two variables in each set of observations

$n$  refers to the number of sets of observations

Kendall's Tau is used to measure the strength association between two sequences by counting the concordant and discordant pairs in two sequences (Kendall, 1938). In this statistical approach, a concordant pair consists of a pair of observations that assume the same ranks in two sequences whereas a discordant pair consists of a pair of observations that do not assume the same ranks in two sequences (Chen & Popovich, 2002). In sum, the calculation of Kendall's Tau focuses on the differences in rankings within pairs of cases. To employ Kendall's Tau, one does not need to meet the assumption the data is statistically normally distributed and is therefore an appropriate statistical operation for the examination of relationships between quantitative variables in this research study. Kendall's Tau is used in the examination of the strength of association between quantitative variables suggestive of knowledge creation and

variables that promote knowledge creation in this research, in order to address the second and third research question. Kendall's Tau is determined using Equation 2:

$$\tau = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\frac{1}{2}n(n-1)} \quad (2)$$

where  $n$  refers to the number of sets of observations

### **3.6 Method for Research Question 4**

To address Research Question 4, this study needs to employ a statistical approach to identify groups of users based on a consideration of the individual variables. A literature review is carried out and focused on prior studies that have analyzed and identified groups of users in online learning environments (see Section 3.6.1). Based on the literature review, this study identifies an appropriate statistical approach – K-means clustering – and in Section 3.6.2, a discussion is carried out to describe the statistical background behind the K-means clustering analysis that has featured in examination of learner behavior in online environments.

#### **3.6.1 Examining Groups in Research of Online Communities**

In MOOC environments, Kizilcec et al. (2013) have examined trajectories of engagement to characterize learners based on their participation patterns within three MOOC courses. In this research, they applied k-means clustering to user parameters indicative of interactions with course content to identify distinct user trajectories of engagement within three MOOC courses. Through their research, they found evidence to

support that there are three main user participation trajectories. An example includes a group of users who view content but never took assessment.

K-means clustering is also used to inform the design of an assessment plug-in (ForumDash) for Blackboard Learning Management System (Speck et al., 2014) and a course recommendation system (Aher & Lobo, 2013). The ForumDash assessment plugin employed k-means clustering to cluster together topics that are indicative of engagement with a variety of situations including questions and answers, syllabus and information about other courses (Speck et al., 2014). Aher and Lobo (2013) used k-means clustering to examine relationships between attributes of users based on learning traces collected from the Moodle system. K-means clustering has also been used to examine pockets of discussion based on activity on discussion thread (Sinha, 2014) based on the variables of topic length, duration, content detail and density. Findings from clustering analysis reveal three main clusters: short responses with low detail, short responses with high detail and lengthy topics with high content density. Based on a review of various studies, k-means clustering is deemed as an appropriate statistical approach for the identification and detection of groups of learners based on variables that are representative of their engagement and participation in the online community.

### **3.6.2 K-means Clustering Analysis**

Clustering is a statistical technique that is used to group entities or individual into a preset number of clusters by attempting to reorganize them into homogenous groups (Gore, 2000). Computational algorithms are often used to optimize the process of reduction of variation within groups and maximizing the variation between groups. One

widely used clustering algorithm is the K-means algorithm. The K-means clustering algorithm is governed by Equation 3:

$$\arg \min_{\mathbf{S}} \sum_{i=1}^n \sum_{\mathbf{x}_j \in S_i} ||\mathbf{x}_j - \mu_i||^2 \quad (3)$$

where  $S_i$  refers to randomly selected  $i$  set of points

$n$  refers to number of clusters

$\mathbf{x}_j$  refers to  $j$  set of observations

$\mu_i$  refers to means of the set of points in  $S_i$

The algorithm for the K-means clustering is operated on the concept of iterative calculation of the distance between nodes on the basis of comparing means or standard deviation of group characteristics. This method is operated by first dividing sampled  $n$  observations into clusters, in which observations will be assigned to a cluster with the closest mean. It starts from a given center point, and group together entities by considering the distance from the center. Iterations are carried out until the best possible distribution is obtained (Gore, 2000). The optimal number of clusters ( $k$ ) is decided by the researcher based on the CCC measure, which is based on a comparison of the  $R^2$  value for a given set of clusters with the  $R^2$  of a uniformly distributed set of points. K-means techniques also find clusters by minimizing intra-cluster variation while maximizing inter-cluster variation (Everitt et al., 2001). Limitations of k-clustering include the need to for the variables to be continuous and an inability to account for missing data.



With the completion of K-means clustering, there is a need to verify if there are significant statistical differences between the clustered groups of individuals through statistical tests. This means that there is a need to verify if the clustered groups are significantly different through the comparison between the medians of individual variables in the clustered groups to determine if the clustered samples belong to different populations. This research will employ the Kruskal-Wallis test (Kruskal & Wallis, 1952), the non-parametric equivalent of one-way ANOVA, to examine the statistical differences between groups of learners to ascertain if the groups are statistically different before the comparison of correlation coefficients classified by clusters is performed.

### **3.7 Synthesis of Research Question, Theory and Methods**

Guided by the knowledge creation metaphor, the outlined research activities are targeted at addressing the overarching research question of describing the degree of knowledge advancement in an online engineering community – All About Circuits. This research focuses on the examination of three set of parameters: linguistic features, use of artifacts and social networks which translate to an examination of the use of engineering terms, engineering artifacts and the formation of social networks among engineering learners in the online community. In order to derive these parameters, the research plans relied on the Python programming platform and software libraries to collect and organize half a million messages to extract learner-generation data to derive interaction logs for each discussion topics.

The outlined descriptive research approach centered on the analysis of interaction logs is expected to be capable of addressing the research focus on evaluating knowledge advancement in the online community. This is particularly so as public discussion is

carried out exclusively on discussion threads and each discussion thread is distinct from the others in terms of its content focus and source of origin. The examination of entire set of 62,509 interaction logs will also allow a focus on use of engineering terms, engineering artifacts and the formation of social networks throughout the entire online community, based on the discussion thread in which discussion took place. This will facilitate the statistical description of the three sets of parameters to describe the state of knowledge creation in the online engineering community.

The research activities present opportunities to understand why some discussion topics enjoy high levels of knowledge creation. Table 3.3 describes the anticipated and potential research activities. As suggested by the research question in Table 3.3, the examination of the state of knowledge creation will provide summary statistical description (mean, mode, variance and range) of the examined variables contained within the interaction logs. The interaction logs are derived from individual discussion topics and classified based on a discussion topic identification number. Therefore, the summative descriptive statistics will allow for the identification of discussion topics with high level of knowledge creation which features rich social networks, and active use of engineering terms and artifacts (see second research question in Table 3.3). Once discussion topics with high levels of knowledge creation are identified, there are opportunities to further examine the role of social dynamics and interactions in knowledge creation.

The overall aim of the research is to assess learning in an online engineering community based on the knowledge creation metaphor. The research objectives involve the descriptive examination of the state of knowledge creation through three sets of

parameters – engineering terms, engineering artifacts and social networks. The focus of this research can be traced to the main arguments outlined by the knowledge creation metaphor in Chapter 2. According to this perspective, learners engaged in social interaction to advance and transform both material and artifacts over time. The sub research questions are summarized together with the corresponding theoretical perspectives and research methods in Table 3.3.

Table 3.3

*Research Questions, Theoretical Basis and Methods*

<b>Research Questions</b>	<b>Theoretical Basis</b>	<b>Methods</b>
1. What is the state of engineering knowledge creation at the topic and individual levels?	Knowledge creation metaphor suggests that conceptual artifacts are transformed over time and that material artifacts are advanced and transformed through collective activities.	Descriptive statistics
2. To what extent are topic length, duration and views associated with participation in knowledge creation activities at the topic level?	Learners of knowledge creating communities engage in extensive social interaction over sustained periods of time.	Non-parametric Correlation Analysis
3. To what extent are individual total interactions, active membership period and total membership period associated with individual participation in knowledge creation activities?	Knowledge creation metaphor stresses on the role of experts in the facilitation of knowledge creation among learners	Non-parametric Correlation Analysis

Research Questions	Theoretical Basis	Methods
4. How can learners be grouped based on their individual total interactions and active membership period? How does correlation compare across groups?	Learners on online communities contribute and participate at varying extents.	K-Clustering Analysis; Non-parametric Correlation Analysis

### 3.7.1 Variables Indicative of Knowledge Creation

For the first research question, descriptive statistics is performed for variables across the individual and topic levels. For the second research question, correlation analysis is conducted to examine the relationship between six variables at the topic level (see Table 3.4) with three variables that may predict knowledge creation (see Table 3.5). The six variables, that may suggest knowledge creation, are Network Size, Quoted Reference, Engineering Terms, Links to Web Resources, Links to Digital Files and Attachments. The three variables, that may predict knowledge creation, are Topic Length, Topic Duration and Topic Views. The distribution of the data has to be considered through the Kolmogorov-Smirnov-Lilliefors Test before applying the relevant correlation statistics. The examination of the strengths of association between variables of non-normal data distribution will require the use of Spearman's Rho and Kendall's Tau.

For the third research question, correlation analysis is conducted to examine the relationship between six dependent variables at the topic level (see Table 3.4 and Figure 3.9) with three variables that may predict engagement with knowledge creation (see Table 3.5 and Figure 3.10). The five variables, that may suggest engagement with knowledge creation, are Network Size, Quoted Reference, Engineering Terms, Links to Web Resources, Links to Digital Files and Attachments. The three variables, that may

predict knowledge creation, are Total Unique Interactions, Active Membership Period and Total Membership Period. The distribution of the data has to be considered through the Kolmogorov-Smirnov-Lilliefors Test before applying the relevant correlation statistics. Variables with non-parametric data distribution will require the use of Spearman's Rho and Kendall's Tau.

Table 3.4

*Variables Suggestive of Knowledge Creation at Topic-Level and Individual-Level*

Variable	Description	Supporting Research
<b><i>Social Interaction</i></b>		
Network Size <sup>2</sup>	Joint work and organized collaboration between different individual foster the processes of knowledge creation and collective advancement of knowledge.	Bielaczyc & Collins, 2006; McLaughlin, 2007; Scardamalia & Bereiter, 2006
Quoted References <sup>1,2</sup>	Knowledge creation is a social product, with edits from contributors. Knowledge creation involves collaborative work that fosters communication to know each other and social practices that develop due to to-and-fro communication between different individuals in the community.	Bielaczyc & Collins, 2006; Paavola et al., 2007; Philip, 2010;
<b><i>Contributing to Conceptual Artifacts</i></b>		
Message Post <sup>1</sup>	Posting messages represent shared objects that can be accessed by any individuals and foster externalization of ideas in a shared collaborative space. Messages represent comments and notes leading to ideas and in knowledge creation, ideas interact with ideas to foster the generation of ideas.	Sha & van Aalst, 2003; Seitamaa-Hakkarainen et al, 2010; Xia et al., 2009
Engineering Terms <sup>1,2</sup>	In knowledge creation, the acknowledged use of engineering terms present building blocks, classifiers and tags of conceptual artifacts, can is used to compose the common ground and to generate relevant conceptual artifacts.	Lakkala et al., 2009; Locoro et al., 2010; Sun et al., 2009;

<b><i>Contributions to Material Artifacts</i></b>		
Links to Web Resources <sup>1,2</sup> (Abbreviated as Resources)	Web links are sources outside the communication that foster generation of multiple perspectives by reading and discussion. Transform ideas taken from different sources to fit situation.	Bielaczyc & Collins, 2006; Paavola et al., 2007; van Aalst, 2009
Links to Digital Files <sup>1,2</sup> (Abbreviated as Files)	Knowledge creation feature object-bound entities such as web links that foster shared use of resources and objects, which can catalyst perspective taking and support student ideas.	Chen et al., 2012; Lee et al., 2008; van Aalst, 2009;
Attachments <sup>2</sup>	Attachments represent artifacts that knowledge is embodied, to be shared to collaborate with others. The attachment versions present counts of content modification and advancement of knowledge	McLaughlin, 2007; Muukkonen & Lakkala, 2009; Paavola et al., 2005

1 = individual-level; 2 = topic-level

Table 3.5

*Variables That Predict Knowledge Creation at Topic-Level and Individual-Level*

<b>Variable</b>	<b>Description</b>	<b>Supporting Research</b>
<b><i>Topic-Level</i></b>		
Topic Length	Edits can represent indexes of participation in knowledge creation. Online publishing favor knowledge creation. Topic length suggests that multiple ideas are generated and discussed. Dense communication represents knowledge creation.	Leinonen et al., 2009; Scardamalia, 2002; van Aalst, 2009; Xia et al., 2009
Topic Duration	Sustained involvement over multiple phases and periods of times in activities that aim to improve the overall community through problems.	Leinonen et al., 2009; Seitamaa-Hakkarainen et al., 2010; Zhang et al., 2007
Topic Views	Views suggest participation and efforts to develop ideas and engage with the knowledge creation process, and holistic exposure to various phases of discussion.	Hong et al., 2010; Seitamaa-Hakkarainen et al., 2010
<b><i>Individual-Level</i></b>		

Individual Total Interactions	Individual connections and interactions with others create opportunities for collaboration, spread ideas and contribute to the formation of social practices	Hakkarainen, 2009; Leinonen et al., 2009; Oshima et al., 2012;
Individual Active Membership Period	Individual extended participation throughout discussion and community can help structure discussion and increase engagement with variety of past ideas.	Lee et al 2008; Muukkonen & Lakkala, 2009; Seitamaa-Hakkarainen et al., 2010
Total Membership Period	Individual unique experiences accumulated throughout engagement with community can bring insights and perspectives on issues and ideas, leading to knowledge creation.	Bielaczyc & Collins, 2006; Palonen & Hakkarainen, 2000; Hakkarainen, 2009

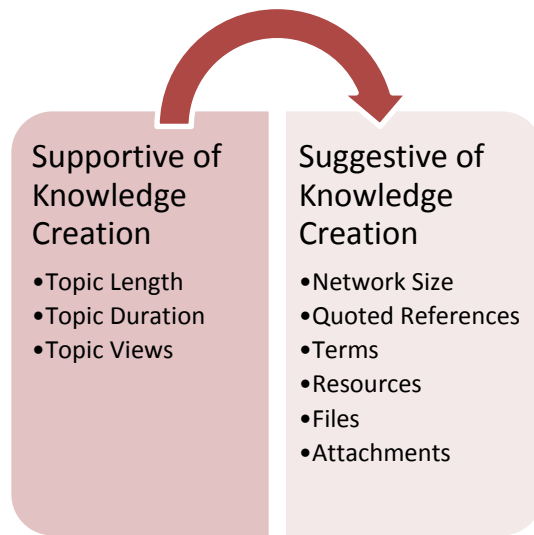


Figure 3.9: Potential Variables for Predictive Relationships Between Topic-Level Variables

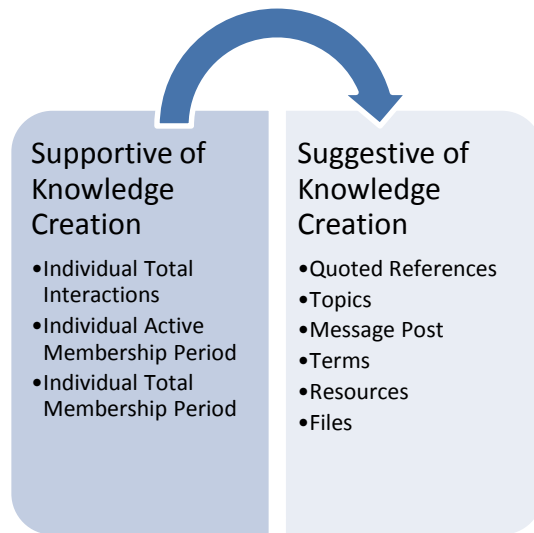


Figure 3.10: Potential Variables for Predictive Relationships Between Individual-Level Variables



## **Chapter 4**

### **Findings**

This chapter presents the findings related to the research questions addressed in this study.

The research questions are:

1. What is the state of engineering knowledge creation at the topic and individual levels?
2. What is the relationship between topic length, duration and views with participation in knowledge creation activities at the topic level?
3. What is the relationship between individual total interactions, active membership period and total membership period with individual participation in knowledge creation activities?
4. How can learners be grouped based on their individual total interactions and active membership period? How do the correlation statistics vary across groups?

Section 4.1 addresses the first research question, with descriptive statistics and frequency counts of variables at both topic and individual level of participation. Section 4.2 addresses the second research question with a focus on identifying the type of association between six variables that may suggest participation in knowledge creation and three variables that are strongly associated with engagement in knowledge creation at the topic-level. Section 4.3 addresses the third research question with a focus on identifying the type of association between five variables that may suggest participation in knowledge creation and three variables that are strongly associated with engagement in

knowledge creation at the individual-level. Section 4.4 addresses the fourth research question with a focus on identifying groups of individuals with similar participation characteristics.

#### **4.1. Participation Demographics**

This section describes the participation demographics of the online community. The sub-sections describe participation demographics related to the topics (Section 4.1.1) and membership (Section 4.1.2). Findings from this section indicate:

1. The 65,208 discussion topics consist of close to 503,908 messages contributed by only 31,219 members out of a total registered 182,987 members over a period of 3,353 days.
2. The annual message post count has seen an increase from 2003 to 2009 but has been at a decline from 2010 to 2012.

##### **4.1.1 Community Participation**

This section provides an overview of the descriptive statistics based on the participation in the online engineering communities. As described in Table 4.1, there are a total of 65,208 discussion topics that consist of 503,908 messages. While the community report a total of registered 182,937 members at the point of data collection, only 31,219 members (17.1% of all members) have posted at least one message to the community whereas 151,178 (82.9% of all members) have not posted any messages to the community. This suggests that certain community features may require membership registration and the search function is one of the community feature that requires membership to access.

Table 4.1

*Participation Parameters at the Community Level*

Participation Parameter	Count
Discussion Topics	65,208
Messages	503,908
Active Users	31,219
Total Members	182,937
Period of Establishment	3,353 days

As described in Table 4.2, the rate of change of messages has been steadily increasing. The increase in annual message count rate for year 2010 and 2011 was 22,136 and 8,723 respectively and they are both substantially lower than the previous high of 32,421: with a decline of 10,285 and 23,698 respectively. Figure 4.1 demonstrated a consistent pattern that increases in annual message count rate exceeded its previous highest levels from the 2003 to 2009.

Table 4.2

*Community Message Count from 9/27/2003 to 11/5/2012*

Year	Cumulative Message Count	Annual Message Count	Annual Change in Message
2003 (From 9/27/2003)	99	99	99
2004	2497	2398	2299
2005	7532	5035	2637
2006	18003	10471	5436
2007	41639	23636	13165

2008	95479	53840	30204
2009	181740	86261	32421
2010	290137	108397	22136
2011	407257	117120	8723
2012 (Till 11/5/12)	503908	96651	-20469

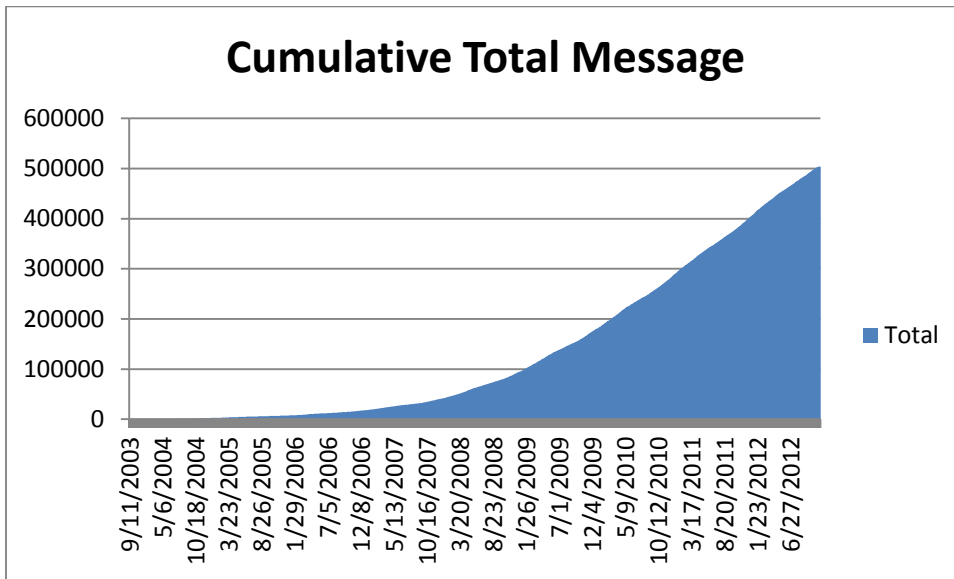


Figure 4.1 Cumulative Message Count in Online Engineering Community

#### 4.1.2 Topic Participation

To identify general trends for participation at the topic-level and to understand the distribution of data, descriptive statistics and frequencies of the topic-level parameters is provided in Table 4.3. They include the means, median, mode, standard deviation, minimum and maximum of the each variable. The descriptive statistics covers the entire collection of topics at  $n = 65,209$ .

Table 4.3:

*Descriptive Statistics of Topic-Level Variables*

Variable (Units)	Mean	Median	Mode	Standard Deviation	Minimum	Maximum
Topic Length (Message Count)	7.73	5.00	2.00	13.3	1.00	1890
Topic Duration (Days)	19.4	1.42	0.00	116	0.00	3190
Topic Views (Count of Views)	1290	803	503	2220	0.00	183000
Network Size (Unique Contributors)	3.69	3.00	2.00	2.53	1.00	187
Quoted Reference (Count)	1.86	0.00	0.00	6.40	0.00	702
Terms/Message (Word Count)	23.9	20.0	11.0	18.5	0.00	524
Links To Online Resources (Frequency Count)	1.06	0.00	0.00	2.42	0.00	130
Links To Digital Files (Frequency Count)	0.534	0.00	0.00	1.42	0.00	92.0
File Attachments (Frequency Count)	0.729	0.00	0.00	2.78	0.00	316

Note: N = 65,209\*

There are 65,209 distinct discussion topics, as demonstrated by their distinct discussion topics and topic identification number.

Variables in the social interaction dimension include network size and quoted references. The mean network size was at least 3 unique participants ( $\bar{x}$  = 3.69) and the standard deviation was 2.53 ( $s$  = 2.53). The median network size was 3 ( $\tilde{x}$  = 3) the range was 187 ( $x_{\min}$  = 0 and  $x_{\max}$  = 188). The mean number of quoted references was 1.86 ( $\bar{x}$  =

1.86) and the standard deviation was 6.40 references ( $s = 6.40$ ). The median links was no links ( $\tilde{x} = 0$ ) and the range was 702 links ( $x_{\min} = 0$  and  $x_{\max} = 702$ ). Variables in the dimension of conceptual artifacts include the use of engineering terms and message posted. The mean engineering terms was 195 per topic or 23.9 per message and the standard deviation was 386 terms per topic or 18.5 per message. The median engineering terms per message was 20 terms ( $\tilde{x} = 20$ ) and the range was 524 ( $x_{\min} = 0$  and  $x_{\max} = 524$ ).

Variables in the dimension of material artifacts include external online resources, digital files and attachments. The mean number of links to external online resources was 1.06 ( $\bar{x} = 1.06$ ) and the standard deviation was 2.42 links ( $s = 2.42$ ). The median external links to external online resources was no links ( $\tilde{x} = 0$ ) and the range was 130 references ( $x_{\min} = 0$  and  $x_{\max} = 130$ ). The mean number of digital files was 0.534 ( $\bar{x} = 0.534$ ) and the standard deviation was 1.42 digital files ( $s = 1.42$ ). The median links was no digital files ( $\tilde{x} = 0$ ) and the range was 92 digital files ( $x_{\min} = 0$  and  $x_{\max} = 92$ ). The mean number of attachments was 0.729 ( $\bar{x} = 0.729$ ) and the standard deviation was 2.78 attachments ( $s = 2.78$ ). The median was no attachments ( $\tilde{x} = 0$ ) and the range was 316 attachments ( $x_{\min} = 0$  and  $x_{\max} = 316$ ).

The mean topic length was more than half a page long ( $\bar{x} = 7.73$ ) and the standard deviation was 13.3 ( $s = 13.3$ ). The median topic length was a third of a page long ( $\tilde{x} = 5$ ) the range was 1889 ( $x_{\min} = 1$  and  $x_{\max} = 1890$ ). The mean topic duration was 19.4 days or 461 hours ( $\bar{x} = 465.1051$ ) with a standard deviation of 116 days 2800 hours. The median topic duration was a 34.1 hours or 1.42 days ( $\tilde{x} = 1.42$ ) and the range was 3190 days ( $x_{\min} = 0$  and  $x_{\max} = 3190$ ).

Figure 4.2 shows the variation in data using a heat-map visualization of the variable of topic duration as an example. There is a higher number of lightly colored boxes with larger sizes which in turn occupied more space in the heat map. This observation indicates the dominance of topics with shorter durations over topics conducted through longer durations. For instance, there are 2,154 topics with duration of 1 day or less as compared to 140 topics with duration of between 178 to 178 days. Topics with extended duration are represented as dark green regions in the heat map and can be observed to occur less frequently by representation of a smaller bottom right hand corner of the heat map.

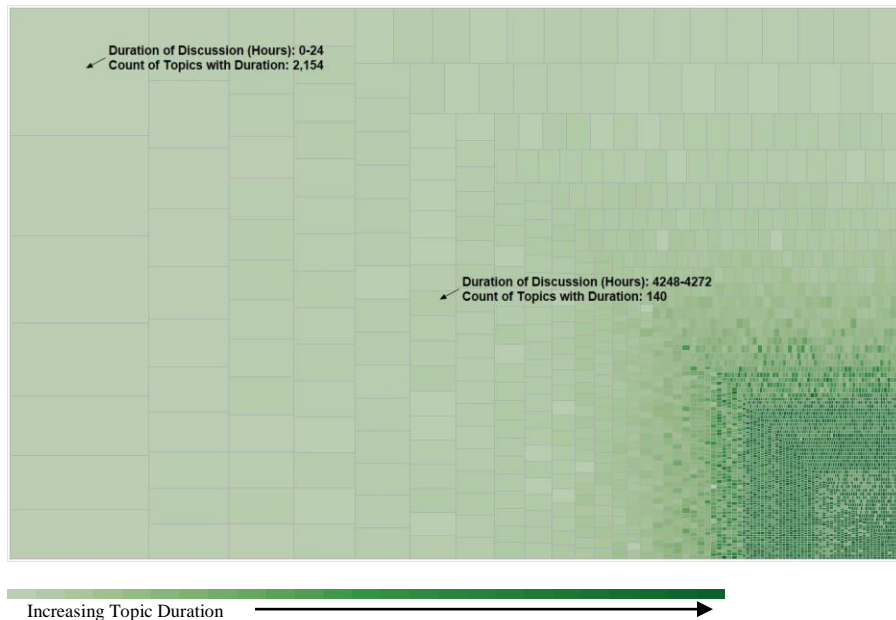


Figure 4.2 Heat Map of Topic Duration in the Online Community

*Note.* Size of square tile indicate the equivalent of the percentage distribution of each duration period. Both horizontal and vertical axes are equal representation of the square tile.

Overall, the computation of descriptive statistics suggests that participation varied at the topic level. With the exception of the network size and engineering terms per message, each variable has a standard deviation that is more than one times of the mean of the variable. The large standard deviation of the variables suggest that there is variation in data at both the topic level. The variation of data can be better understood under the consideration that data analysis was performed over a 10 year period rather than a cross-sectional approach with selection of topics of similar size.

#### **4.1.3 Individual Participation**

To identify general trends in individual participation and to understand the distribution of data, descriptive statistics of the individual level parameters is provided in Table 4.4. They include the means, median, mode, standard deviation, minimum and maximum of the each variable. The descriptive statistics cover the entire collection of topics at  $N = 31,219$ .



Table 4.4

*Descriptive Statistics of Individual Participation in Community*

Variable	Mean	Median	Mode	Standard Deviation	Minimum	Maximum
Network Degree (Interactions)	19.9	3.00	1.00	158	0.00	10400
References (Count)	3.80	0.00	0.00	80.3	0.00	10100
Engineering Terms (Count)	495	41.0	8.00	8450	0.00	890000
Message Post (Count)	16.1	2.00	1.00	243	1.00	21900
Web Resources (Count)	2.18	0.00	0.00	53.4	0.00	5710
Digital Files (Count)	1.10	0.00	0.00	28.2	0.00	3340
Active Membership (Days)	76.5	1.00	1.00	244	1.00	3230
Total Membership (Days)	1090	1020	1360	706	1.00	3350

Notes: N= 31,219\*. There are 31,219 active individuals, as demonstrated by their contribution of posts and user identification number.

Variables in the dimension of conceptual artifacts include the creation of topics and messages, and the use of engineering terms. The mean message posted was more than a page long ( $\bar{x} = 16.1$ ) and the standard deviation was 243 ( $s = 243$ ). The median message posted was 2 ( $\tilde{x} = 2$ ) and the range was 21900 ( $x_{\min} = 0$  and  $x_{\max} = 21900$ ). The mean engineering terms posted was 495 ( $\bar{x} = 495$ ) and the standard deviation was 8450 ( $s = 8450$ ). The median message engineering terms posted was 41 ( $\tilde{x} = 41$ ) and the range was 890000 ( $x_{\min} = 0$  and  $x_{\max} = 890000$ ).

The large variation in the data can be observed in a “packed bubbles” visualization of individual post count (see Figure 4.3). Learners with long membership periods, as indicated by smaller user identification numbers, do not necessarily record

large post counts as indicated by the size of the bubble representative of each user. Only one of the most prolific contributor (labeled bubbles on Figure 4.3) was an early registrant with an identification number of 60. Learners who have shorter total membership periods may have made considerable post counts. The largest amount contributions (represented by the largest bubble) was made by a learner who registered as the 12669<sup>th</sup> member.

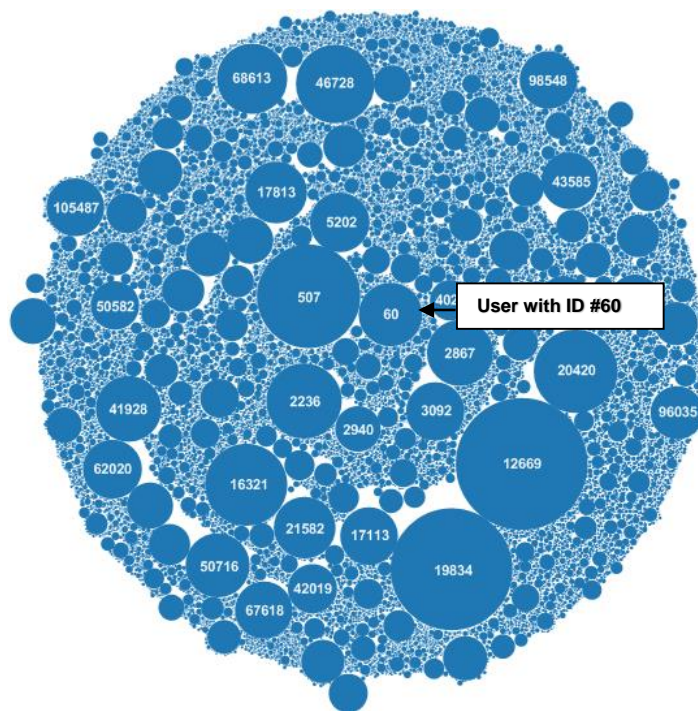


Figure 4.3 Packed Bubbles Visualization of Individual Post Count

*Note: Larger bubbles correspond to larger number of individual post count*

Variables in the dimension of material artifacts include web resources and digital file links. The mean web resources posted was 2.18 ( $\bar{x} = 2.18$ ) and the standard deviation was 53.4 ( $s = 53.4$ ). The median web resources posted was a 0 ( $\tilde{x} = 0$ ) and the range was 5710 ( $x_{\min} = 0$  and  $x_{\max} = 5710$ ). The mean digital files posted was 1.10 ( $\bar{x} = 1.10$ ) and the

standard deviation was 28.2 ( $s = 28.2$ ). The median digital files posted was zero ( $\tilde{x} = 0$ ) and the range was 3340 ( $x_{\min} = 0$  and  $x_{\max} = 3340$ ).

Variables in the dimension of social interaction include network degree and quoted references. The mean network degree was 19.9 ( $\bar{x} = 19.9$ ) and the standard deviation was 158 ( $s = 158$ ). The median network degree was 3 ( $\tilde{x} = 3$ ) and the range was 10400 ( $x_{\min} = 0$  and  $x_{\max} = 10400$ ). The mean direct references made to others was less than 4 ( $\bar{x} = 3.80$ ) and the standard deviation was 80.3 ( $s = 80.3$ ). The median direct reference made was 0 ( $\tilde{x} = 0$ ) and the range was 10100 ( $x_{\min} = 0$  and  $x_{\max} = 10100$ ).

The mean individual active membership period was 76.5 days ( $\bar{x} = 76.5$ ) and the standard deviation was 244 days ( $s = 244$ ). The median active membership period was 1 ( $\tilde{x} = 1$ ) and the range was 3229 days ( $x_{\min} = 1$  and  $x_{\max} = 3230$ ). The mean total membership period was 1090 days ( $\bar{x} = 1090$ ) and the standard deviation was 706 days ( $s = 706$ ). The median total membership period was 1020 day ( $\tilde{x} = 1020$ ) and the range was 3249 days ( $x_{\min} = 1$  and  $x_{\max} = 3250$ ).

Overall, the computation of descriptive statistics suggests that participation varied substantially at the individual level. With the exception of the network size and engineering terms per message, each variable has a standard deviation that is more than one times of the mean of the variable. The large standard deviation of the variables suggest that there is variation in data at both the individual level. The variation of data can be better understood under the consideration that data analysis was performed over a 10 year period rather than a cross-sectional approach with selection of individuals of similar contribution levels.

## 4.2 Association between Topic-Level Variables

This sub-section describes correlation statistics between variables suggestive of knowledge creating activities with topic level parameters. Kolmogorov–Smirnov–Lilliefors (KSL) tests showed that distributions of all the variables are not normal distributions. Each test is carried out with the null hypothesis as the data being normally distributed ( $H_0$  = The data is from the Normal distribution). For example, the P-value for the variable of Topic Duration is less than the level of significance of 0.001 which rejects the null hypothesis and indicates that the data is not normally distributed.

A further examination of the distributions of variables through histograms (see Figure 4.4) verified that the assumptions of normality are not observed. This reaffirms the previous assertion that the data is not normally distributed. Based on the consideration of non-normality of the variables, non-parametric correlation analysis is selected over than parametric correlation analysis. In this analysis, the Spearman's  $\rho$  and Kendall's  $\tau$  are selected as measures of the strength of associations.

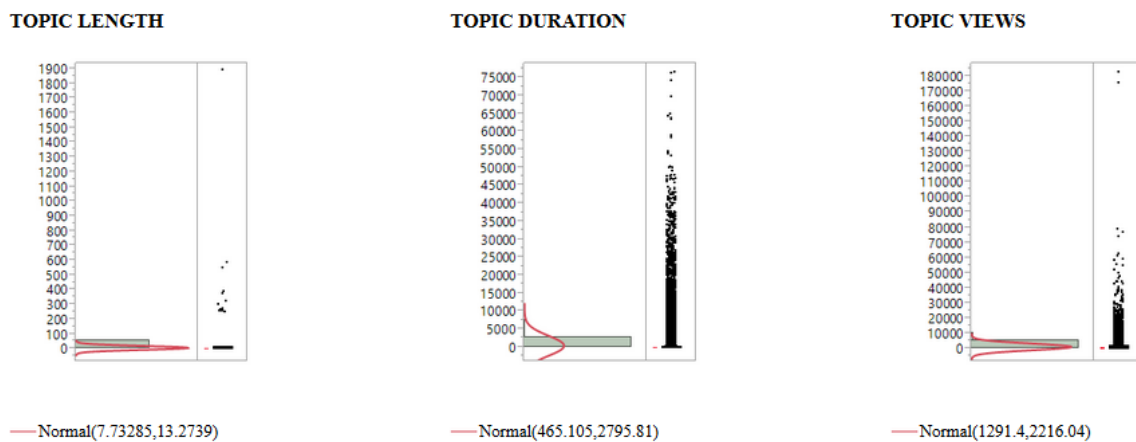


Figure 4.4 Histogram of Select Topic Level Variables

*Note:*  $H_0$  = The data is from the Normal distribution. Small p-values reject  $H_0$ .

#### 4.2.1 Nonparametric Correlation Analysis

This subsection provides an examination of the association between the topic-level variables used in the study of the online communities. For non-parametric distributions, the Spearman's  $\rho$  and Kendall's  $\tau$  represents the strength of association between variables and has a value range of -1 to 1. Strong correlation values have a range of 0.5 to 1.0, whereas moderate correlation values have a range between 0.3 to 0.49 and weak correlation values have a range of 0 to 0.29 (Cohen, 1988).

The correlation analysis between six topic-variables variables and Topic Length are presented in Table 4.5. The Spearman rho values and Kendall tau values indicate that the correlations are positive and statistically significant. The average Spearman rho is 0.571, which suggests an overall strong association between Topic-Level Variables and Topic Length. According to this statistical measure, the strongest correlation is between network size and topic duration at  $\rho = 0.836$  and the weakest correlation is between Files and Topic Length at  $\rho = 0.323$ . The average Kendall tau is 0.472, which suggests an overall moderate association between Topic-Level Variables and Topic Length. According to this statistical measure, the strongest correlation is between Network Size and Topic Length at  $\tau = 0.721$  and the weakest correlation is between Files and Topic Length at  $\tau = 0.276$ .

Table 4.5

*Correlation Between Topic-Level Variables with Topic Length*

Variable	Spearman's $\rho$	p-value	Kendall's $\tau$	p-value
1. Network Size	0.836	<0.001	0.721	<0.001
2. Quoted References	0.692	<0.001	0.579	<0.001
3. Terms	0.821	<0.001	0.653	<0.001
4. Resources	0.418	<0.001	0.339	<0.001
5. Files	0.323	<0.001	0.266	<0.001
6. Attachments	0.335	<0.001	0.276	<0.001

The correlation analysis between six topic-level variables and Topic Duration are presented in Table 4.6. The Spearman rho values and Kendall tau values indicate that the correlations are positive and statistically significant. The average Spearman rho is 0.391, which suggests an overall moderate association between Topic-Level Variables and Topic Duration. According to this statistical measure, the strongest correlation is between Network Size and Topic Duration at  $p = 0.597$  and the weakest correlation is between Attachments and Topic Duration at  $p = 0.216$ . The average Kendall tau is 0.298, which suggests an overall moderate association between Topic-Level Variables and Topic Duration. According to this statistical measure, the strongest correlation is between Network Size and Topic Duration at  $\tau = 0.464$  and the weakest correlation is between Attachments and Topic Duration at  $\tau = 0.171$ .

Table 4.6

*Correlation Between Topic-Level Variables with Topic Duration*

Variable	Spearman's $\rho$	p-value	Kendall's $\tau$	p-value
1. Network Size	0.597	<0.001	0.464	<0.001
2. Quoted References	0.442	<0.001	0.341	<0.001
3. Terms	0.560	<0.001	0.398	<0.001
4. Resources	0.292	<0.001	0.227	<0.001
5. Files	0.236	<0.001	0.187	<0.001
6. Attachments	0.216	<0.001	0.171	<0.001

The correlation analysis between six topic-level variables and Topic Views are presented in Table 4.7. The Spearman rho values and Kendall tau values indicate that the correlations are positive and statistically significant. The average Spearman rho is 0.285, which suggests an overall weak association between Topic-Level Variables and Topic Views. According to this statistical measure, the strongest correlation is between Terms and Topic Views at  $p = 0.407$  and the weakest correlation is between Attachments and Topic Views at  $p = 0.323$ . The average Kendall tau is 0.212, which suggests an overall weak association between Topic-Level Variables and Topic Views. According to this statistical measure, the strongest correlation is between Terms and Topic Views at  $\tau = 0.280$  and the weakest correlation is between Attachments and Topic Views at  $\tau = 0.117$ .

Table 4.7

*Correlation Between Topic-Level Variables with Topic Views*

Variable	Spearman's $\rho$	p-value	Kendall's $\tau$	p-value
1. Network Size	0.356	<0.001	0.261	<0.001
2. Quoted References	0.340	<0.001	0.259	<0.001
3. Terms	0.407	<0.001	0.280	<0.001
4. Resources	0.236	<0.001	0.181	<0.001
5. Files	0.220	<0.001	0.173	<0.001
6. Attachments	0.149	<0.001	0.117	<0.001

**4.3 Association between Individual-Level Variables**

This subsection describes an examination of the association between the individual-level variables that are suggestive of participation in knowledge creating activities in the online community. Kolmogorov–Smirnov–Lilliefors (KSL) tests showed that distributions of all the variables are not normal distributions. Each test is carried out with the null hypothesis as the data being normally distributed ( $H_0$  = The data is from the Normal distribution). For example, the P-value for the variable of Individual Total Interactions is less than the level of significance of 0.001 which rejects the null hypothesis and indicates that the data is not normally distributed. A further examination of the distributions of variables through histograms (see Figure 4.5) verified that the assumptions of normality are not observed. This reaffirms the previous assertion that the data is not normally distributed. Based on the consideration of non-normality of the variables, non-parametric correlation analysis is selected over than parametric correlation



analysis. In this analysis, the Spearman's  $\rho$  and Kendall's  $\tau$  are selected as measures of the strength of associations.

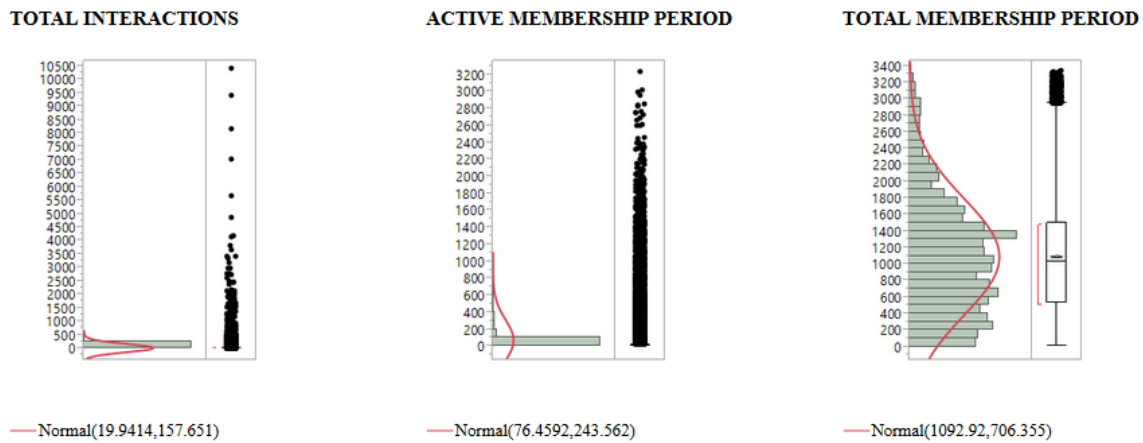


Figure 4.5 Histogram of Select Individual Level Variables

*Note:* Ho = The data is from the Normal distribution. Small p-values reject Ho.

For non-parametric distributions, the Spearman's  $\rho$  and Kendall's  $\tau$  represents the strength of association between variables and has a value range of -1 to 1. Strong correlation values have a range of 0.5 to 1.0, whereas moderate correlation values have a range between 0.3 to 0.49 and weak correlation values have a range of 0 to 0.29 (Cohen, 1988).

#### 4.3.1 Nonparametric Correlation Analysis

The correlation analysis between five individual-level variables and Individual Total Interactions are presented in Table 4.8. The Spearman rho values and Kendall tau values indicate that the correlations are positive and statistically significant. The average Spearman rho is 0.389, which suggests an overall moderate association between Individual-Level Variables and Individual Total Interactions. According to this statistical measure, the strongest correlation is between Replies and Individual Total Interactions at

$p = 0.543$  and the weakest correlation is between Files and Individual Total Interactions at  $\rho = 0.287$ . The average Kendall tau is 0.313, which suggests an overall moderate association between Individual-Level Variables and Individual Total Interactions. According to this statistical measure, the strongest correlation value is between Replies and Individual Total Interactions at  $\tau = 0.439$  and the weakest correlation is between Files and Individual Total Interactions at  $\tau = 0.241$ . Scatter plots illustrates that Individual Total Interactions increases with individual engagement with activities suggestive of knowledge creation (see Figure 4.6).

Table 4.8

*Correlation Between Individual-Level Variables with Individual Total Interactions*

Variable	Spearman's $\rho$	p-value	Kendall's $\tau$	p-value
1. Quoted References	0.397	<0.001	0.332	<0.001
2. Terms	0.394	<0.001	0.290	<0.001
3. Replies	0.543	<0.001	0.439	<0.001
4. Resources	0.314	<0.001	0.261	<0.001
5. Files	0.287	<0.001	0.241	<0.001

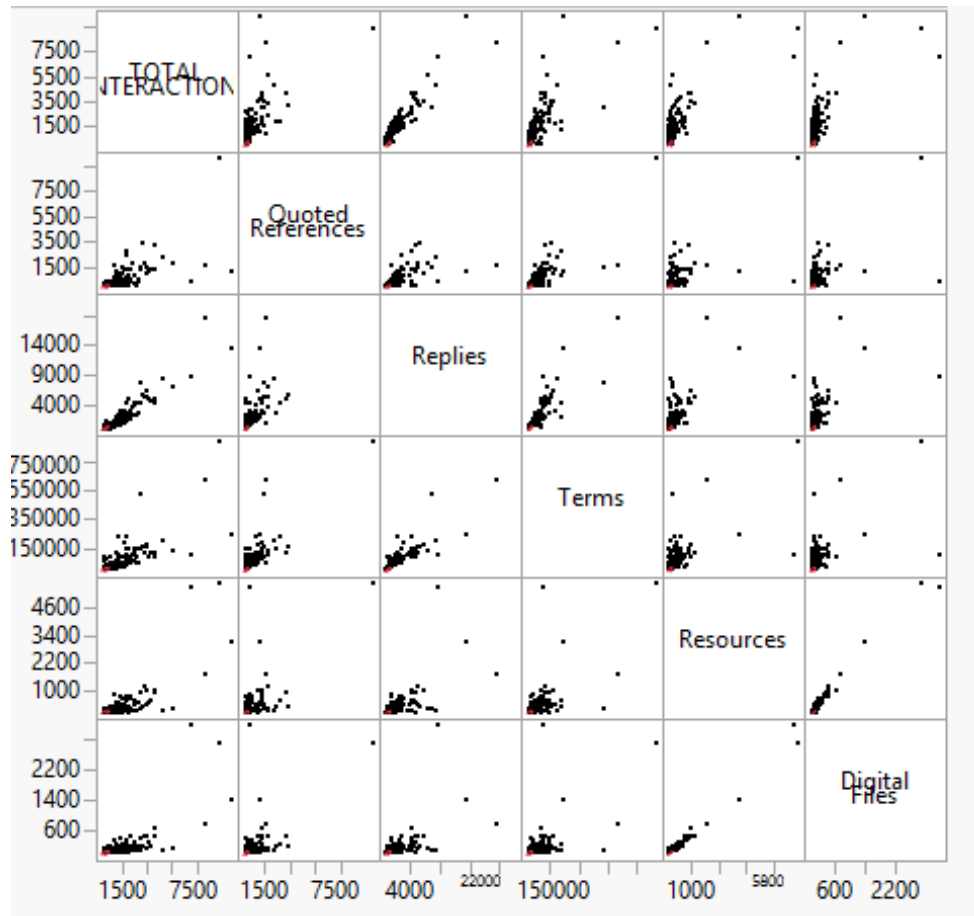


Figure 4.6 Scatter Plots of Individual Total Interactions with Individual Variables Suggestive of Knowledge Creation

The correlation analysis between five individual-level variables and Individual Active Membership Period are presented in Table 4.9. The Spearman rho values and Kendall tau values indicate that the correlations are positive and statistically significant. The average Spearman rho is 0.514, which suggests an overall moderate association between Individual-level Variables and Individual Active Membership Period. According to this statistical measure, the strongest correlation is between Replies and Individual Active Membership Period at  $p = 0.772$  and the weakest correlation is between Files and Individual Active Membership Period at  $p = 0.373$ . The average Kendall tau is 0.437,

which suggests an overall moderate association between Individual-level Variables and Individual Active Membership Period. According to this statistical measure, the strongest correlation is between Replies and Individual Active Membership Period at  $\tau = 0.658$  and the weakest correlation is between Files and Individual Active Membership Period at  $\tau = 0.334$ .

Table 4.9

*Correlation Between Individual-Level Variables with Individual Active Membership Period*

Variable	Spearman's $\rho$	p-value	Kendall's $\tau$	p-value
1. Quoted References	0.424	<0.001	0.377	<0.001
2. Terms	0.603	<0.001	0.468	<0.001
3. Replies	0.772	<0.001	0.658	<0.001
4. Resources	0.396	<0.001	0.350	<0.001
5. Files	0.373	<0.001	0.334	<0.001

The correlation analysis between five individual-level variables and Individual Total Membership Period are presented in Table 4.10. The Spearman rho values and Kendall tau values indicate that the majority of the association is positive and statistically significant. The average Spearman rho is -0.0470, which suggests an overall weak association between Individual-level Variables and Individual Total Membership Period. According to this statistical measure, the strongest correlation is between Terms and Individual Total Membership Period at  $p = -0.0730$  and the weakest correlation is

between Resources and Individual Total Membership Period at  $\rho = -0.0200$ . The average Kendall tau is -0.033, which suggests an overall weak association between Individual-level Variables and Individual Total Membership Period. According to this statistical measure, the strongest correlation is between Terms and Individual Total Membership Period at  $\tau = -0.0490$  and the weakest correlation is between Resources and Individual Total Membership Period at  $\tau = -0.0140$ .

Table 4.10

*Correlation Between Individual-Level Variables with Individual Total Membership Period*

Variable	Spearman's $\rho$	p-value	Kendall's $\tau$	p-value
1. Quoted References	-0.000108	0.9848	-0.000270	0.995
2. Terms	-0.0734	<0.001	-0.0485	<0.001
3. Replies	-0.0198	<0.001	-0.0140	<0.001
4. Resources	-0.0474	<0.001	-0.0370	<0.001
5. Files	-0.00175	0.7573	-0.00129	0.775

#### 4.4 K-means Cluster Analysis

The process of identifying groups of users in All About Circuits is described in Figure 4.7. K-means clustering is performed using two parameters – Individual Active Period and Individual Total Membership Period to arrive at the optimal number of clusters based on the comparison of diagnostic value of Cubic Clustering Criterion (CCC). As described in Table 4.11, K-means clustering was performed from a k range from 3 to 20, and a number of 8 clusters yield the best clustering results with the largest

CCC value of 13.9. Hence, clusters with  $k = 8$  is further examined for the sample of 31,219 individuals. In subsequent sections, statistical differences between groups are examined by Kruskal-Wallis tests before each cluster is examined for similarities in descriptive statistics.

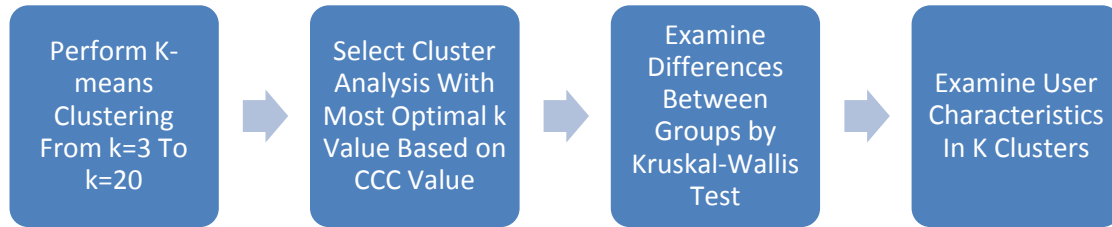


Figure 4.7 Identification of Groups of Learners Through K-Means Clustering Analysis

Table 4.11

*Iterative K-means Cluster Comparison*

Number of Clusters	CCC Value	Number of Clusters	CCC Value
3	-146	12	-61.5
4	6.75	13	-76.3
5	-21.8	14	-90.0
6	-51.7	15	-108
7	-77.1	16	-64.0
8	13.9	17	-132
9	-122	18	-140
10	-27.0	19	-100
11	-45.7	20	-110

*Note.* \* = Largest CCC Value

Clustering analysis indicates that a group of three clusters (2, 3 and 8) have similar properties – low active membership period means with much higher total membership period means (See Table 4.12). Engagement ratio, which is a result of Active Period Mean divided by Total Membership Period Mean, is less than 2 for these clusters. As suggested by very low membership active period mean values, these clusters consist of online learners that have participated very sparsely in the online community. Their sparse participation ranges from 8.40 days to 9.26 days, which suggest their contributions to the online community is unlikely to go beyond the topics that they have initiated. Looking at the means of Total Membership Period, it is noted that they can be organized in decreasing order of membership registration with the following order: 2, 3 and 8. Learners from cluster 2 have registered for an average of 2370 days, which suggests that they have joined at an early period when the community was forming, but have never returned to engage with others after a brief period of active membership.

Table 4.12

*Descriptive Statistics of Active Participation and Total Membership Period Classified by Cluster*

Cluster	Count	Active Period Mean	Active Period SD	Total Membership Period Mean	Total Membership Period SD	Engagement Ratio
1	1705	252	76.9	1024	558	24.6
2	3851	9.26	27.7	2370	387	0.390
3	11632	7.85	19.7	1310	259	0.600
4	339	1400	174	1730	391	81.0
5	595	926	121	1450	484	64.0
6	950	549	95.1	1200	527	45.6
7	82	2210	346	2500	393	88.1
8	12065	8.42	21.3	433	247	1.94

Furthermore, clustering analysis demonstrates that a group of three clusters (4, 5 and 7) have similar properties in terms of their persistence in participation in the community. This is determined by comparing their active membership period means with their total membership period means in terms of engagement ratio. For instance, the engagement ratio for these clusters (4, 5 and 7) are above 64. This comparison suggests that learners grouped into these three clusters have extensively participated throughout their membership period and persisted as a participant in the online community. Finally, clustering analysis also point to the presence of two clusters of members (1 and 6) that have participated moderately in the forums but have ceased their participation after a



significant period of time. Engagement ratio, which is a result of Active Period Mean divided by Total Membership Period Mean, is between 24.6 and 45.1 for these two clusters of users. The group of two clusters consists of members that have actively engaged with the community for close to a year but have since left the community. An observed characteristic about these two clusters of members is that they have been registered members years after the formation of the community, which suggests that they are not founding members who have left the community after a period of participation and engagement with the community.

The Kruskal-Wallis test is employed to examine whether the means of the population representing the variable are similar. With the Kruskal-Wallis test, the chi-square statistic evaluates the presence of differences in mean ranks to examine if the null hypothesis that the medians are equal across the clusters. Kruskal-Wallis tests suggest there are significant differences between clusters pertaining to each individual-level variables ( $p < 0.001$ ). For instance, test for the variable of Quoted References yielded a p-value of less than the level of significance of 0.001 with a  $\chi^2$  value of 4020 and rejects the null hypothesis that there are no significant differences between the 8 clusters. Kruskal-Wallis score means provides support to the previous argument that clusters 2, 3 and 8 have similar properties. As demonstrated in Table 4.13, these three clusters have consistently lower Score Means for the individual level variables (Quoted References, Terms, Replies, Resources, Files and Attachments). Similarly for the clusters 4, 5 and 7, they have Score Means that range from 22,600 to 29,400 across the individual-level variables. For the clusters 1 and 6, these two clusters have score means that range from 19,700 to 27,100.

Table 4.13

*Kruskal-Wallis Score Means For Individual-Level Variables, Classified by Cluster*

Cluster	Quoted References ( $\chi^2=4020$ )	Terms ( $\chi^2=5570$ )	Replies ( $\chi^2=7090$ )	Resources ( $\chi^2=4350$ )	Files ( $\chi^2=4260$ )
2	14900	12000	12800	13800	14400
3	14600	14000	14000	14600	14900
8	15100	15200	14600	15300	15100
1	19700	24200	25900	20000	19100
6	21100	25800	27100	22200	21200
4	24700	28300	28900	25700	25100
5	22600	27100	27900	23500	22700
7	27300	29000	29400	27200	27400

Note. \* = not significant at .01 level

#### 4.4.1 Nonparametric Correlation Analysis Based on Cluster Classification

Classified by clusters, correlation analysis was conducted for Individual Variables with Individual Total Interactions as well as Individual Variables with Individual Active Membership Period. Individual Total Membership Period is not considered due to weak correlation statistics (see Table 4.10) Non-parametric correlation statistics (Spearman's  $\rho$  and Kendall's  $\tau$ ) was used as KSL tests indicated non-normality of data distribution.

All p-values are below 0.01 and indicates statistical significant associations for Individual Variables and Individual Total Interactions for all clusters. As demonstrated in Table 4.14, Spearman's  $\rho$  values suggest that the majority of  $\rho$  values between Individual Variables with Individual Total Interactions are above 0.5 for Cluster 4, 5 and 7. This

indicates that there is overall strong and positive association between Individual Variables and Individual Total Interactions for Clusters 4, 5 and 7. On the other hand, Spearman's  $\rho$  values suggest the majority of the  $\rho$  values are below 0.5 between Individual Variables with Individual Total Interactions for Cluster 1 and 6. This suggests an overall positive and moderate association between Individual Variables and Individual Total Interactions. Also, the majority of Spearman's  $\rho$  values are below 0.3 between Individual Variables with Individual Total Interactions for Cluster 2, 3 and 8. This in turn demonstrates an overall weak and positive association between Individual Variables with Individual Total Interactions for Cluster 2, 3 and 8.

Table 4.14

*Non-parametric Correlation Analysis For Individual Variables with Individual Total Interactions, Classified by Cluster (Spearman's  $\rho$ )*

Cluster	Quoted References	Terms	Replies	Resources	Files
2	0.298	0.202	0.359	0.142	0.135
3	0.263	0.218	0.398	0.175	0.140
8	0.329	0.305	0.472	0.187	0.153
1	0.419	0.562	0.628	0.436	0.372
6	0.512	0.689	0.749	0.570	0.519
4	0.684	0.852	0.879	0.778	0.745
5	0.576	0.721	0.770	0.591	0.573
7	0.732	0.835	0.865	0.758	0.743

Note. \* = not significant at .01 level

The majority of the p-values are below 0.01 and indicates statistically significant associations for most bivariate correlations between Individual Variables and Individual Active Membership Period for all clusters except for Cluster 6 and 7. As demonstrated in Table 4.15, Spearman's  $\rho$  values suggest that the majority of  $\rho$  values between Individual Variables with Individual Active Membership Period are below 0.3 for Cluster 4 and 5. This indicates that there is overall weak and positive association between Individual Variables and Individual Active Membership Period for Clusters 4 and 5.

On the other hand, Spearman's  $\rho$  values suggest the majority of the  $\rho$  values are below 0.3 between Individual Variables with Individual Active Membership Period for Cluster 1 and 6. This suggests an overall positive and weak association between Individual Variables and Individual Total Interactions. Also, the majority of Spearman's  $\rho$  values are above 0.3 between Individual Variables with Individual Active Membership Period for Cluster 2, 3 and 8. This in turn demonstrates an overall moderate and positive association between Individual Variables with Individual Active Membership Period for Cluster 2, 3 and 8. Overall, these  $\rho$  values indicates that the relationships between variables are consistent across the three groups of clusters. In the next chapter, further discussion will be carried out to point out how the clusters can be understood as three groups termed as "Engaged", "Transient" and "Disengaged".

Table 4.15

*Non-parametric Correlation Analysis For Individual Variables with Individual Active Period, Classified by Cluster (Spearman's  $\rho$ )*

Cluster	Quoted References	Terms	Replies	Resources	Files
2	0.319	0.473	0.724	0.251	0.236
3	0.302	0.496	0.699	0.258	0.223
8	0.341	0.508	0.696	0.268	0.254
1	0.0706	0.107	0.106	0.102	0.0897
6	0.0579*	0.0767	0.0775	0.0731	0.103
4	0.122	0.135	0.141	0.126	0.152
5	0.119	0.164	0.152	0.154	0.160
7	0.196*	0.120*	0.193*	0.185*	0.198*

*Note.* \* = not significant at .01 level

## **Chapter 5**

### **Discussion**

The first research goal of this research is to describe the state of knowledge creation in the online communities through descriptive statistics of variables suggestive of engagement with knowledge creation and variables that may influence knowledge creation at the topic and individual level. Secondly, this research sought to examine the relationships between topic and individual characteristics that are supportive of knowledge creation and variables that are suggestive of engagement in knowledge creation. As most studies reviewed in this research has been carried in the context of classrooms and courses, this study aims to contribute to the knowledge pool by making connections between the existing knowledge pool with the findings from much larger groups of learners from online engineering communities. Thirdly, the research aims to group learners based on their membership and active participation in the online community, and to examine the relationships between individual characteristics. The following sections discuss the research findings with relevant literature in knowledge creation to address its significance and contribution to the betterment of the academic literature surrounding knowledge creation and online communities.

#### **5.1 State of Knowledge Creation**

In this study, a statistical descriptive study was carried out to examine the state of knowledge creation by looking at a collection of variables at the community level, topic-level and individual-level. The selection of these variables aligns with research notions and

empirical findings as suggested by a review of literature to be predictors of knowledge creation (see Table 3.4 and Table 3.5).

Overall, the state of knowledge creation in this online community is vibrant when considering the total cumulative numbers at the community level. There are a total of 65,209 topics and a total of 503,908 messages in this forum. A total of 66,072 links to web resources and a total of 33,379 links to digital files were shared. Furthermore, a total of 45579 unique file attachments were contributed to the online communities. 1,493,090 engineering terms were used from a list of 12,379 engineering terms predetermined from the IEEE Dictionary of Electrical and Electronics Terms (See Appendix B). The described frequencies of count demonstrate that the community overall is accumulating engineering knowledge through shared products of engineering collaboration which are suggestive of engagement with knowledge creation. This concurs with the general research notion that online communities enable conditions for sustained work around shared artifacts and collaboration in pursuit of shared purpose (Hong et al., 2010; Xia et al., 2009).

At the topic-level, results showed that on average, each topic will consist of approximately 8 messages, a network size that indicates the presence of 4 unique participants, 23.886 engineering terms per message, 1 link to online resources, 1 link to online digital files and 1 attachment. This result provides support that discussion carried out on topics received rich contributions from learners in the online participants. However, findings from this research do not suggest that the majority of the discussion topics demonstrate intense advancement of material artifacts. This contrasts against prior research of knowledge creation who found the advancement of material artifacts is substantial in formal learning environments. In a study of a science classroom that adopted the

knowledge creation pedagogy, hundreds of designs and experiments were conducted over a period of 13 months (Seitamaa-Hakkarainen et al, 2010).

However, descriptive examination of participation activities at the topic level reveals the presence of a small number of topics that received a larger portion of contributions over extended periods of time. One observation is that there is large standard deviation across most variables which suggest that there is significant variation in the descriptive data. This suggests that while some topics may receive greater contributions from users, others may not receive a single contribution. One potential explanation for this observation is that there are numerous topics that have similar themes and learners are likely to be referred to inactive threads that may contain the information they need. Similar research notions have been brought up by van Alst (2009) who found that duplicate messages are written in an online knowledge creation environment where students unknowingly contribute duplicate notes to assigned questions and furthermore, Kornish and Ulrich (2011) who found redundancy and repetition in idea generation in a classroom setting.

The median topic view is high as compared to the number of unique members participating in the topic at a ratio of 350:1. The relatively high number of topic views as compared to unique participants in the topic provides support that lurking is prevalent in this online community. This concurs with the research notion that lurking is a common and prevalent activity in many online communities (Preece et al, 2004; Nonnecke et al., 2006). The presence of lurkers on online communities is captured by the topic views metric on the forum software and despite of the lack of participation by lurkers, their views of discussion topic add to the cumulative views of discussion topics. Furthermore, this finding



support the research notion that the discussion topics not only serve as discussion avenues but act as an archival source of information for learners who have similar learning interests but have not showed up at the active period of discussion (Bian et al., 2008).

At the individual level, descriptive statistics reveal that individuals will on average initiate approximately 2 topics, 16 messages, 4 quoted references, 20 unique interactions with other learners, 495 engineering terms, 2 links to online resources, 1 link to online resources. These numbers are not high in particular, as contrasted against researcher findings that suggest that individual engaged in knowledge creation tend to participate over periods of several months and contribute hundreds of messages (Sun et al., 2009; van Aalst, 2009). The mean and median frequency count does not provide detailed information about the distribution of participation. An examination of the standard deviations of participation variables that are suggestive of knowledge creation at the individual level indicated large variability and suggests that the selection of the variables do not accurately measure participation in knowledge creation for all learners in the online community. This has been a challenge for several researchers and echoes the calls for more accurate measures of learning that can represent learning at all types of learner (Siemens & Baker, 2012; Shum & Ferguson, 2012).

As proposed by learning scientists, instructors and experts are key players in the process of knowledge creation and often play a pivotal role in advancing knowledge (Seitamaa-Hakkarainen, Viilo & Hakkarainen, 2010). In this online community, the varied distribution of participation attests to the presence of a group of highly active members that is akin to the presence of instructors in online discussion forums and formal classroom environments. Specifically, the large standard deviations across most variables are

suggestive of moderate variation in the variables. The variation in the data suggests to a certain extent that a smaller group of users are contributing in high levels of frequencies to the online communities. This group of learners is driving the process of knowledge creation through sharing resources and digital files, by consistent use of engineering terms, by being points of contact and getting in/out interactions, and providing strong levels of presences throughout the community.

## **5.2 Strengths of Associations Between Topic-level Variables with Topic Length, Topic Duration and Topic Views.**

Non-parametric correlation analysis demonstrates that there are strong associations only between the Topic Length and topic-level variables that suggest group engagement with knowledge creation. There is overall moderate correlation between Topic Duration and topic-level variables that suggest engagement with knowledge creation. In addition, there is weak correlation between Topic Views and topic-level variables that suggest engagement with knowledge creation. These findings provide support for prior research that sustained engagement and community activity are difficult parameters to measure and to account for in learning environments (Bishop, 2007; Preece, 2001). While researchers have noted that sustained engagement with online discussion and artifact advancement over time are indications of rich knowledge creation processes (Lakkala et al, 2009; Paavola et al., 2007), this research contributes a deeper understanding of the limitations of using continuous variables as part of characterizing participation over time as a quantitative parameter.

The relationships between the three variables of Topic Length, Topic Duration and Topic Views and the six topic-level variables are of interest, as this research looks for associations that can assist with understanding the occurrence of activities suggestive of knowledge creation. There is evidence to support that the relationships between Topic Length and the topic-level variables are the strongest, with the half of pair-wise correlations being strong associations. The relationships between Topic Duration and topic-level variables are overall moderate, with half of the pairwise correlations strong but the other half being weak associations. On the other hand, the relationship between Topic Views and the topic-level variables are the weakest, with half of the pair-wise correlations being weak associations. Overall, there is reason to suggest that Topic Length is an accurate variable that can be used to provide data about engagement in knowledge creation, with sufficient accuracy to test hypotheses relating summative variables with topic-level variables that are suggestive with online knowledge creation. This finding points to the importance of putting in place measurement systems that can inform assessment of online learning but also outlines the challenge of answering calls to develop accurate and reproducible approaches to examine relationships between simple user behavior and complex learning objectives (United States Department of Education, 2013).

### **5.2.1 Strengths of Associations Between Topic Length and Topic-Level Variables**

The strong association between Topic Length and the Network Size ( $\rho = 0.84$ ) suggests that longer discussion topics are likely to attract the participation of greater numbers of unique individual in support of advancing conceptual artifacts and material artifacts. This finding supports the research position that longer discussion is likely to be associated with the engagement amongst a larger number of unique learners, who in turn contribute unique expertise and perspective to the discussion. As argued by Bielaczyc et al. (2006), the involvement of experts and insiders are likely to bring about ideas that are “cobbled together” from past ideas and further advance current ideas encased in discussion topics.

The strong statistical correlation between Topic Length and Terms ( $\rho = 0.82$ ) indicates that the number of messages in each topic is associated with the use of engineering terms. This finding outlines the practice of the forum in the sense that informal chit chat or socialization is unlikely to feature heavily in the discussion topics as compared to technical discussion. This provides further support that the overall community focus is on advancement of technical ideas. The correlation analysis also indicates that lengthy discussion topics are likely to be accompanied by widespread use of engineering terms and they in turn present a rich source of engineering knowledge for learners who did not initially participate in the discussion but subsequently access the discussion topic for information. This finding provides further support to the notion that lengthy discussion archives are rich sources of information that are sought after by information seekers (Adamic et al., 2007; Zhang et al., 2009).

The strong association between Topic Length and Quoted References ( $p = 0.69$ ) indicates that discussion topics that are consisted of more messages are likely to be accompanied by in-topic references between learners. What this suggests also is that longer discussion provides the problem space to work with, which fosters social interactions between learners in form of quoted references between individual learners. This provides support to the notion that more sustained participation is likely to be linked with engaged social interactions (Paavola & Hakkarainen, 2005; 2009). As quoted references are optional features that are targeted replies between an individual learner and another, the strong relationships between Topic Length and Quoted References also outline the tendency for targeted responses to intensify in discussion threads that are intended to draw the opinions of many others rather than general truths (Aikawa et al., 2011), which in turns contribute to the length of the discussion.

There are moderate correlations between Topic Length and variables suggestive of the advancement of material artifacts. Findings suggest that topic length has moderate strengths of associations between the length of discussion topics and the occurrence of use of external resources, digital files and contribution of unique attachments. Lengthy discussion is unlikely to attract the advancement of these material artifacts. This provides support to the research notion that specific technology affordances can better facilitate the process of manipulating material artifacts such as sketches, schematics and engineering files in online forums (Hakkarainen et al., 2006). KP-Lab is one example of an integrative environment that can better foster co-development of learning artifacts (Hakkarainen et al., 2006). In this environment, learners are able to view knowledge objects from different perspectives, and access tools to create and edit various kinds of

diagrams. These features are not available on the vBulletin software platform used by the All About Circuits online community.

### **5.2.2 Strengths of Associations Between Topic Duration and Individual Variables**

One of the two strong correlations gathered from this cross-section of correlational analysis was between Topic Duration and Terms, in which Spearman Correlation indicates strong association between the amount of time elapsed for online discussion and the use of engineering terms in support of advancing conceptual artifacts ( $\rho = 0.54$ ). The second strongest correlations gathered from this cross-section of correlational analysis was between Topic Duration and Terms, in which Spearman Correlation indicates strong association between the amount of time elapsed for online discussion and the use of engineering terms in support of advancing conceptual artifacts ( $\rho = 0.54$ ). These findings resonates with the research idea that sustained engagement with learning activities over time is related to making learning moves to advance ideas and concepts in a community (Seitamaa-Hakkarainen et al., 2010). Particularly, researchers have consistently stated that efforts and commitment to utilize knowledge creation pedagogy as guidance for teaching was critical to the success of knowledge creation pedagogy (Paavola & Hakkarainen, 2009).

There are weak correlations between Topic Duration and topic-level variables suggestive of the advancement of material artifacts including the use of online resources, online digital files and user generated attachments. The Spearman rho ranges from 0.22 (between Topic Duration and Attachments) to 0.29 (between Topic Duration and Resources). These findings does not seen to concur with learner participation in advancing material artifacts is facilitated over sustained periods of participation. As

identified by researchers (Gan, 2008; Seitamaa-Hakkarainen et al., 2010), prolonged participation over extended periods of time increases the likelihood of student use and manipulation of material artifacts in a learning environment. However, this research appears to suggest that other variables may come into consideration when examining the advancement of material artifacts in online environments. One potential explanation for the lack of strong relationships between Topic Duration and variables suggestive of the advancement of material artifacts can be attributed to the presence of duplicate topics (of a similar content area) that spurred user references to topics that has been created by members at an earlier time and researchers have deemed such behavior as a challenge to analyzing text corpus data on online forums (Claridge, 2007).

### **5.2.3 Strengths of Associations Between Topic Views and Topic-Level Variables**

Based on correlation analysis, findings suggest that there is moderate association between Topic Views and variables that are suggestive of social engagement in knowledge creation processes. The Spearman's  $\rho$  between Topic Views and Network Size is at 0.36 whereas the Spearman's  $\rho$  between Topic Views and Network Size is at 0.34. The moderate association indicates that Topic Views is not strongly associated with the two parameters, and concur with the concerns of researchers that Topic Views can be influenced by lurkers and unregistered learners- who do not contribute to the advancement of the topic but are reading the discussion messages (Nonnecke & Preece, 2003).

Based on correlation analysis, findings suggest that there is moderate association between Topic Views and variables that are suggestive of advancing material artifacts as part of knowledge creation. The Spearman's  $\rho$  between Topic Views and Resources is at

0.236 whereas the Spearman's  $\rho$  between Topic Views and Files is at 0.340. The Spearman's  $\rho$  between Topic Views and Network Size is at 0.356. The weak to moderate association indicates that Topic Views is expected, as viewing the topic is a distinctively less demanding activity than posting uniquely generated graphics and links to resources, both demonstrating involved engagement with the content area. This suggests that Topic Views can be influenced by lurkers and unregistered learners who do not contribute to the advancement of the topic by sharing resources and information.

### **5.3 Strengths of Associations Between Individual-level Variables with Individual Total Interactions, Individual Total Membership Period and Individual Active Membership Period**

The relationships between the three variables of Individual Total Interactions, Active Membership Period and Total Membership Period and the five individual-level variables are examined, with a focus on associations that can assist with understanding the individual participation in activities suggestive of knowledge creation. Non-parametric correlation analysis suggests that the relationships between Individual Active Membership Period and the individual-level variables are the strongest, with the half of pair-wise correlations being strong associations. The relationships between Individual Active Membership Period and individual-level variables are overall strong, with two of the pairwise correlations strong but the other three being moderate associations. On the other hand, the relationship between Individual Total Membership Period and the topic-level variables are the weakest, with all the pair-wise correlations being weak associations. In the following sub-sections, a discussion is carried out to determine the



relevance of the findings as compared to previous work carried out in investigation of knowledge creating communities.

### **5.3.1 Strengths of Associations Between Active Membership Period and Individual Variables**

There is strong association between Individual Active Membership Period and Replies ( $\rho = 0.77$ ) and between Individual Active Membership Period and Terms ( $\rho = 0.60$ ). These strong relationships provide support for prior research that the facilitation of knowledge creation will require extensive periods of participation in the community (Leinonen et al., 2009; Zhang et al., 2007). While researchers have noted that expert guidance is critical for knowledge creation processes (Bielaczyc & Collins, 2006; Scardamalia & Bereiter, 2006), it appears that individual active period of participation is not strongly associated with sustained engagement. This conflict against the literature outlines a pitfall of using the membership period as a predictor. Possible explanations for this have been highlighted in various literature including users taking a vacation, sporadic participation and the school seasonal effects. According to descriptive findings of the state of knowledge creation in the online community, a wide variation in individual participation can explain for the weak correlation findings.

There is moderate association between Individual Active Membership Period and Quoted References ( $\rho = 0.42$ ), between Individual Active Membership Period and Resources ( $\rho = 0.40$ ) and between Individual Active Membership Period and Files ( $\rho = 0.37$ ). The moderate relationships between Individual Active Membership Period and the two variables that suggest involvement with manipulating material artifacts (Resources and Files) suggest that individual period of participation in the online forums is not the

only contributor to the engagement with activities that seek to advance material artifacts. This is in line with the findings by Muukkonen et al. (2009) who found that experienced learners, who possessed more relevant experiences acquired outside the learning activities, are more capable of projecting their learning towards a process that focuses on advancing objects with other learners. In particular, data collected in this forum does not include personal or educational information that can be used to provide answers to further the correlational analysis. Active participants without significant experiences in the content area may not have the necessary knowledge or resources to extensively engage with content areas that have been identified as problems of understanding by other learners in the forums.

The moderate association between Individual Active Membership Period and Quoted References provides a nuanced conflict with research findings that are outlined by other researchers. As suggested by Philip (2010), not all participants in a knowledge creation environment sought to interact with others but for those who did, social interaction measures are capable of highlighting the strength of ties between individuals. The moderate relationship suggests that active participation period is not the only contributors to the acts of quoting and referencing other learners. This research opens up the idea that the learner traces originating from quoting and referencing other learners is one measure that informs but does not consistently represent social interactions in the online community.

### **5.3.2 Strengths of Associations Between Individual Total Interactions and Individual Variables**

There is strong association between Individual Total Interactions and Replies ( $\rho = 0.54$ ). This finding suggests that learner engagement with posting messages increases as their individual total interaction increase. The finding seems to support the call by educational researchers for online communities to adopt knowledge practices that serve to guide the contributions of collective efforts to advance shared artifacts (Paavola & Hakkarainen, 2009; Seitamaa-Hakkarainen, Viilo & Hakkarainen, 2010). In particular, this strong association between individual interactions with other learners and individual posting of messages to the online community supports research findings that approval and acceptance by other communities is one potential motivation for learners persisting in the online community.

The correlation between Individual Total Interactions and other variables suggestive of individual participation in engineering knowledge creation are moderate at best. It is expected that Individual Total Interactions is only moderately associated with Terms ( $\rho = 0.40$ ). This is particularly so as Quoted Interactions will require users to access another button to reply to other students and is an optional function that has to be used when users refer to one another. However, it is unexpected that Individual Total Interactions is only moderately associated with Terms ( $\rho = 0.39$ ). As suggested by literature, a learner will become more knowledgeable about the community practices upon extended communication and participation in the community (Brown & Duguid, 1991; Lave & Wegner, 1991). It appears that from this research other factors can contribute to the using and speaking in engineering words, and some examples of

enhancing one's professional abilities to communicate in language that align with the community of one's personal interests by co-teaching and engaging in activities that mimic the way an expert will communicate (Roth & Tobin 2002).

### **5.3.3 Strengths of Associations Between Individual Total Membership Period and Individual Variables**

There are very weak relationships between individual-level variables and Individual Total Membership Period. The average non-parametric correlation is very weak and negative ( $\rho = -0.03$ ) which highlights that the measure of an individual period of official membership stay in the online community are weakly associated with individual engagement with the advancement of knowledge objects in the online community. Researchers have provided one potential explanation can help understand the relationships revealed by the correlation analysis in this study: membership is required to access several functionalities in online software (Zhang et al., 2010). Several members are not active participants in the online community despite registering for an account. Registration for an account will provide an individual to unrestricted use of searching and authorship functionalities. In addition to registering to ask a question by initiating a topic, it is possible that these sparsely active members registered an account to subscribe to community mail-lists and register their preferred username for future use. Overall, the very weak associations between individual variables and Individual Total Membership Period suggest that measures of participation has to take consideration the length of time in which learners make contributions to knowledge creation rather than the length of membership period.

#### **5.4 Clustering Analysis and Correlation Analysis Based on Classification by Clusters**

Based on k-means clustering analysis of participation tendencies of 31,219 learners in the All About Circuits online community, this research uncovers the presence of eight clusters of learners classified into three groups of learners - Disengaged, Transient and Engaged. The findings attest to the research notion that the online communities are supported by various groups of users who can be classified based on their level of participation and quality of contributions in the online community. This view has been identified in studies of open source software development team by Crowston and Howison (2005). The teams of software developers make up an onion-like social structure where at the center lie people who participate in highest numbers, the next circle is the participants who are posting in medium range, and then the ones in low range and farthest from the core are the peripheral members who contribute very little in these teams. Such social organization can be seen in the open source software technical support communities (Lakhani & von Hippel, 2003) as well as the studied online engineering community suggesting that there is a core group of users that are highly proficient with assisting other learners through dyadic engagement and making contribution posts in very high frequencies.

Specifically, this research found that a relatively large number of learners supported by a much smaller group of core members who have persisted in their level of engagement and participated actively in a diverse range of discussion. The core group of 1,016 members (3.25% of all members) – Engaged – have consistently participated in the forums have consistently make contributions to the forums by responding and providing assistance. The presence of a core group of active learners is critical to a knowledge

creating community because they represent a long-term engagement with the community and their volunteer efforts sustain brief but intense interactions with a relatively larger number of learners. It also confirms the research notion that online communities are sustained by core groups of help-givers (Yang et al., 2008). The presence of a small group of experts supporting a much larger number of less active members have been documented in organizational science literature in as early as a decade ago (Lakhani & von Hippel, 2003; Crowston & Howison, 2005) and echoed even in a recent study which reports that 1% of top users contributes more than a quarter of answers in Stack Exchange, an online community for programmers (Mamykina et al., 2012; Treude et al., 2011). Since the act of providing assistance and help to members is key to the advancement of discussion in this online community, the contributions of a small select group of users yet again prove to be instrumental in the support of a proportionally larger number of users seeking help in an online engineering community.

This research illuminates the presence of a group of 2,655 learners (8.50% of all members) – named the Transient – attest to the finding that the ability to provide consistent help is not merely a time commitment and effort, but also reliant on the one's expertise level within the pertinent discussion area. For instance, studies by Singh and Twidale (2008) suggest that the process of help-giving in these communities is many-to-many and often iterative. On the other hand, Lakhani and von Hippel (2003) found that active participants on online communities are proficient at helping others: they spend only 2% of their time answering questions and 98% of their time reading questions to craft potential answers to the questions. Posnett et al (2012) examined user answering activities over time on Stack Exchange and found that the ability of users to users to

answer and score points decrease over time and that the amount of time spent on site does not correlate with answering ability. Findings from this research of the All About Circuits online community brings to the floor another perspective that advances the understanding of knowledge creating communities in the sense that a small group of online learners disengages with the online community even if they have actively participated for extended periods of time that allowed them to enhance their expertise levels.

Based on the classification of learners based on eight clusters obtained from k-means clustering analysis, correlation analysis was further performed between the two variables (Individual Total Interactions and Active Membership Period) and the variables that are indicative of individual engagement in knowledge creation. The two major findings suggest that Individual Total Interaction exhibit the strongest associations with variables for the group of Engaged whereas Active Membership Period exhibits the strongest association with knowledge creation variables for the group of Disengaged. These findings have implications for prior research conducted in the context on online communities catered towards learners in the technical domain areas. For instance, Hanrahan, Convertino and Nelson (2012) examined Stack Exchange and found little correlation between problem difficulty, as indicated by on the duration of specified question and the average expertise of involved answerers. Their findings echoes the views of other researchers who found evidence that suggests that community experts demonstrate a bias in providing assistance and prefer to answer questions where they may have a higher chance of making a contribution (Pal & Konstan, 2010). The unique contribution made by this research of All About Circuits is that there is evidence that

suggest highly active users from the Engaged group exhibit consistent engagement in knowledge creation activities as contrasted against other users from the groups of Disengaged and Transient.



## **Chapter 6**

### **Conclusions**

This research considered online communities as productive avenues for self-driven informal learning, especially around content areas such as open source software, application development and programming languages. These communities bring together participants with a range of interests and expertise level to engage in information sharing and collaboration, and learn from one another (Crowston et al., 2012). Researchers have mostly examined online communities that do not form with the primary intent to educate others. Many of them exist with the sole intent to provide help to learners and enthusiasts who are facing challenges with their task on hand. In this sense, these online help forums offer an accessible platform for discussion and collaborative learning for members of all levels of expertise (Mamykina et al., 2011; Yang et al., 2008). The research findings stand out due to its contrast against prior research work whereas the others in ways concur with what other researchers have found. They provide the basis for comparison of online engineering communities with online communities of other disciplines, which in turn deepens our understanding of the pertinence of online engineering communities in supporting engineering education.

#### **6.1 Implications of Research**

The descriptive studies of individual and topic-level variables uncover large variation in data distribution. To evaluate the normality of data, KSL tests are carried out and diagnostic plots were examined. Interpretation of the KSL tests and diagnostic plots indicate that the data distribution is non-normal and data analysis will benefit from non-

parametric rather than parametric approaches. Analysis of data indicates that not all individuals are equally responsive and active in participation and learning through online communities. Based on the descriptive examination of learning activities conducted in support of knowledge creation, deeper insights are gained into how learner interactions are shaped and how assessment practices can be developed to be response to learning traces from multimodal sources. The high number of off-topic messages raises a concern whether online discussion forums are productive avenues for significant acquisition of domain-specific knowledge and echoes the concerns raised by researchers in the domain of learning sciences (Stegmann et al., 2007). This research found that while voluntary and open online discussion in this educational setting are rich in discourse and knowledge advancement held over a long period of time, they seldom feature advanced levels of knowledge construction resulting from the integration of multiple perspectives and argumentations. This is, however, contrary to other studies who found high levels of knowledge creation within focused students groups on online discussion forums in support of formal instructional settings (Palonen & Hakkarainen, 2000; van Aalst, 2009). On the other hand, consistent with the research notion that individual participation is unequal and variable in online communities (Dearman & Truong, 2010; Lampe et al., 2010), this research also uncovered a huge variation in individual participation activity levels in the online engineering community.

Based on non-parametric correlation analysis, the study suggests potential variables that can contribute to the development of predictive analytics and operationalized variables in assessment of learning creation outcomes. At the topic level, *Topic Length* is a strong measure that has high correlations with the advancement of

*conceptual artifacts* and *material artifacts*. Due to its strong association with variables that are suggestive of engagement with online learning, the use of Topic Length as a measure can play an important role when the interest is in identifying parameters that are closely associated with topic-level engagement with learning and knowledge creation. This has been identified as a fundamental building block capable of identifying learning gains and pinpointing factors that can hamper knowledge creation. With increased awareness of learning and development over time, educators can develop intervention systems that to provide consistent levels of support to online discussion and learner discourse. This is similar to the idea as suggested by the learning analytics system, Signals, developed by Purdue University (Arnold, 2010).

At the individual level, findings from correlational analysis suggest that *Individual Total Membership Period* is a poor associative measure to be used to understand the advancement of knowledge artifacts and the engagement with social interactions. Although this research presents a preliminary step in establishing the relationship between engagements with knowledge creation with topic-level and individual level variables, the findings have suggested the avoidance of individual membership information to assess relationships with acts to create knowledge. When considered with other factors, it seems reasonable to state that the measure of *Individual Total Interactions* and *Individual Active Membership Period* will serve to be useful in characterization of predictors that can improve the ability to detect knowledge creation, and to identify risk factors that hamper knowledge creation in online communities. These measures may reveal changes in tendency to participate in knowledge creation, and detect

potential social by-products and knowledge artifacts of learner collaboration in online engineering communities.

Online communities represent a significant avenue to understand many online learning and education related issues. One fertile area of research is focused on the social structure of online communities such as newcomer socialization (Ducheneaut, 2005), free user-to-user help (Lakhani & von Hippel, 2003), and distribution of tasks (von Krogh et al., 2003). This research contributes to this research in social organization – into understanding the different types of learners and participants in online communities with a focus on their underlying contributions and social interactions. Through k-means clustering analysis, this research attests to the common research notion that the online communities are supported by the presence of three groups of members – Disengaged (88% of all individuals), Transient (10%) and Engaged (2%). Overall, this research found that a relatively large number of learners are supported by a much smaller group of core members who have persisted in their engagement and participated actively in a diverse range of discussion. The core group of members – Engaged – have consistently participated in the forums and never ceased to contribute to the forums by responding to other learners. The presence of a core group of active learners is critical to a knowledge creating community because they represent a long-term engagement with the community and their volunteer efforts sustain brief but intense interactions with a relatively larger number of learners. It also confirms the research notion that online community is sustained by a core group of help-givers (Yang et al., 2008).

## **6.2 Contributions of Research**

Measures of knowledge creation are developed in this research that will assist engineering educators with the evaluation of online discourse and online artifact advancement in online knowledge creating communities. The recent popularity of online education initiatives such as MOOCs have relied heavily on online discussion to support learners with little personalization (Daniels, 2012) and this has led to challenges in catering to the differing needs of subpopulations of learners (Kizilcec et al., 2013). Findings from this research further underscores the relevance and the usefulness of clustering analysis to better understand user structures in online discussion forums. Many online communities function under a basic principle in which that a small number of core volunteers will be able to support a large number of learners (Mamykina et al., 2011). The significance of this study is best understood under the research setting of this study in which learner contributions to online discussion is carried out voluntarily and that learners can choose to disengage from the community without a significant impact on their learning and educational experiences.

Overall, this research also highlights an opportunity for community leaders to address the design of tasks and roles to address the contrasting participation tendencies between the persisting learners and learners who are likely to dropping out as well as to foster more engaged learning experiences. Online communities have much to gain by encouraging participation from the larger pool of learners as it will in turn encourage higher levels of intellectual contributions and idea diversity – crucial to the collective advancement of knowledge creation in this online community. The central contributions of this research are to the knowledge pool surrounding the knowledge creation

framework, with a focus on how artifacts, both conceptual and physical, play a role in knowledge creation through sustained discussion on online engineering communities. Taking into account that learners are volunteering their time and offering their expertise, very specific practices have been established to allow them to maximize their voluntary efforts to foster knowledge creation in this online community. As such, help-seekers who are accustomed to these practices are more likely to attain productive learning experiences. It is also observed that many learners, either due to lack of information or their status as newcomers to the community, were often not cognizant of the community practices such as using code-tags to organize their program codes or scientific formulas, nor do they provide sufficient contextual information required to advance understanding of ideas and problems.

The research ideas outlined in this research are targeted at deepening the empirical understanding how interactional dynamics and linguistics tendencies play a role in fostering knowledge creation in online community communities. The outlined research work contributes to the deepening of knowledge base targeted at understanding the functions of online communities that are supported by discussion forum software. The research questions presented in this research, however, allow for a broader contribution, particularly useful for the engineering education research community. Particularly, this study draws attention to the use of online discussion forums beyond the engineering classroom to further engineering students' knowledge and skills.

### 6.3 Limitations and Assumptions

One major limitation of this study is that data used in the analysis covers only one major online engineering community (i.e., AllAboutCircuits.com out of the many online communities in the World Wide Web. Further, the focus of this work is on electrical and electronics engineering content areas. Hence, it is not expected that the findings can be generalized to other online engineering communities. This is especially so as the discussion platform used in other online communities may differ from the one being studied and therefore affords varying forms of participation from users. Forums may also have structural differences as the use of reputation systems, guest accounts and restricted archives that can undermine generalizability. However, the employed data processing and analytical approaches can be applied to other online communities supported by an asynchronous messaging system.

Several assumptions are made regarding the representations of conceptual and material artifacts in knowledge creation. Social interactions in the online community are also assumed to be taking place within publicly accessible messages contained within discussion topics. In this study, it is assumed that the use of engineering terms is suggestive of engagement with conceptual artifacts. It is also assumed that the use of external online resources and file attachments are suggestive of engagement with material artifacts. Discussion taking place through other means such as private messages between individuals are not accounted for.

There are several assumptions for the statistical approaches (Spearman's rho and Kendall's tau) utilized in this study. At first, the data is examined if it can be used for parametric correlation through Pearson correlation. If the assumptions for Pearson

correlation are not met, non-parametric versions of the Pearson correlation – the Spearman’s rho and Kendall’s tau – are used. These assumptions are normality, linearity and homoscedasticity (Creswell, 2009). The assumption of normality means that the data conforms to normal distribution whereas linearity refers to the assumption that there is a linear relationship between the variables examined. Homoscedasticity on the other hand refer to the expectation that the dependent variable will demonstrate equal or similar levels of variance across the range of values for an independent variable. Furthermore, correlation analysis assumes that individual observations are statistically independent and that correlation does not imply casual relationships between the variables.

#### **6.4 Future Research Directions**

This research presents the initial results of a broader research agenda that seeks to contribute quantitative analysis of large datasets to the theoretical and analytical frameworks in knowledge creation. Future research plans include the collection of data sets from other online communities in the electrical engineering domains. Potential research sites include Electro-Tech-Online.com and Electronics-Lab.com. In this research, generalization of findings to other online communities is limited because of the constraint of data collection and analysis to one online community. There is pressing need for more data to seek generalization of findings to the engineering online communities, as a whole, to strengthen the capacity to understand knowledge creation on online communities and provide actionable information to develop decision mechanisms in future designs of online systems.

The approaches adopted in this study are highly accessible with open source programming language such as Python, which means that further research can be



extended in subsequent studies by other researchers. My future research plans include the extension of statistical analysis with a focus on multivariate modeling of individual participation parameters based on the independent predictor variables that are examined in this research. This research also seeks to contribute to the development of predictive tools that identify the onset of knowledge creation in online discussion and to provide actionable information for educators to supporting learners in knowledge creation more quickly. Future research will also consider the examination of varying degree of learner participation in the advancement of knowledge artifacts and the degree of social engagement amongst learners. This will shed light about definitive characteristics of various types of learners in online communities and the differences in their quality of contributions. An envisioned approach will be a content analysis of learner discourse that focuses on different groups of learners identified in clustering analysis. While quantitative analysis of large datasets allow for exploration and validation, a small-scale qualitative approach is likely to provide more detailed information on specific users in the online community and extend the research scope in this research.

As the knowledge creation perspective is built on a prominent theoretical perspective in computer supported collaborative learning known as knowledge building, the outlined research ideas have the potential to contribute to the prior work on knowledge building. This is particularly so as the knowledge building is predominantly a technology-mediated pedagogy and knowledge building is fostered in a classroom environment and facilitated by online discussion software. The focus of this research is however on analyzing and examining informal online communities where there is limited presence of educational practitioners such as teachers guiding and fostering knowledge advancement.

This research draws greater attention to informal settings as knowledge creating communities by focusing on the extent of knowledge advancement in informal online communities.

## References

- Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008). Knowledge sharing and Yahoo! Answers: Everyone knows something. In *Proceedings of the International World Wide Web Conference 2008*, (pp. 665-674), Beijing, China.
- Aikawa, N., Sakai, T., & Yamana, H. (2011). Community QA question classification: Is the asker looking for subjective answers or not. *IPSJ Online Transactions*, 4, 160-168.
- Allen, I. E., & Seaman, J. (2007). *Online nation: Five years of growth in online learning*. Newburyport, MA: The Sloan Consortium.
- Anderson, K. J. (2001). Internet use among college students: An exploratory study. *Journal of American College Health*, 50(1), 21–26.
- Arnold, K. E. (2010). Signals: Applying academic analytics. *Educause Quarterly*, 33(1).
- Bakharia, A. & Dawson, S. (2011). SNAPP: A Bird's-Eye View of Temporal Participant Interaction. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, (pp. 168-173), Alberta, Canada.
- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of International AAAI Conference on Weblogs and Social Media*, San Jose, CA.
- Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah, NJ: Lawrence Erlbaum Associates.

- Bian, J., Liu, Y., Agichtein, E., & Zha, H. (2008, April). Finding the right facts in the crowd: factoid question answering over social media. In *Proceedings of the 17th international conference on World Wide Web* (pp. 467-476). ACM.
- Bielaczyc, K., Collins, A., O'Donnell, A. M., Hmelo-Silver, C. E., & Erkens, G. (2006). Fostering knowledge-creating communities. *Collaborative learning, reasoning, and technology*, 37-60.
- Bird, S., Loper E. & Klein, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media Inc.
- Bishop, J. (2007). Increasing participation in online communities: A framework for human-computer interaction. *Computers in human behavior*, 23(4), 1881-1893.
- Boguraev, Branimir & Christopher Kennedy (1999). Salience-based content characterisation of text documents. In Inderjeet Mani & Mark T. Maybury (Eds.), *Advances in Automatic Text Summarization*, pp. 99–110. Cambridge, MA: MIT Press.
- Bourne, J., Harris, D., & Mayadas, F. (2005). Online engineering education: Learning anywhere, anytime. *Journal of Engineering Education*, 94(1), 131–146
- Brown, J. S., & Duguid, P. (1991). Organizational learning and communities of practice: toward a unified view of working, learning and innovation, *Organization Science*, 2(1), 40-57.
- Bruckman, A. (2006). Learning In Online Communities. In R. K. Sawyer Ed., *The Cambridge handbook of the learning sciences*. pp. 461–472. New York, NY: Cambridge University Press.

- Calabrese Barton, A. & Tan, E. (2010). We be burnin: Agency, Identity and Learning in a Green Energy Program. *Journal of the Learning Sciences*. 19(2), 187-229
- Cambridge, D. & Perez-Lopez, K. (2012). First steps towards a social learning analytics for online communities of practice for educators. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, (pp. 69-72), Vancouver, British Columbia.
- Cantor, D. I. (2009). Discussion Boards as Tools in Blended EFL Learning Programs. *PROFILE Issues in Teachers Professional Development*, 11(1), 113-121.
- Chen, B., Scardamalia, M., Resendes, M., Chuy, M., & Bereiter, C. (2012). Students' intuitive understanding of promisingness and promisingness judgments to facilitate knowledge advancement. In J. van Aalst, K. Thompson, M. J. Jacobson, & P. Reimann (Eds.), *The future of learning: Proceedings of the 10<sup>th</sup> international conference of the learning sciences*, (pp. 111-118). Sydney, Australia.
- Cheung, W. S. & Hew, K. F. (2004). Evaluating the extent of ill-structured problem solving process among pre-service teachers in an asynchronous online discussion and reflection log learning environment. *Journal of Educational Computing Research*, 30(3), 197-227.
- Chubin, D., Donaldson, K., Olds, B., & Fleming, L. (2008). Educating generation net – Can U.S. engineering woo and win the competition for talent? *Journal of Engineering Education*, 97(3), 245–257.
- Claridge, C. (2006). Constructing a corpus from the web: message boards. *Language and Computers*, 59(1), 87-108.

- Cohen, L., Manion, L. & Morrison, K. (2007). *Research Methods in Education (6th ed)*. London & New York: Routledge.
- Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches (3rd ed.)*. Thousand Oaks, CA: Sage.
- Crowston, K., & Howison, J. (2005). The social structure of free and open source software development. *First Monday*, 10(2).
- Crowston, K., Li, Q., Wei, K., Eseryel, U. Y., & Howison, J. (2007). Self-organization of teams for free/libre open source software development. *Information and software technology*, 49(6), 564-575.
- Crowston, K., Wei, K., Howison, J., & Wiggins, A. (2012). Free/Libre open-source software development: What we know and what we do not know. *ACM Computing Surveys (CSUR)*, 44(2), 7.
- De Liddo, A., Shum, S. B., Quinto, I., Bachler, M. & Cannavacciuolo. (2011). Discourse-Centric Learning Analytics. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, Alberta, Canada.
- Dearman, D., & Truong, K. N. (2010, April). Why users of yahoo!: answers do not answer questions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 329-332). ACM.
- Dennen, V. P. (2005). From message posting to learning dialogues: Factors affecting learner participation in asynchronous discussion. *Distance Education*, 26, 127-148.

- Ducheneaut, N. (2005). Socialization in an open source software community: A socio-technical analysis. *Computer Supported Cooperative Work (CSCW)*, 14(4), 323-368.
- Duval, E. (2011). Attention please!: learning analytics for visualization and recommendation. *In Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, Alberta, Canada.
- Engeström, Y. (1987). *Learning by expanding. An activity-theoretical approach to developmental research*. Helsinki: Orienta-Konsultit.
- Engeström, Y. (1999). *Innovative learning in work teams: Analyzing cycles of knowledge creation in practice*. In Y. Engeström, R. Miettinen & R.-L. Punamäki (Eds.) *Perspectives on activity theory*, (pp. 377- 404). Cambridge: Cambridge University Press
- Ferguson, R. & Buckingham Shum, S. (2012). Social learning analytics: five approaches. *In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, Vancouver, Canada.
- Foehr, U. G. (2006). *Media multitasking among American youth: Prevalence, predictors and pairings*. Menlo Park, CA: Henry J. Kaiser Family Foundation
- Fournier, H., Kop, R. & Sitlia, H. (2011). The Value of Learning Analytics to Networked Learning on a Personal Learning Environment. *In Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, Alberta, Canada.
- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2011). *How to design and evaluate research in education (8th edition ed.)*. New York: McGraw-Hill.

- Garrison, D. R. (2003). Cognitive presence for effective asynchronous online learning: The role of reflective inquiry, self-direction, and metacognition. In J. Bourne and J. C. Moore (Eds.) *Elements of Quality Online Education: Practice and direction*. (p. 47-58). Needham, MA: Sloan-C.
- Garrison, D. R. (2007). Online community of inquiry review: Social, cognitive, and teaching presence issues. *Journal of Asynchronous Learning Networks*, 11(1), 61-72.
- Garrison, D. R., Anderson, T., & Archer, W (2000). Critical inquiry in a text-based environment: Computer conferencing in higher education, *The Internet and Higher Education*, 2, pp. 87–105.
- Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education*, 15, 7–23.
- Gazan, R. (2006). Specialists and synthesists in a question answering community. *Proceedings of American Society for Information Science and Technology*, 43(1), 1-10.
- Granovetter, M. S. (1977). The Strength of Weak Ties. *American Journal of Sociology* 78, 1360-1380.
- Hakkarainen, K. (2009). A knowledge-practice perspective on technology-mediated learning. *International Journal of Computer-Supported Collaborative Learning*, 4(2), 213-231.
- Hakkarainen, K., Ilomäki, L., Paavola, S., Muukkonen, H., Toiviainen, H., Markkanen, H., & Richter, C. (2006). Design principles and practices for the knowledge-



- practices laboratory (KP-Lab) project. In *Innovative Approaches for Learning and Knowledge Sharing* (pp. 603-608). Springer Berlin Heidelberg.
- Hanrahan, B. V., Convertino, G., & Nelson, L. (2012, February). Modeling problem difficulty and expertise in stackoverflow. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion* (pp. 91-94). ACM.
- Harasim, L.M. (Ed.). (1990). *Online education: Perspectives on a new environment*. New York: Praeger.
- Harper, F. M., Raban, D., Rafaeli, S., & Konstan, J. A. (2008). Predictors of answer quality in online Q&A sites. In *Proceedings of CHI*, (pp. 865-874).
- Haythornthwaite, C. & Gruzd, A. (2007). A Noun Phrase Analysis Tool for Mining Online Community. In Steinfield, C., Pentland, B., Ackerman, M. & Contractor, N. (Eds.), *Communities and Technologies 2007: Proceedings of the Third Communities and Technologies Conference*, pp. 67-86.
- Haythornthwaite, C. (2006). Learning and knowledge exchanges in interdisciplinary collaborations. *Journal of the American Society for Information Science and Technology*, 57, 8, 1079-1092.
- Haythornthwaite, C. (2008). Learning relations and networks in web-based communities. *International Journal of Web Based Communities*, 42, 140-15
- Haythornthwaite, C. (2011). Learning networks, crowds and communities. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, Alberta, Canada.

- Hew, K. F., Cheung, W. S. & Ng, C. S. L. (2010). Student contribution in asynchronous online discussion: A review of the research and empirical exploration. *Instructional Science*, 38(6), 571-606.
- Hewitt, J. (2005). Toward an understanding of how threads die in asynchronous computer conferences. *Journal of the Learning Sciences*, 14(4), 567–589.
- Hong, H. Y., Scardamalia, M., & Zhang, J. (2010). Knowledge Society Network: Toward a Dynamic, Sustained Network for Building Knowledge. *Canadian Journal of Learning & Technology*, 36(1).
- Howe, N., & Strauss, W. (2000). *Millennials rising: The next great generation*. New York: Vintage.
- Internet Brands. (2013). *vBulletin*. Retrieved from <http://www.vbulletin.com>
- Jamieson, L. H., & Lohmann, J. R.. (2009). *Creating a culture for scholarly and systematic innovation in engineering education*. Washington, DC: American Society for Engineering Education.
- Johnson, L., Adams, S., & Haywood, K. (2011). *The NMC horizon report: 2011 K-12 edition*. Austin, Texas: The New Media Consortium.
- Johri, A., Teo, H., Lo, J., Dufour, M., & Schram, A. B. (2014). Millennial engineers: Digital media and information ecology of engineering students. *Computers in Human Behavior*, 33, 286-301.
- Kornish, L. J., & Ulrich, K. T. (2011). Opportunity spaces in innovation: Empirical analysis of large samples of ideas. *Management Science*, 57(1), 107-128.

- Kvavik, R. B. (2005). Convenience, communications, and control: How students use technology. In D. G. Oblinger & J. L. Oblinger (Eds.). *Technology*, 86(7), 1–20. Educause.
- Kvavik, R. B., Caruso, J. B., & Morgan, G. (2004). *ECAR study of students and information technology, 2004: Convenience, connection and control*. Boulder, CO: EDUCAUSE Center for Applied Research. R
- Lakhani, K. R., & von Hippel, E. (2003). How open source software works: "Free" user-to-user assistance. *Research Policy*, 32(6), 923-943
- Lakhani, K. R., & Wolf, R. G. (2005). Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects. In J. Feller, B. Fitzgerald, S. A. Hissam & K. R. Lakhani (Eds.), *Perspectives on Free and Open Source Software*. Cambridge: MIT Press.
- Lampe, C., Wash, R., Velasquez, A., & Ozkaya, E. (2010, April). Motivations to participate in online communities. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1927-1936). ACM.
- Lave, J., & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*, Cambridge, Cambridge University Press
- Lee, M. J., McLoughlin, C., & Chan, A. (2008). Talk the talk: Learner-generated podcasts as catalysts for knowledge creation. *British Journal of Educational Technology*, 39(3), 501-521.
- Leinonen, T., Vadén, T., & Suoranta, J. (2009). Learning in and with an open wiki project: Wikiversity's potential in global capacity building. *First Monday*, 14(2).

- Lipponen, L., Hakkarainen, K. & Paavola, S. (2004). Practices and orientations of CSCL. In J.W. Strijbos, P. Kirschner & R. Martens (Eds.), *What we know about CSCL: And implementing it in higher education*. Kluwer Academic Publishers, 3, 31-50.
- Locoro, A., Mascardi, V. & Scapolla, A. M. (2010). NLP and ontology matching: a successful combination for triological learning In Proceedings of International Conference for Agents and Artificial Intelligence.
- Macfadyen, L.P., & Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers and Education*, 542, 588-599.
- Mak, S., Williams, R., & Mackness, J. (2010). Blogs and forums as communication and learning tools in a MOOC. In *Networked Learning Conference*, University of Lancaster, Lancaster, 275-285.
- Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G. & Hartmann, B. (2011). Design lessons from the fastest Q&A site in the West. In *Proceedings of CHI* (pp. 2857–2866).
- Marcus, M., Santorini, B., & Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English : the Penn Treebank. *Computational linguistics*, 19(2), 313–330
- Minna, L., Sami, P., Kari, K., Hanni, M., Merja, B., & Hannu, M. (2009, June). Main functionalities of the Knowledge Practices Environment (KPE) affording knowledge creation practices in education. In *Proceedings of the 9th international conference on Computer supported collaborative learning-Volume 1* (pp. 297-306). International Society of the Learning Sciences.

- Muukkonen, H., & Lakkala, M. (2009). Exploring metaskills of knowledge-creating inquiry in higher education. *International Journal of Computer-Supported Collaborative Learning*, 4(2), 187-211.
- Naidu, S., & Järvelä, S. (2006). Analyzing CMC content for what? Computers & Education, 46(1), 96-103.
- Nam, K. K., Ackerman, M. S., & Adamic, L. A. (2009). Questions in, knowledge in?: a study of naver's question answering community. *In Proceedings of CHI* (pp. 779-788).
- Nasir, N. S., & Hand, V. (2008). From the court to the classroom: Opportunities for engagement, learning and identity in basketball and classroom mathematics. *Journal of the Learning Sciences*, 17(2), 143-180.
- Nonaka, I. & Takeuchi, H. (1995). *The knowledge-creating company*. New York, Oxford: Oxford University Press.
- Nonnecke, B., & Preece, J. (2003). Silent participants: Getting to know lurkers better. In *From usenet to CoWebs* (pp. 110-132). Springer London.
- Nonnecke, B., Andrews, D., & Preece, J. (2006). Non-public and public online community participation: Needs, attitudes and behavior. *Electronic Commerce Research*, 6(1), 7-20.
- Oblinger, D. (2003). Boomers, Gen-Xers and Millennials: Understanding the new students. *EDUCAUSE Review*, 38, 37–47.
- Oblinger, D., & Oblinger, J. (2005). Is it age or IT: First steps towards understanding the net generation. In D. O. A. J. Oblinger (Ed.), *Educating the net generation* (pp. 2.1–2.20). Educause Publication.

- Oshima, J., Oshima, R. & Matsuzawa, Y. (2012). Knowledge building discourse explorer: A social network analysis application for knowledge building discourse', *Educational Technology Research and Development*, 60(5), 903–921.
- Paavola, S. & Hakkarainen, K. (2009). From meaning making to joint construction of knowledge practices and artefacts: A triological approach to CSCL. In C. O'Malley, D. Suthers, P. Reimann, & A. Dimitracopoulou (Eds.), *Computer Supported Collaborative Learning Practices: CSCL2009 Conference Proceedings* (pp. 83–92).
- Paavola, S., & Hakkarainen, K. (2005). The knowledge creation metaphor – An emergent epistemological approach to learning. *Science & Education*, 14, 535-557.
- Paavola, S., Lipponen, L., & Hakkarainen, K. (2002). Epistemological Foundations for CSCL: A Comparison of Three Models of Innovative Knowledge Communities. In G. Stahl (Ed.), *Computer Support for Collaborative Learning: Foundations for a CSCL community. Proceedings of the Computer-supported Collaborative Learning 2002 Conference* (pp. 24–32).
- Paavola, S., Lipponen, L., & Hakkarainen, K. (2004). Models of Innovative Knowledge Communities and Three Metaphors of Learning. *Review of Educational Research*, 74(4), 557–576.
- Pal, A., & Konstan, J. A. (2010). Expert identification in community question answering: exploring question selection bias. *Proceedings Of International Conference On Information And Knowledge Management* (pp. 1505-1508).

- Palonen, T. & Hakkarainen, K. (2000). Patterns of interaction in computer-supported learning: A social network analysis. *Proceedings of the Fourth International Conference on the Learning Sciences* (pp. 334-339).
- Pena-Shaff, J. & Nicholls, C. (2004). Analyzing student interactions and meaning construction in Computer Bulletin Board (BBS) discussions. *Computers and Education*, 42, 243-265.
- Philip, D. N. (2010). Social network analysis to examine interaction patterns in knowledge building communities. *Canadian Journal of Learning and Technology*, 36(1).
- Piaget, J. (1970). *Science of education and the psychology of the child*. New York: Viking.
- Poole, D. M. (2000). Student participation in a discussion-oriented online course: A case study. *Journal of Research on Computing in Education*, 33(2), 162–177.
- Popper, K. R. (1972). *Objective knowledge: An evolutionary approach*. London: Oxford University Press.
- Posnett, D., Warburg, E., Devanbu, P., & Filkov, V. (2012). Mining Stack Exchange: Expertise is Evident From Initial Contributions. *Proceedings of ASE International Conference on Social Informatics*.
- Preece, J. (2001). Sociability and usability in online communities: Determining and measuring success. *Behaviour & Information Technology*, 20(5), 347-356.
- Preece, J., Nonnecke, B., & Andrews, D. (2004). The top five reasons for lurking: improving community experiences for everyone. *Computers in human behavior*, 20(2), 201-223.

- Raymond, E. S. (1999). *The Cathedral and the Bazaar*. Sebastopol, CA: O'Reilly & Associates.
- Rideout, V. J., Foehr, U. G., & Roberts, D. F. (2010). *Generation M2: Media in the lives of 8- to 18-year-olds*. Menlo Park, CA: The Henry J. Kaiser Family Foundation.
- Roth, W. M., & Tobin, K. (2002). *At the elbow of another*. New York: Peter Lang.
- Sami, P., & Kai, H. (2009, June). From meaning making to joint construction of knowledge practices and artefacts: A trialogical approach to CSCL. In *Proceedings of the 9th international conference on Computer supported collaborative learning-Volume 1* (pp. 83-92). International Society of the Learning Sciences.
- Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 97–115). New York, NY: Cambridge University Press.
- Scardamalia, M., Bransford, J., Kozma, B., & Quellmalz, E. (2012). New assessments and environments for knowledge building. In *Assessment and teaching of 21st century skills* (pp. 231-300). Springer Netherlands.
- Schellens, T. & Valcke, M. (2006). Fostering knowledge construction in university students through asynchronous discussion groups. *Computers & Education*. 46(4), 349-370.
- Seitamaa-Hakkarainen, P., Viilo, M., & Hakkarainen, K. (2010). Learning by collaborative design: technology-enhanced knowledge practices. *International Journal of Technology and Design Education*, 20, 109-136.



- Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher*, 27(2), 4–13.
- Sha., L., & van Aalst, J. (2003). "Probing individual, social, and cultural aspects of knowledge building", Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 21-25, 2003
- Shum, S. B., & Ferguson, R. (2012). Social Learning Analytics. *Journal of educational technology & society*, 15(3).
- Siemens, G. (2012). Learning Analytics: Envisioning a Research Discipline and a Domain of Practice. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, Vancouver, Canada.
- Siemens, G., & d Baker, R. S. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254). ACM.
- Singh, V. & Twidale, M.B. (2008). The Confusion of Crowds: non-dyadic help interactions. *Proceedings of Computer Supported Cooperative Work*, San Diego, CA.
- Snelson, C. (2008). YouTube and beyond: Integrating web-based video into online education. In K. McFerrin, R. Weber, R. Carlsen & D. A. Willis (Eds.), *Proceedings of Society for Information Technology and Teacher Education International Conference 2008* (pp. 732-737). Chesapeake, VA: AACE.

- Sowe, S. K., Stamelos, I., & Angelis, L. (2008). Understanding knowledge sharing activities in free/open source software projects: An empirical study. *Journal of Systems and Software*, 81(3), 431-446.
- Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge*. Cambridge, MA: MIT Press.
- Stegmann, K., Weinberger, A., & Fischer, F. (2007). Facilitating argumentative knowledge construction with computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning*, 2(4), 421–447.
- Sun, Y., Zhang, J., & Scardamalia, M. (2010). Knowledge building and vocabulary growth over two years, Grades 3 and 4. *Instructional Science*, 38(2), 147-171.
- Suthers, D., & Chu, K. (2012). Multi-mediated community structure in a socio-technical network. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 43-53), Vancouver, Canada.
- Suthers, D., Hoppe, H. U., Laat, M. & Simon Buckingham, S. (2012). Connecting levels and methods of analysis in networked learning communities. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, (pp. 11-13), Vancouver, Canada.
- Tapscott, D., & Williams, A. (2008). *Wikinomics: How mass collaboration changes everything*. New York, NY: Atlantic
- Teo, H. J., & Johri, A. (2013). Experts learn more: A exploratory study of argumentative knowledge construction in *Proceedings of Computer Supported Collaborative Learning (CSCL) 2013*, Madison, WI, USA.

- Teo, H. J., & Johri, A. (2014). Fast, functional, and fitting: Organizing characteristics of help-giving on a newcomer online forum. *Proceedings of Computer Supported Collaborative Work (CSCW) 2014*, Baltimore, MD.
- Teo, H. J., Johri, A., & Mitra, R. (2013). Visualizing and analyzing productive structures and patterns in online communities using multilevel social network analysis. *Proceedings of Computer Supported Collaborative Learning (CSCL) 2013*, Madison, WI, USA.
- Thomas, M. J. W. (2002). Learning within incoherent structures: The space of online discussion forums. *Journal of Computer Assisted Learning*, 18 (3), 351-366.
- Treude, C., Barzilay, O., & Storey, M. A. (2011, May). How do programmers ask and answer questions on the web? In *Software Engineering (ICSE), 2011 33rd International Conference on* (pp. 804-807). IEEE.
- Unal, Z. & Unal, A. (2011). Evaluating and Comparing the Usability of Web-Based Course Management Systems. *Journal of Information Technology Education*, 10, 19–38.
- United States Department of Education. (2010). *National Education Technology Plan 2010. Transforming American education: Learning powered by technology*. Washington, DC: US Department of Education.
- United States Department of Education (2013). *Expanding evidence approaches for learning in a digital world*. United States Department of Education. Office of Educational Technology.

- van Aalst, J. (2009). Distinguishing knowledge-sharing, knowledge-construction, and knowledge-creation discourses. *International Journal of Computer-Supported Collaborative Learning*, 4(3), 259-287.
- van de Sande, C. & Greeno, J. G.. (2012). Achieving alignment of perspectival framings in problem-solving discourse. *Journal of the Learning Sciences*, 21(1), 1-44.
- van de Sande, C. & Leinhardt, G. (2007). Help! Active student learning and error remediation in an online calculus e-help community. *Electronic Journal of e-Learning*. 5(3), 227-238,
- van de Sande, C. (2009). Grassroots open, online, calculus help forums. In *Proceedings of the 8th International Conference on Computer Supported Collaborative Learning* (pp. 43–47). International Society of the Learning Sciences
- van de Sande, C. (2010). Free, open, online, mathematics help forums: the good, the bad, and the ugly. In *Proceedings of the 9th International Conference of the Learning Sciences*, Chicago, IL.
- van de Sande, C. (2011). A description and characterization of student activity in an open, online, mathematics help forum. *Educational Studies in Mathematics*, 77(1), 53 –78.
- Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari, R., & Duval, E. (2011). Dataset-driven research for improving recommender systems for learning. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, (pp. 44-53), Alberta, Canada.

- Von Krogh, G., Spaeth, S., & Lakhani, K. R. (2003). Community, joining, and specialization in open source software innovation: a case study. *Research Policy*, 32(7), 1217-1241.
- Vonderwell, S. (2003). An examination of asynchronous communication experiences and perspectives of students in an online course: *A case study. Internet and Higher Education*, 6, 77–90
- Vygotsky, L. (1978). *Mind in Society*. London: Harvard University Press.
- Wang, G. , Gill, K., Mohanlal, M., Zheng, H., & Zhao, B. (2013). Wisdom in the Social Crowd: an Analysis of Quora. In *Proceedings of The 22nd International World Wide Web Conference*, Rio de Janeiro, Brazil, May 2013.
- Wasserman, S. & Katherine F. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.
- Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46(1), 71–95.
- Wellman, B. (2012). Networked individualism: how the personalized internet, ubiquitous connectivity, and the turn to social networks can affect learning analytics. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, Vancouver, Canada
- Wise, A. F., & Chiu, M. M. (2011). Analyzing temporal patterns of knowledge construction in a role-based online discussion. *International Journal of Computer-Supported Collaborative Learning*, 61, 445-470.

- Worsley, M. & Blikstein, P. (2010). *Towards the Development of Learning Analytics: Student Speech as an Automatic and Natural Form of Assessment*. Paper presented at Annual Meeting of the American Education Research Association, Denver, CO.
- Xia, H., Luo, S., & Yoshida, T. (2010). Exploring the knowledge creating communities: an analysis of the Linux Kernel developer community. *Managing Knowledge for Global and Collaborative Innovations. Series of Innovation and Knowledge Management*, 8, 385-398.
- Yang, J. & Wei, X. (2009). Seeking and offering expertise across categories: A sustainable mechanism works for baidu knows. In *Proceedings of International AAAI Conference on Weblogs and Social Media 2009*, San Jose, CA.
- Yang, J., Adamic, L. A., & Ackerman, M. S. (2008). Competing to share expertise: The taskCn knowledge sharing community. In *Proceedings of International AAAI Conference on Weblogs and Social Media 2008*, Seattle, WA.
- Zhang, J., Scardamalia, M., Lamon, M., Messina, R., & Reeve, R. (2007). Socio-cognitive dynamics of knowledge building in the work of nine- and ten-year-olds. *Educational Technology Research and Development*, 55(2), 117–145.
- Zhang, J., Scardamalia, M., Reeve, R., & Messina, R. (2009). Designs for collective cognitive responsibility in knowledge-building communities. *The Journal of the Learning Sciences*, 18(1), 7-44.
- Zhuadar, L., & Yang, R. (2012). Cyberlearners and learning resources. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, (pp. 65-68), Vancouver, Canada.

## Appendices

### Appendix A: File Types of Digital Files

*Table A.1* File Types and Corresponding Extensions

File-types	File extensions
Text Documents	doc, docx, odt, ott, oth, odm txt, rtf, pdf
Presentation	pot, ppt, pptx odp, odg, otp
Spread Sheets	xls, xlsx ods, ots
Database files	dat, db, m, mdb, sql,
Compressed Files	7z, zip, rar, tgz, tar
Image files	bmp, gif, jpg, png

## **Appendix B: Sample Engineering Terms from IEEE Dictionary of Electrical and Electronics Terms**

### **Alphabet A**

A64

AAC

AACSR

ABASIC

ABET

Absorption

AC

Accelerometer

Accumulator

Active

Actuator

Adapter

Adder

Adiabatic

Admittance

Afterpulse

Agile

Alarm

Algorithm

Algorithmic

Alias

Allocation

Allotment



## **Alphabet B**

Backscatter

Bandwidth

Basic

Bias

Bignum

Binary

Bipolar

Bistable

Bits

Blip

Board

Bolometer

Boltzmann's

Boolean

Booster

BPAM

Bps

Branch

Breadboard

Breaker

Broadband

Buffer

Buildup

Bus

## **Alphabet C**

Cables

Cache

CAD

Calibrate

Calorimeter

CAM

Centering

Centrifugal

Channel

Charge

Chip

Chronometer

Circuit

Clock

Coaxial

Codec

Coupling

CRC

Crossband

Cryptography

Current

Cyclotron

Cylinder

## **Alphabet D**

DAC

Damping

Database

Datapath

Deassembler

Decay

Decompiler

Decoupling

Dielectric

Diffusion

Diode

Directivity

Dirichlet

Discharge

Doping

Downstream

Drain

Ducibility

Dumper

Duplex

Duration

Duster

Dynatron

## **Alphabet E**

EAROM

ECCM

Echo

Eddy

EDR

EEPROM

Electricity

Electrode

Electrolyte

Emissivity

Emittance

Emulation

Encode

Encryption

Ensemble

Epilog

Equalize

Equiphase

Equipotential

Equipotential

Excitation

Executable

Exponential

## **Alphabet F**

Fadeout

Fanout

Faraday

Fault

F-bits

Feedback

Fidelity

Field-coil

Filament

Filter

Firmware

Flameproof

Flip-flop

Flowchart

Fluorescent

Flux

Fluxmeter

FM

FORTRAN

Fourier

Frequency

Fresnel

Functionality

Fuse

## **Alphabet G**

Gain

Galvanic

Gamma

Gaussian

Generator

GHz

Glossmeter

GUI

Gyration

Grid

Grating

Glow

Grazing

Gether

Gal-min

Gauge

Geiger-Mueller

Granular

Gyromagnetic

Ground

Groove

Garbage

Gimbal

## **Alphabet H**

Halftone

Hall-effect

Harmonic

Hash

HDAM

Headloss

Heavy-duty

Helical

Helmholtz

Hemispherical

Hertz

Heterojunction

Hexadecimal

HMOS

Hodoscope

Homodyne

Horsepower

Housing

Human-centered

HVDC

Hydraulic

Hydroelectric

Hypertext

Hysteresis

## **Alphabet I**

IACK

IC

Iconoscope

Identifier

Idle

IEEE

IF

IGFET

Illuminance

Immitance

Impedance

Incandescence

Inductance

Inert

Infrared

Input

Insulation

Intensifier

Interchannel

Interconnect

Interference

Interflectance

Isotropic

Iteration



## **Alphabet J**

Johnson

Join

Joined

Joint

Jointly

Joints

Jordan

JOSEF

JOSS

Joule

Joule's

Journal

Joystick

Jpd

J-scope

JTAG

Jukebox

Jump

Jumper

Junction

Junction-to-ambient

Junctor

Jurisdiction

Justification

Justified

Justify

Just-release

## **Alphabet K**

Karabiner

Karabiner

Kelvin

Kenotron

Kernel

keV

Keystroke

Keyword

kHz

Kick-sorter

Kilobyte

Kinematic

Kinescope

Kirchhoff

Klydonograph

Klystron

Knob-and-tube

Knot

KOPS

Ku-band

kV

KWAC

## **Alphabet L**

Lagrange

Laminate

LAN

Last-in-first-out

Lattice

Lead-acid

LED

Leveling

Lexicon

Life-cycle

Lifetime

Limits

Linearity

Linkage

LISP

Listener

Lithium

Load

Logarithm

Loop

Loss

Lumen

## **Alphabet M**

Macroprocessor

Magnet

Magnetron

Matrix

Maxwell

Megahertz

Memory

Messenger

Metadata

MFLOPS

Microcomputer

Microwave

Milliwatt

MIMD

Modulation

Monostable

Motherboard

Motor

Mueller

Multiplex

Mux

## **Alphabet N**

NAND

Nano

N-channel

Near-field

Nesting

Neuroelectricity

NMOS

Node

NOT

Null

Nominal

Netlist

Neutral

Navigator

Nomenclature

Noise

Network

Neuroelectricity

Newline

Negate

Narrow-band

## **Alphabet O**

Octant

Octave

Occam

Odd

Odometer

OEM

Offline

Ohmic

Online

Opacity

Optimization

Optoelectronic

Oscilloscope

Output

Overcurrent

Overlay

Outage

Oxidation

Ozone

Overvoltage

Outlet

Operating

Omnibearing

OCR

## **Alphabet P**

Packet

PAL

Parabola

Parser

Partition

PCB

p-channel

Peak

Peripheral

Periscope

Permittivity

Perturbation

pH

Perveance

Phase

Phasor

Photo

Photosensitive

Piezoelectric

Piggyback

Pins

Planck

Plasma

Plate

## **Alphabet Q**

QA

QAM

Q-Channel

Q-Switch

Quadrature

Quadruplex

Qualifier

Quality

Quantization

Quantum

Quartz

Quench

Query

Queue

Quiescent

Quill

Quinary

Quinhydrone

Quisition

Quotation

Quote

Quotient

Qwerty



## **Alphabet R**

Re-engineering

Register

Resonance

Response

Retardant

Retrieval

Rheostat

Rhombic

RIF

Ripple

RISC

RJ-11

RLC

RMS

Roadband

Robot

ROM

Rotary

Rotator

Rotor

Routine

Routine

Run-time

## **Alphabet S**

Sampling

Sawtooth

Semiconductor

Sensor

Shunt

Simplex

Solder

Solenoid

Spectrum

Splitter

Stepped

Stochastic

Storage

Stroboscope

Substrate

Superconducting

Surge

Sweep

Switches

Synchronism

Syntax

System

## **Alphabet T**

Tachometer

TEM

Terminal

Thermojunction

Thread

Throughput

Thyristor

Timer

Token

Tolerance

Toroid

Torque

Tracer

Transceiver

Transfer

Transformer

Transistor

Trigger

Triplex

Truncate

TTL

Tuner

Tunneling

## **Alphabet U**

UHF

Ultrasonic

Ultraviolet

Unbiased

Unbiased

Undervoltage

Undirected

Unidirectional

Uniformity

Uniprocessor

UNIRAM

Unity

Univalent

Unloaded

Unodecimal

Unquenched

Unsigned

UPC

Uplift

Uplink

Upstream

USASCII

UVH

## **Alphabet V**

Vacuum

Valence

Valve

Vapor

Varactor

Varhour

Variable

Variometer

VBA

V-beam

VCI

Vector

Velocity

Verify

Vernier

Version

Vertex

VHF

Vibration

Volt

Voltage

Volt-ampere

VRAM

## **Alphabet W**

Wading

Warmup

Watts

Wavelength

Wavetail

Weight

Welder

Wheatstone

Whistler

White-box

Wien

Wigwag

Winch

Windage

Window

Wire

Wireless

Withstand

Work

Workload

Wound

Write-enable

Wye

## **Alphabet X**

X

X200

X25

X3

X400

X75

X-address

X-axis

X-band

X-datum

Xerographic

Xerography

Xeroprinting

XID

XNOR

X-on-X-off

XOR

X-position

X-R

X-radiation

X-ray

X-series

XXXX

X-Y

## **Alphabet Y**

Y2K

Y-address

Yagi-Uda

Yaw

Y-axis

Y-connected

Years

Yield

Yields

Yoke

Y-position

Y-T



## **Alphabet Z**

Z-address

Z-axis

ZBASIC

Zeeman

Zeitgebers

Zener

Zero-address

Zero-based

Zero-byte

Zero-error

Zero-latency

Zero-level

Zero-modulation

Zero-period

Zero-phase

Zero-phase-sequence

Zeros

Zero-sequence

Zero-suppression

Zonal

Zonal-cavity

Zone

Zone-plate

Z-position

## Appendix C: Scatterplots for Correlation Analysis

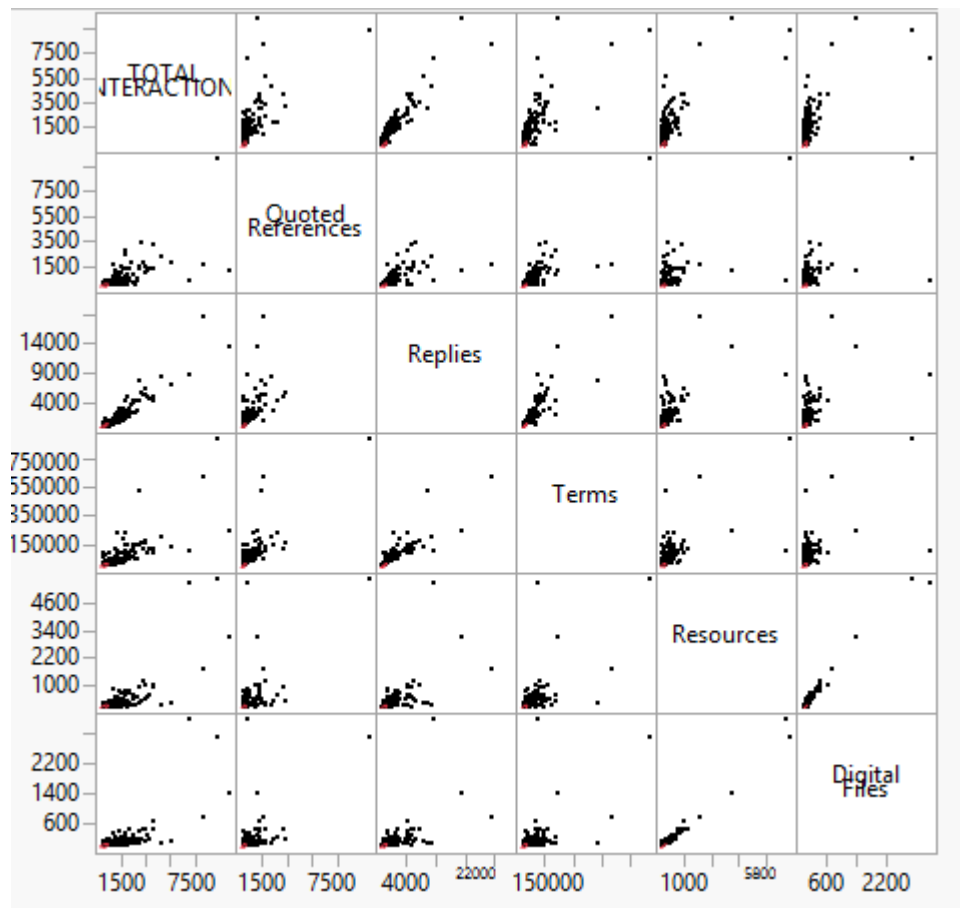


Figure C-1: Scatter Plots for Individual Total Interactions and Variables Suggestive of Engagement with Knowledge Creation

*Note:* The units for the variables across the two axes are described in Table 4.4

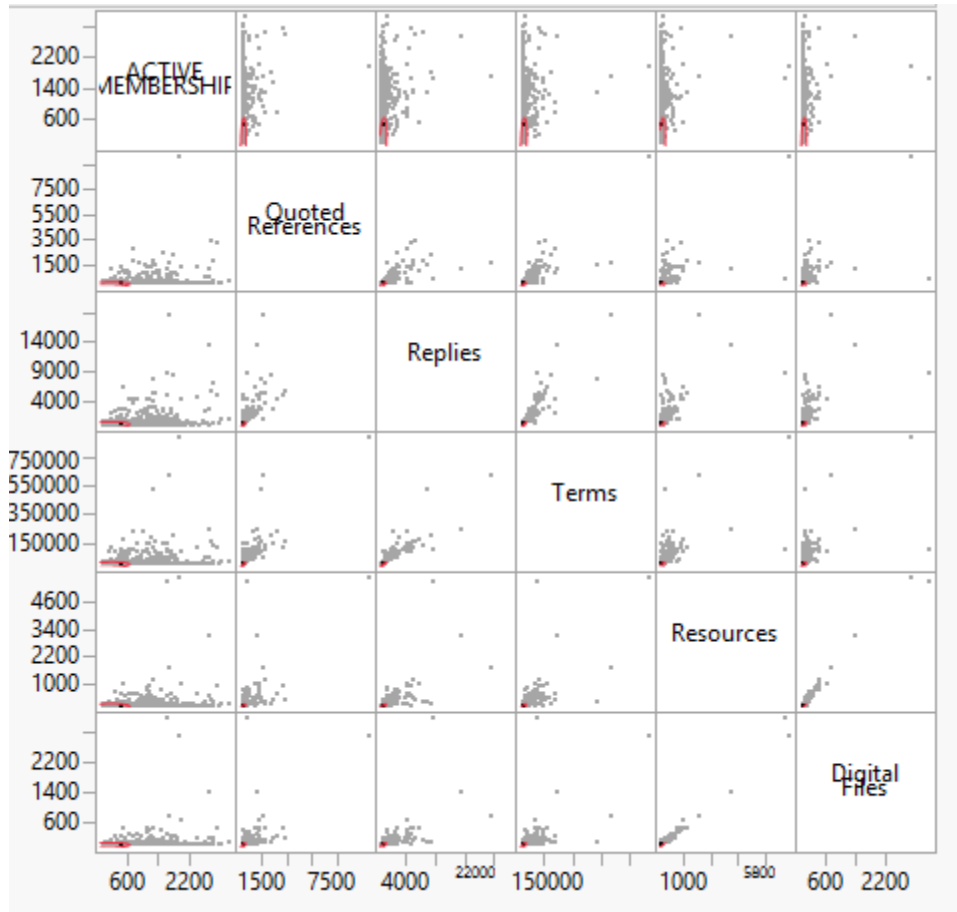


Figure C-2: Scatter Plots for Individual Active Membership Period and Variables Suggestive of Engagement with Knowledge Creation

*Note:* The units for the variables across the two axes are described in Table 4.4

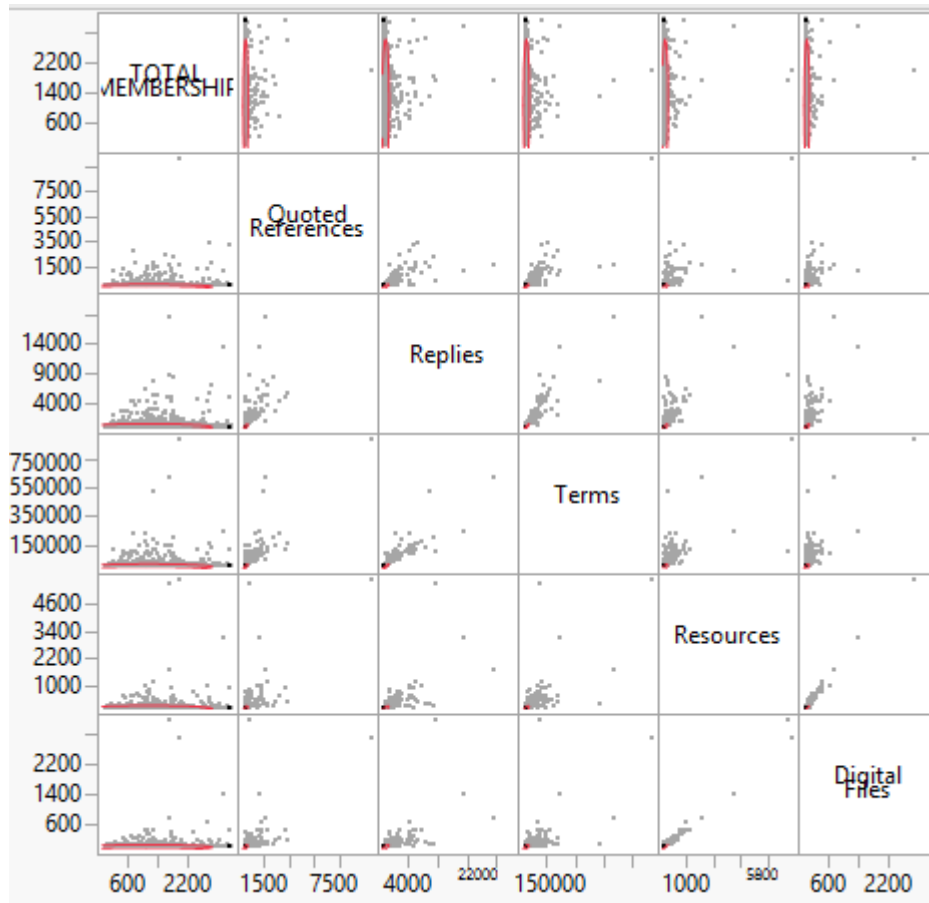


Figure C-3: Scatter Plots for Individual Total Membership and Variables Suggestive of Engagement with Knowledge Creation

*Note:* The units for the variables across the two axes are described in Table 4.4

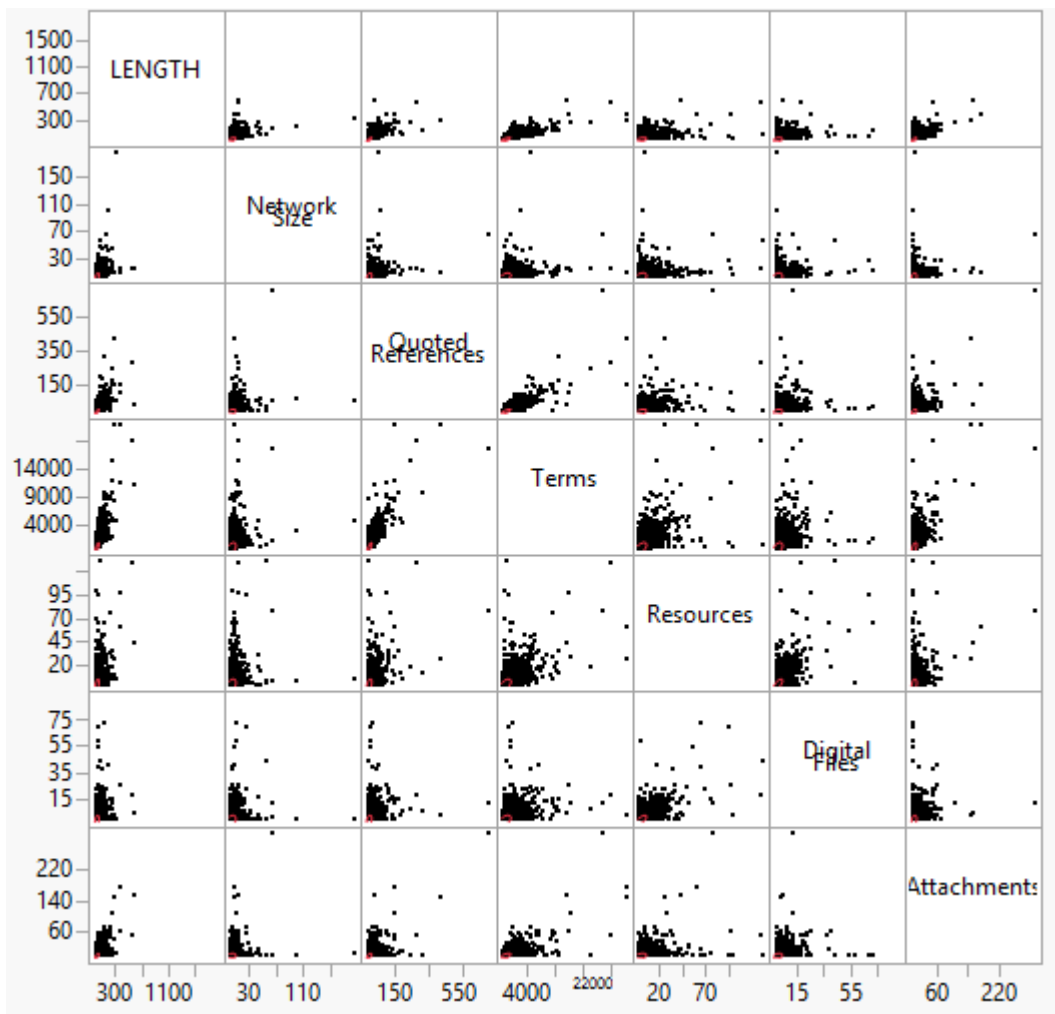


Figure C-4: Scatter Plots for Topic Length and Variables Suggestive of Engagement with Knowledge Creation

*Note:* The units for the variables across the two axes are described in Table 4.3

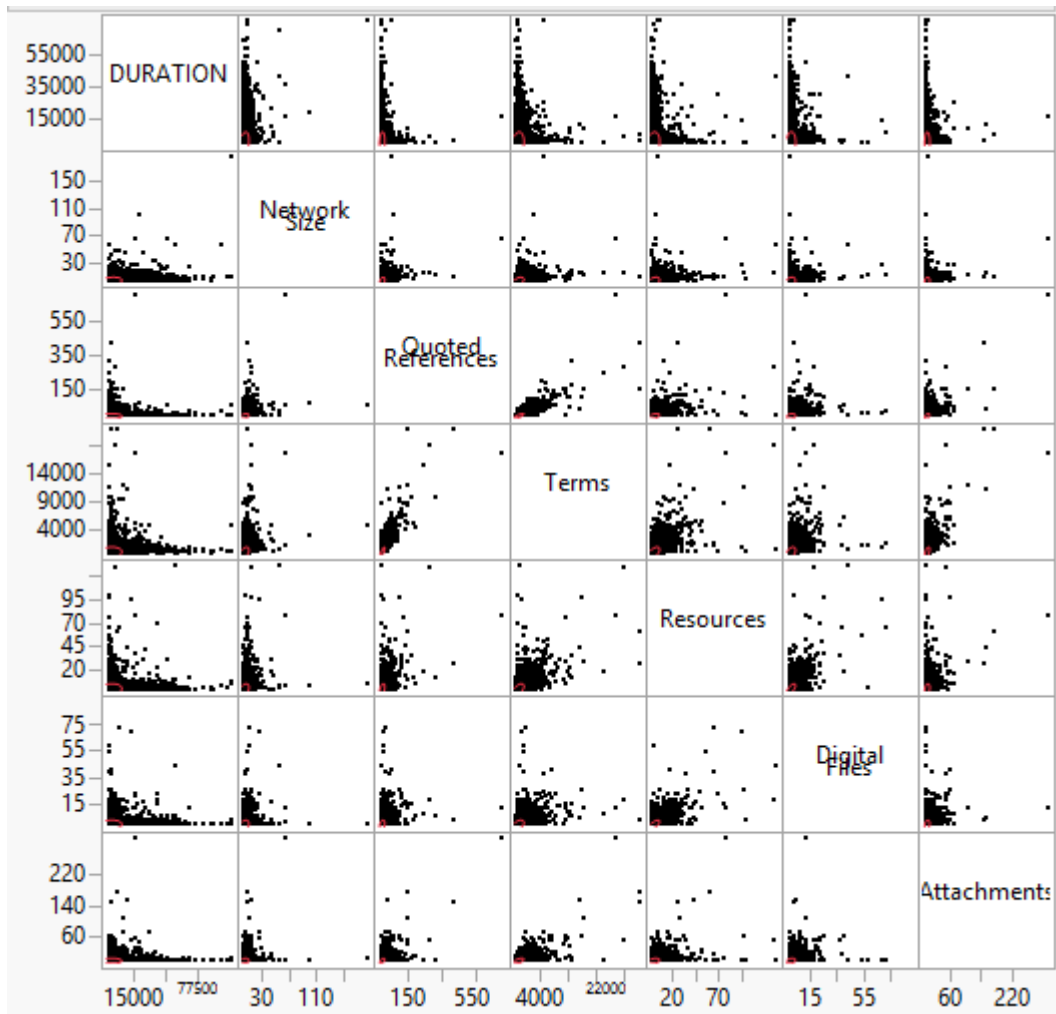


Figure C-5: Scatter Plots for Topic Duration and Variables Suggestive of Engagement with Knowledge Creation

*Note:* The units for the variables across the two axes are described in Table 4.3

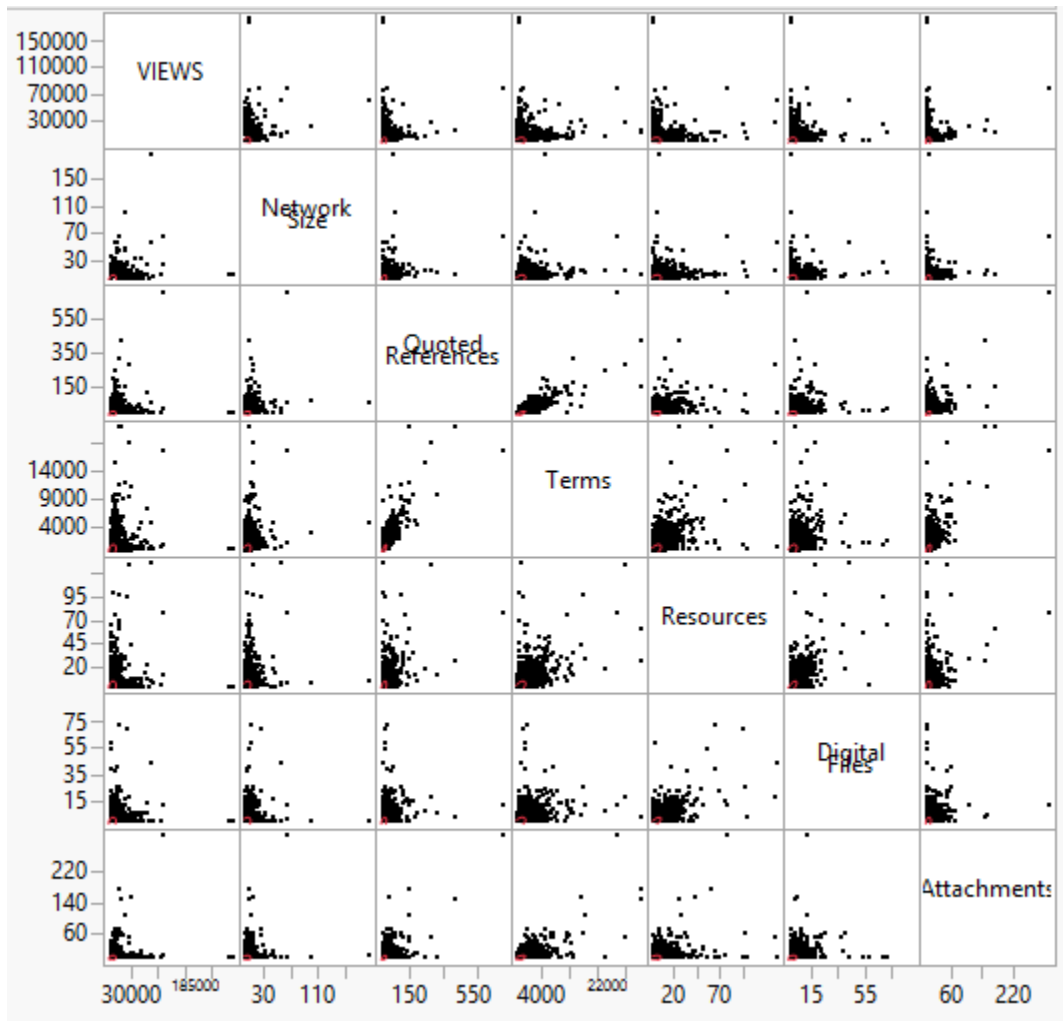
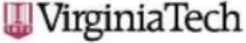


Figure C-6: Scatter Plots for Topic Views and Variables Suggestive of Engagement with Knowledge Creation

*Note:* The units for the variables across the two axes are described in Table 4.3

## Appendix D: IRB Approval Letter and Application Documentation

	<b>Office of Research Compliance</b> Institutional Review Board 2000 Kraft Drive, Suite 2000 (0497) Blacksburg, VA 24060 540/231-4606 Fax 540/231-0959 email <a href="mailto:irb@vt.edu">irb@vt.edu</a> website <a href="http://www.irb.vt.edu">http://www.irb.vt.edu</a>
---	---

**MEMORANDUM**

**DATE:** December 6, 2012

**TO:** Hon Jie Teo, Aditya Johri

**FROM:** Virginia Tech Institutional Review Board (FWA00000572, expires May 31, 2014)

**PROTOCOL TITLE:** Study of Online Discussion Forums

**IRB NUMBER:** 12-989

Effective December 6, 2012, the Virginia Tech Institutional Review Board (IRB) Chair, David M Moore, approved the New Application request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

<http://www.irb.vt.edu/pages/responsibilities.htm>

(Please review responsibilities before the commencement of your research.)

**PROTOCOL INFORMATION:**

Approved As:	<b>Exempt, under 45 CFR 46.110 category(ies) 4</b>
Protocol Approval Date:	<b>December 6, 2012</b>
Protocol Expiration Date:	<b>N/A</b>
Continuing Review Due Date*:	<b>N/A</b>

\*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

**FEDERALLY FUNDED RESEARCH REQUIREMENTS:**

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

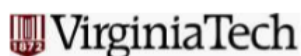
**Invent the Future**  
VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY  
*An equal opportunity, affirmative action institution*



Date*	OSP Number	Sponsor	Grant Comparison Conducted?

\* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this IRB protocol is to cover any other grant proposals, please contact the IRB office (irbadmin@vt.edu) immediately.



## Institutional Review Board Existing Data Research Protocol

Note: complete this application only if this research project only involves the collection or study of **existing data**. Once complete, upload this form as a Word document to the IRB Protocol Management System: <https://secure.research.vt.edu/irb>

### 1. DO ANY OF THE INVESTIGATORS OF THIS PROJECT HAVE A REPORTABLE CONFLICT OF INTEREST? (<http://www.irb.vt.edu/pages/researchers.htm#conflict>)

- ☒ No  
☐ Yes, explain:

### 2. WILL THIS RESEARCH INVOLVE COLLABORATION WITH ANOTHER INSTITUTION?

- ☒ No, go to question 3  
☐ Yes, answer questions within table

IF YES
Provide the name of the institution [for institutions located overseas, please also provide name of country]:
Indicate the status of this research project with the other institution's IRB: <input type="checkbox"/> Pending approval <input type="checkbox"/> Approved <input type="checkbox"/> Other institution does not have a human subject protections review board <input type="checkbox"/> Other, explain:
Will the collaborating institution(s) be engaged in the research? ( <a href="http://www.hhs.gov/ohrp/policy/engage08.html">http://www.hhs.gov/ohrp/policy/engage08.html</a> ) <input type="checkbox"/> No <input type="checkbox"/> Yes
Will Virginia Tech's IRB review all human subject research activities involved with this project? <input type="checkbox"/> No, provide the name of the primary institution: <input type="checkbox"/> Yes
<i>Note: primary institution = primary recipient of the grant or main coordinating center</i>

### 3. IS THIS RESEARCH FUNDED?

- ☒ No, go to question 4  
☐ Yes, answer questions within table

IF YES
Provide the name of the sponsor [if NIH, specify department]:
Is this project receiving federal funds? <input type="checkbox"/> No <input type="checkbox"/> Yes

If yes,

Does the grant application, OSP proposal, or "statement of work" related to this project include activities involving human subjects that are not covered within this IRB application?

- ☐ No, all human subject activities are covered in this IRB application
- ☐ Yes, however these activities will be covered in future VT IRB applications, these activities include:
- ☐ Yes, however these activities have been covered in past VT IRB applications, the IRB number(s) are as follows:
- ☐ Yes, however these activities have been or will be reviewed by another institution's IRB, the name of this institution is as follows:
- ☐ Other, explain:

Is Virginia Tech the primary awardee or the coordinating center of this grant?

- ☐ No, provide the name of the primary institution:
- ☐ Yes

**4. DOES THIS STUDY INVOLVE CONFIDENTIAL OR PROPRIETARY INFORMATION (OTHER THAN HUMAN SUBJECT CONFIDENTIAL INFORMATION), OR INFORMATION RESTRICTED FOR NATIONAL SECURITY OR OTHER REASONS BY A U.S. GOVERNMENT AGENCY?**

*For example – government / industry proprietary or confidential trade secret information*

- ☒ No
- ☐ Yes, describe:

**5. DOES THIS STUDY INVOLVE SHIPPING ANY TANGIBLE ITEM, BIOLOGICAL OR SELECT AGENT OUTSIDE THE U.S.?**

- ☒ No
- ☐ Yes

**6. DESCRIBE THE BACKGROUND, PURPOSE, AND ANTICIPATED FINDINGS OF THIS STUDY:**

In education research, there has been a rising interest in informal learning environments outside the traditional classroom. This study answers this call by examining learning in an informal learning environment where learners engage in discussion for a variety of purposes including homework help, engineering projects and deepening their conceptual understanding of engineering-related topics. The goal of this study is to examine the interactions of non-identified users in a online discussion forum or bulletin board using network analysis, content analysis and modeling techniques. The study builds on previous work in education data mining to look at specific user tendencies that arise in an online learning community. The anticipated findings for this study include a visualized social network of the community indicative of discussion quality and a statistical model that describe users' tendencies to contribute to discussion in the community.

**7. EXPLAIN WHAT THE RESEARCH TEAM PLANS TO DO WITH THE STUDY RESULTS:**

*For example - publish or use for dissertation*

This study results will be used for a dissertation project as well as to be used for academic conferences or publications. The findings will be anonymous and the researcher will not divulge identifying information of the users.

**8. WILL PERSONALLY IDENTIFYING STUDY RESULTS OR DATA BE RELEASED TO ANYONE OUTSIDE OF THE RESEARCH TEAM?**

*For example – to the funding agency or outside data analyst, or participants identified in publications with individual consent*

- ☒ No  
☐ Yes, to whom will identifying data be released?

**9. WILL ANY STUDY FILES CONTAIN PARTICIPANT IDENTIFYING INFORMATION (E.G., NAME, CONTACT INFORMATION, VIDEO/AUDIO RECORDINGS)?**

- ☒ No, go to question 10  
☐ Yes, answer questions within table

IF YES
Describe if/how the study will utilize study codes:
If applicable, where will the key [i.e., linked code and identifying information document (for instance, John Doe = study ID 001)] be stored and who will have access?
<i>Note: the key should be stored separately from subjects' completed data documents and accessibility should be limited.</i>

**10. WHERE WILL DATA BE STORED?**

Discussion pages downloaded from the forums will be in HTML (hyper-text markup language) format. After the pages have been downloaded, a computer program will be run to extract data such as the textual content and numerical content such as the data of post, order of post and numerical user identification number. The collected content will be stored in a password-secured database established in a password-secured Windows PC.

**11. WHO WILL HAVE ACCESS TO STUDY DATA?**

Only PI and co-PIs

**12. DESCRIBE THE PLANS FOR RETAINING OR DESTROYING THE STUDY DATA:**

The downloaded HTML files will be destroyed after the textual and numerical data have been extracted from the discussion pages will be downloaded from the discussion forum. The extracted data will be stored in a password-secured database within a password-secured Windows PC.

**13. FROM WHERE DOES THE EXISTING DATA ORIGINATE?**

An online discussion forum (allaboutcircuits.com)

**14. PROVIDE A DETAILED DESCRIPTION OF THE EXISTING DATA:**

The existing data is a collection of HTML web pages in a discussion forum that is open to public. There is no need to register for an account to log into the system in order to gain access to these HTML web pages. The individual web page consists of a discussion between users who contribute to the discussion with text or

images. Each post is stamped by a post identification number, date of post and the order of post in the discussion. The existing data do not contain any email addresses, telephone number or names that identifies the user.

**15. IS THE SOURCE OF THE DATA PUBLIC?**

- ☐ No, go to question 16  
☒ Yes, you are finished with this application

**16. WILL ANY INDIVIDUAL ASSOCIATED WITH THIS PROJECT (INTERNAL OR EXTERNAL) HAVE ACCESS TO OR BE PROVIDED WITH EXISTING DATA CONTAINING INFORMATION WHICH WOULD ENABLE THE IDENTIFICATION OF SUBJECTS:**

- Directly (e.g., by name, phone number, address, email address, social security number, student ID number), or
  - Indirectly through study codes even if the researcher or research team does not have access to the master list linking study codes to identifiable information such as name, student ID number, etc
- or
- Indirectly through the use of information that could reasonably be used in combination to identify an individual (e.g., demographics)

- ☒ No, collected/analyzed data will be completely de-identified  
☐ Yes,

**If yes,**

*Research will not qualify for exempt review; therefore, if feasible, written consent must be obtained from individuals whose data will be collected / analyzed, unless this requirement is waived by the IRB.*

**Will written/signed or verbal consent be obtained from participants prior to the analysis of collected data?**  
-select one-

*This research protocol represents a contract between all research personnel associated with the project, the University, and federal government; therefore, must be followed accordingly and kept current.*

*Proposed modifications must be approved by the IRB prior to implementation except where necessary to eliminate apparent immediate hazards to the human subjects.*

*Do not begin human subjects activities until you receive an IRB approval letter via email.*

*It is the Principal Investigator's responsibility to ensure all members of the research team who collect or handle human subjects data have completed human subjects protection training prior to handling or collecting the data.*

-----END-----

## Appendix E: Python Software Routine

```
import nltk                                #import the nltk tool kit for natural language processing
import inspect                             #import the inspect module to obtain class information to feed into code
import os                                  #import the os module to allow the program to access computer directories
import glob                                #import the glob module to identify files of similar pathnames
import csv                                 #import the csv module to read and write processed data into csv spreadsheets
import codecs                              #import the codecs module to recognize Unicode characters
from bs4 import BeautifulSoup              #import BeautifulSoup module to extract characters from HTML Files
```

```
# this loop looks for characters obtained from all identifiable HTML tags of tobeparsed1 iteratively in all HTML Files in a folder
```

```
with codecs.open('dump1.csv', "w", encoding="utf-8") as csvfile:
    for file in glob.glob('D:\PythonFiles\Data\*html*'):
        print 'Processing', file
        soup = BeautifulSoup(open(file).read())
        rows = soup.findAll('tr')
        iii = 0

        for tr in rows:
            aols = tr.findAll('div', {"id": "tobeparsed1"})
            for aol in aols:
                print >> csvfile, aol.get_text("|", strip=True)
```

```
# this code calculate the frequencies of various parameters from an organized text corpus
```

```
from nltk.corpus import terms
nltk.corpus.terms.words()
eterm_text = terms.words()
filtered_words = [w for w in eterm_text if not w in
nltk.corpus.stopwords.words('english')]
fdist2 = nltk.FreqDist(filtered_words)
fdist2.plot(50, cumulative=True)
```