AUTOMATED DISCOVERY OF SOCIAL NETWORKS IN ONLINE LEARNING COMMUNITIES

BY

ANATOLIY ANATOLIYOVYCH GRUZD

B.S., Dnipropetrovsk National University, 2002
M.S., Dnipropetrovsk National University, 2003
M.S., Syracuse University, 2005

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2009

Urbana, Illinois

Doctoral Committee:

Professor Caroline Haythornthwaite, Chair and Director of Research
Associate Professor P. Bryan Heidorn
Associate Professor Michael B. Twidale
S.D. Clark Professor Barry Wellman, University of Toronto

ABSTRACT

As a way to gain greater insights into the operation of online communities, this dissertation applies automated text mining techniques to text-based communication to identify, describe and evaluate underlying social networks among online community members. The main thrust of the study is to automate the discovery of social ties that form between community members, using only the digital footprints left behind in their online forum postings. Currently, one of the most common but time consuming methods for discovering social ties between people is to ask questions about their perceived social ties. However, such a survey is difficult to collect due to the high investment in time associated with data collection and the sensitive nature of the types of questions that may be asked. To overcome these limitations, the dissertation presents a new, content-based method for automated discovery of social networks from threaded discussions, referred to as 'name network'. As a case study, the proposed automated method is evaluated in the context of online learning communities. The results suggest that the proposed 'name network' method for collecting social network data is a viable alternative to costly and time-consuming collection of users' data using surveys. The study also demonstrates how social networks produced by the 'name network' method can be used to study online classes and to look for evidence of collaborative learning in online learning communities. For example, educators can use name networks as a real time diagnostic tool to identify students who might need additional help or students who may provide such help to others. Future research will evaluate the usefulness of the 'name network' method in other types of online communities.

ACKNOWLEDGEMENTS

This project would not have been possible without the support of many people. Many thanks to my adviser, Caroline Haythornthwaite, who guided me through this research and read my numerous revisions. Also thanks to my committee members, Bryan Heidorn, Michael Twidale and Barry Wellman, who offered guidance and support. Many thanks to Elizabeth Liddy from Syracuse University, who introduced me to Natural Language Processing, a key component of my thesis. Also many thanks to my fellow doctoral students for their stimulating discussions. Thanks to the Graduate School of Library and Information Science for awarding me a Dissertation Completion Fellowship, providing me with the financial means to complete this project and share my findings with the research community at numerous conferences. And finally, thanks to my partner, parents, and friends, who endured this long process with me, always offering support and love.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 INTRODUCTION

As social creatures, our daily life is intertwined with others within a wide variety of social networks involving our relatives, friends, co-workers, and a vast array of acquaintances and strangers. It is only natural that our digital life on the Internet is also made up of various social structures and networks. As Wellman (2001) noted, "computer networks are inherently social networks, linking people, organizations, and knowledge" (p. 2031). From this perspective, the Internet is more than just a means for people to support their existing social relationships; it also allows people to create new, exclusively virtual relationships through their membership in online groups and communities. This is made possible through the abundance of free and easy-to-use web-based information, communication, and community building technologies including communication technologies such as emails, web forums, chats, instant messaging, and twitter[1]; information dissemination and exchange technologies of web pages, wikis[2], blogs[3] and video blogs; social networking technologies such as Facebook, Myspace, and LinkedIn; online courseware such as Moodle, Blackboard, and Sakai; and virtual environments such as SecondLife and World of Warcraft.

Each reply to an email, link to a web page, posting of a blog, or comment on a Youtube video, leaves a digital trace, a record that explicitly or implicitly connects the

[1] Twitter is "a service for friends, family, and co–workers to communicate and stay connected through the exchange of quick, frequent answers to one simple question: What are you doing?" (http://twitter.com)

[2] Wiki is "a page or collection of web pages designed to enable anyone who accesses it to contribute or modify content" (http://en.wikipedia.org/wiki/Wiki)

[3] Blog is "a web site, usually maintained by an individual with regular entries of commentary, descriptions of events, or other material such as graphics or video" (http://en.wikipedia.org/wiki/Blogs)

poster to another online participant. Each of these recorded actions creates a network of attention around topics of interest, common affiliation, communities of practice, or collective action. Online contributions and interconnections are growing daily, reflected in the increasing volume of texts and a growing number of participants. The numbers are impressive: a 2008 Wikipedia compilation of sources estimated 4.6 terabytes of data posted daily on Usenet[4]; Technorati's 2008 'State of the Blogosphere'[5] indicates 900,000 blogs are created each day, with 184 million people worldwide who have started a blog (23-26 million in the U.S.), and 346 million blog readers (60-94 million in the U.S.); various estimates suggest something in the order of 100 billion emails sent per day (Leggatt, 2007).

This abundance of online data being captured and stored over the years offers a unique opportunity for social scientists and Internet researchers to study the inner workings of online communities. Researchers can now more closely scrutinize these recorded interactions and answer questions about what group's interests and priorities are, how and why one online community emerges and another dies, how people reach agreement on common practices and rules in an online community, and how knowledge and information are shared among group members. Answers to these and other related questions will allow us to understand not only how people meet, communicate and establish social relationships online, but also how the Internet can be used to develop new technologies that better serve the information needs of members of online communities. For instance, social networking websites like Facebook and MySpace are good examples

[4] http://en.wikipedia.org/wiki/USENET, retrieved March 30, 2009

[5] http://technorati.com/blogging/state-of-the-blogosphere, retrieved October 30, 2008

of how advancements in information technology can help people to meet online, and form and support a much larger number of online relationships than it has been possible before.

However, many researchers and the public at large are finding it practically impossible to keep up with the vast amount of accumulated online data. This problem holds true even in smaller closed online communities. For example, in a single online class consisting of between 17-29 students from one of the datasets examined here, the members of the class easily generated hundreds of public discussion board postings in a short 15 week period, and that excludes postings in private discussions, email correspondence, and chat. But data volume alone does not fully speak to the enormity of the task associated with analyzing online data. To exploit an online dataset fully, researchers often need to examine the same set of data in multiple passes, each time looking for different things. Some of the many things that they can be looking for may include common patterns of exchange, development of shared language and understanding, and emergence of roles and positions that may be unique to online interactions. Each one of these kinds of passes takes a substantial amount of time to accomplish when managed by hand. Thus, it is not surprising that there is an increasing interest in the ability to retrieve and analyze online social networks automatically.

Discovering details about online social networks automatically has already proven useful in deciding what information is relevant on the Internet, identifying credible websites, finding popular resources, and sharing information within a network of trust. Other uses of social network data include conducting viral marketing, identifying and tracking terrorist cells on the Internet, analyzing consumers' perceptions of products, and

measuring the effectiveness of political campaigns in online and offline media.

One of the reasons automated discovery of social networks has become so popular is that it tends to be unobtrusive, scalable and fast. Because the type of data typically analyzed comes from the public domain, it avoids the difficulties of obtaining respondent compliance in completing the often burdensome social network questionnaires. By avoiding human responses, automated network data collection is also not encumbered with many of the shortcomings related to the subjectivity of traditional data collection techniques, e.g., that, in a sociometric survey, respondents may provide partial answers, respond in ways they believe make their behavior look better, exaggerate interactions, forget people and interactions, or perceive events and relationships differently from other network members (see discussions by Dillman, 2000 and Lange et al., 2004).

This dissertation focuses on new ways of discovering online social networks automatically. In particular, the work introduces and evaluates an automated approach for social network discovery from threaded discussions. Bulletin board style threaded discussions were chosen due to their wide acceptance and usage by various online communities.

The most common automated method used to collect information on social networks in online communities is to gather 'who talks to whom' data which counts the number of messages exchanged between individuals from their recorded interactions. A higher number of messages exchanged is usually interpreted as a stronger tie between people. This method is often used with email-type data. In online communities that use threaded discussions, researchers have relied on information in posting headers about the chain of people who had previously posted to the thread (further referred to as 'reference

chain') to gather 'who talks to whom' data. For logical and practical reasons, researchers have generally assumed that the reference chain may reveal the addressee(s). More specifically, it is usually assumed that a poster is replying to the previous poster in the reference chain. (For the remainder of the dissertation, I refer to any social network that is built using information in the reference chain as a 'chain network'.) Unfortunately, this assumption is not always true in highly argumentative and collaborative communities such as online classes. A previous poster is not always the addressee of the posting. A poster may address or reference other posters from earlier in the thread, from another thread, or even from other channels of communication (e.g., emails, chats, face to face meetings, etc). So, while the use of reference chains provides some mechanism to approximate 'who talks to whom' data for threaded discussions, such approximation is not very accurate and is likely to undercount the number of possible connections, thereby underestimating the connections among members of the network. To overcome the inherent flaws associated with gathering 'who talks to whom' data from threaded discussions, this dissertation proposes a new approach for inferring social networks based on the actual content of postings. The social networks built based on this approach will be referred to as *name networks* (for reasons that will be explained in Chapter 3).

As a case study, the proposed 'name network' approach is evaluated in the context of online learning using data collected from six different graduate level online classes taught at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign in Spring 2008. The online learning context was chosen because the online classes relied primarily on bulletin boards to conduct their discussions, each class represented a closed community with a finite number of online participates, and the

time frame for data collection had a clear beginning and end.

The end result of the research is the development of a low-cost and time-efficient set of techniques for on-the-fly content analysis of online communities and a new schema for using these techniques to discover social networks automatically. The following section outlines the whole dissertation and describes the research questions. A summary of the main steps in the study are also presented in Figure 1.1 (Figures can be found at the end of each chapter).

1.2 RESEARCH OUTLINE AND QUESTIONS

1.2.1 Building Name Networks

In order to develop an automated method for building name networks, the following research question needs to be addressed:

> **Question 1: What content-based features of postings help to uncover nodes and ties between group members?**

To answer this question, Chapter 2 first discusses several possible ways of building social networks from text, drawing on examples from the existing literature in Computational Linguistics. Then Chapter 3 proposes and describes the new method called 'name network'. In general, to build the name network, the method starts by automatically finding all mentions of personal names in the postings and uses them as nodes. To discover ties between nodes, the method connects a poster to all names found in his/her postings.

1.2.2 Comparing Name Networks with Those Derived from Other Means

To evaluate the proposed method of building social networks and identify what will be gained from using this more elaborate method, social networks derived using the

'name network' method are compared to those derived from other means, specifically (1) chain (reply-to) networks and (2) students' self-reported (perceived) social networks. The chain networks are built automatically using students' posting behaviors, and the self-reported social networks are built based on the data collected via online questionnaire from students in the six online courses that participated in the study. Datasets and procedures used to evaluate the proposed 'name network' method are discussed in more detail in Chapter 4.

Chapter 5 addresses the following research question:

Question 2: How is the proposed name network associated with the chain network and with the self-reported network?

To answer this question, the study relies on QAP correlations and exponential random graph models (p* models) to perform a comprehensive comparison between these networks. The supposition is that (1) the name network is capable of identifying communication patterns between people more accurately than the chain network, and (2) the name network more closely matches perceived structures of online participants than the chain network.

As mentioned earlier, the 'name network' method is evaluated here in the context of an online learning environment. Below is a quick explanation of why uncovering perceived social structures of online participants is especially important in the learning context.

Traditionally, it is presumed that observed social networks can more accurately reflect relations between group members compared to individual perspectives and thus provide a better representation of what is really going on in an online community. But for

online learning environments, due to the individualized nature of learning outcomes, it may be more important to identify and understand perceived social networks in order to study collaborative learning. This is because what is deemed as important or relevant to one student may only be marginally valued by another student. Until now, the only reliable way to collect perceived data has been through surveys which are difficult to collect due to the sensitive and resource-intensive nature of network questions. Therefore, it would be a methodological breakthrough if an automated method for mimicking perceived social networks is devised.

1.2.3 Identifying Types of Social Relations in the Name Network

To further explore the nature of the name network and evaluate its usefulness in the evaluation of collaborative learning, the types of social interactions and relations embedded in the name network are examined. This is discussed in Chapter 5. Specifically, Section 5.3 looks for relations that are known to be crucial in shared knowledge construction and community building such as 'help', 'work together' and 'socialize', and thus important in achieving successful collaborative learning. The presence of these kinds of relations in the name network would signify its ability to reflect collaborative learning processes and perhaps even predict learning outcomes. Therefore, the last question is:

Question 3: What types of social relations does the name network include?

To address this question, a detailed, manual exploration of social networks and postings is conducted. This exploration uses a web-based text mining and visualization system called Internet Community Text Analyzer (ICTA), available at http://textanalytics.net.

ICTA was built as part of this dissertation research. It can be used to analyze a wide variety of other types of text-dependent online communities. A description of ICTA's capabilities and interface is given in Chapter 4. Since its inception, ICTA has already proved to be useful in at least two separate studies. One was a LEEP language study by Haythornthwaite and Gruzd (2007). Another study used ICTA to analyze communal discourse produced by members of the "i-neighbors.org" website by Keith N. Hampton (the publication titled "Glocalization: Internet Use and Collective Efficacy", in preparation). In this dissertation, ICTA was used as a visualization and assessment tool for presenting results derived from the 'name network' method.

Chapter 6 presents the overall summary of the findings and specific examples on how the name network can be used to study online classes and assess collaborative learning. The chapter also addresses limitations of the 'name network' method and presents ideas for future research.

1.3 MAIN CONTRIBUTIONS OF THE RESEARCH

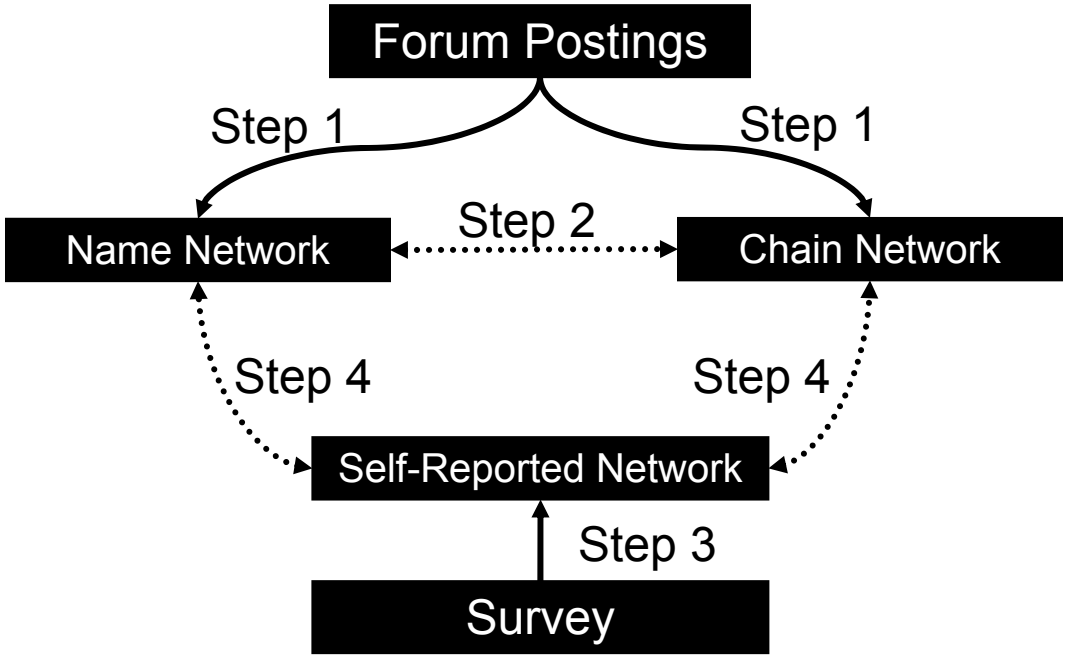
This work makes the following contributions to the research on online communities, social network analysis, computer-mediated communication and collaborative learning:

- Development of a novel approach (name network) for content-based, automated discovery of social networks from threaded discussions in online communities and a framework for evaluating this new approach,
- Demonstration of the proposed automated approach for collecting social network data as a viable alternative to the costly and time-consuming collection of users' data on self-reported networks,

- Demonstration of how the name network can be used to study online classes and assess collaborative learning,
- Development of the ICTA web-based system (http://textanalytics.net) for content and network analysis.

1.4 FIGURES

Figure 1.1: Study outline



Main steps:

- **Step 1**: Using postings from actual online classes, derive social networks using the 'name network' method (as proposed in this dissertation) and the 'chain network' method.
- **Step 2**: To determine whether social networks discovered by the 'name network' method are any different from those discovered by the 'chain network' method, compare them against each other. This is done, to ensure

that there are enough differences to justify the time and expense to build an alternative to chain networks. Both methods are described in Chapter 3, and the results of the comparison are described in Chapter 5, Section 2.

- **Step 3**: Collect information about actual social interactions in the class and build so-called self-reported networks by asking students to complete the online survey about their personal networks. This step is explained in more details in Chapter 4.
- **Step 4**: Compare the self-reported network against the name network and the chain network in order to check which of the two types of networks (name or chain) better resembles actual social interactions in the class. To perform these comparisons, the study relies on QAP correlations, exponential random graph models (ERGMs or p* models), and the manual exploration of social networks and postings using Internet Community Text Analyzer (ICTA). This step and the results are described in Chapter 5, Section 3.

CHAPTER 2: LITERATURE REVIEW

This chapter provides a brief overview of Social Network Analysis (SNA) and explains why it is an effective method for the study of online communities, and provides a review of common approaches to automated discovery of social network data.

2.1 SOCIAL NETWORK ANALYSIS

Social Network Analysis (SNA) is a commonly used method to study social interactions of online groups at an individual level as well as group level (e.g., Haythornthwaite et al., 1996; Haythornthwaite, 1998; Wasserman & Faust, 1994; Wellman, 1996). According to the social network perspective, individual behavior is defined by others. Thus, to understand individual behavior, we need to "describe patterns of relationships between actors, analyze the structure of these patterns, and seek to uncover their effect on individual behavior" (Nurmela et al., 1999; n.p.). SNA seeks to represent datasets in a form of social networks. In a social network, there are nodes which represent group members, and edges (often referred to as ties) that connect people by means of various types of relations. The strength of the relations is usually conveyed via a weight assigned to each tie.

A network representation of social interactions provides researchers with an effective mechanism for studying collaborative processes in online communities, such as shared knowledge construction, information sharing and exchange, influence, support. Because the case examined in this dissertation is online learning communities, the three examples below demonstrate how SNA can be used to study social interactions in online classes. These are just a few of the many examples of studies that have relied on SNA to evaluate individual participation based on the position of individual nodes in a network,

and group cohesion based on general properties of a network.

Using centrality, density and QAP correlation measures, Haythornthwaite (2001) compared class interactions across four different self-reported networks: Collaborative Work, Exchanging Advice, Socializing, and Emotional support. One of the main research questions posed in that study, that is also relevant to this dissertation work, was "Do the four relations describe similar structures or do they capture different aspects of student interaction?" The author found some similarity between the Collaborative Work and Exchanging Advice networks, but other networks described different network configurations. Furthermore, all but one network, Emotional Support, moved toward a team structure over time suggesting that by the end of the semester students interacted mostly with team members, "perhaps as a means to reduce communication load and complete the course" (p. 219).

In another study, Reyes and Tchounikine (2005) demonstrated how a tutor could rely on participants' centrality and group cohesion in a social network to assess collaborative learning. In one of the threaded discussions that they studied, group cohesion was extremely low. Upon further investigation, it was discovered that the two participants with the highest centrality were dominating the conversation, a condition that may be undesirable for learning communities where wider spread contribution is intended, and thus of value to identify.

Among most recent work, Cho et al. (2007) explored different social network properties such as *degree*, *closeness*, *betweenness*, and *structural holes* to find the relationship between students' positions in the social network and their success in the class, and to see which measures correlate with final grades. For example, they found that

"*closeness centrality* was significantly [and positively] associated with students' final grades" (p. 322). Additionally, the authors used a novel *change propensity* measure to find "the degree to which an individual renewed his/her social and intellectual capital as the person participates in a new learning environment" (p. 318). This measure takes into account students pre-existing social networks with class members. They found that "students' initial network positions had negative effects on change propensity," meaning that "those who were central in the pre-existing network were more likely to stay in their initial social circles, whereas peripheral actors were more likely to alter their network compositions, since they are not bound to pre-existing networks" (p. 322).

These studies demonstrate that SNA is a promising method that can be used to study group dynamics among online learners, and offers a new window into understanding the various social connections that develop within online communities. However, to study online communities using SNA, social network data about connections among members of a particular community needs to be collected. The following two sections describe common approaches to gathering data about social networks focusing on the newly emerging area of automated techniques.

2.2 CURRENT METHODS TO COLLECT SOCIAL NETWORK DATA

A traditional way to collect information about social networks is to ask group members themselves about their ties with others. However, this method is very time consuming and prone to a high rate of non-response. In survey research in general, there are two main reasons for high rates of non-response: asking questions that are highly sensitive, and conducting surveys that have a burdensome quantity or type of questions (Dillman, 2000). Furthermore, as shown in most of the studies on the accuracy of the

acquired social networks, “individual reports about social interactions differ substantially from the objective observations of these interactions” (Lange et al., 2004, p.354). Responders may lie, forget and/or simply perceive the events and relationship with other group members differently. As a result of these inherent flaws with survey data, many researchers are turning to cheaper and more objective, automated methods for collecting data on social networks. Some of these automated methods include using movement tracking devices (e.g., Matsuo et al., 2006), log analysis (e.g., Nurmela et al., 1999), and co-citation analysis (e.g., White et al., 2004). In online communities the most common automated method is based on finding ‘who talks to whom’ data, which counts the number of messages exchanged between individuals based on their recorded interactions. A higher number of messages exchanged is usually interpreted as indicating a stronger tie between people.

While this method has some advantages, there is an important shortcoming of this approach. It has to do with the ability to count accurately the messages exchanged between people in common forms of computer-mediated communication such as chats and threaded discussion forums. Unlike email headers, chat messages or posting headers in discussion forums do not usually contain information about the addressee(s). What makes this even more challenging is that group chats and threaded discussions tend to be open-ended and free flowing communication that is accessible to every member of a group; as a result, it is difficult to determine automatically how much influence, if any, a particular student or posting might have on another member of the community. As a work around for this problem associated with threaded discussions, some researchers rely on information in posting headers about the chain of people who had previously posted to

the thread (referred to here as 'reference chain'). For logical and practical reasons, it is generally assumed that the reference chain may reveal the addressee(s). More specifically, it is usually assumed that a poster is replying to the immediately previous poster in the reference chain. Unfortunately, the above mentioned assumption is not always true in highly active, argumentative, or collaborative communities such as online classes, or where many discussion topics may be in play at one time. A poster may also address or reference other posters from earlier in the thread, from another thread, or even from other channels of communication (e.g., emails, chats, face to face meetings, etc). Further, an individual may seem to respond to one post, but in the text refer to several others, synthesizing and bringing together comments of many posters. So, while the use of reference chains provides some mechanism to approximate 'who talks to whom' data for threaded discussions, such approximation is not very accurate and is likely to cause an undercounting of possible connections. Further, it is unclear to what extent chain networks represent 'real' social networks, i.e. those defined by the multi-relational set of interactions and perceptions.

In the examples below if we were to rely on just the chain network to discover ties, we would miss some important ties. In Example 1, the chain network only finds one connection between Sam and Gabriel. But there are actually four possible connections with Sam. This is because except for Gabriel, the other addressees (Nick, Ann and Gina) in the sample message below were not among the people who had previously posted to the thread.

In Example 2, Fred is the first person who posted to the thread, thus the reference chain is empty. As a result, the 'chain network' method finds no connections in this

posting. However, upon closer examination, there is actually one potential connection between the poster Fred and a person named Dan, who has not posted to the current thread.

> Example1:
>
> FROM: **Sam**
> REFERENCE CHAIN: **Gabriel**
>
> **Nick**, **Ann**, **Gina**, **Gabriel**: I apologize for not backing this up with a good source, but I know from reading about this topic that libraries […]

> Example 2:
>
> FROM: **Fred**
> REFERENCE CHAIN: <empty>
>
> I wonder if that could be why other libraries around the world have resisted changing – it's too much work, and as **Dan** pointed out, too expensive.

Based on the preceding discussion, SNA seems to be an effective method for the analysis of social interactions in online communities. However, SNA procedures in their current form can not accurately build social networks from a non-email type of communication such as chats or threaded discussions which are often favored in online communities. As mentioned earlier, the current work is focusing specifically on threaded discussions. However, this work is easily adaptable to accommodate other online data types such as non-threaded discussion lists, chats, wikis and blogs. The next section discusses a number of modern text mining techniques that can be used to overcome these inherent flaws in the methodological approach currently used in SNA for inferring social network data from threaded discussions.

2.3 AUTOMATED DISCOVERY OF SOCIAL NETWORKS FROM TEXTUAL DATA

Text mining techniques have been gaining in sophistication over the past decade. These techniques now offer ways to discover social networks from documents published on the Internet and text-based online communication. In general, to discover social networks from textual data, the following steps are taken:

- *Node Discovery*, when all references to people such as names, pronouns, and email addresses are identified.
- *Coreference and Alias Resolution*, which resolves ambiguities about people, e.g., differentiating between people with the same name and creating a single identity for those with multiple aliases.
- *Tie Discovery*, which determines whether or not there are social connections between people identified in the first two steps.

The following describes each of these three steps and provides examples from the literature.

2.3.1 Common Practice: Node Discovery

Node discovery from text is usually conducted through the discovery of personal names and other references to people in the text. It is part of a broader task in Computational Linguistics (CL), called Named Entities Recognition (NER). NER itself is a set of text mining techniques designed to discover named entities, connections and the types of relations between them (Chinchor, 1997). In NER, a named entity is defined very broadly. It may be a person, organization, or even a geographic location. NRE is commonly used in various Natural Language Processing (NLP) applications such as

machine translation, information extraction, and question answering systems. An example of an application that deals specifically with people's names is anonymization or pseudonymization for the purpose of hiding sensitive data in private or secret documents such as personal medical records and vital government documents (e.g., Medlock, 2006; Sweeney, 2004; Uzuner et al., 2007).

Since it is relatively easy to find pronouns and email addresses in text[6], the following review focuses on the discovery of personal names. There are two primary approaches to finding personal names in the text. The first and easiest approach is to look up each word in a dictionary of all possible personal names. If a word is in the dictionary of names, then it is considered to be a name. Examples of electronic dictionaries with English names include the Dictionary of First Names (Hanks et al., 2006), the publicly accessible U.S. Census[7], a commercial InfoSphere Global Name database from IBM[8], and a web resource 'Behind the Name'[9]. Researchers who have relied on this approach include Harada et al. (2004), Patman and Thompson (2003), Sweeney (2004). This approach is easy to implement and run; however, it tends to leave out names that are not already found in the dictionary. These may be names of non-English origin, informal variations of names, or nicknames. Additionally, this approach does not take into account that in different sentences a word may be a name or just a noun, e.g., "*Page* asked for my

[6] Pronouns can be found by comparing each word in the text with a list of possible pronouns, and email addresses can be found by matching each word with a string pattern like [part1]@[part2].[part3]. Even as people begin to disguise their email addresses to avoid web crawlers, it is not much more of a programming task to remove blanks or search for 'at' instead of '@'.

[7] US Census - http://www.census.gov/genealogy/names/dist.all.last

[8] IBM InfoSphere Global Name - http://www-306.ibm.com/software/data/ips/products/masterdata/globalname

[9] 'Behind the Name' website - http://www.behindthename.com

help" and "Look at *page* 23". To make sure that an algorithm finds the name 'Page' in the first sentence above and ignores the word 'page' in the second, some researchers may consider only capitalized words as potential candidates for personal names and ignore others. However, this restriction is not very practical with informal texts such as computer-mediated communication where names are often not capitalized.

An alternate approach to finding personal names can be used that does not require using a dictionary of names. This approach applies linguistic rules or patterns to the content and sentence structure to identify potential names. The linguistic rules and patterns are often built based on characteristic attributes of words such as word frequencies, context words, word position in the text. Some work in this direction includes that of Chen et al. (2002), Bikel et al. (1997), Nadeau et al. (2006), Sekine and Nobata (2004), and Yang and Hauptmann (2004).

In practice, these two approaches are usually used together; for example, finding all names based on the dictionary first, and then using linguistic rules/patterns to find names that are not in the dictionary. Using such a hybrid approach, Minkov et al. (2005), for example, reported a 10-20 per cent improvement in accuracy. The downside of a hybrid approach is that it tends to increase the time needed to process the textual data. For a more detailed survey of modern NER techniques, see Nadeau and Sekine (2007).

For the 'name network' method, I use a hybrid approach to find and extract personal names, similar to the one described above (see Section 3.1.1 for more details).

2.3.2 Common Practice: Coreference and Alias Resolution

Once names and other words that refer to people (e.g., titles, pronouns, email addresses) are identified, the next step is coreference and alias resolution. The goal of this

step is two-fold: to group all mentions of the same person together (for example, 'you', 'John', 'Mr. Smith' and 'j.smith@mail.net'), but at the same time to distinguish between two or more people with the same name. Similar to the previous step of identifying named entities, for coreferencing, Computational Linguistics (CL) relies on a more general approach based on Machine Learning (ML) techniques and tries to link not just names, but any coreferring noun phrases across sentences and documents; where noun phrases may refer to people, organizations or any other objects in the world. In this instance, CL uses ML techniques to determine the likelihood that a set of noun phrases might refer to the same entity. The likelihood is measured based on the unique characteristic attributes of noun phrases such as the distance between noun phrases in the text, lexical similarities, how often phrases collocate with each other, their agreement in gender and semantic meanings, etc. For recent work in this area, see Culotta et al, 2007; Luo et al, 2004; Soon et al, 2001; Yang et al, 2008.

In practice, to discover a social network from transcripts of computer-mediated communication (CMC), there is usually no need to perform a full coreference and alias resolution. Quite often only resolution among personal names, email addresses and sometime pronouns is sufficient to complete the task. That is why researchers working with CMC-type data such as emails often rely on simple rule-based or string-matching approaches. For example, McArthur and Bruza (2003) approached a pronoun coreference resolution in an email archive by simply replacing pronouns '*I*', '*my*', '*me*' with the sender's name and pronouns '*you*' and '*your*' with the receiver's name. There are also a number of simple but effective methods that match variations of names and/or email addresses by relying on phonetic encoding and/or pattern matching techniques (e.g., Bird

et al., 2006; Christen, 2006; Feitelson, 2004; Patman and Thompson, 2003). For example, a simple rule may state that an email address belongs to a person if it contains either his/her first or last name and the initial of the other. According to this rule, some examples of emails that will be attributed to *John Smith* are *john.smith@mail.org*, *jsmith@mail.net*, *john@smith.net*, *s.john@mail.net*. For a more in-depth review of personal name matching techniques, see the recent survey conducted by Reuther and Walter (2006).

The second part of this step, alias resolution, requires special attention. Alias resolution can be performed as part of general NER, but it can also be conducted as a standalone procedure; it has a broad range of applications in research on authorship, citation analysis, spam detection, author disambiguation in digital libraries, and more. The purpose of the various approaches to alias resolution is to distinguish between two or more people with the same name by identifying the unique 'signature' that can be associated with each person. These approaches often rely on either unique linguistic characteristics of a person's writing (e.g. common writing styles, punctuation marks, average length of sentences, expertise keywords, etc.; e.g., Fleischman and Hovy, 2004; Hsiung, 2004; Hsiung et al., 2005; Mann and Yarowsky, 2003; Pedersen et al., 2006) or network-based patterns of interactions (e.g., common senders and recipients; e.g., Hölzer et al., 2005; Malin et al., 2005). When extracting social networks from web pages, alias resolution is often addressed by automatically assigning a set of expertise keywords (see Bollegala et al., 2006; Matsuo et al., 2006; Matsuo et al., 2007) or summaries of several contextual sentences (Phan et al., 2006) to each name in the text. The assumption here is that two different people (even with the same name) are usually mentioned in different

contexts in the text. So, the task is reduced to finding a set of discriminating words and/or semantic features to describe uniquely a particular person.

However, a content-based approach on its own is not likely to work well for online classes. This is because all students are usually engaged in the discussion of the same topics; thus, they all are using more or less similar keywords in their postings. As a result, many students are likely to have similar content-based 'signatures'. Therefore, for the 'name network' method to work effectively, the work presented here has added the use of traffic-based features such as information from the reference chains and not just content-based features. Specifically, to perform coreference and alias resolution, names extracted from the content of messages (content-based feature) are associated with unique identifiers in the form of posters' email addresses found in the posting headers (traffic-based feature). All names that are associated with the same email address are considered to belong to the same person. (See Section 3.1.2 for a detailed description of the algorithm.)

2.3.3 Common Practice: Tie Discovery

After all network nodes are identified and grouped to represent unique people, the next step is to uncover if and how these nodes are interconnected. There are two main methods in the literature for automated discovery of ties based on textual information. One is based on similarity between users' profiles. A profile is either created manually by a person on their own (e.g. a Facebook profile) or pulled out automatically from information on the Internet (e.g., a person's homepage, emails, or parts of the text written about that person by others). A simple way to measure the similarities is to count how many profile items two people have in common (e.g., Adamic and Adar, 2003). Another

common approach includes measuring the semantic similarity between words extracted from the profiles. According to this method, two people are connected when the value of semantic similarity between their profiles is higher than a predefined threshold. In other words, people are considered to be connected when there is a substantial overlap of words and phrases found in their profiles. There are many different sources on how to measure semantic similarity, e.g., Kozima and Furugori (1993), Maguitman et al. (2005), and Resnik (1999). A variation of this method is often used for expert or cooperator identification. For example, Campbell et al. (2003) relied on keywords submitted by users and the content of the users' emails to form what they call an expertise graph which connects people based on their self-professed expertise.

Another method is to use some sort of co-occurrence metric to calculate the number of times two names co-occur in close proximity within the text. This approach is especially popular among researchers who use web pages to build networks. This is because search engines make it easy to count the co-occurrence of two people on web pages. Matsuo et al. (2006) counted the number of hits from an Internet search engine in response to a query consisting of two names joined via the boolean operator 'AND'. Kautz et al. (1997) used this approach in their application called ReferralWeb for visualizing and searching social networks on the Web.

Out of the two possible approaches described above to build the name network, the co-occurrence-based approach was chosen for this dissertation work. Specifically, the co-occurrence between posters' email addresses found in the posting headers and names found in the body of posters' messages were used to reveal social ties in the online classes (See Section 3.1.2). The main reason for rejecting the approach that measures

similarity between users' profiles is because it is unclear whether this approach finds 'real' social ties. For example, just because two students in a class have similar interests as evidenced by the similarities outlined in their users' profiles, it does not necessarily mean that they share a social tie or even talk to each other in the class[10].

2.4 SUMMARY

This chapter provided a brief overview of Social Network Analysis and explained why it is an effective method for studying online communities in general and learning communities in particular. The chapter also reviewed several possible ways of collecting social network data with a focus on automated techniques. The literature review in this chapter was designed to prepare the grounds for addressing the first research question: What content-based features of postings help to uncover nodes and ties between group members? The next chapter directly addresses this research question by proposing and describing a new method called 'name network' for inferring social networks from postings in threaded discussion.

[10] However, this is an interesting subject of a separate empirical study.

CHAPTER 3: METHOD

This chapter describes the novel content-based approach for inferring a social network from postings in threaded discussions, referred to as 'name network,' that was derived for this dissertation. The approach starts by finding all mentions of personal names in the discussion postings. These become the nodes in the name network. Once all nodes are identified, the next step is to discover how these names/nodes are connected to each other in order to infer a social network. These two basic steps for building the name network are discussed in greater detail in the following sections. But, before proceeding to the discussion of these steps, it is important to consider some of the challenges associated with building the name network. As shown in Figure 3.1 (given at the end of this chapter): (1) some names found in the postings do not belong to group members (e.g., kurt, dewey); (2) some words that may be personal names are not used as names in a particular context (e.g., mark); (3) some names belong to names of buildings or organizations (e.g. Santa Monica); and (4) group members may refer to each other using informal names (e.g., Chris instead of Christopher) or western style names for international students (e.g., Kevin instead of Kwang). For the 'name network' technique to be workable and successful, it had to address these challenges.

3.1 NAME NETWORK METHOD

To develop the 'name network' method, the best practices from the literature on Computational Linguistics (described in Section 2.3 above) were relied on as the basis for deriving the method described here. The following sections describe the actual implementation of the 'name network' method as proposed by the author and used in this work to discover social networks from the threaded discussions.

3.1.1 Name Network: Node Discovery

Currently there are many software packages that can perform NRE-related tasks[11]. However, most of them are weak in terms of execution speed and accuracy. Furthermore, these packages are often trained on documents from newspaper or medical domains which make them unsuitable for working with the idiosyncratic spellings, capitalizations, and grammar in CMC-type data. To address these limitations, a hybrid approach to personal name discovery was used. The approach attempts to satisfy the following two criteria: (1) to process messages in real-time and (2) to understand informal online texts.

The algorithm works as follows. First, to avoid redundancy and wasted processing time, any texts that belong to previous messages are removed. This is done automatically using a string matching mechanism called *regular expressions*. Specifically, the algorithm accomplishes this by removing all lines from the messages that appeared after a pattern ' <name> wrote: ' and start with a greater-than character '>', the specific designator used to indicate reply messages in this system (other systems may use other characters, e.g., ':' in which case those would be used to identify previous message text). Second, it removes 'stop-words ', such as *and*, *the, to*, *of*, etc. There are many different versions of 'stop-word' lists freely available on the Internet. The one used in this work is part of the Natural Language Toolkit[12], and it includes 571 words. Third, the algorithm normalizes all remaining words by stripping all special symbols from the beginning and end of any word, including possessives (e.g '--*Nick*' or '*Nick's*' becomes *Nick*).

---

[11] A comprehensive list of available Named Entity Recognition packages - http://alias-i.com/lingpipe/web/competition.html

[12] Natural Language Toolkit - http://nltk.sourceforge.net

For all remaining words, to determine whether a word is a personal name, the algorithm relies on a dictionary of names and a set of general linguistic rules derived manually. To find first names, the procedure uses a dictionary containing over 5,000 of the most frequently used first names in the United States as reported by the 1990 US Census[13]. To find last names in the text, the algorithm looks for any capitalized word that follows a first name. If a capitalized word is found in that context, the name is classified as a middle or last name.

In addition to the dictionary, two additional sources of personal names were considered: a class roster (list of all class participants) (e.g., Matsuo et al., 2006) and the 'From' field in the message header (e.g., Culotta et al., 2004). These options are possible because of the known set participants, and the information about posters associated with threaded discussion. However, the use of the class roster was not as effective as it was originally thought it would be. This is primarily because students often did not use formal names from the roster to refer to each other, but instead used nicknames and informal names (e.g., *Ren* for *Karen*, *Dan* for *Daniel*). Furthermore, the use of the class roster limits the ability of the algorithm to perform well on texts produced by groups with unknown membership. Thus, the use of the class roster for determining names was abandoned as an option for the 'name network' model.

The second additional source of names (the 'From' field) proved to be more useful. In some cases, in addition to a poster's email address, the 'From' field also includes his/her name enclosed within a set of parentheses. To recognize names from the 'From' field of the message header, the algorithm uses a simple string matching pattern

[13] 1990 US Census - http://www.census.gov/genealogy/names

that looks for only words found within the round brackets (if any). For example, the following record "*agruzd2@uiuc.edu (Anatoliy)*" will produce *Anatoliy*. The 'name network' method retains the added step of checking for names in the message header formed in this way.

To recognize names that are not likely to be found in the dictionary, such as nicknames, abbreviated names, unconventional names, etc. – for example *CH* or *CarolineH* – the algorithm relies on context words that usually indicate personal names such as titles (e.g., *Professor*, *Major*, *Ms.*) and greetings (e.g., *Hi* or *Dear*) (See Appendix A). In the future, post-thesis research, other types of context words can be considered as well, such as communication and motion verbs that usually express actions associated with humans (e.g. *say*, *tell*, *warn*, *walk*, *run*, etc). Such verbs can be obtained from various lexical resources such as VerbNet, EVCA, and VerbOcean (Chklovski & Pantel, 2004; Klavans & Kan, 1998).

Once all names are identified, it is then necessary to remove names of those who are not part of the social network. To exclude personal names that are part of a building or organization name, such as the *Ronald Reagan Presidential Library*, the algorithm first ignores all sequences of more than three capitalized words, and second removes phrases in which the last word was included in a pre-compiled list of prohibited words such as *Street* or *Ave* (See Appendix A).

Finally, for all words that are identified as potential names, the algorithm attempts to determine the confidence level that the particular word is actually being used as a person's name in the text. This is accomplished by factoring in the commonality of a name in the US Census (if applicable) and whether or not the first character of a word is

capitalized. For example, consider the word '*page*'. According to the US Census, the name '*Page*' is possessed by 0.034 per cent of the population sample. Therefore, its confidence level assigned by the algorithm will be $\frac{0.034}{par}$; where *par* is a parameter that will take a value of 1 if the word is capitalized or a value greater than 1 otherwise. This is done to "punish" non-capitalized words and reduce their confidence level of being a name. The current version of the algorithm adopts a conservative approach by setting the value of *par* to 10 for all non-capitalized words and 1 for capitalized words. (The threshold was set manually, based on the observed quality of the results obtained from a few sample datasets.) That is, the final score for non-capitalized '*page*' will be $\frac{0.034}{10} =$ 0.003. Since it is less than a pre-set threshold of 0.0099 the word will be removed from the further consideration.

While the algorithm described above is very thorough, it is still not capable of achieving 100 per cent accuracy. This is because at this point in the process, incorrectly spelled names may be missed and some possible false-positive words may still be on the list. However, since accurate name extraction is a vital foundational building block in automated inference of social networks, the resulting accuracy should be as close to the 100 per cent level as possible. In cases where it is applicable, the ICTA tool has been designed to allow human intervention to bridge the gap between automated name discovery and the final list of names. This allows those with knowledge of the group members (e.g., knowing nicknames and full names) and/or of the subject matter (knowing that references to Reagan may be to a library or airport in the current context) to fine-tune the name list.

Thus, to increase the final level of accuracy, the ICTA tool includes a web-enabled interface where researchers can manually review and edit the list of extracted names created by the algorithm (see Figure 3.2). After running the name extractor, a researcher can use this interface to add names that were missed by the extractor or delete false-positive words. The algorithm will remember these words for future runs as well. To improve the readability of the extracted names, all names are displayed in the form of a *tag cloud*. The larger font size in the tag cloud indicates the higher frequency of occurrences of a particular name in the dataset. Clicking on any name from the tag cloud returns a list that shows all instances where that name was found along with 2-3 words preceding and following the name (see Figure 3.3), and from there a user can also go to the exact location in the text where a potential name was found. This is especially helpful for uncovering false-positive results. For example, in one of the experiments, it was possible to verify that a word '*Mark*', a common name in the English language was not actually a name in that instance, but part of the term '*Mark* Up language'. To ensure the 100 per cent accuracy of the final results discussed later in Chapter 5, all names found by the system were manually inspected using this interface.

The end result of this semi-automated name extraction exercise is a list consisting of all occurrences of personal names in the postings.

3.1.2 Name Network: Evaluation of Node Discovery

To evaluate the accuracy and effectiveness of the proposed name extraction algorithm (further referred to as *Local name extractor*), results from this technique were compared with those from another automated name extractor constructed based on Alias-

I LingPipe[14], a state of the art toolkit for linguistic analysis, normally applied to formal texts. After selecting two non-intersecting sample datasets from the online learning discussions, Subset A (853 postings) and Subset B (534 postings), both extractors were used to find personal names from within the postings. The results were then compared using evaluation measures traditionally used in NER tasks: *P* precision and *R* recall. These measures are calculated in the following manner. Precision $P = \frac{T1}{T1 + F}$, defined as a ratio of all correctly identified names (T1) to all words labeled by the program as names (T1+F); where F is the number of false-positive results (words that were incorrectly labeled as names). And recall $R = \frac{T1}{T2}$, is defined as a ratio of all correctly identified names (T1) to all names in the dataset (T2). T2 is calculated by counting all distinct names found by both algorithms.

The results demonstrate that the *Local name extractor* returned far fewer false-positive results, than LingPipe: 12 and 16 per cent versus 40 and 57 per cent of the total number of extracted names in Subset A and B respectively (see Table 3.1). In other words, a user would have to remove fewer incorrectly labeled words when using the Local name extractor than with LingPipe. This fact is also supported by the higher values of precision *P* for the Local name extractor: 1.46 times higher for Subset A and almost two times higher for Subset B than for LingPipe. A more detailed examination of the results shows that a larger number of mislabeled words by LingPipe are capitalized words such as names of software products (e.g., '*Adobe Acrobat*', '*Dreamweaver*'), words of exclamation and amazement such as '*Aha*', '*Yeah*', '*Duh*', '*Wow*', and greetings such as

[14]Alias-I LingPipe toolkit for linguistic analysis - http://www.alias-i.com/lingpipe

'*Hi pg*', '*Hey*', '*Hello all*'. This is probably because LingPipe was originally trained on newswire corpora where words of exclamation and amazement as well as greetings are rare. For the Local name extractor, the most common reason for false-positive results was the selection of words from the name dictionary such as '*major*', '*long*', and '*mark*' that were not used as names in the text. Future research on this project will focus on reducing the number of false-positive results in the Local name extractor.

When examining recall values, both algorithms showed comparable results around 0.65 - 0.70. Recall indicates how many more words need to be added manually. Among the names that were missed by LingPipe were group names and nicknames (e.g., *dw*, *ed*) which are difficult to detect for any algorithms. But there were also missed names that should have been easy to find such as Wendy, Vincent, Scot, and Robert. As for the Local name extractor, the most frequently missed names were solitary last names that were not preceded or succeeded by other contextual words. This can be explained by the fact that the Local name extractor was not designed to recognize solitary last names because group members in these online communities were observed to refer to each other by their first names or nicknames. Since this is not necessarily true for other types of datasets, the future versions of the Local name extractor will include an option to look for solitary last names as well. The Local name extractor also missed four names of group members in Subsets A and three in Subset B due to the names' foreign origins. For the purposes of this study, these names were later added manually using the web interface. Despite this drawback, the substantially higher precision makes the Local name extractor a very effective and efficient tool for personal name extraction.

3.1.3 Name Network: Tie Discovery and Alias Resolution

After all network nodes consisting of previously extracted personal names are identified, the next step is to uncover if and how these nodes are interconnected. The method used relied on both the content of messages and the reference chain as provided in the threaded discussion data (i.e., the history of who posted before the current post) to infer ties between people.

The algorithm works under the assumption that the chance of two people sharing a social tie is proportional to the number of times each of them mentions the other in his/her postings either as an addressee or a subject person. As a way to quantify this assumption, the algorithm adds a nominal weight of 1 to a tie between a poster and all names found in each posting. To demonstrate how the algorithm works, a sample posting below is used:

> From: wilma@bedrock.us
> Reference Chain: tank123@gl.edu, hle@gl.edu
>
> Hi **Dustin**, **Sam** and all, I appreciate your posts from this and last week […]. I keep thinking of poor **Charlie** who only wanted information on "dogs". […] Cheers, **Wilma**.

Note: The 'Reference Chain' listed above is present in the stored data associated with each message so the threaded discussion software can reproduce the threads at any time. However, these data are not normally visible in this way to those engaging in threaded discussions.

As indicated in the header, this posting is from *wilma@bedrock.us*, and it is a reply to the post by *hle@gl.edu*. And *tank123@gl.edu* is a person who actually started the thread. There are four names in the posting: *Dustin*, *Sam*, *Charlie*, and *Wilma*. According to the algorithm, there will be connections between the poster *wilma@bedrock.us* to each

name in the postings:

wilma@bedrock.us - Dustin
wilma@bedrock.us – Sam
wilma@bedrock.us – Charlie
wilma@bedrock.us – Wilma

However, there are a few problems with this approach. First, *Wilma* is a poster; so there is no need for the *wilma@bedrock.us – Wilma* connection. Second, what will happen if more than one person has the same name? For example, suppose that there is more than one *Sam* in the group, how would we know which *Sam* is mentioned in this posting? Conversely, there could be situations where many different names can belong to one person. Furthermore, in the example above, *Charlie* is not even a group member; he is just an imaginary user. Ideally, the poster should not be connected to *Charlie*. To address these problems, it is typical practice to use an *alias resolution* algorithm.

To disambiguate name aliases, the algorithm adopts a simple but effective approach that relies on associating names in the postings with email addresses in the corresponding posting headers (further referred as *name-email* associations). By learning name-email associations, the algorithm knows that there are, for example, two *Nicks* because of the existence of two associations for *Nick* with two different email addresses. The easiest way to discover such name-email associations is to use a class roster or University's online phonebook directory. However, because students often did not use formal names from the roster to refer to each other (but nicknames and informal names), such resources were not used to complete this task in this case. Instead, a general approach was developed that learns all associations automatically.

The derived algorithm relies on an assumption, similar to one used by Hölzer et al. (2005), that the higher number of collocations of two objects generally indicates a stronger association between them. In this particular case, the two objects are (1) a personal name from the body of the posting and (2) an email from the posting header. To improve the accuracy of associations, instead of counting collocations for all names and all emails, the algorithm associates a name with either a poster's email or with an email of potential addressee(s) (emails from the *reference chain*). As a point of clarification, the association between a name and poster's email will be called *Association type P* (or just *Association P*), where *P* stands for *poster*. And the association between a name and addressee's email will be called *Association type A* (or just *Association A*), where *A* stands for *addressee*. In addition to counting the number of collocations, the five-step alias resolution algorithm also assesses the confidence level for each association. The confidence level is assigned based on two criteria: the position of a name in the posting (e.g., at the beginning, middle or end of the posting), and a list of context words as described below.

Step 1. Determining Associations Type P. A poster's name usually appears near the end of a posting in the signature. To find the Association P and estimate its confidence level, the algorithm first calculates how far a name is from the end of the posting using the following formula: $\frac{\text{pos}}{100}$, where *pos* is a relative position of a name inside the posting (in per cent). This value is taken as an initial value of the confidence level of Association P.

Step 2. Determining Associations Type A. Next, the algorithm estimates the confidence level of Association A. To calculate the initial value of the association, it uses the complement of the formula from Step 1: $1-\frac{pos}{100}$. This is because the closer the name is to the beginning of the posting, the more likely it is to be a part of the greeting. And if a word is part of the greeting, then it is more unlikely that the name belongs to a poster.

Step 3. Taking Into Account Context Words. Next, the algorithm checks if a name appears in close proximity (1-2 words) to one or more words or phrases that are commonly used in the signature such as 'thank you', 'best regards', 'cheers', or a character indicating a new line. (A dictionary of these words/phrases was compiled manually before running the algorithm. See Appendix A.) If yes, the value of the confidence level of Association P is increased by a factor of *m*. And if a name appears in close proximity to words or phrases that commonly appear with addressees, then the confidence level of Association A is also multiplied by *m*. The words that commonly appear with addressees include those used in greetings – 'hi', 'hello', 'dear', etc; those used to state agreement – 'agree with', 'disagree'; or words that refer to others – 'according to', 'said that', etc. In the current version of the algorithm, *m* was set to 2 for reasons described below.

One of the main reasons for using a multiplier *m* is to resolve situations where Association A is equal to Association P (within a 10 per cent margin of error). This situation may arise when a name is found near the middle of the posting. For example, when a short message is followed by a long signature, this might cause the name in the signature to appear in the middle of the message. In such situations, by relying only on

information about the position of a name in the message, it would be impossible to determine whether a name belongs to a poster or an addressee. But if it is known that a name appears in close proximity to one or more words that are commonly used in the signature, for example, then we can say with a higher level of confidence that the name belongs to the poster. To reflect this logic, the multiplication factor *m* was introduced. The value of 2 was chosen through a simple iterative procedure of parameter optimization. In this procedure, the algorithm was run on Subset A and B described in Section 3.1.2 with different values of *m* ranging from 1.5 to 3 (increased at an increment of 0.5 at each subsequent step). When a subsequent step did not produce better results, the iterative procedure was then stopped. The results showed that when *m* was less than 2, there were 9 postings in Subset A and 4 postings in Subset B where associations P and A were improperly identified. Upon closer examination of the improperly identified postings, it was discovered that all of these postings had some text that came after the poster's signature. However, when *m* was greater or equal to 2, all posters and addressees were identified correctly. That is why it was decided to set the value of *m* to 2 for the current study. However, for other types of CMC like mailing listservs, where postings may include more noise and where the position of a name has less to do with its role as a poster or an addressee, a higher value of *m* may be required.

To demonstrate Steps 1-3, the sample posting given above is used again. The result of running Step 1, 2 and 3 is shown in Table 3.2. For each word that represents a personal name in the posting, the table includes information about (1) two context words preceding and following the word, (2) its relative position from the beginning of the posting in per cent, and (3) whether a name appears in close proximity (1-2 words) to one

or more words that are commonly used in the signature or with addressees. The last two columns in Table 3.2 show the estimated confidence level for associations P and A measured as described above. For example, the value of Association A (addressee) for Dustin is computed as $1-\frac{pos}{100} = 1-\frac{0}{100} = 1$. To achieve the final estimate of Association A, 1 is then multiplied by 2 because the word 'Dustin' appears near one of the pre-defined words commonly used with addressees - 'Hi'.

The next step explains the procedure of selecting the strongest association for each name. For example, this procedure will help to decide whether *Sam* is a poster and thus, should be associated with the poster's email or is he a recipient of the posting and therefore, should be associated with an addressee's email.

Step 4. Choosing Between Association P or A: In step 4, the algorithm compares and selects the association – P for Poster or A for Addressee – with the highest confidence level. When the difference between values for P and A is insignificant (less than 10 per cent), it rejects both associations due to insufficient information to make the determination. If P is greater than A, then the algorithm assigns that particular name to the poster's email found in the 'From' field of the header. Otherwise, the name is assigned to an addressee's email. In the example above, *Charlie* will be ignored due to the lack of evidence to support one association or the other. *Wilma* will be associated with the poster's email *wilma@bedrock.us*.

Dustin and *Sam* will be considered to be addressees. However, as noted earlier, there is no information on addressees' emails in the posting headers. To work around this, the algorithm assumes that addressee(s) are likely to be somebody who posted in the

thread previously; therefore, there is a good chance that their emails are in the reference chain. But because it is not known to which email to associate each name, names are associated with all emails in the reference chain using different weights. The algorithm distributes weights based on an email's position in the reference chain.

Examination of the threaded discussions shows that the earlier an email appears in the reference chain, the less likely its owner is to be referred to in the current posting. Thus, it should get the least weight. A rectangular hyperbola function is a good candidate for weight assessment. The current version of the algorithm uses the following variation of the rectangular hyperbola function $w = 1 - \frac{1}{pos + 1}$, where *pos* is the email's position in the reference chain. Following the formula above, when the value of *pos* is increasing, indicating that we are moving from the first person in the chain to the most recent one, the weight ***w*** will be also increasing from 0.5 to close to 1. In the sample posting above, Association A, between *Dustin* and *tank123@gl.edu* (thread starter), will get a weight of $1 - \frac{1}{1+1} = 0.5$ and between *Dustin* and *hle@gl.edu* will get a weight of $1 - \frac{1}{2+1} = 0.67$.

After processing all postings, the result is a list of name-email pairs and corresponding confidence levels. Because each message is unique, the confidence levels calculated based on different postings will be different across postings even for the same name-email pair. To combine evidence from different postings, the algorithm calculates the overall value of the confidence level based on the confidence values of all occurrences of each unique name-email pair. Below is a formula that was devised to accomplish this:

[OVERALL CONFIDENCE LEVEL] $_{\text{for each unique name-email pair}} = N_P \bullet M_P \bullet Par + N_A \bullet M_A$

M_P, M_A are the median confidence level values for associations type P and A correspondently for each unique name-email pair. Note: The reason *median* and not *average* function is used is to reduce the effect of possible outliers that may appear due to the variations in the posting formatting.

N_P, N_A represent the number of occurrences of each unique name-email pair for associations type P and A correspondently. Note: The reason medians M_P and M_A are multiplied by N_P and N_A is to reflect the fact that the overall confidence level should grow proportionally to the number of the observed postings with that name-email pair.

Par is an experimentally defined parameter (in the current version, it is set to 2). ***Par*** is used to give more weight to the M_P-component of the formula. This is because there is less uncertainty in identifying associations of type P than associations of type A.

To exclude 'weak' associations that might have appeared due to an error or those associations that do not have enough supportive evidence, the algorithm removes all associations where the value is less than *0.001* (defined experimentally). This is also an effective way to remove all names of people who have never posted to the bulletin board. Since each student in the class posted at least one message to the bulletin board, and the only people who did not post any messages but were mentioned by somebody in a class were non-class members, it is safe to remove non-posters from the cases studied. However, it may not be safe to remove non-posters from other online communities. For example, employees may be discussing their boss' initiative using his name on the company's online forum, but the boss has never posted to that online forum himself. Since the boss is also part of the social network, all connections to him discovered from the forum postings should be also included into the resulting name network. To address

such situations, ICTA provides the user with an option of whether to search for and include non-posters into the name network.

Finally, to achieve the highest level of accuracy on this task, a semi-automated approach was adopted. More specifically, a web interface was developed to allow a manual correction of the extracted associations (see Figure 3.4). For each email address that had at least one name associated with it, ICTA displays a list of choices for possible aliases sorted by their confidence levels. Using this interface, a researcher can easily remove and/or add a new name-email association by selecting a name from a list of all names found in the dataset from a drop down menu. The manual examination of the resulting name-email pairs for the two sample datasets from Section 3.1.2, Subset A (853 postings) and Subset B (534 postings), showed that the algorithm was able to associate all names with the corresponding email addresses correctly.

Using the formula described above, the end result of Step 4 is a list of email-candidates with their corresponding overall confidence level values for each distinct name found in the dataset. This list of email-candidates is then used during the final step of the alias resolution algorithm.

Step 5. Disambiguating Personal Names: After learning all possible name-email associations and their overall confidence levels, the algorithm goes through all postings once again to replace those names mentioned in the body of the postings that have been associated with at least one email. If a name has more than one email-candidate, then the algorithm uses the email with the highest level of confidence. However, in some cases selecting an email with the highest level of confidence may produce an incorrect result. For example, in the sample dataset, there were two *Wilmas*: *wilma@bedrock.us* with the

confidence level set to 27.45 and *wm2@iso.edu* with the confidence level set to 18.83. If we were to select an email with the highest confidence level, then all mentions of *Wilma* in all postings would be attributed to only *wilma@bedrock.us*. But, of course, this would be wrong since in some instances it might be *wm2@iso.edu*.

To ensure that the algorithm identifies the right *Wilma*, the following fail-safe measure was implemented. If there are more than one email-candidate, the algorithm then relies on an additional source of evidence – the reference chain. First, it identifies an overlap between email-candidates for a name (from Step 4) and emails from the reference chain. If the overlap is empty, then the algorithm proceeds as usual and uses the email with the highest confidence level (further referred to as the *strongest candidate*). When the overlap is not empty, it means that one or more email-candidates have previously posted to the thread. Based on the manual analysis of the dataset, the name mentioned in the posting is more likely to belong to an email-candidate that is also in the reference chain than to an email-candidate that is not. Taking this observation into consideration, if there are two possible email-candidates, as in case with Wilma, and the strongest candidate (*wilma@bedrock.us*) is not present in the reference chain, but the other candidate (*wm2@iso.edu*) is, then the algorithm uses the one that is also in the reference chain. In cases, when both email-candidates have previously posted to the thread, the algorithm takes the candidate who has posted the most recent posting to the thread.

This section concludes the description of the 'name network' algorithm which consists of two main steps: node discovery and tie discovery. The next section below briefly explains the procedure for building chain networks, an alternative automated method for collecting network data from the bulletin board postings.

3.2 CHAIN NETWORK METHOD

Chain networks are built automatically using information from the posting headers, specifically reference chains. (A reference chain refers to a running list of group members who previously posted to a particular discussion thread.) There are at least four distinct options for building chain networks (other options may be considered as variations of those listed below):

- Option 1: Connecting a poster to the last person in the post chain only
- Option 2: Connecting a poster to the last and first (=thread starter) person in the chain, and assigning equal weight values of 1 to both ties
- Option 3: Same as option 2, but the tie between a poster and the first person is assigned only half the weight (0.5)
- Option 4: Connecting a poster to all people in the reference chain with decreasing weights

Originally all four options were considered for this study. However, in the end only Option 1 was used. This is because according to the results described in Section 5.2.1, the other three options were judged to be unreliable and would have introduced far more false-positive connections; thus, reducing the accuracy of the chain networks.

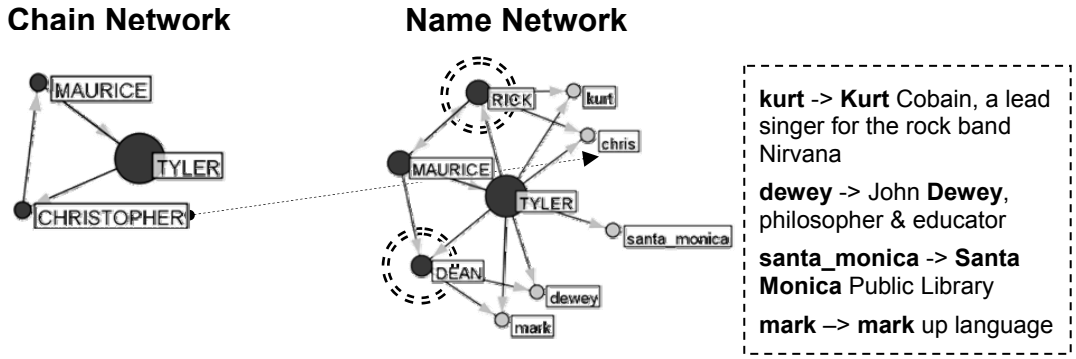
3.3 SUMMARY

This chapter outlined the implementation of both the 'name network' and 'chain network' methods used in this study. The chapter also highlighted some of manual override features available in ICTA that researchers can use to further improve the accuracy of the node and tie discovery steps. The next step in the research is to evaluate the effectiveness of the two automated methods for building social networks and to

determine whether anything new or useful can be gained from using the new automated 'name network' method. As a standard for the evaluation, manually derived self-reported networks are used, built using surveys given to the students who have elected to participate in the study. The next chapter, Chapter 4 describes data collection procedures and tools for building students' self-reported social networks. Chapter 5 reports on the procedures and results of the comparative analysis of the social networks derived using the 'name network' method against chain networks and self-reported social networks.

3.4 FIGURES AND TABLES

Figure 3.1: A personal network for Tyler from a sample dataset



Notes:

- The size of each node represents out-degree (the number of ties to others)
- The nodes with the dotted circles around them were added by the author to highlight some of the differences between the two networks.
- The light grey nodes in the Name Network were identified by the system as non-posters.

Figure 3.2: A web interface for editing extracted names: Top 30 names automatically extracted from the Internet Researchers' listserv for messages posted during October 2002

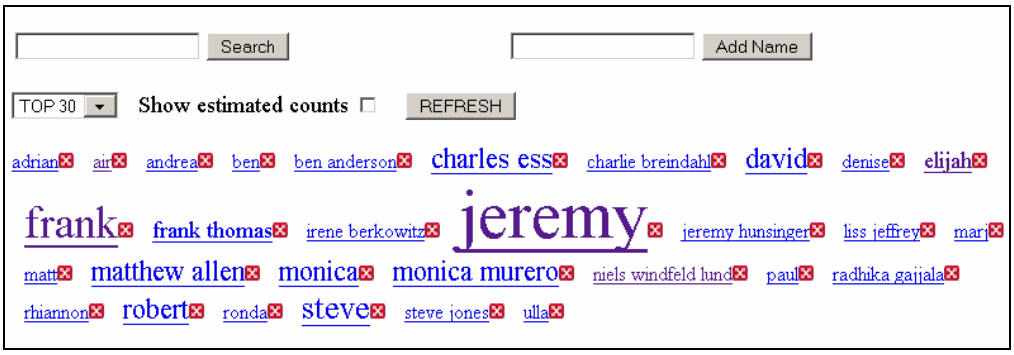


Figure 3.3: A list of messages containing "Jeremy"

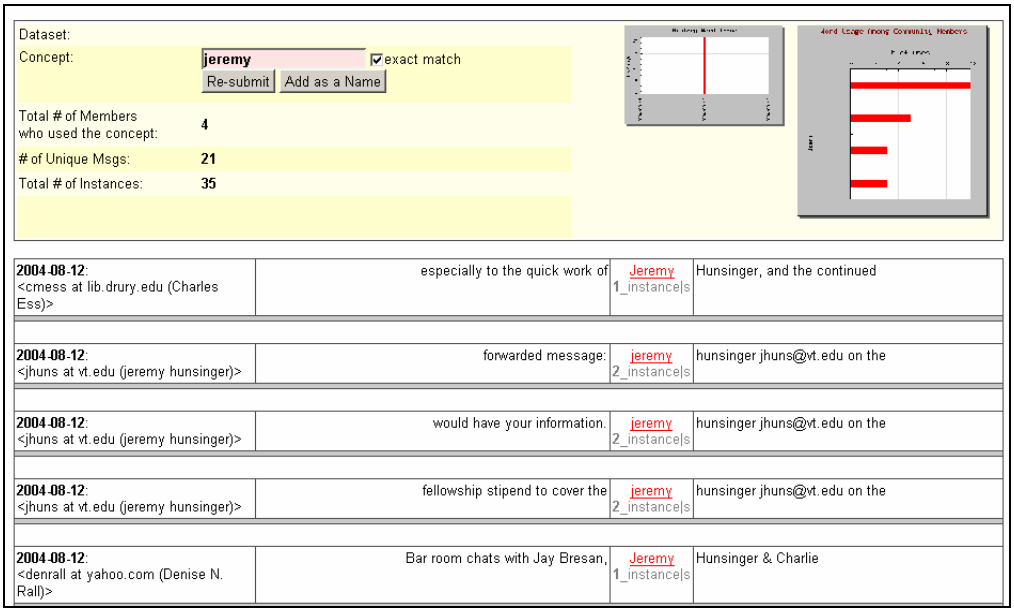
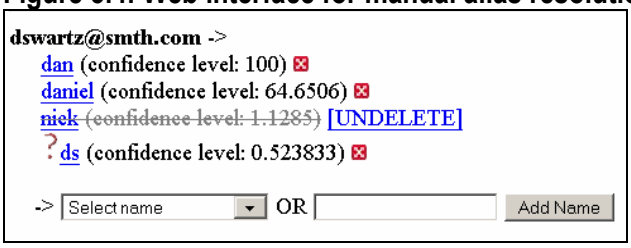


Figure 3.4: Web interface for manual alias resolution



Note: In the example above, the system assigned four names Dan, Daniel, Nick, and DS to dswartz@smth.com. However, after a manual examination of the results, a user deleted Nick who was incorrectly associated with this email due to his frequent collocations in postings with Dan. Also a question mark next to DS indicates a small confidence level (less than 1).

Table 3.1: Comparing Local and LingPipe name extractors

	Subset A		Subset B	
	Local	**LingPipe**	**Local**	**LingPipe**
Total # names	1459	997	929	577
Total # correct distinct names discovered (**T1**)	331	340	195	176
False-Positive (incorrectly identified) (**F**)	45(12%)	227(40%)	39(16%)	238(57%)
# names in common	171		99	
# of missed names found by the alternate algorithm (**D**)	160	165	74	96
# of missed names of group members (**M**)	4	0	3	0
Total names in the dataset **T2**=T1+M+D	500	500	272	272
Precision **P** = T1/(T1+F)	0.88	0.60	0.83	0.43
Recall **R** = T1/T2	0.66	0.68	0.71	0.65

Table 3.2: The results of running the 'name network' algorithm

Words to the Left	**Name**	**Words to the Right**	**Position, %** (pos)	**Context word?**	**Association P** (Poster?) $\frac{pos}{100} \cdot 2^{\dagger}$	**Association A** (Addressee?) $(1-\frac{pos}{100}) \cdot 2^{\dagger}$
Hi	**Dustin**	Sam and	0	Yes ('Hi')	0	1 • 2 = 2
Hi Dustin	**Sam**	and all	1	Yes ('Hi')	0.01	0.99 • 2 = 1.98
Of poor	**Charlie**	who only	50	No	0.50	0.50
Cheers *	**Wilma**		88	Yes ('Cheers' and a new line)	0.88 • 2 = 1.76	0.12

Notes:

'*' indicates a new line

† - multiply by 2 only if a name appears in close proximity (1-2 words) to one or more words that are commonly used in the signature (for Association P) or with addressees (for Association A).

CHAPTER 4: DATA COLLECTION

This chapter begins with a description of two datasets used for training and evaluation of name networks, and then moves into a discussion of data gathering procedures for building students' self-reported social networks. The chapter ends with a description of a web-based system called ICTA that was developed and used in this research for data management, manipulation, exploration and visualization.

4.1 DATASETS

The main dataset for this study consists of bulletin board postings and students' responses to an online questionnaire from six graduate level online classes at the Graduate School of Library and Information Science (GSLIS), University of Illinois at Urbana-Champaign (UIUC). See Table 4.1 for more details about these classes. The data was collected in Spring 2008 as part of a larger study on online learning in collaboration with Caroline Haythornthwaite.

Instructors in these classes primarily relied on Moodle (an open source course management system) to make announcements, distribute class materials and facilitate weekly discussions using bulletin boards. Once a week, students met online using a locally implemented chat facility integrated into the Moodle environment. During these live sessions, the instructor delivered the lecture via a live audio feed. During the lecture, students could ask questions or answer instructor's questions by typing in the chat room. During some live sessions, the instructors divided students into smaller discussion groups, with each group using a separate chat room for their discussions. The data used here includes only those bulletin board discussions that were public to the class as a whole.

A different, but similar dataset collected as part of an earlier study was used for training and iterative improvements of the 'name network' method. This dataset consists of bulletin board postings from eight iterations of the same online class, given by the same instructor at the GSLIS from 2001 to 2004. These discussions were conducted using a local implementation of the online learning environment developed at GSLIS. See Table 4.2 for more details. For the pilot study, two samples from this dataset were used: Sample A consists of 853 postings in 12 bulletin boards, and Sample B consists of 534 postings in 5 bulletin boards.

These two datasets were selected because they possess a few unique characteristics that make them ideal candidates for conducting these experiments. These characteristics include: being a close community (with known membership), having a clear beginning and end to the time frame for data collection, and a non-intrusive data collection capability.

Prior to the beginning of the recent data collection, permission was obtained from the University of Illinois Institutional Review Board and the instructors of the selected classes. Next, an alert message was posted to the online 'news and announcement' bulletin board for each class. This alert message (see Appendix B) informed students that their public postings and chat room logs would be made available for analysis after the class is over. If students did not want their text quoted in any way, they were asked to contact the researchers. (A similar procedure had been followed for the earlier data collection.) All students' names have been anonymized to protect their privacy.

4.2 COLLECTING SELF-REPORTED SOCIAL NETWORKS

Students' self-reported social networks were collected via an online questionnaire

administered once near the end of the semester (See Appendix C). The questionnaire was designed based on Haythornthwaite's 1999 LEEP study protocol. To encourage participation in the questionnaire, all respondents were automatically enrolled in a random drawing for one of three iPod Shuffles. The questionnaire consisted of three main sections: (1) students' perceived social structures, (2) prominent members of the class, and (3) interactions in the class as a whole. The first two sections were specifically designed to collect information on self-reported social networks. The first section asked students to indicate the frequency of their associations with each classmate on a scale from 1 to 5 (with [5] indicating a more frequent association) with respect to three different relations: learning something new about the subject matter from another student, working together, and friendship. The second section asked students to nominate 5 to 8 prominent students that best fit the following four criteria: "influential in one's learning", "important in promoting discussion", "help with understanding a topic or assignment" and "made class fun". Each question in section 1 and 2 was designed to discover one of the many possible social relations (e.g., learn, work, help, etc) that might exist between the students. The third and final section was designed to assess learning at a group level by asking students questions about knowledge building and sense of community dimensions as proposed by Law (2005) and Lin et al. (2007). Completion of the questionnaire was voluntary. Students were provided with a consent form online. A list of all students in each class was included and the students answered by ranking and/or selecting individuals who meet the criterion.

Special consideration was given to decide when exactly to invite students to fill out the questionnaire. If asked too early in the semester, students could not have been

together long enough to make judgments about the strength of their relationships to each other. On the other hand, the end of the semester is not a good time because it is known to be very busy period in students' lives. Further, if asked to fill out the questionnaire long after the class is over, there is a risk that students' reflections of class interactions would not be as accurate. Taking these considerations into account, it was decided to invite participants approximately 3-4 weeks prior to the end of the semester and then send a reminder to complete the questionnaire right after the end of the semester. Out of 128 students who were invited to complete the questionnaire, 81 students (63 per cent) completed the survey.

A self-reported network was built to capture the 'strong' ties among classmates using the following procedure. First, the procedure added a tie between each respondent and his or her nominees. For the questions from section 1 of the survey, only nominations with an association level of three or higher were considered. This is because a nomination with an association level less then three means that a student "doesn't know" the nominee or the nominee is "just another member of class". Such responses indicate no or a very weak personal connection between the student and the nominee, and therefore they were not used to establish a tie in the self-reported network. For the questions from section 2 of the survey, all nominations were considered since the questions in this section were already designed to solicit answers only about strong connections in the class.

The next step was to assign weights to each tie. The weights were assigned based on how many times each nominee was selected by the same respondent in different questions of the survey. To better reflect actual social relationships between students, the procedure removed all 'weak' ties with a weight of less than three. This design decision

reflects the principle of relational multiplexity from Sociology that tells us that people with closer relationships are likely to maintain ties based on more than one type of relation (see, for example, Haythornthwaite, 2008). Since the procedure only kept so-called 'strong' ties, it is very likely that they will be symmetric. To help restore some ties missing due to the non-respondents, the resulting network was symmetrized.

Open source software called phpESP[15] was used to conduct the survey. A Social Network Analysis tool called ORA[16] v.1.9.5.2.6 was used for storage and basic manipulations of the network data. Internet Community Text Analyzer (ICTA)[17] described in the following section was used to build name and chain networks automatically and to explore postings in the datasets manually.

4.3 INTERNET COMMUNITY TEXT ANALYZER (ICTA)

As part of this dissertation work, I developed a web-based system for content and network analysis and visualization called Internet Community Text Analyzer (ICTA). This section describes some background information about the development process of ICTA, its infrastructure and user interface in more details.

At the beginning of my PhD program in 2006, my advisor, Professor Caroline Haythornthwaite, came to me with a challenge. She had a large archive of bulletin board postings from eight online classes collected over a period of four years. Each class in this archive generated on average about 1500 postings. The challenge was that other than looking through the data manually one message at a time, there was no quick or easy way

[15] phpESP - http://phpesp.sourceforge.net

[16] ORA - http://www.casos.cs.cmu.edu/projects/ora

[17] Internet Community Text Analyzer (ICTA) - http://textanalytics.net

to analyze and make sense of such a large amount of data. Since I came to the Library and Information Science field with a Computer Science background, I immediately realized that a possible solution to this challenge is to develop some kind of automated system that will allow researchers like my adviser to not only explore and manipulate users' generated online textual data but to also visualize and analyze such data and ultimately derive some wisdom from it.

The main goal in developing ICTA is to provide researchers and other interested parties with an automated system for analyzing text-based communal interactions with the help of various interactive visualizations. ICTA's web-based architecture will stimulate collaborative research by allowing researchers to access and analyze datasets remotely from anyplace where there is web access and quickly share their results with their collaborators anywhere in the world. Another benefit of a web-based software implementation is the ability to outsource data processing and have all the heavy computing be done on a speedier remote server. For example, once data has been entered on a stand-alone website, it can then be sent to ICTA to be analyzed in real-time, and then immediately returned and presented using useful visualizations to a community's web space.

The first version of ICTA v1.0 was developed and presented at the Communities and Technologies conference in 2007. ICTA v1.0 made it much easier to explore and analyze the dataset. It facilitated searching the stored versions of the text from these eight classes. The main screen of this tool provided the user with a means to select the class and bulletin board(s) to be analyzed and perform the analysis. During the analysis, the system looked for the top 100 noun phrases in the selected bulletin boards based on their

frequency counts. A tag cloud was then generated to give an immediate visual representation of all of the noun phrases that were identified during the analysis phase (see Figure 4.1). The size of a noun phrase within the cloud correlated with its frequency count, the higher the frequency, the larger the word would appear within the tag cloud. By clicking on any noun phrase from the initial list of 100 noun phrases, ICTA v1.0 returned a list of all instances where that particular noun phrase was located within the dataset (see Figure 4.2). And by clicking on any of the instances the researcher could then see the full posting, with the selected noun phrase highlighted. As an alternative approach, the user could also search by an individual noun phrase, instead of performing the top 100 words analysis. In this case, the user would simply type the desired phrase into a text box, and ICTA then returned a list of all instances where that particular noun phrase was found (if it is present in the dataset) as described earlier.

In the 2006-2007 LEEP language study conducted together with Professor Caroline Haythornthwaite, ICTA v1.0 allowed us to study word use in eight online classes and answered questions like what the community interests and priorities are and what are the patterns of language that characterize a particular community. For example, we were able to discover an interesting pattern in the use of the topic of databases that is of particular relevance and importance to the field of Library and Information Science (LIS) and to the students in the classes. It turned out that the use of words like "database(s)" and "RDB"[18] actually declined over time in terms of the per cent of messages in the semester containing at least one occurrence of the word. At first this did not seem to be appropriate given the importance of the term; however, the percentage

[18] RDB stands for "Relational Data Base"

reduction in use may reflect the increased familiarity with databases that students have in general as these become more part of the overall curriculum. (See more details in Haythornthwaite & Gruzd, 2007)

The analysis of these eight online classes with ICTA v1.0 showed that the system was useful in the preliminary exploration of large datasets and in the identification of important issues/topics being discussed by group members and their changes over time. However, this first study with ICTA also warranted two important improvements. First, ICTA v1.0 did not have capabilities to upload a new dataset for the analysis automatically which made it less valuable for other researchers who want to analyze their own datasets. As a result, the first improvement to the next version of ICTA was to add an interface where anybody can create their own account and upload their own data for further analysis with ICTA. Second, ICTA v1.0 primarily focused on the text analysis of online interactions. Although useful, text analysis alone does not provide a complete picture of an online community. Specifically, it does not take into account relationships between group members that may also provide important insights into the internal operation of an online community. For example, using a simple automated text analysis, we can easily tell that there are many disagreement-type postings from a particular dataset; however, this information alone does not tell us whether the postings are coming from just a few members who tend to disagree with each other or is it the general characteristic of this particular community as a whole. In other words, when it comes to studying online communities via their textual exchanges, it is important not only to know what they are all talking about, but also to whom they are talking. To increase the range of the types of research questions that a researcher could address with ICTA and

hopefully provide researchers with an additional view into the inner working of an online community, I added a social network discovery and visualization component to the system. These new components in ICTA can use both traffic-based (who talks to whom) and content-based data to automatically extract social networks information and offer various visual representations of the analysis.

There are some other projects on the Internet that broadly share a few similar functionalities with ICTA, however they are designed for other fields and have different implementation and goals in mind. These are visualization tools like Swivel[19] and IBM's Many Eyes[20] that allow anybody to upload some data and then visualize it by selecting one of the available visualizations types such as graphs, charts, histograms, etc. There are three main differences between these pure visualization tools and ICTA. First, these tools are not tuned to work with computer-mediated communication (CMC)-type data such as emails, forum postings or chat transcripts, etc. They mostly work with data that is already organized in a table format such as a table of top 50 US companies that made most money in 2008 and their corresponding revenues. Second, these online tools provide only top-level visualizations without interactive features that would allow researchers to be better engaged with their data by being able to explore and delve into their datasets at different levels of granularity. Finally, the visualization tools mentioned here lack some basic security features. Most researchers are working with private datasets; at the very least they all want some control over who can have access to their dataset and view the results. Overall, the data visualization tools mentioned above are easy to use and good for

[19] Swivel - http://www.swivel.com

[20] IBM's Many Eyes - http://manyeyes.alphaworks.ibm.com

the masses since they allow for basic visualizations, sharing and discussion over the Internet. But for the reasons mentioned above, these tools are not satisfactory for researchers whose needs are very different. The newest version of ICTA tries to address all these problems.

In its current state, ICTA is a fully functional prototype designed to test and evaluate the effectiveness of different text mining and social network discovery techniques. The goal at this stage is to identify a range of optimal values for various parameters that control automated procedures. Eventually, these optimal values will be used as default settings in a future simplified single-step 'one-button' version of ICTA. Below is a brief description of ICTA's current multi-step interface and functionalities.

First, a user starts by importing a dataset. To do this, he/she can upload a file or specify the location of an external repository (See Figure 4.3). Currently, ICTA can parse computer-mediated communication (CMC) that has been stored in one of three data formats: XML (e.g., RSS feeds), MySQL database or CSV text file. After the data is imported, the second step is to remove any text that may be considered as noise (See Figure 4.4). This is an optional step that is primarily designed to remove redundant or duplicate text that has been carried forward from prior messages. To accomplish this, ICTA simply removes all lines that start with a symbol commonly indicating quotation such as ">" or ":". But a user is not restricted to just these two symbols. In fact, in the 'expert' mode, it is possible to remove almost any text patterns such as URLs or email addresses from messages using a mechanism called *regular expression*.

After the data importing and cleansing steps are completed, the data is then ready to be analyzed. In this stage, ICTA uses capabilities from ICTA v1.0 described above to

build concise summaries of the communal textual discourse. This is done by extracting the most descriptive terms (usually nouns and noun phrases) and presenting them in the form of interactive concept clouds and semantic maps (see Figure 4.5) or stacked graphs that show the use of important topics over time (see Figure 4.6). With a summary in hand, a researcher or a member of an online group can quickly identify emerging community interests and priorities as well as patterns of language and interaction that characterize a community. (See Haythornthwaite & Gruzd, 2007, for more details on this type of text analysis.)

Another feature that is available in the 'text analysis' step is to define different groups or categories of words/phrases/patterns (so-called linguistic markers) and then count how many instances of each category are in a dataset and then display them in a form of a treemap view (See Figure 4.7 & 4.8). Using this functionality, a researcher, for example, can define and use categories consisting of various linguistic markers that have been shown to be useful in identifying instances of social, cognitive and/or meta-cognitive processes such as decision-making, problem-solving, question-answering, etc (see, for example, Alpers et al., 2005; Corich et al., 2006; Pennebaker & Graybeal, 2001). For demonstration purposes, ICTA comes with several commonly used categories such as 'agreement', 'disagreement', 'uncertainty', 'social presence', etc. Users can modify the existing categories or create their own that will better reflect their research questions.

The final part of the analysis stage consists of building chain networks and name networks as described in previous sections of this work. This part is the focal point of this dissertation. When building these networks from a CMC-type dataset, there are a lot of different parameters and thresholds choices to select from. To find the most optimal

configuration for a particular type of datasets, ICTA's interface allows users to fine tune many of the available parameters and thresholds. For example, one of the choices that is likely to influence network formation is how to estimate tie strengths between individuals. ICTA provides a range of options for doing this estimation: from a simple count of the number of messages exchanged between individuals to an estimation based on the amount of information exchanged between individuals. For ease of use and a quicker start, users can also use the default options.

After networks are built, they can be visualized and explored using a built-in network visualization tool. Users also have the option of exporting the resulting networks to other popular social network analysis programs such as Pajek[21] or UCINET[22]. In addition to a number of basic visualization features such as scaling, changing graph layouts, selecting cut off points to hide 'weak' nodes or ties, ICTA can also display excerpts from messages exchanged between two individuals to show the context of their relations. The ability to call up and display excerpts from messages makes it a lot easier to 'read' a network and understand why a particular tie exists. This feature is activated by moving a mouse over an edge connecting two nodes (see Figure 4.9). ICTA is also capable of simultaneously displaying two different types of networks of the same group on the same graph using different colors to display edges from different networks. The latter makes it easier to study the quality of and differences/similarities between different networks.

[21] Pajek - http://vlado.fmf.uni-lj.si/pub/networks/pajek

[22] UCINET - http://www.analytictech.com/ucinet

To implement interactive visualizations, I used the Flare Prefuse[23] library developed at the University of California in Berkeley (Heer et al., 2005). Flare Prefuse is an open-source ActionScript library for creating visualizations that run in the Adobe Flash Player. The main reason for using Flare Prefuse is two-fold. First, the library makes it much easier for developers to create professional looking, interactive visualizations that are very easy to customize for a particular application. Flare Prefuse has also been successfully used in other web applications such as GrOWL - a tool for editing and visualizing ontologies for the semantic web (Krivov et al., 2007), PhotoArcs - a tool for creating and sharing photo-narratives (Ames & Manguy, 2006), and by Medynskiy et al. (2006) to visualize online software development communities.

The second reason for using Flare is that it is fully integrated with Adobe Flash which is a leading multimedia platform on the Internet. Adobe Flash Player is now undisputedly the de facto method for playing videos, animations and visualizations on the Internet. As a result, most web browsers and other Internet-enabled devices (mobile phones, play stations, etc) are capable of playing Flash applications. Thus, by choosing Flash as the main engine for visualizations, ICTA is more accessible to the end users regardless of their hardware platform or Internet browser choice.

4.4 SUMMARY

This chapter described two datasets used for training and evaluation of name networks, and data gathering procedures for building students' self-reported social networks. The chapter also introduced a web-based system called ICTA that was developed and used in this research for data management, manipulation, exploration and

[23] Flare Prefuse library - http://flare.prefuse.org

visualization. The next chapter describes how the data collected for this research was analyzed to address the research questions set forth in Section 1.2 and what was discovered during the analysis.

4.5 FIGURES AND TABLES

Figure 4.1: First version of ICTA. Main screen: here user can select the specific course and bulletin board(s) to be analyzed, and the system then looks for the top 100 noun phrases found in the selected bulletin boards based on their frequency counts.

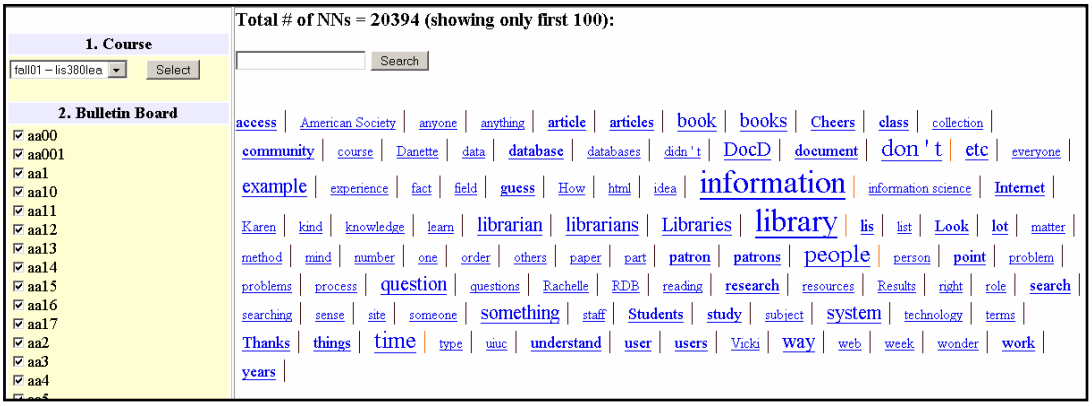


Figure 4.2: First version of ICTA: "Information" in context

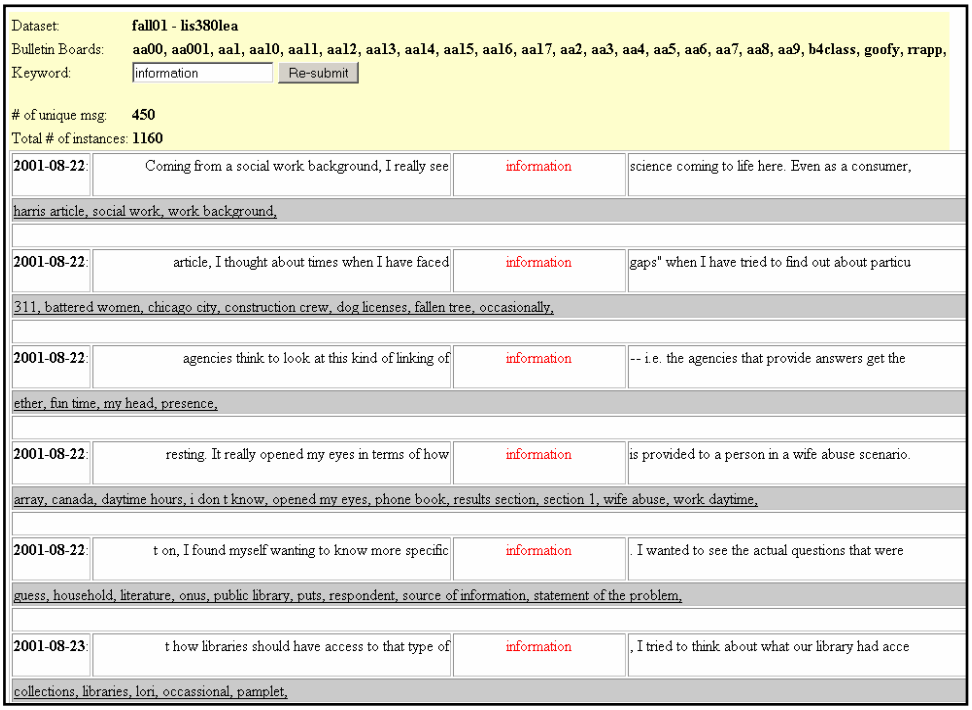


Figure 4.3: ICTA: Importing dataset

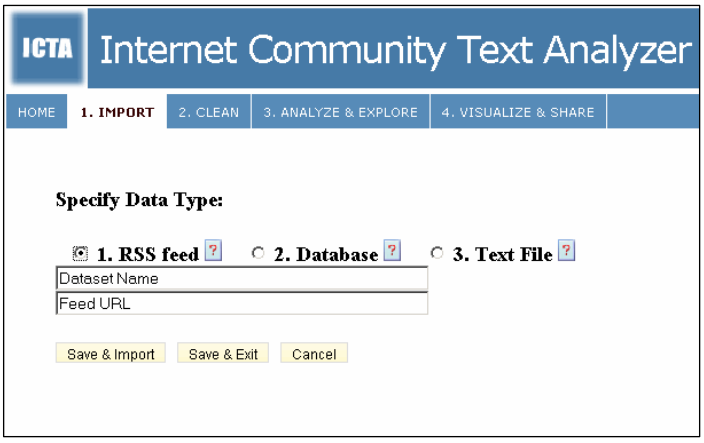


Figure 4.4: ICTA: Cleansing dataset

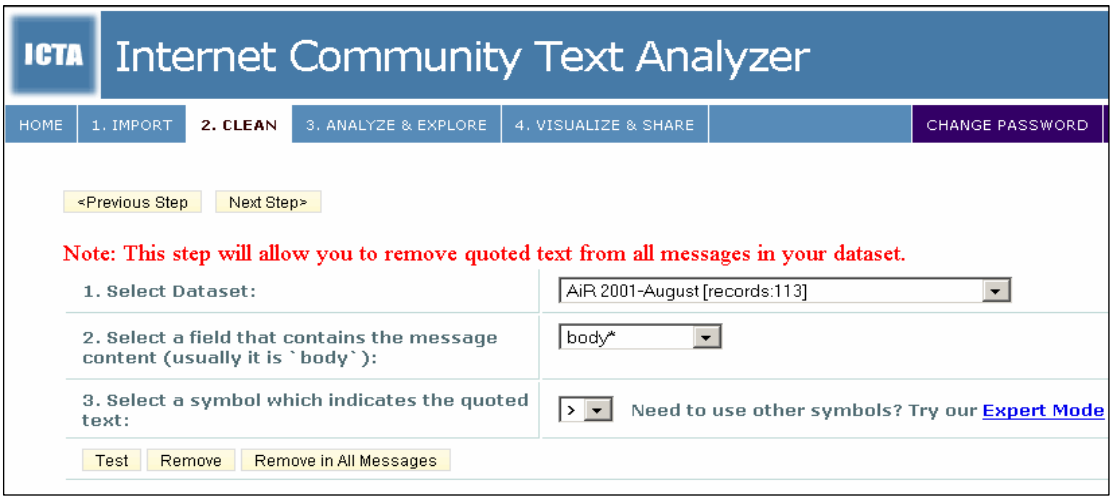
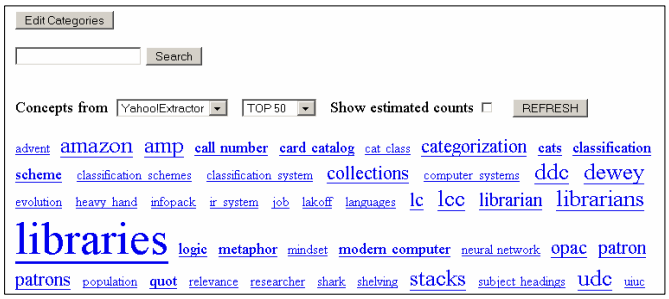
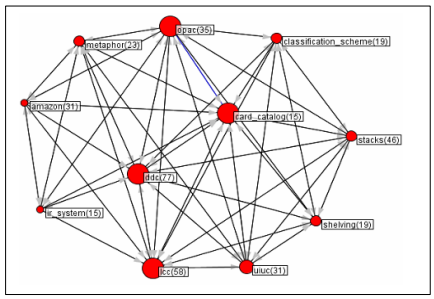


Figure 4.5: ICTA: An example of a concept cloud (a) and a semantic map (b)



(a)



(b)

Figure 4.6: ICTA: Stack graph showing the use of important topics over time

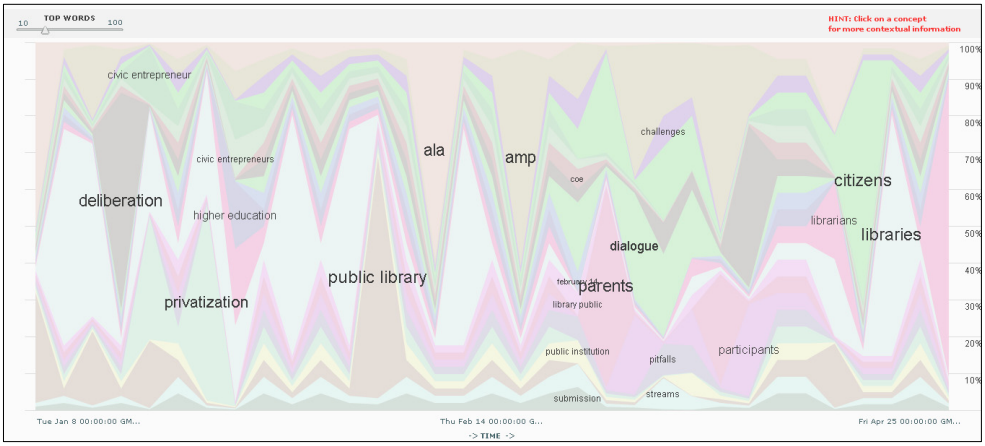
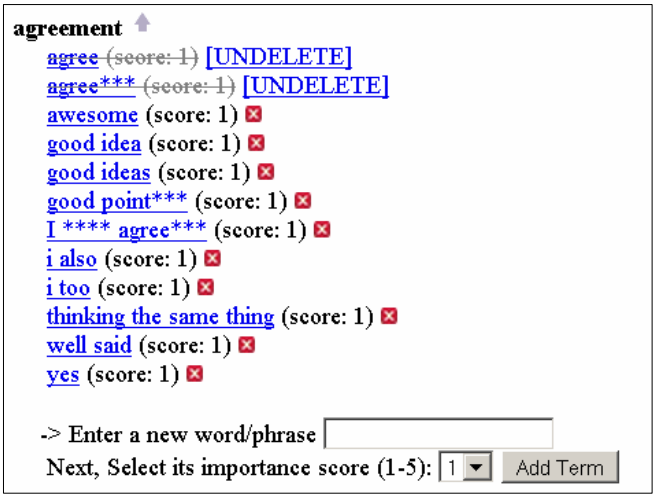
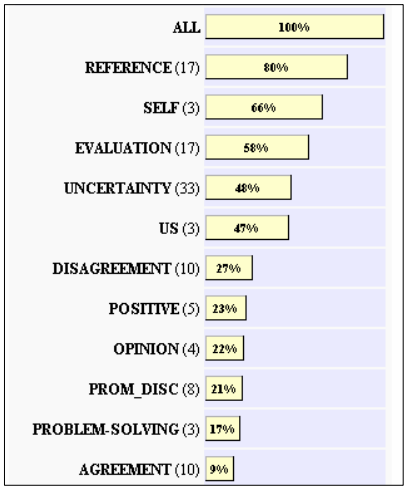


Figure 4.7: ICTA: Sample linguistic markers in the category called "Agreement" (a) and the proportion of messages in each predefined social, cognitive or meta-cognitive category (b)

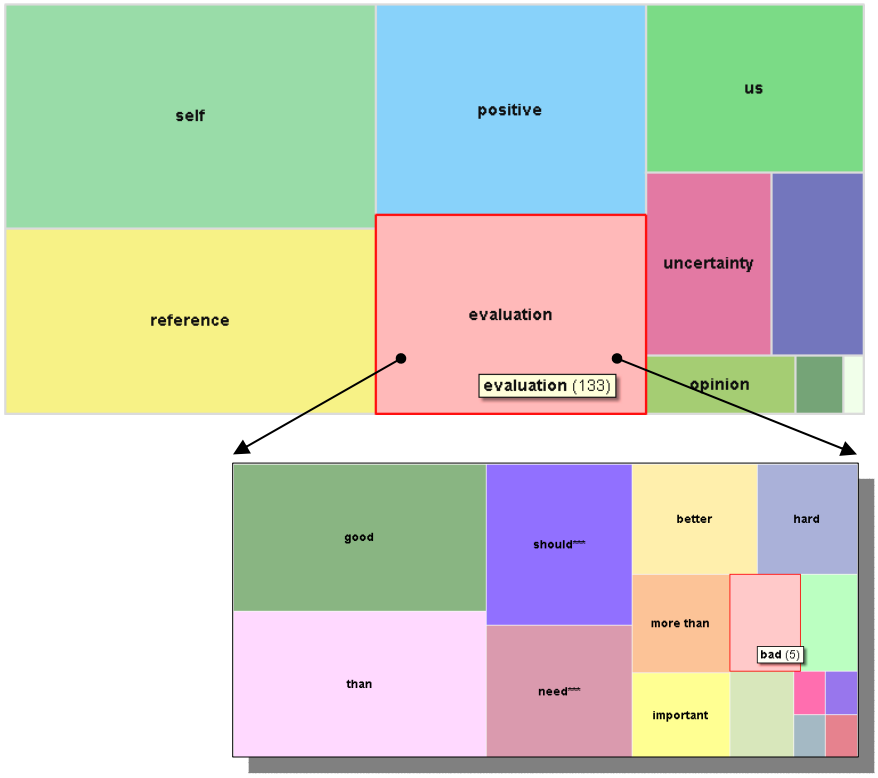




(a) **(b)**

Note: *** (three stars) at the end of a word means that a word may have different endings. For example, "read***" will be equal to "read", "reads", "reading", "readings".
* (one star) between words in a phrase is a placeholder for "any word". For example, "from * post" will be equal to "from my post", "from his post", and "from Nick's post".

Figure 4.8: ICTA: A treemap view of various predefined social, cognitive and meta-cognitive categories found in a sample dataset



Note: Each box represents a certain social, cognitive or meta-cognitive category as defined by ICTA's user. The size of a box corresponds to the proportion of postings in the dataset that match each category. By clicking on any of the found categories, a user will be presented with a treemap representing found linguistic markers in each category.

Figure 4.9: ICTA: Social network visualization

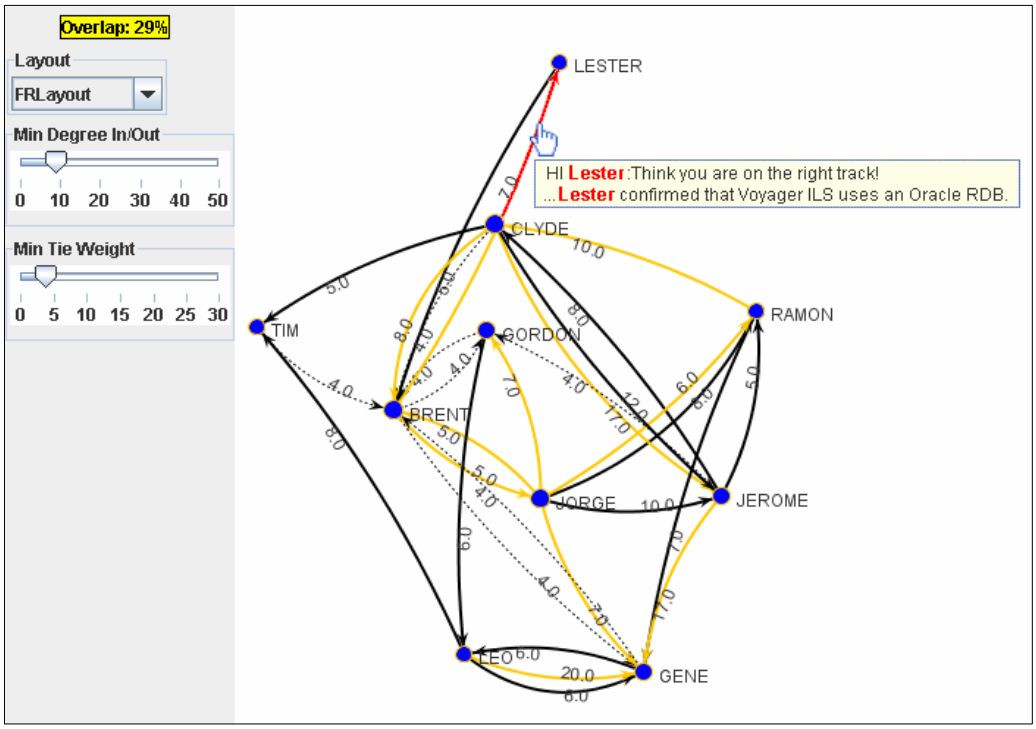


Table 4.1: Spring 2008 LEEP classes: Basic statistics for class-wide bulletin board postings

	Class #1	Class #2	Class #3	Class #4	Class #5	Class #6
Number of Messages	608	855	1,502	164	412	497
Number of Participants	28	20	25	21	19	15
Number of Bulletin Boards	46	36	29	9	14	12
Avg. Number of Characters per Message	1352	886	1090	1047	1310	863

Table 4.2: 2001-2004 LEEP classes: Basic statistics for class-wide bulletin board postings

	F01A	F01B	F02A	F02B	F03A	F03B	F04A	F04B
Number of Messages	1207	1581	1469	1895	1279	1242	1493	2157
Number of Participants	38	47	47	54	54	46	54	52
Number of Bulletin Boards	22	22	28	28	25	24	28	27
Avg. Number of Characters per Message	1073	1056	864	898	1286	953	967	1058

CHAPTER 5: ANALYSIS AND FINDINGS

This chapter provides a detailed analysis of the data collected for this research and offers a set of findings by addressing each of the three research questions set forth in Section 1.2.

5.1 BUILDING NAME NETWORKS

The first research question is

Question 1: What content-based features of postings help to uncover nodes and ties between group members?

This question was discussed in detail in Chapter 3. In short, personal names in postings were used to find nodes and ties between people in the name network. Personal names were chosen as the main input into building the name network because they are good indicators of social ties. Linguistically speaking, the use of personal names performs two main communicative functions as identified by Leech (1999): (1) addressee identification and attention-getting, (2) social bond maintenance function. The first function is self-explanatory, when calling somebody by his/her name, a person identifies somebody among others to talk to and at the same time tries to get that person's attention. The main purpose of social bond maintenance is to sustain and reinforce social relationships. For example, when someone uses formal names and titles, it might be to indicate subordination in the relationship. Or, someone might use informal names or nicknames to show the same social status or to emphasize friendship. The social bond function of naming is especially important in online groups. Since names are "one of the few textual carriers of identity" in discussions on the web (Doherty, 2004, p. 3), their use is crucial for the creation and maintenance of a sense of community (Ubon, 2005) and of

social presence (Rourke et al, 2001). As Ubon (2005) put it, by addressing each other by name, participants "build and sustain a sense of belonging and commitment to the community" (p.122). In sum, by focusing on personal names, the 'name network' method can quickly identify addressees of each message and thus automatically discover "who talks to whom" in one-to-many types of online communication such as threaded discussions and chats. Furthermore, the social bond function of personal names suggests that the discovered ties between people will not just reflect communication patterns, but are also likely to reflect real social relationships between people.

5.2 NAME NETWORKS VERSUS CHAIN NETWORKS

For ease of discussion, the second research question is split into two separate questions:

Question 2.1: How is the proposed name network associated with the chain network?

Question 2.2: How is the proposed name network associated with the self-reported network?

This section only addresses Question 2.1; Question 2.2 is answered below in Section 5.3. In addition to using quantitative analysis, qualitative content analysis of the postings was also used to provide a better understanding of differences between the name network and the chain network. This provided the information necessary to address the third research question:

Question 3: What types of social relations does the name network include?

The analysis began with comparing the name network and the chain network using QAP correlations (Krackhardt, 1987). This was done to determine the level of overlap between these two types of networks. QAP correlation relies on Pearson's

correlation coefficient to compare relational data. It was chosen as the method of measurement for this work because "it presumes neither random sampling of cases from a population […] nor independence of observations" (White et al., 2004, p. 116). Software called ORA was used to compute the QAP correlations.

The results of the comparison are presented in Table 5.1. All tests were significant ($p<=0.05$). In all classes, pairs of name and chain networks demonstrated moderate correlations between 0.45 and 0.69 (See the 'QAP' column in Table 5.1). As expected, there is some overlap between posting behavior as represented by the chain network and 'naming' behavior as represented by the name network. However, there are also substantial differences in what is revealed by each of these networks. To better understand these differences and assess the accuracy of both networks, the next section compares all connections that make up each tie from the name network with those from the chain network. More specifically, the next step in the analysis determines how many connections discovered by the 'name network' method were not discovered by the 'chain network' method and vice versa.

5.2.1 Connections Missed by the Chain Network

The chain network is built based on the information in the reference chains; as a result, it will fail to connect a poster to poster's addressee whose email is not yet in the reference chain. This situation can arise in one of two ways: (1) when it is a first posting of a new thread or (2) when an addressee has not posted anything to an existing thread. Since all of the names extracted for building the name network were manually inspected for accuracy in the study, it is fair to use these names as actual addressees of postings or people who are somehow connected to the poster. Using an automated script, the number

of instances was counted for each of the two situations described above. The counts revealed some pleasantly unexpected results (See Table 5.2). On average, the 'chain network' method missed about 33 per cent of the potentially important connections as compared to the 'name network' method. Of the missed connections, about 70 per cent (or 23 per cent from the total count) came from postings that were the thread starters (Column A) and about 30 per cent (or 10 per cent from the total count) came from subsequent messages in a particular thread (Column B). What stands out about this finding is that the majority of missed addressees (70 per cent) were found in the thread starting messages. This discovery is especially interesting given that most previous research on interactivity in online discussions only uses 'reply' messages to estimate class interactivity (see, for example, Bonnett et al., 2006). However, these findings suggest that thread starting postings should not be ignored and can also be very interactive in nature. This is because many of these types of postings tend to include references to discussions that occurred among the community members in different threads. Therefore, thread starting postings should also be seriously considered when estimating interactivity of online discussions. Since the 'name network' method is capable of capturing connections to other group members even in the thread starting messages, the method also might be a good model for studies on group interactivity. Future research is needed to confirm this observation.

Another 7 per cent of connections that were missed by the 'chain network' method (Column D) were connections that occurred when an actual addressee or a 'reference' person was the author of a previous posting in the thread, but not the most recent one. This happened because the 'chain network' method, as presented in this work,

connects a message sender to the most recent poster in the thread. Generally speaking, it is easy to revise the 'chain network' method to include these 7 per cent of missed connections. The revised method would connect a poster to all previous posters or to two most recent posters in the thread. However, this will likely also introduce more false-positive connections; thus, adversely affecting the overall accuracy of the 'chain network' method. This conclusion comes from examining the relationship between an actual addressee and his/her position in the reference chain (see Table 5.2). Specifically, when examining all cases where an addressee is in the reference chain (COLUMN C and D in Table 5.2), in 90 per cent of those cases, the addressee is the most recent poster in the reference chain (COLUMN C). Thus, if a person in the reference chain who is not the most recent poster were considered as an addressee of a posting, this would be right in only 10 per cent or less of the times.

To determine the exact nature of connections that were missed by the chain network, all postings that correspond to columns A and B in Table 5.2 were examined for all six classes. This content analysis, described next, helped address the third research question by revealing what types of social interactions the name network includes.

Situation 1: First Posting of a Thread. The semi-automated content analysis of postings using ICTA revealed that among the most commonly used names in the first posting of a new thread was the instructor's name. Specifically, instructor's name was used to

- Ask the instructor about something (e.g., "[Instructor's name] if you see this posting would you please clarify for us"),
- Ask peers to clarify something that the instructor said during the lectures (e.g., "I

remember [Instructor's name] asking us to email her with topics [...] I wonder if that is in replacement of our bb question?"), or

- Share information with classmates obtained from the instructor via some other personal communication such as email. (e.g., "I just got a reply from [Instructor's name], and she said that […]")

This type of postings, and the ties derived from them, is very important in the context of learning. This is because 'student-instructor' ties derived from these messages can be used to identify students who are repeatedly asking for instructor's help. For example, a high weight for a tie between a student and the instructor suggests that a student is uncertain about class content or procedure, and an indication that extra attention may be needed from the instructor. However, if many students are connected to the instructor via these types of messages, then it may indicate that lectures or other class materials are unclear to not just one student and thus either the materials or a delivery method might need to be reconsidered by the instructor.

Another common category of messages involves an instructor mentioning a student. These were usually announcements from the instructor containing names of students responsible for leading a class discussion. For example, "Dan, [...] Since you have studied [Topic], would you get our discussion going on the forum for this week". Sometimes an instructor praises a student for some good work in the class. This suggests that if there is a tie from an instructor to a student based on this kind of postings, it is very likely that this student is doing well in the class. Identifying reliable and successful students in a class may be of importance for an instructor or for school's administration, especially when formal grading information is unavailable. For example, an instructor

can use such information to assign students into more effective groups, e.g., depending on pedagogic need, they might assign at least one student who is doing above average in the class to a group, or they might assign the high achievers to a group of their own.

Another common type of message in this category involves an instructor listing groups with their individual members for smaller group discussions. After examining these postings, I concluded that the ties derived from them do not necessarily reflect relationships between the instructor and a student. Instead, these postings can be used to automatically identify students who were assigned to work together, thus potentially creating 'work' ties. 'Work' ties are especially important for studying online groups since they are often precursors of even closer ties between online participants (See, for example, Haythornthwaite, 2002). This was confirmed by several students in the comment section of the online survey. They viewed the breakdown into smaller groups during live sessions as a good way to get to know their peers.

The last category of messages involved a student mentioning other student(s). In these cases, the poster often took a leadership role in a group, for example, by summarizing other group members' postings or assigning roles for a project as demonstrated in the following excerpt:

> "*Some quick poking around shows that Steve and myself are here in Champaign, [...] and Nicole is in Chicago. [...] does anyone have a strong desire to be our contact person to the administrators*"

This type of messages is useful in identifying active group members and group leaders and would be very useful when studying collaborative learning. However, a lot of messages like this from the same person may be perceived negatively by other group members. For example, in a related study, when analyzing a large collection of Usenet

newsgroup messages, Fiore et al. (2002) found that online participants who dominated the conversations were often viewed unfavorably. Nevertheless, a more detailed analysis is needed to study the influence of this type of connections in the online learning environment.

Situation 2: Subsequent Posting in a Thread. The detailed examination of subsequent postings revealed three main types of references/relations:

- A reference to an event or interaction that happened outside the bulleting board (e.g., "Dan and I have been corresponding via e-mail and he reminded me that we should be having discussion here"). This type of messages is likely to connect people who work together. It is also suggestive of stronger personal ties. This is because according to the idea of media multiplexity, stronger ties tend to communicate via more communication channels (See, for example, Haythornthwaite & Wellman, 1998; Haythornthwaite, 2001).
- A reference to someone as part of a group when providing a feedback to the whole group or posting on behalf of the whole group and signing the names of all group members (e.g., "Angela and Natasha, I couldn't wait to see your site. I knew it was going to [be] awesome!"). This is another type of messages that will likely indicate 'work'-related ties.
- A reference to somebody who presented or posted something a while ago or via different communication channel (e.g., "[…] it made me think of the faceted catalogs' display that Susan posted"). These postings appear to be useful for identifying 'learning' ties. This is because they show that a poster was not just commenting on the previous post, but rather on something that was said a while

ago. This means that the poster was following the class discussion, and a student mentioned in the posting made some significant contribution to the discussion that resonated with the current poster. All these activities can be categorized as evidence of learning.

In sum, based on the discussion in this section, the name network is shown to be well adept at detecting three of the social relations that are considered by many researchers to be crucial in shared knowledge construction and community building: 'help', 'work' and 'learning'.

5.2.2 Connections Missed by the Name Network

The previous section summarized common types of connections that were missed by the chain network. However, because on average only 25 per cent of all postings include personal names, it is important to also check what types of connections were missed by the name network. For this analysis, I randomly selected one of the bulletin boards in Class #1.This sample set consisted of 71 postings. Out of 71, 43 postings did not mention any personal names. The manual content analysis of these 43 postings revealed that the majority of the postings (31 postings, 72 per cent) did not address any particular person in the class, but rather addressed the class as whole. Usually, these messages expressed the opinion of the author about a matter that is being discussed and/or attempted to summarize what has been said in the forum without referring to any particular person or posting in the class. In some cases, a student posted a link that he or she believed to be relevant to the class discussion/topic. The results described above suggest that it is possible to ignore messages without names since these messages, for the most part, do not address any particular person in the class. An alternative view is to

develop a hybrid approach of using the chain network and the name network to complement each other. To make a final recommendation, this issue will be a subject of a post-dissertation research. Further research might also explore how the balance between these more general postings and the name-using postings affect perceptions of class functioning.

5.3 NAME NETWORKS VERSUS SELF-REPORTED NETWORKS

This section covers the second half of the second research question:

Question 2.2: How is the proposed name network associated with the self-reported network?

In order to answer this question, results from the chain network and the name network were compared with results from the self-reported network to determine which of the two derived networks is a better approximation of the self-reported social network (if any). Until now, the only reliable way to collect perceived data has been through surveys that demand both the time of the researcher and of the participants. Therefore, it would be a methodological breakthrough if an automated method for mimicking perceived social networks were devised.

For this analysis, pair wise comparisons of the three types of networks were conducted using statistical network models and specifically Exponential Random Graph models (p* models; Robins, in press, 2007). To build p* models, I used XPNET software (Wang et al., 2006). There are a few important reasons why p* models were selected to conduct this comparison and not other statistical models or QAP correlations. First, since some students did not participate in the survey, some possible ties were probably missing in the self-reported networks. As a result, QAP correlations would likely produce

inadequately lower results. Second, parameters estimated by a p* model are easy to interpret and compare across different pairs of networks. Finally, a p* model is the only statistical model that is capable of modeling different network structures as well as individual characteristics of the group members (Snijders, 2008).

Using p* models, for each class I estimated the parameter EdgeAB for a pair of the chain network and the self-reported network first and then for a pair of the name network and the self-reported network. The parameter EdgeAB indicates the likelihood of two networks sharing ties not by a chance alone. The results are shown in Table 5.3. The model was converged (t–statistics<0.1 for all estimated parameters) and the model was found to be significant (the goodness of fit for EdgeAB was less than 0.1 and between 1 and 3 for all other parameters) for all classes, except the case of a pair of the name and self-reported networks for Class #6.

The results show that for four out of six classes, the name network is consistently more likely to share ties with the self-reported network than the chain network (more than just by a chance alone). This supports my general expectation that the name network is more reflective of students' perceived relationships. However, for two smaller classes, Class #5 and Class #6, the name network was less likely to match the self-reported network than the chain network. (For Class #6, the model was not significant.) This was a very puzzling but intriguing result. It led to a separate investigation which is described in subsequent sections. The results of this investigation, as laid out in Section 5.3.1, provided some plausible explanations for the underperformance of the 'name network' method and suggested some concrete steps on how to further improve the 'name network' method in the future.

To find out why the name networks for Class #5 and Class #6 were less likely to share ties with the self-reported networks than the chain networks, I analyzed the network signatures for two students, Nick and Anna, from Class #5. These two students were selected because their network signatures were the most different in each of the two types of networks. One student, Nick, had several ties in the self-reported network that were missing in the name network. The second student, Anna, had a couple of ties in the name network that were missing in the self-reported network. For these two students, I examined all of their ties that exist in the self-reported network but not in the name network and vice versa. One of the main goals of this analysis was to identify what caused the 'name network' method to miss some self-reported ties and to include some ties that are not in the self-reported network. Furthermore, the analysis helped to identify any additional clues from the content of postings that can be used to improve the 'name network' extractor. For this examination, I used ICTA.

5.3.1 Why Did the Class #5 Name Network Miss Some Self-Reported Ties?

A student named Nick from Class #5 was selected by seven other students in the self-reported survey, but strangely in the name network, Nick was not connected to any of these seven individuals. After a brief investigation, it was determined that Nick only posted three messages to the bulletin board for the whole semester. There was simply not enough evidence on the bulletin board for the name network to discover ties to other individuals. So, on the surface, it is not clear what the basis was for these seven nominations from his fellow students. A posting from the instructor can shed some light on this mystery. The instructor mentioned Nick on the bulletin board once, when assigning students into smaller discussion groups for the chat sessions. It turned out that

the other two students who were assigned to work with Nick were among those who nominated Nick in the survey. This suggested an important future improvement to the 'name network' method. In addition, to connecting a poster with all people who are mentioned in the body of his or her posting, the 'name network' method should also connect any people whose names co-occur in close proximity in the same messages. With such a modification, Nick would gain two more additional ties in the name network to the two students who nominated him in the survey. As a proof of the concept, I re-built the name network for this class using co-occurrence of names in the text as an additional indicator of personal ties and re-run the comparison analysis between the name network and the self-reported network for Class #5. This time the likelihood of sharing ties between these two networks increased from 0.96 to 1.50 (t-statistics = 0.067) which is higher than the corresponding value from the chain network (1.03). This result led to adding an additional option for building the name network in ICTA. This new option allows a researcher to select whether or not to connect people in the name network based on the co-occurrence of their names in the text.

The instructor's message also suggested that the transcripts of chat conversations may contain additional evidence in support of ties disclosed in the survey. Having access to all class-wide chat transcripts, I imported them into ICTA and examined the use of the name 'Nick' in the chat messages. I discovered that all seven students who selected Nick in the survey were either mentioned by Nick, mentioned in the same context with Nick (worked together) and/or they mentioned Nick in the chat conversations. For example, "Nick and Phil, I'm thinking that the framing tool might be useful for your projects". This anecdotal evidence suggests that the analysis of any additional communication media

used in a class will improve the discovery of social networks within a class and will better reflect self-reported social networks. This observation is in line with the previous research on media multiplexity which asserts that people with stronger ties tend to use more media than those with weaker ties (See, for example, Haythornthwaite & Wellman, 1998; Haythornthwaite, 2001).

The approach of examining multiple data sources to discover social networks is not new and has been employed by social scientists for decades; for example, researchers often combine different data collection instruments such as interviews and on-site observations to improve the reliability of their social network data. But this approach has gained more popularity in the last decade due to the wide use of Internet by different groups and communities as their primary communication and publishing media. For instance, Stefanone and Gay (2008) relied on both email networks and forum networks to study social interactions of undergraduate students. Matsuo et al. (2006) used self-declared FOAF (Friend-of-a-Friend) networks, web-mined collaborator networks, and face-to-face meeting networks to build Polyphonet, a community support system for two different conferences. Aleman-Meza et al. (2006) used two social networks, FOAF (Friend-of-a-Friend) extracted from pages on the Semantic Web and a co-authorship network of authors from the DBLP Computer Science Bibliography[24] to determine the degree of Conflict of Interest among potential reviewers and authors of scientific publications. It is also important to note that although there seems to be an increasing interest in this approach, it is not always feasible or possible to collect social network data from multiple sources for any particular online communities. Some groups may only

[24] DBLP Computer Science Bibliography - http://www.sigmod.org/dblp/db

use one channel of communication. Also, there is still the ongoing research question of how to combine evidence of social relationships from different types of data.

In my post-thesis analysis, I am planning to devise and evaluate a method for collecting and combining evidence from both bulletin boards and chat transcripts to build the name network.

5.3.2 Why Did the Class #5 Name Network Include Some Ties That Were Not in the Self-Reported Network?

Anna is a well connected student in the self-reported network. However, she only had three strong ties in the name network. For the purposes of this section, I only focus on two ties from the name network that are missing in the self-reported network. (The third tie was reported in the self reported network and thus is not relevant to this part of the discussion.) The two ties in questions are with fellow students Rick and Mark.

The tie between Anna and Rick resulted from Rick posting three different messages to Anna thanking her for "insights", "thoughtful comments", and "all the wonderful posts and information". However, surprisingly there was no tie between these two students in the self-reported network. After a detailed investigation, it turned out that Rick did select Anna in the survey as a person who influenced his learning and helped the most in the class. (Rick was not nominated by Anna.) But because all ties with a weight less than 3 were removed (See Section 4.2), a tie of 2 between Anna and Rick also disappeared. As an experiment, I built a 'learning' network based on the students' responses to only one of the question in the survey about 'learning'. In this learning network, there was a tie between Anna and Rick. Next I compared this 'learning' network with the original name network (without using co-occurrences). The resulting likelihood

has slightly increased from 0.96 to 1.17 (t-statistics = -0.062). This suggests that the name network was a bit more similar to the 'learning' network than to the overall self-reported network for this particular class. Therefore, the continuation of this study will be to compare the name network with each type of the self-reported networks to determine if the name network is better in predicting 'learning' ties than others. However, it is possible that for some other class, depending on the prevalence of one type of interactions over the other, the name network can better reflect other types of self-reported networks such as 'friendship' or 'work' networks. Therefore, as a future improvement, the 'name network' method should be able to not just discover ties but also categorize them into different relations. This can be done by using information about roles of participants (e.g., student, guest speaker, instructor, etc), a position of a message in the thread as suggested in Section 5.2.1, and/or the context words where particular names are mentioned in a posting. For example, words like "thank you", "help", "assistance" may indicate that a student helping another student, thus they are connected via the 'help' relation. With such an algorithm in hand, it will be possible to build the name network that reflects only 'help' relations, only 'learning' relations, only friendship or some other relation that is important to members of a certain online community.

The discussion above also suggests that attention needs to be paid to the weight selected for the tie cutoff. This may vary by sample, relation, discussion, etc. Therefore, to make ICTA as flexible as possible, it was decided to keep all ties discovered by the 'name network' algorithm regardless of the weight, and allow a researcher to set a threshold for the tie cutoff in the final visualization.

The tie between Anna and Mark resulted from Mark posting two messages with Anna's name in them. The first posting from Mark was a question directed at Anna, "Anna -- what did you mean by [word] in paragraph 3 of your reply?" The second message was a thank you message from Mark to Anna for posting an interesting article to the bulletin board. (There were no messages from Anna mentioning Mark's name.) But regardless, this may be enough to suggest a tie between Mark and Anna. Unfortunately, because Mark did not participate in the survey, the self-reported network did not include a tie between them. In such case, a researcher can rely on tools like ICTA to conduct a semi-automated content analysis of messages to make the final decision about the accuracy of the 'name network' method.

5.3.3 Accounting for Agreement and Disagreement in the Postings

This section describes another possible reason for why the name network sometimes includes 'false-positive' ties (ties that are not part of the self-reported network). It should be noted that for practical and ethical reasons the questionnaire used to collect the self-reported networks avoided asking about negative ties and concentrated on neutral to positive types of relationships like friendship. As a result, by definition, the self-reported network does not include any 'dislike' or 'negative' types of ties. However, the 'name network' method does not have any predisposition towards filtering or ignoring negative ties. Negative ties can be inferred from postings that are considered to be confrontational in tone; for example, when one student disagrees with another on some issue. And the 'name network' method does not distinguish between negative, neutral or positive types of postings. For this reason, it is possible that the 'name network' method will include negative ties while the self-reported network will ignore those same ties

entirely.

To verify this supposition, I conducted a manual content analysis of postings from one of the 6 classes, Class #5. Specifically, I wanted to find out whether the expression of agreement in a posting might lead to a declaration of a tie between a poster and an addressee in a self-reported network. Or conversely, the expression of disagreement might lead to no declaration of a tie between a poster and an addressee in a self-reported network. The analysis started with the examination of all 156 instances from 125 postings when a personal name was used. First, I identified the tone of each posting, specifically the level of poster's agreement or disagreement with an addressee of the posting. The 'agreement'-type of postings included phrases like "Rick, I completely agree with you", "love your idea", "great post", "impressed with Ann's posting", and some examples of the 'disagreement'-type of postings are "Stephen, yes - but what I was referring to […] " or "Mark, […] I would take issue with this because I think […]". Second, for all students who participated in the survey, I counted (1) how many times a name was used to intensify agreement with an addressee when there was also a tie between the poster and the addressee in the self-reported network and (2) how many times a name was used to express disagreement with an addressee and there was not a tie between the poster and this addressee in the self-reported network.

The results are presented in Table 5.4. As it turned out, in Class #5, personal names were used more often to intensify agreement than disagreement. This suggests the supportive and non-confrontational nature of this class. In another study of eight LIS online classes, Haythornthwaite and Gruzd (2007) found similar results showing that 5 to 12 per cent of messages expressed agreement, compared to less than 1 per cent

expressing disagreement. In another study of online collaborative learning, Nguyen and Kellogg (2005) also found that naming occurred more often in expressions of agreement than disagreement. As an explanation for the prevalent use of personal names within agreement postings, Savignon and Roithmeier (2004) proposed that their use helps to sustain the collaboration, achieve so-called neutral footing and to avoid "strong declarations of fact".

Next, I examined whether there was a relationship between the tone of a posting and the existence of a self-reported tie between a poster and an addressee. Of the 49 ties discovered by the name network, 21 came from postings where a name was mentioned at least once in agreement, 3 came from postings that expressed disagreement, and the remaining 24 postings were neutral in tone. Of the 21 ties that expressed agreement, 16 of those ties were also found in the self-reported network. As for the 3 ties that expressed disagreement and discovered by the name network, none of them were found in the self-reported network. This observation coupled with the previous research by others in this area (described below) suggests that the 'name network' method may be better at predicting self-reported ties from postings that express agreement than disagreement. This phenomenon can be explained by a simple fact that positive messages tend to leave positive impressions on the reader/addressee of the message and negative messages tend to leave negative impressions. This is inline with the empirical findings by Mabry (1997) who also discovered that "[m]essages seeking positive or negative coalescence on an issue were significantly related to a message's perceived emotional tone" (n.p.). Following this argument, it is expected that any name network (as described in this work) built based on the dataset with a lot of confrontational messages will not resemble

perceived social networks unless these perceived social networks seek out representation of 'dislike' relationships. Further research with datasets where both agreement and disagreement postings are common is needed to reach a more definitive conclusion.

Although the number of messages expressing disagreement was very low in Class #5 and does not appear to strongly influence the 'name network' method, the literature review shows that confrontational messages are not that uncommon in online learning forums. For instance, Williams and Humphrey (2007) studied interactivity in threaded discussions of seven graduate-level courses in the Teaching English as a Second Language program. In some cases Williams and Humphrey (2007) observed that names were used to increase confrontation. This usually happened when names were used as intensifiers with so-called FTAs (face-threatening speech acts) such as "disagreement or dissatisfaction with a previous posting". However, Williams and Humphrey also did further suggest that "FTAs are important for meaningful class discussions, because it is partly through disagreement and its resolution that meaning is constructed" (p.139).

In conclusion, the proportion of postings expressing agreement versus disagreement depends on a variety of factors such as the polarizing nature of a discussion topic, individual beliefs, group's willingness to achieve a common goal, forum's policies, the performance of a moderator to mitigate a conflict and so on. There is also a possibility that a group that was initially in complete agreement may develop conflicting views on some topic later. As stated by Mabry (1997), "[d]ialogues of all sorts often turn from platforms for agreement to the exchanging of claims (contentions) and counterclaims" (n.p.). Therefore, in the future post-thesis research, to even further improve the 'name network' method, it will be helpful to recognize situations when a

name is used to express friendliness/agreement versus confrontation/disagreement with an addressee. This can be done in an automated fashion by using sentiment analysis. Sentiment analysis is the field of study that deals with automated techniques for identifying emotional polarity of text. (The recent review of the state-of-the-art techniques in the sentiment analysis is published that of Pang & Lee, 2008). A simple approach could be developed to first determine a poster's attitude toward an addressee of the posting and then only select neutral or friendly postings to build the name network. On the other hand, it may be useful to know if a particular topic is becoming controversial which can be a signal to an instructor to preemptively intervene and to manage a conflict before it escalates.

5.4 SUMMARY

This chapter described a comprehensive analysis of the 'name network' method using both quantitative and qualitative approaches. The analysis revealed that there are clear and critical differences between the social networks discovered by the 'name network' method versus those discovered by the 'chain network' method. The analysis also revealed that the name network tends to better resemble self-reported social interactions in the class than the chain network. Finally, the analysis of postings used to derive the name network suggested a number of important improvements for the 'name network' method. The next and last chapter summarizes and presents these and other important findings and suggestions for future research that came out of this study.

5.5 TABLES

Table 5.1: QAP correlations between pairs of the name and chain networks for six online classes

	# of Students	Chain Network Density	Name Network Density	QAP correlations*
Class #1	28	0.23	0.13	0.50
Class #2	20	0.48	0.35	0.51
Class #3	25	0.48	0.28	0.58
Class #4	21	0.08	0.1	0.45
Class #5	19	0.22	0.15	0.53
Class #6	15	0.39	0.17	0.69

* The number of random permutations used for the analysis was 5,000

Table 5.2: The relationship between an actual addressee and his/her position in the reference chain

Class	# of all postings*	# of found instances of named addressees	# of times an addressee is NOT in the reference chain when found in …		# of times when an addressee is IN the reference chain as …	
			a first posting of a new thread	a subsequent posting in a thread	the most recent poster	other
			COLUMN A	COLUMN B	COLUMN C	COLUMN D
Class #1	608	149	50	11	81	7
Class #2	855	271	59	30	153	29
Class #3	1,502	306	37	21	232	16
Class #4	164	96	17	16	51	12
Class #5	412	156	46	26	76	8
Class #6	497	107	27	4	73	3
Average (%)		100%	**23%**	**10%**	60%	**7%**

* On average, about 25% of all postings included personal names

Table 5.3: EdgeAB - the likelihood of two networks to share ties not by a chance alone

Class	Chain* & Self-Reported Networks		Name* & Self-Reported Networks	
	Estimated parameter EdgeAB	t-statistics	Estimated parameter EdgeAB	t-statistics
Class #1	0.81	0.075	1.73	-0.085
Class #2	0.99	0.044	1.52	0.031
Class #3	1.17	-0.057	1.31	0.001
Class #4	0.61	-0.007	1.11	0.064
Class #5	1.03	-0.004	0.96	-0.071
Class #6	1.33	0.053	0.82	Not significant

* Because self-reported networks likely include only strong ties (Bernard et al. 1981), all weak ties (with weights less than 2) were removed from all chain and name networks (except those for Class #4 due to its low network density). Following the requirements of XPNET, both chain and name networks were then binarized, a process where all weights of existing ties were set to 1. Finally, all networks were symmetrized using the following procedure: if there is a connection between one student to another, then it was assumed that for strong ties there is also a connection in the opposite direction.

Table 5.4: A relationship between a tone of a posting and the existence of a self-reported tie

	In the self-reported network	Not in the self-reported network
Agreement	16	5
Disagreement	0	3

CHAPTER 6: CONCLUSIONS AND FUTURE RESEARCH

6.1 CONCLUSIONS

The ‘name network’ method as proposed and evaluated in this work provides one more option for understanding and extracting social networks from online discussion boards, and it is a viable alternative to costly and time-consuming collection of users’ data on self-reported networks.

For the cases studied, the name network provided on average 40 per cent more information about social ties in a group as compared to the chain network. This additional information is available mostly because the name network can account for instances when a poster addresses or references somebody who has not previously posted to a particular thread. The ability of the name network to find message addressees based on the content of the message makes it possible to also discover social networks from Internet data where message addressees can not be determined from the information in the message header. Some examples of such data include chats, blogs, news stories, Youtube comments.

Furthermore, there is evidence that the name network provides a better reflection of self-reported ties than the chain network. This is primarily because the name network is well adept at detecting three of the social relations that are considered by many researchers to be crucial in shared knowledge construction and community building: ‘learning’, ‘work’ and ‘help’. For example, ‘learning’ relations were often discovered by the ‘name network’ method in postings that refer to somebody else who has presented or posted something earlier. By referencing ideas or comments from earlier postings the poster demonstrates that he or she was following the class discussion and learning with

and from others. ‘Work’ relations were commonly revealed by the ‘name network’ method through postings that refer to an interaction that happened outside the bulleting board or postings that mention members of the same study group in the same context. Finally, ‘help’ relations were often discovered in postings from students asking for the instructor’s help or postings that mention classmate’s name in the context of words like “thank you”, “help”, “assistance” indicating that a student is helping another student. These characteristics of the name network make the method a useful diagnostic tool for educators to evaluate and improve teaching models from the student’s perspective.

Some specific examples of how ties in the name network can be interpreted and used in the assessment of e-learning are listed below.

(1) To identify students who might need extra attention from the instructor, we can look for students with more connections (abnormally higher tie weight) to the instructor (especially those with connections that were derived from the thread starting postings). Or if we discover that many different students are connected to the instructor over a short span of time, this may indicate that lectures or other class materials were unclear. This can serve as a signal to the instructor to either cover the materials again, adjust the materials or change the delivery method.

(2) To find students who are doing well in the class and may be good candidates to provide help to their peers and aid in the learning process of their fellow students, we can look for students who have connections going from the instructor to them. On the surface, it might not be obvious as to how one can use such connections to identify potential student-helpers. However, when you couple this with the knowledge that most instructors tend to reprimand bad students in private and praise good students in

public (over the bulletin board) such as complementing a student for some good work in the class, then it becomes clear why such connections are good identifiers of successful students in a class.

(3) To identify students who tend to or would likely to work together on projects, we can look for strong 'student' to 'student' ties. If two students are connected in the name network, it usually means that they are either already working together on or tend to positively view each other postings and/or class presentations. The instructor can use this information to find students who share same interests and to group them together for more successful group projects.

(4) To find active group members who often take a leadership role in a group, we can look for nodes that have many connections to other students, especially those connections that were derived from postings mentioning more than one student. These types of nodes tend to be initiated by students who have taken the initiative and organized their fellow students to complete a particular task or objective. The instructor can use this information about potential group leaders to select a contact person in a group or find a person who would be good in leading a class discussion or organizing other class-related activities.

The study also suggested the following improvements for ICTA that have already been implemented: (1) using co-occurrence of names as an additional indicator of possible ties between people, (2) allowing ICTA's users to select whether or not to include non-posters who are mentioned by group members into the name network, (3) providing ICTA's users with an option to set a cutoff value for the tie weight filter.

Furthermore, the results of this research also suggested at least three additional future improvements that might help to increase the accuracy of social network discovery using the 'name network' method. The suggested improvements include (1) identifying types of different relations based on the context words used in the postings, (2) using multiple data sources to increase confidence in the existence of dyadic relationships, and finally (3) using techniques from sentiment analysis to determine the level friendliness of the relationships.

In conclusion, the 'name network' method for social network discovery proposed in this dissertation can be used to transform even unstructured Internet data into social network data. With the social network data available, it is much easier to analyze and make judgments about social connections between community members. The 'name network' method can be used where more traditional methods for data collection on social networks such as surveys are too costly or not possible, or they can be used in conjunction with traditional methods.

6.2 LIMITATIONS OF THE METHOD

This section briefly describes three main limitations of the 'name network' method and suggests possible solutions.

First, the 'name network' method is more expensive computationally then the 'chain network' method. This limitation can be addressed by integrating the 'name network' method directly into a system used by an online community. This way, new postings can be processed on the fly as they come into the system, instead of analyzing all postings at once which takes longer. Also with the improvements in the hardware and the efficiency of text mining techniques, this limitation will be less of an issue in the

future.

The second limitation of the method is its implementation of the alias resolution method. Currently it uses an email address as a unique identifier of a participant. This approach assumes that each student uses only one email address to post messages to the bulletin board. This assumption is true for the datasets in the study. However, for other online groups this may not be the case. Thus, the alias resolution procedure in the 'name network' method needs to be modified to work properly with datasets from other domains. This can be accomplished by using one of the alternative alias resolution methods described in Section 2.3.2.

Finally, the 'name network' method relies only on postings that include personal names. However, based on the empirical evidence from Section 5.2, on average, only about 25 per cent of all postings contain personal names. This means that we are potentially missing a chance to discover possible social ties from the remaining 75 per cent of the postings. This may be especially problematic for small size datasets like Class #5. A possible solution to address this limitation is to rely on multiple data sources if available, as suggested in Section 5.3.1, and/or use the 'chain network' method to extract ties from postings that do not include personal names. The future research will address this question by conducting a more detailed content analysis of postings that do not include personal names to find out their functions in online discussions and decide how to use these types of message in deriving social networks of class participants if at all.

6.3 FUTURE RESEARCH

Many online classes are now using multiple types of electronic communication methods such as forums, chats, and wikis to carry on their discussions. It is important to

know how we can capture and combine network information from these various data streams to build a more comprehensive view of an online community. In my post-thesis research, I am planning to devise and evaluate a method for collecting and combining evidence from bulletin boards, chat transcripts and wikis to build the name network. Some of the challenges here include matching names that people use across different communication mediums. For example, a system needs to know that 'AnneT' on the bulletin boards is the same person as 'Anne2' in the chat room and 'Anne Tolkin' on a wiki page. Another challenge is how to use pages on a class wiki, which are essentially web pages, to discover social ties between people. This is because in their raw form, wiki pages or web pages in general provide very little information about dyadic relationships. For example, discovering that two people are mentioned on the same web page as attending the same presentation is not sufficient, on its own, to make judgments about their social relationship. Nevertheless, with the proper text mining techniques, web pages may still reveal explicit or implicit declarations of relationships between two or more people. However, from a programming point of view, the latter is a more challenging task compared to analyzing threaded discussions, because the majority of web pages is essentially unstructured text that requires a lot more automated processing to discover relational declarations.

Another future direction involves applying the 'name network' method to datasets in other domains. So far the 'name network' method has only been tested in the e-learning environment. The future research will include the testing of the 'name network' method with data generated by other types of online communities such as health support groups, communities of political bloggers, emerging communities of the creators and

viewers of Youtube videos (a popular video sharing site) and rapidly growing and diverse communities in the Second Life (a 3-dimensional virtual world). As mentioned earlier, to be effective in datasets of other domains, the 'name network' method will require additional modifications to the name alias algorithm. For example, currently the 'name network' method resolves name aliases by assuming that each group member uses only one unique email address to post messages to the forum. However, this is not always the case in online forums. A participant may post messages using different email addresses. Another challenge is that in chat transcripts, there is no information about users' email addresses.

In the future research, I also plan to even further improve the 'name network' method. For instance, one way to make the name network to more closely resemble perceived social interactions is to ignore postings that express confrontation or disagreement. This is because an initial data analysis suggests that ties derived from postings expressing confrontation/disagreement are less likely to be reported by people. Another way to improve the 'name network' method is not only to identify automatically that two people are connected, but also to find out automatically how they are connected, what types of social relations they share, and what roles they have in a group. For example, while the 'name network' method provides data that two people are connected, it does not reveal the nature of their relationships automatically. Their relationships may be a friendship, but it may equally be strictly formal (e.g., supervisor-subordinate). The future research will need to rely on additional techniques for automated role and relationship identification. For example, a method can closely examine the context in which personal names are mentioned in the text. For example, if two names appear in the

same sentence with a word like "committee", then an assumption can be made that these two people are members of the same organization, oriented to the same responsibilities. Some initial work in this direction has been done by Matsuo et al. (2007) and Mori et al. (2005) on web pages and by Diehl et al. (2007), Carvalho et al. (2007) and McCallum et al. (2005) on email datasets.

In conclusion, this is an exciting, new area of research. More and more web applications[25] (often called *social web apps*) are using information about users' personal networks to help users find more relevant information, share information with friends or make better decisions. As text mining techniques become more accessible, more web applications will take full advantage of online social network data. Industry led initiatives in this direction by Facebook[26] and Google[27] need to be noted due to their significance and reach. Each company has built a free web interface which gives web developers access to the personal networks information of their and their partners' users. This has resulted in an explosion of various useful social web apps. While most of this newly available data is already pre-organized into a network form, additional processing of the texts produced by these communities as proposed and described in this dissertation could reveal even more details about the nature of social ties between their members.

[25] An up-to-date list of popular social web applications that utilize information about online social networks - http://www.programmableweb.com/tag/social

[26] Facebook Developer's API - http://developers.facebook.com

[27] Google Open Social - http://code.google.com/apis/opensocial

REFERENCES

Adamic, L.A. and Adar, E. (2003). Friends and neighbors on the web. *Social Networks,* 25(3), 211-230.

Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Ding, L., Kolari, P., Sheth, A. P., Arpinar, I. B., Joshi, A., and Finin, T. (2006). Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection. *Proceedings of the 15th International Conference on World Wide Web* (pp. 407-416). New York, NY: ACM Press.

Alpers, G.W., Winzelberg, A.J., Classen, C., Roberts, H., Dev, P., Koopman, C. and Barr Taylor, C. (2005). Evaluation of Computerized Text Analysis in an Internet Breast Cancer Support Group. *Computers in Human Behavior,* 21(2), 361-376.

Ames, M. and Manguy, L. (2006). Photoarcs: A Tool for Creating and Sharing Photo-Narratives. *Proceedings of CHI '06 Extended Abstracts on Human Factors in Computing Systems* (pp. 466-471). New York, NY: ACM Press.

Bernard, H.R., Killworth, P. and Sailer, L. (1981). Summary of Research on Informant Accuracy in Network Data and the Reverse Small World Problem. *Connections,* 4(2), 11-25.

Bikel, D.M., Miller, S., Schwartz, R. and Weischedel, R. (1997). Nymble: A High-Performance Learning Name-Finder. *Proceedings of the 5th Conference on Applied Natural Language Processing* (pp. 194-201). Morristown, NJ: Association for Computational Linguistics.

Bird, C., Gourley, A., Devanbu, P., Gertz, M. and Swaminathan, A. (2006). Mining Email Social Networks. *Proceedings of the 2006 International Workshop on Mining Software Repositories* (pp 137-143). New York, NY: ACM Press.

Bollegala, D., Matsuo, Y. and Ishizuka, M. (2006). Extracting Key Phrases to Disambiguate Personal Names on the Web. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 223-234). Berlin: Springer.

Bonnett, C., Wildemuth, B.M. and Sonnenwald, D.H. (2006). Interactivity between Proteges and Scientists in an Electronic Mentoring Program. *Instructional Science,* 34, 21–61.

Campbell, C.S., Maglio, P.P., Cozzi, A. and Dom, B. (2003). Expertise Identification Using Email Communications. *Proceedings of the 12th International Conference on Information and Knowledge Management* (pp. 528-531). New York, NY: ACM Press.

Carvalho, V.R., Wu, W. and Cohen, W.W. (2007). Discovering Leadership Roles in Email Workgroups. *Proceedings of 4th Conference on Email and Anti-Spam* (Mountain View, CA). Retrieved October 30, 2008, from http://www.ceas.cc/2007/papers/paper-08.pdf

Chen, Z., Wenyin, L. and Zhang, F. (2002). A New Statistical Approach to Personal Name Extraction. *Proceedings of the 19th International Conference on Machine Learning* (pp. 67-74). San Francisco, CA: Morgan Kaufmann Publishers.

Chinchor, N. (1997). MUC-7 Named Entity Task Definition. *Proceedings of the 7th Message Understanding Conference.* Retrieved October 30, 2008, from http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html

Chklovski, T. and Pantel, P. (2004). Verbocean: Mining the Web for Fine-Grained Semantic Verb Relations. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Barcelona, Spain). Retrieved October 30, 2008, from http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Chklovski.pdf

Cho, H., Gay, G., Davidson, B. and Ingraffea, A. (2007). Social Networks, Communication Styles, and Learning Performance in a CSCL Community. *Computers & Education,* 49(2), 309-329.

Christen, P. (2006). Comparison of Personal Name Matching: Techniques and Practical Issues. *Proceedings of the Workshop on Mining Complex Data (MCD) held at the IEEE International Conference on Data Mining* (pp. 290-294). Washington, DC: IEEE Computer Society.

Corich, S., Kinshuk and Hunt, L.M. (2006). Measuring Critical Thinking within Discussion Forums Using a Computerised Content Analysis Tool. *Proceedings of the 5th International Conference on Networked Learning 2006* (Lancaster, UK). Retrieved October 30, 2008, from http://www.networkedlearningconference.org.uk/past/nlc2006/abstracts/pdfs/P07%20Corich.PDF

Culotta, A., Bekkerman, R. and McCallum, A. (2004). Extracting Social Networks and Contact Information from Email and the Web. *Proceedings of the 1st Conference on Email and Anti-Spam* (Mountain View, CA). Retrieved October 30, 2008, from http://www.ceas.cc/papers-2004/176.pdf

Culotta, A., Wick, M. and McCallum, A. (2007). First-Order Probabilistic Models for Coreference Resolution. *Proceedings of Human Language Technologies* (pp. 81-88). Rochester, NY: Association for Computational Linguistics.

Diehl, C., Namata, G. and Getoor, L. (2007). Relationship Identification for Social Network Discovery. *Proceedings of the 22nd National Conference on Artificial Intelligence* (pp. 546-552). Menlo Park, CA: AAAI Press.

Dillman, D.A. (2000). *Mail and Internet Surveys: The Tailored Design Method.* New York: Wiley.

Doherty, C. (2004). Naming the Trouble with Default Settings. *Proceedings of "SFL Ripples in the 21st Century" Australian Systemic Functional Linguistics Association Conference* (Brisbane, Australia). Retrieved October 30, 2008, from http://eprints.qut.edu.au/archive/00001084

Feitelson, D.G. (2004). On Identifying Name Equivalences in Digital Libraries. *Information Research,* 8(4), paper 192.

Fiore, A.T., Tiernan, S.L. and Smith, M.A. (2002). Observed Behavior and Perceived Value of Authors in Usenet Newsgroups: Bridging the Gap. *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves* (pp. 323-330). New York, NY: ACM Press.

Fleischman, M.B. and Hovy, E. (2004). Multi-Document Person Name Resolution. *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Reference Resolution Workshop.* Retrieved October 30, 2008, from http://www.mit.edu/~mbf/ACL_04.pdf

Hanks, P., Hardcastle, K. and Hodges, F. (2006). *Dictionary of First Names.* Oxford, UK: Oxford University Press.

Harada, M., Sato, S. and Kazama, K. (2004). Finding Authoritative People from the Web. *Proceedings of the Digital Libraries and the 2004 Joint ACM/IEEE Conference* (pp. 306-313). New York, NY: ACM Press.

Haythornthwaite, C. (1996). Social Network Analysis: An Approach and Technique for the Study of Information Exchange. *Library and Information Science Research,* 18(4), 323-342.

Haythornthwaite, C. (1998). A Social Network Study of the Growth of Community among Distance Learners. *Information Research,* 4(1), 4-1.

Haythornthwaite, C. (2001). Exploring Multiplexity: Social Network Structures In A Computer-Supported Distance Learning Class. *The Information Society,* 17(3), 211-226.

Haythornthwaite, C. (2002). Strong, Weak, and Latent Ties and the Impact of New Media. *The Information Society,* 18(5), 385-401.

Haythornthwaite, C. (2008). Learning Relations and Networks in Web-Based Communities *International Journal of Web Based Communities,* 4(2), 140-159.

Haythornthwaite, C. and Gruzd, A.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* (pp. 67-86). Springer.

Haythornthwaite, C. and Wellman, B. (1998). Work, Friendship, and Media Use for Information Exchange in a Networked Organization. *Journal of the American Society for Information Science,* 49(12), 1101-1114.

Heer, J., Card, S.K. and Landay, J.A. (2005). Prefuse: A Toolkit for Interactive Information Visualization. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 421-430), New York, NY: ACM Press.

Hölzer, R., Malin, B. and Sweeney, L. (2005). Email Alias Detection Using Social Network Analysis. *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 52-57), New York, NY: ACM Press.

Hsiung, P. (2004). *Alias Detection in Link Data Sets*. Master's thesis, Carnegie Mellon University, Pittsburgh, PA. Retrieved October 30, 2008, from: http://www-2.cs.cmu.edu/~hsiung/misc/masters.pdf

Hsiung, P., Moore, A., Neill, D. and Schneider, J. (2005). Alias Detection in Link Data Sets. *Proceedings of the 3rd International Workshop on Link Discovery at the International Conference on Intelligence Analysis* (pp. 52-57). New York, NY: ACM Press.

Kautz, H., Selman, B. and Shah, M. (1997). Referral Web: Combining Social Networks and Collaborative Filtering. *Communications of the ACM*, 40(3), 63-65.

Krivov, S., Williams, R. and Villa, F. (2007). GrOWL: A Tool for Visualization and Editing of OWL Ontologies. *Web Semantics,* 5(2), 54-57.

Klavans, J. and Kan, M.Y. (1998). Role of Verbs in Document Analysis. *Proceedings of the 17th International Conference on Computational Linguistics* (pp. 680 - 686). Morristown, NJ: Association for Computational Linguistics.

Kozima, H. and Furugori, T. (1993). Similarity between Words Computed by Spreading Activation on an English Dictionary. *Proceedings of the 6th Conference on European Chapter of the Association for Computational Linguistics* (pp. 232 - 239). Morristown, NJ: Association for Computational Linguistics.

Krackhardt, D. (1987). QAP Partialling as a Test of Spuriousness. *Social Networks,* 9(2), 171-186.

Lange, D.D., Agneessens, F. and Waege, H. (2004). Asking Social Network Questions:A Quality Assessment of Different Measures. *Metodološki Zvezki,* 1(2), 351-378.

Law, N. (2005). Assessing Learning Outcomes in CSCL Settings. *Proceedings of the 2005 Conference on Computer Support for Collaborative Learning: Learning 2005: the Next 10 Years!* (pp. 373- 377). International Society of the Learning Sciences.

Leech, G. (1999). The Distribution and Function of Vocatives in American and British English Conversation. In H. Hasselggård and S. Oksefjell (Eds.), *Out of Corpora: Studies in Honour of Stig Johansson*. Amsterdam/Atlanta, GA: Rodopi.

Leggatt, H. (2007, April 12). Spam Volume to Exceed Legitimate Emails in 2007. *BizReport : Email Marketing*. Retrieved October 30, 2008, from http://www.bizreport.com/2007/04/spam_volume_to_exceed_legitimate_emails_in_2007.html

Lin, H., Fan, W., Wallace, L. and Zhang, Z. (2007). An Empirical Study of Web-Based Knowledge Community Success. *Proceedings of the 40th Annual Hawaii International Conference on System Sciences* (pp. 178-188). Retrieved October 30, 2008, from http://doi.ieeecomputersociety.org/10.1109/HICSS.2007.65

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N. and Roukos, S. (2004). A Mention-Synchronous Coreference Resolution Algorithm Based on the Bell Tree. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (Article No. 135). Morristown, NJ: Association for Computational Linguistics.

Mabry, E.A. (1997). Framing Flames: The Structure of Argumentative Messages on the Net. *Journal of Computer Mediated Communication,* 2(4).

Maguitman, A.G., Menczer, F., Roinestad, H. and Vespignani, A. (2005). Algorithmic Detection of Semantic Similarity. *Proceedings of 14th International Conference on World Wide Web* (pp. 107-116). New York, NY: ACM Press.

Malin, B., Airoldi, E. and Carley, K.M. (2005). A Network Analysis Model for Disambiguation of Names in Lists. *Computational & Mathematical Organization Theory,* 11(2), 119-139.

Mann, G. and Yarowsky, D. (2003). Unsupervised Personal Name Disambiguation. *Proceedings of Conference on Computational Natural Language Learning* (pp. 33-40). Morristown, NJ: Association for Computational Linguistics.

Matsuo, Y., Hamasaki, M., Nakamura, Y., Nishimura, T., Hasida, K., Takeda, H., Mori, J., Bollegala, D. and Ishizuka, M. (2006). Spinning Multiple Social Networks for Semantic Web. *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, (pp. 1381-1387). Menlo Park, California: AAAI Press.

Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K. and Ishizuka, M. (2007). Polyphonet: An Advanced Social Network Extraction System from the Web. *Web Semantics: Science, Services and Agents on the World Wide Web,* 5(4), 262-278.

McArthur, R. and Bruza, P. (2003). Discovery of Social Networks and Knowledge in Social Networks by Analysis of Email Utterances. *Proceedings of ECSCW 03 Workshop on Moving From Analysis to Design: Social Networks in the CSCW Context* (Helsinki, Finland). Retrieved October 30, 2008, from http://www.ischool.washington.edu/mcdonald/ecscw03/papers/mcarthur-ecscw03-ws.pdf

McCallum, A., Corrada-Emanuel, A. and Wang, X. (2005). Topic and Role Discovery in Social Networks. *Proceedings of International Joint Conference on Artificial Intelligence.* Retrieved October 30, 2008, from http://www.cs.umass.edu/~mccallum/papers/art-ijcai05.pdf

Medlock, B. (2006). An Introduction to NLP-Based Textual Anonymisation. *Proceedings of the 5th International Conference on Language Resources and Evaluation* (Genoa, Italy). Retrieved October 30, 2008, from http://www.benmedlock.co.uk/anonLREC.pdf

Medynskiy, Y.E., Ducheneaut, N. and Farahat, A. (2006). Using Hybrid Networks for the Analysis of Online Software Development Communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 513-516). New York, NY: ACM Press.

Minkov, E., Wang, R.C. and Cohen, W.W. (2005). Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text. *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 443-450). Morristown, NJ: Association for Computational Linguistics.

Mori, J., Sugiyama, T. and Matsuo, Y. (2005). Real-World Oriented Information Sharing Using Social Networks. *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work* (pp. 81-84). New York, NY: ACM Press.

Nadeau, D., Turney, P. and Matwin, S. (2006). Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity, in J. G. Carbonell and J. Siekmann (Eds.), *Lecture Notes in Computer Science: Advances in Artificial Intelligence* (pp. 266-277). Berlin / Heidelberg: Springer.

Nguyen, H.T. and Kellogg, G. (2005). Emergent Identities in On-Line Discussions for Second Language Learning. *The Canadian Modern Language Review,* 62(1), 111-136.

Nurmela, K., Lehtinen, E. and Palonen, T. (1999). Evaluating CSCL Log Files by Social Network Analysis. *Proceedings of the 1999 Conference on Computer Support for Collaborative Learning*. Palo Alto, California: International Society of the Learning Sciences.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval,* 2(1-2), 1-135.

Patman, F. and Thompson, P. (2003). Names: A New Frontier in Text Mining. In H. Chen, R. Miranda, D. D. Zeng, C. Demchak, J. Schroeder, T. Madhusudan (Eds.), *Proceedings of the Intelligence and Security Informatics: First NSF/NIJ Symposium* (pp. 27-38). Berlin / Heidelberg: Springer.

Pedersen, T., Kulkarni, A., Angheluta, R., Kozareva, Z. and Solorio, T. (2006). An Unsupervised Language Independent Method of Name Discrimination Using Second Order Co-Occurrence Features. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 208-222). Berlin / Heidelberg: Springer.

Pennebaker, J.W. and Graybeal, A. (2001). Patterns of Natural Language Use: Disclosure, Personality, and Social Integration. *Current Directions in Psychological Science,* 10(3), 90-93.

Phan, X.-H., Nguyen, L.-M. and Horiguchi, S. (2006). Personal Name Resolution Crossover Documents by a Semantics-Based Approach. *IEICE Transactions on Information and Systems,* E89-D(2), 825-836.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial IntelligenceResearch,* 11, 95-130.

Reyes, P. and Tchounikine, P. (2005). Mining Learning Groups' Activities in Forum-Type Tools. *Proceedings of the 2005 Conference on Computer Support for Collaborative Learning: Learning 2005: the Next 10 Years!* (pp. 509-513). International Society of the Learning Sciences.

Reuther, P. and Walter, B. (2006). Survey on Test Collections and Techniques for Personal Name Matching. *International Journal of Metadata, Semantics and Ontologies,* 1(2), 89-99.

Robins, G. (In press). Exponential Random Graph (P*) Models for Social Networks. In R. Myers (Ed.), *Encyclopaedia of Complexity and System Science*, Berlin / Heidelberg: Springer.

Robins, G., Pattison, P., Kalish, Y. and Lusher, D. (2007). An Introduction to Exponential Random Graph (P*) Models for Social Networks. *Social Networks,* 29(2), 173-191.

Rourke, L., Anderson, T., Garrison, D. R. and Archer, W. (2001). Methodological Issues in the Content Analysis of Computer Conference Transcripts. *International Journal of Artificial Intelligence in Education,* 12, 8-22.

Savignon, S.J. and Roithmeier, W. (2004). Computer-Mediated Communication: Texts and Strategies. *Computer Assisted Language Instruction Consortium Journal,* 21(2), 265-290.

Sekine, S. and Nobata, C. (2004). Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *Proceedings of the Language Resources and Evaluation Conference* (Lisbon, Portugal). Retrieved October 30, 2008, from https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2004/LREC/pdf/65.pdf

Snijders, T.A.B. (2008). Models for Social Networks. *Course Lecture.* University of Oxford.

Soon, W.M., Ng, H.T. and Lim, D.C.Y. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics,* 27(4), 521-544.

Stefanone, M.A. and Gay, G. (2008). Structural Reproduction of Social Networks in Computer-Mediated Communication Forums. *Behaviour and Information Technology,* 27(2), 97-106.

Sweeney, L. (2004). Finding Lists of People on the Web. *Computer Science Technical Report CMU-CS-03-168*. Carnegie Mellon University, Pittsburg, PA. Retrieved October 30, 2008, from http://reports-archive.adm.cs.cmu.edu/anon/2003/CMU-CS-03-168.pdf

Ubon, A.N. (2005). *Social Presence in Asynchronous Text-Based Online Learning Communities: A Longitudinal Case Study Using Content Analysis*. Doctoral dissertation, Department of Computer Science, The University of York, UK.

Uzuner, O., Luo, Y. and Szolovits, P. (2007). Evaluating the State-of-the-Art in Automatic De-Identification. *Journal of the American Medical Informatics Association,* 14(5), 550-563.

Wang, P., Robins, G. and Pattison, P. (2006). Pnet: Program for the Estimation and Simulation of P* Exponential Random Graph Models, *User Manual*. Department of Psychology, University of Melbourne.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis*. Cambridge, MA: Cambridge University Press.

Wellman, B. (2001). Computer networks as social networks, *Science* 293(5537), 2031-2034.

Wellman, B. (1996). For a Social Network Analysis of Computer Networks: A Sociological Perspective on Collaborative Work and Virtual Community. *Proceedings of the 1996 ACM SIGCPR/SIGMIS Conference on Computer Personnel Research* (pp. 1-11). New York, NY: ACM Press.

White, H.D., Wellman, B. and Nazer, N. (2004). Does Citation Reflect Social Structure?: Longitudinal Evidence from the Globenet Interdisciplinary Research Group. *Journal of the American Society for Information Science and Technology,* 55(2), 111-126.

Williams, R.S. and Humphrey, R. (2007). Understanding and Fostering Interaction in Threaded Discussion. *Journal of Asynchronous Learning Networks,* 11(2).

Yang, J. and Hauptmann, A.G. (2004). Naming Every Individual in News Video Monologues. *Proceedings of the 12th Annual ACM International Conference on Multimedia* (pp. 580-587). New York, NY: ACM Press.

Yang, X., Su, J., Lang, J., Tan, C.L., Liu, T. and Li, S. (2008). An Entity-Mention Model for Coreference Resolution with Inductive Logic Programming. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 843-851). Morristown, NJ: Association for Computational Linguistics.

APPENDIX A: CONTEXT WORDS FOR NAME DISCOVERY

Context words left to a **sender**	'*', '-', '/', '~', 'cheers', 'thanks', 'thank you', 'yours', 'sincerely', 'regards', 'wishes', 'take care', 'see you soon'
Context words right to a **sender**	'*', 'ps', 'p.s.', 'ps.', 'ps,'
Context words left to an **addressee**	'i', 'hi', 'hello', 'dear', 'hey', 'replay', 'to', 'cc', 'from', 'as', 'with', 'agree', 'disagree', 'and', 'p.s.', 'ps', 'ps.', 'ps,'
Context words right to an **addressee**	'i', 'wrote', ',', 'pointed', 'said', 'and'
Context words left to any **name**	'hi', 'hello', 'dear', 'hey', 'prof', 'professor', 'ms', 'mr', 'mrs', 'dr', 'gov', 'sen', 'lt', 'col', 'cheers', 'thanks', 'yours', 'regards', 'wishes', 'care', 'soon'
Context words right to **non-name**	'street', 'ave', 'st', 'association', 'foundation', 'award', 'university', 'school', 'department', 'conference', 'drive', 'institute', 'system', 'college', 'method'

Note: '*' indicates a new line.

APPENDIX B: GENERAL ALERT LETTER FOR ONLINE CLASSES

General Alert Letter for LEEP classes
To: GSLIS Students in LEEP class
From: Caroline Haythornthwaite
Re: Alert re use of transcripts of class-wide online discussions

This message is to alert you to research being conducted by Caroline Haythornthwaite, Associate Professor in Graduate School of Library and Information Science, and to provide you with details about the study and your rights as a participant. This work is being conducted with Anatoliy Gruzd, a doctoral student at GSLIS, for whom this research comprises part of his doctoral dissertation work. In the future, other Research Assistants working with Professor Haythornthwaite may also be involved in this study.

In our research, we are looking at how students learn and interact online. This work is being undertaken to help us understand learner behaviors in online settings, and is part of an ongoing research program by Caroline Haythornthwaite that examines online learning. Results from this study will help faculty and administration to understand online learning processes and to design effective programs for future students.

The research examines the transcripts of the class chat (main room only, not including whispers or chat in other rooms), and postings to Moodle forums (those open to the whole class) in order to study how students communicate online. The only transcripts being examined are those that are already recorded as part of class records and which are public to the class as a whole.

In any reports or publications, all data will be anonymized. While looking at trends, something that has been written in the class chat room or in discussions may provide a good example to show to others. Should such examples be used in reports or publications, all names of class members will be changed to pseudonyms.

If you do not want any text of yours to be used as examples, please email, or write to Caroline Haythornthwaite (contact information below) and your online text will not be used in this way. Only information aggregated with use by others will be used. Only members of the research team will know that you have asked for your text not to be quoted. Please send such requests as soon as possible, and preferably by no later than one month after the end of the semester. Later requests will be honored as well, but research reports may already have been given or submitted which cannot be altered.

As part of this study, we are also interested in students' perceptions of class interaction. To gather that information, we will be asking you at a later date to complete a short online survey. We anticipate that this will take you no longer than 15 minutes to complete, and details will follow about this. In most cases we will ask you

to complete this once at the end of the semester. For some classes we will ask for cooperation in tracking perceptions of interaction over the semester, and would ask you to complete a survey three times during the semester. Again, details will follow about this.

Participation in any part of this study is voluntary. Risks associated with the research are very low, and are considered no greater than those of everyday life. Minimal risk is associated with the impact on reputation if text from class-wide LEEP discussions, or diagrams of class interaction, even with anonymization, reveal information that may be considered to affect an individual's reputation.

Results of this research will be disseminated in academic venues, including working papers, conference papers, journal publications, book chapters, and dissertation work in print or online.

Students may discontinue participation at any time, with no negative consequences. Your instructor will have no knowledge of your agreement to participate or your decision to withdraw. The decision to participate, decline, or withdraw from participation will also have no effect on your grades at, status at, or future relations with the University of Illinois.

If you have any other questions or concerns about this research, please contact Caroline Haythornthwaite (haythorn@uiuc.edu or 217-244-7453).

If you have any questions about your rights as a participant in this study, please contact the University of Illinois Institutional Review Board at 217-333-2670 (collect calls accepted if you identify yourself as a research participant) or via email at irb@uiuc.edu. The Institutional Review Board is the office at the University of Illinois responsible for protecting the rights of human subjects involved in studies conducted by University of Illinois researchers.

Please print a copy of this informed consent document for your records.

Thank you for your participation.

Caroline Haythornthwaite (haythorn@uiuc.edu)
Associate Professor, Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign, 501 East Daniel St., Champaign, IL, 61820

APPENDIX C: ONLINE QUESTIONNAIRE

SECTION 1: Please answer the first THREE questions for EACH student in the class. With [5] indicating a *more frequent association*, please indicate on a scale from [1] to [5]

1. HOW OFTEN YOU LEARNED SOMETHING NEW ABOUT THE CLASS SUBJECT MATTER FROM EACH STUDENT IN YOUR CLASS:

With [5] indicating a *more frequent association*, please indicate on a scale from [1] to [5],

2. HOW OFTEN YOU WORKED WITH EACH STUDENT IN THE CLASS:

With [5] indicating *a closer relationship*, please indicate on a scale from [1] to [5],

3. YOUR FRIENDSHIP RELATIONSHIP WITH EACH OF THE OTHER MEMBERS OF THE CLASS (considering both work for this class and for other classes or associations)

1 - don't know this person
2 - just another member of class
3 - a slight friendship
4 - a friend
5 - a close friend

SECTION 2: Please answer the remaining FOUR questions by SELECTING ONLY 5 TO 8 STUDENTS from the list.

4. From the list of all students, SELECT THE 5 TO 8 STUDENTS WHO HAVE BEEN MOST INFLUENTIAL IN YOUR LEARNING IN THIS CLASS.
5. From the list of all students, SELECT THE 5 TO 8 STUDENTS WHO HAVE BEEN MOST IMPORTANT IN PROMOTING DISCUSSION IN THIS CLASS.
6. From the list of all students, SELECT THE 5 TO 8 STUDENTS WHO MOST OFTEN GAVE YOU OR FELLOW CLASSMATES HELP WITH UNDERSTANDING A TOPIC OR ASSIGNMENT IN THE CLASS:
7. From the list of all students, SELECT THE 5 TO 8 STUDENTS WHO MOST OFTEN MADE CLASS FUN AND ENJOYABLE.

SECTION 3:

With [5] indicating *greater agreement* with the statement, please indicate on a scale of [1] to [5] how much you agree with each of the following statements about your class.

1 - never

2 - rarely

3 - for some of the course

4 - during most of the course

5 - throughout the whole course

I learned a great deal about the subject in this class

I felt that the class worked together throughout the class

I felt that class members welcomed new ideas

I felt that class members were supportive

I felt satisfied with class interaction

I enjoyed this class

LAST PAGE

Our research explores how interaction in online classes relates to student learning and experiences. If you have further comments about interaction in this class, please leave us a note here.

Thank you very much for your participation. Please click SUBMIT to complete the questionnaire.

AUTHOR'S BIOGRAPHY

Anatoliy Anatoliyovych Gruzd was born in Dnipropetrovsk, Ukraine, on November 25, 1980. During the course of his academic career, Anatoliy was immersed in a multidisciplinary program that had a strong emphasis on Information Technology and involved faculty from both the Computer Science and Library & Information Science (LIS) schools. After successfully completing a Bachelor's and Master's degree in Computer Science from Dnipropetrovsk National University in Ukraine, Anatoliy was awarded the US Department of States Edmund S. Muskie Graduate Fellowship to pursue a Master's in LIS in the United States. The Master's program from Syracuse University equipped him with the practical and theoretical knowledge in Library and Information Science. While the LIS doctoral program at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign taught him how to be a better researcher and allowed Anatoliy to apply his advanced knowledge of Computer Science to tackle important LIS related issues.