

Understanding and Augmenting Expertise Networks

by

Jun Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in The University of Michigan
2008

Doctoral Committee:

Associate Professor Mark Steven Ackerman, chair

Professor Atul Prakash

Assistant Professor Lada A. Adamic

Professor Volker Wulf, University of Siegen

Copyright Jun Zhang

All rights reserved

2008

ACKNOWLEDGEMENTS

This dissertation bears my name, but it could not possibly have happened without tremendous support and dedication from many individuals. Although I have space to thank only a few of these people here by name, every one of them should know that I am immensely grateful.

First, I thank all of my dissertation committee members. This work would not have been possible without the chair and my advisor Mark S. Ackerman. His support, guidance and feedback have been invaluable in carrying out the work described here and in learning more generally how to conduct rigorous research and have a successful career. Deepest gratitude is also due to Lada A. Adamic. She helped me in numerous ways and the close cooperation with her was a great experience. I am also extremely appreciative of the wise advice and feedback from Atul Prakash and Volker Wulf.

In this thesis, chapter 2 is based on the Group'05 conference paper co-authored with Mark S. Ackerman. Chapter 3 is based on the WWW'07 conference paper co-authored with Mark S. Ackerman and Lada A. Adamic. Chapter 4 is based on the working paper co-authored with Lada A. Adamic, Eytan Bakshy, and Mark S. Ackerman. I would like to thank all my co-authors for helping me with these papers.

In addition to my committee members and paper co-authors, George Furnas, Marshall Van Alstyne, Michael Cohen, Judy Olson, and Paul Resnick at SI, Alison Lee and Catalina Denise at IBM Research, have given me patient guidance and encouragement during various stages of my academic life.

I also wish to thank the support staff at the School of Information, particularly Sue Schuon, who has been constant and invaluable sources of support and assistance in

navigating the University bureaucracy. In addition, I thank all my friends and fellow graduate students for their valuable feedback and advice; and for making my time in Ann Arbor so enjoyable.

Lastly, and most importantly, I wish to thank my parents and my wife. To them I dedicate this thesis.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
LIST OF FIGURES	vi
LIST OF TABLES	viii
LIST OF APPENDICES	ix
ABSTRACT.....	x
CHAPTER 1 INTRODUCTION.....	1
EXPERTISE AND EXPERTISE SHARING.....	3
RESEARCH FOCUS -- EXPERTISE NETWORKS.....	5
RESEARCH FRAMEWORK.....	7
CHAPTER 2 SEARCHING FOR EXPERTISE IN SOCIAL NETWORKS	14
INTRODUCTION	14
SEARCHING IN SOCIAL NETWORKS.....	15
SIMULATION.....	19
DATA ANALYSIS.....	29
SENSITIVITY ANALYSIS	38
SUMMARY.....	41
CHAPTER 3 EXPERTISE NETWORKS IN ONLINE COMMUNITIES	47

INTRODUCTION	47
EXPERTISE NETWORK IN ONLINE COMMUNITIES	50
EMPIRICAL STUDY OF AN ONLINE COMMUNITY	52
EXPERTISE RANKING ALGORITHMS	57
EVALUATION.....	61
SIMULATIONS	67
SUMMARY AND FUTURE WORK	76
CHAPTER 4 EXAMINING KNOWLEDGE SHARING ON YAHOO ANSWERS.....	81
INTRODUCTION	81
PRIOR WORK.....	82
YAHOO ANSWERS AND DATASET	85
CHARACTERIZING YA CATEGORIES.....	87
EXPERTISE AND KNOWLEDGE ACROSS CATEGORIES.....	97
PREDICTING BEST ANSWERS.....	105
CONCLUSIONS.....	108
CHAPTER 5 CONCLUSION.....	112
SUMMARY.....	112
DESIGN IMPLICATIONS.....	114
FUTURE WORK.....	120
SUMMARY AND CONTRIBUTION	123
APPENDICES.....	125

LIST OF FIGURES

Figure 1-1: Method for converting a topic thread into a network.....	6
Figure 1-2: SWIM search and refer process	8
Figure 2-1: Enron Email Network	21
Figure 2-2: Cumulative Out-Degree Distribution of Enron email network.....	22
Figure 2-3: Percentage of succeed queries within different search length using various strategies	30
Figure 2-4: Distribution of Number of People used using various strategies.....	31
Figure 2-5: Distribution of depth of query chain using various strategies	33
Figure 2-6: Distribution of a user’s frequency of being used using various strategies.....	34
Figure 3-1: We map a replying relationship into a directed graph.	50
Figure 3-2: The web is a bow tie	54
Figure 3-3: The Java Forum network is an uneven bow tie.....	54
Figure 3-4: Degree distribution of the Java Forum network.....	55
Figure 3-5: The correlation profile of the Java Forum network. The color corresponds to the logarithm of the frequency of such degree pairings.....	57
Figure 3-6: The performance of various algorithms in different statistical metrics	64
Figure 3-7: Box plots of algorithm rankings vs. human ratings	66
Figure 3-8: Snapshot of the network simulator interface.....	68
Figure 3-9: Simulated degree distributions with ‘best preferred’ helpers	70
Figure 3-10: Simulated degree distributions with a growing network	70
Figure 3-11: Degree correlation profile of the “best preferred” network	71

Figure 3-12: Performance of expertise-detection algorithms on the ‘best preferred’ network	71
Figure 3-13: ‘best preferred’ network’	73
Figure 3-14: ‘just better’ network	73
Figure 3-15: Simulated degree distributions with ‘just better’ helpers.....	74
Figure 3-16: Correlation profile of the ‘just-better network’	74
Figure 3-17: Performance of expertise ranking algorithms in the ‘just better’ network ..	75
Figure 3-18: A case where a high expertise node has low authority	76
Figure 4-1 Post length vs. thread length	88
Figure 4-2 Clustering of categories by thread length and overlap between askers and repliers.....	90
Figure 4-3 Indegree distributions for different categories.	92
Figure 4-4 Outdegree distributions.	92
Figure 4-5: Sampled ego network of three selected categories:	93
Figure 4-6 Motif profiles of selected categories	95
Figure 4-7 Similarities between categories.....	96
Figure 4-8 Illustration of the hierarchical entropy calculation,	100
Figure 4-9 The distribution of entropy and percent best answer across users who had answered at least 40 questions	101
Figure 4-10 Relationships between focus and best.....	103
Figure 4-11 Answer length and best answer selection.....	107
Figure 5-1 The system structure of QuME	118
Figure 5-2 Screenshots of QuME interface	119
Figure 5-3 An interface for new users	120

LIST OF TABLES

Table 2-1 Evaluated Algorithms (* indicates algorithms that we proposed based on related works).....	24
Table 2-2 : The success rate of various algorithms.....	30
Table 3-1: Comparison of bow tie analysis between Web and the Java Forum network .	54
Table 3-2: Basic ExpertiseRank algorithm.....	60
Table 3-3: Five levels of expertise rating	62
Table 3-4: Bow tie structure of the ‘best preferred’ network	70
Table 4-1 Summary statistics for selected QA networks.....	94
Table 4-2 Correlating category entropy and % best answers	103
Table 4-3 Correlation between focus and score*.....	104
Table 4-4 Predicting the best answer	105

LIST OF APPENDICES

A. Literature Interview	126
B. Community Network Simulator.....	171

ABSTRACT

This thesis investigates large scale knowledge searching and sharing processes in online communities and organizations. It focuses on understanding the relationship between social networks and expertise sharing activities. The work explores design opportunities of these social networks to bootstrap knowledge sharing, by using the specific social characteristics of social networks which can lead to sizeable differences in the way expertise is searched and shared. The potential impact of this approach was examined in three related studies using data from Java Forum, Yahoo Answers, and Enron.

The Java Forum study investigated how people asked and answered questions in this online community using advanced social network analysis metrics. Furthermore, it explored algorithms that made use of the network structure to evaluate expertise levels. It also used simulations to explore possible social structures and dynamics that would affect the interaction patterns and network structure in online communities. The Yahoo Answers study extended the Java Forum study into a more general community setting and covered much more diverse knowledge sharing dynamics. It analyzed both content properties and social network interactions across sub-forums with different types of knowledge, as well as examined the range and depth of knowledge that users share across these sub-forums. The Enron study, on the other hand, investigated how social network structure could affect the expertise searching process in organizational communication networks using simulations and social network analysis. Based on findings in these studies, a novel expertise sharing system, QuME, was proposed and developed.

This thesis provides a network theoretical foundation for the analysis and design of knowledge sharing communities. It explores new opportunities and challenges that arise in online social interaction environments, which are becoming increasingly ubiquitous and important. This work also has direct implications for practitioners. The ability to add the level of expertise would be a major step forward for expertise finding systems, and would likely open up a range of new application possibilities.

CHAPTER 1

INTRODUCTION

Asking questions and seeking help from other people is one of the most traditional ways for people to solve real life problems. The inventions of paper, printing, and computer technologies allowed people to access large volumes of information easily and quickly. Especially with the recent development of web and information retrieval technologies, it would seem as though people could search and access any information on the Internet in a few seconds, as Bush (Bush 1945) predicted 60 years ago. However, in our personal lives, social networks are still one of the most frequently used channels in many situations when people need to search for information. The sought knowledge ranges from advice on medical treatments, programming, building a computer from scratch, to repairing the kitchen sink. By talking to a person with the needed expertise, one can reach other people's implicit knowledge and interactively clarify problems.

Developing systems to support people sharing expertise has been a research topic for at least 15 years. At first, it was largely studied within organizational settings. Systems that help find people with appropriate expertise are called expertise finders or expertise location engines. These have been explored in a series of CSCW studies (Ackerman et al. 2002). Newer systems, which use social networks to help find experts, have also been explored, most notably in Yenta (Fonder, 1997), Referral Web (Kautz et al., 1997), and most recently commercial systems from Tacit and Microsoft. These

systems attempt to leverage social networks within an organization or community to help find and reach the appropriate others. However, while many systems have been built, there is still a lack of understanding of the characteristics of the social networks through which knowledge is searched and shared, as well as how these characteristics can be exploited in the development of these systems.

Recently, Internet-scale expertise sharing has become a topic of considerable interest. Various online communities have been built to support people sharing their knowledge with others. For instance, The Sun Java Forum has thousands of Java developers coming to the site to ask and answer questions related to Java programming everyday. The Microsoft TechNet newsgroup is a major place for programmers to seek help for programming questions related to Microsoft products. Yahoo Answers had approximately 23 million resolved questions in 25 broad categories within two years of launch. In these communities, people help strangers voluntarily for various motivations, e.g. altruism, incentives to support one's community, reputation-enhancement benefits, expected reciprocity, contributors' sense of efficacy, and the most recently proposed "direct learning benefits" (Lakhani et al. 2003). Unfortunately, the very large size of these communities may impede an individual's ability to find relevant answers or advice. Which replies were written by experts and which by novices? As these help-seeking communities are also often primitive technically, they often cannot help the user distinguish between expert and novice advice. We would therefore like to find mechanisms to augment their functionality and social activities.

The ultimate goal of my research was to develop systems to augment expertise sharing activities in both local and online communities. Ackerman and Halverson (2003) suggested that systems to help people share expertise in their social networks must emphasize social aspects like "structural, shared cognitive, and relational dimensions" because they are the key dimensions to allow knowledge and expertise to be shared among people. Thus, in my dissertation, I focus on studying social network patterns in

organizational email networks and online communities, as well as how we can use them to develop mechanisms to better support expertise sharing activities.

The thesis is organized around three related studies that focus on different perspectives of the thesis topic. This introduction chapter presents the key research problems, why I care, and the “map” as to how these three studies address the problems.

EXPERTISE AND EXPERTISE SHARING

Expertise is defined differently in different disciplines. In the field of psychology, where expertise is defined as human *cognitive skill* acquired by repeatedly performing a task (Anderson 1999), people who have a kind of expertise in a particular topic are called experts. Many early expert databases systems were designed according to this definition. The experts who input into the database are publicly recognized people who are the best (or close to the best) in a certain domain. However, according to this definition, few people can claim themselves as experts in reality, although most will agree that they have expertise in some areas. In many knowledge seeking tasks, finding a person with sufficient expertise instead of an optimal expert is a more practical solution, especially when the former usually bears less cost than the latter. It is close to what March and Simon (1958) suggested: people seldom make fully informed decisions but rather satisfied decisions.

My research focuses on helping people share expertise through their social networks, which emphasizes making use of locally available expertise instead of finding the optimal ones. Thus, in this thesis, I adopted a more practical view of expertise proposed by Ackerman and Halverson (2003), in which “Expertise connotes relative levels of knowledge in people”. According to this definition, an individual can have different levels of expertise on different topics. Such expertise is arranged and valued by the social and organizational settings where the individuals are evaluated. Based on this

definition of “expertise,” expertise sharing aims to help people share what they know, to provide information seekers access to knowledge held by people. A significant difference between the newer expertise finding systems and a traditional expert database is that they allow everyone to contribute as they can.

Expertise sharing is viewed as the next step of knowledge management for organizations by many scholars (e.g. Ackerman and Halverson 2003). First generation knowledge management focused on a repository approach of using information technology to manage organizational knowledge (Ackerman, Wulf et al. 2002). Its key idea was to externalize knowledge from individuals and place it into shared repositories, such as an information database or knowledge base, as documents for later retrieval and use. Its theoretical foundation was a “knowledge creation model” proposed by Nonaka [Nonaka et al., 1995]. In this model, knowledge creation is a spiraling process of interactions between explicit and tacit knowledge, which includes processes of socialization, externalization, combination, and internalization of knowledge. Based on this model, knowledge management systems tend to emphasize gathering, storing, providing, and filtering available explicit knowledge. Such repository view of knowledge management has its advantages. By using standard technology and controlled input, the information put into the repository is easy to search, access, and transfer. By externalizing individuals’ knowledge, it also makes organizations less vulnerable to employee turnover (Argote 1999). However, this approach is limited and is difficult to apply in some situations. For instance, Lave and Wenger (1991) suggested that expertise is usually embedded in some particular situations and environments and is hard to extract. Hinds and Pfeffer (2003) found that it is difficult for people to use the de-contextualized information that is stored in the knowledge base as well as transfer the same knowledge into other contexts.

Expertise sharing aims to help people share their expertise, to provide information seekers access to knowledge held by people directly, which complements the limitations

of accessing information from documents. For instance, by enabling two-way interactions between askers and experts, it is easier for people to build common ground, understand the asker's context and needs, and transfer tacit knowledge. By not requiring experts to totally externalize their knowledge but instead help others in a case by case basis, it may also make them less concerned with losing their power (Hinds and Pfeffer 2003).

Appendix 1 reviews the related work on expertise sharing.

RESEARCH FOCUS -- EXPERTISE NETWORKS

There are many forms of social networks. As Wasserman and Faust point out,

“In the network analytic framework, the ties may be any relationship existing between units; for example, kinship, material transactions, flows of resources or support, behavioral interaction, group co-membership, or the affective evaluation of one person by another.” (Wasserman et al, 1994, p. 8)

The main goal of social network analysis is detecting and interpreting patterns of these connections and their implications. Accordingly, while the term "social network" usually implies affinity networks, there are different types of social networks and the meanings attached to them are different.

I call a network reflecting people's expertise-sharing activities an expertise network (Ackerman, 1993). When people use email to ask and answer questions in an organization, we can view this email network as an expertise network. Such a network indicates what expertise exists within an organization, as well as how it is distributed in practice. In an organizational expertise network, people usually know each other and social relationships may play an important role in the establishment of expertise transactions.

There are also expertise networks in online communities. Online communities containing discussion or question/answer forums usually have a thread structure like what is shown in figure 1(a). A user posts a topic or question, and then some other users post

replies to either participate in the discussion or to answer a question posed in the original post. Using these posting/replying threads in a community, we can create a post-reply network by viewing each participating user as a node, and linking the ID of a user starting a topic thread to a replier's ID, as shown in Figure 1-1.

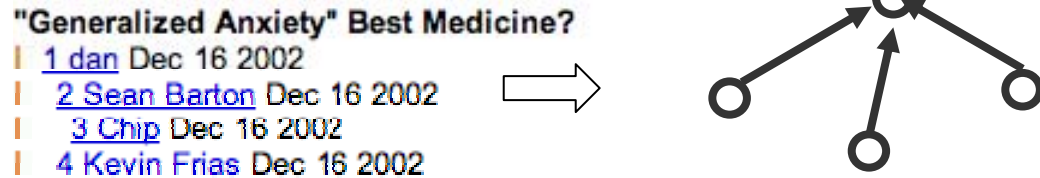


Figure 1-1: Method for converting a topic thread into a network

This post-reply network reflects community members' shared interests. Whether it is a community centered around questions and answers, social support, or discussion, the reason that a user usually replies to a topic is because of an interest in the topic. This indirectly reflects that shared interest between the original poster and the repliers (although the repliers' sentiment about the topic may differ). Furthermore, in a question and answer community, a user's replying to another user's question usually indicates that the replier has superior expertise on the subject than the asker.

All organizations and communities have their own community expertise network. We might imagine, however, that expertise networks have differing characteristics among organizations, communities of practice, communities of interest, and online communities; that is, they may differ more between types of collectivities than within. Understanding expertise networks and their differences is critical for knowing how to provide better technical support through online communities, facilitate the flow of technical or knowledge transfer within organizations and communities, and construct effective online communities of practice.

Specifically, my dissertation focuses on using various social network analysis techniques to characterize expertise networks. It seeks to understand the relationship between the network structure and the knowledge sharing process, thus in turn allowing

me to explore new algorithms and applications to augment expertise sharing in organizations and communities. To do this, I must integrate ideas and knowledge from various fields like expertise sharing, online community and social network studies. I further develop the concept of expertise networks based on previous studies on expertise sharing and online communities. For instance, I adopt social network models and theories (e.g., the small world model in Watts et al, 2001) to guide my research on the relationship between expertise sharing and social networks. I investigate existing community expertise networks through a mixture of methods of empirical observation, data analysis, and simulations. The data analysis focuses on finding meaningful social network metrics to characterize these networks as well as analyzing their impacts on the expertise sharing and searching processes. Simulations, based on empirically examinations of these networks, are used to explore the possible variations of networks, the dynamics of a network, as well as the performance of various expertise searching algorithms on these networks.

RESEARCH FRAMEWORK

The SWIM Prototype and Research Questions

My whole dissertation research could be viewed as my pursuing answers for questions being raised when I developed the Small World Instant Messenger (SWIM) system (Zhang and Van Alstyne, 2004).

SWIM is a novel instant messaging (IM) system that I developed in 2004. It focuses on fostering information search through social networks. It has all the functions that a general IM system has to support questions asking and answering. Two advanced functions are added to support social network based search process. First, SWIM maintains an advanced user profile. Besides letting a user input his expertise and interests manually, SWIM can automatically mine a user's homepage and browser bookmarks to

construct a keyword vector to represent the user's information identity. Second, SWIM has a built-in referral agent that handles the information-querying process automatically.

Figure 1-2 shows how SWIM works. A user starts the information search by sending the query to his referral agent, who broadcasts this query to all his buddies' agents. A referral agent in the buddy's SWIM then searches its own information identity vector first. If no match is found, it either returns empty results or forwards the query to its buddies, depending on its owner's control settings. The query will be passed along in the IM social network until it finds a match, or stops forwarding, or exceeds a number of hops. If a match is found, the path to reach the target person is returned to the searcher. The query and the path are shown to the target person who possibly knows the answer. Then these two persons can either start chatting immediately or discuss the questions asynchronously later if the answering person prefers not to be disturbed at that time.

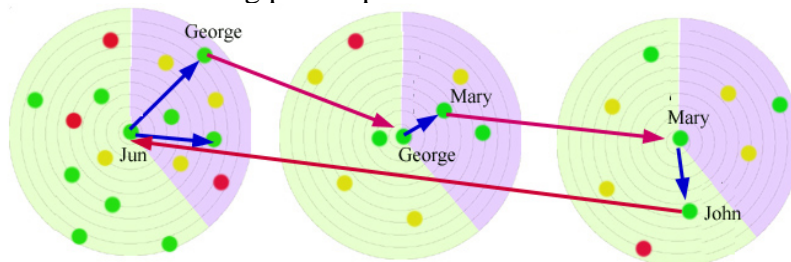


Figure 1-2: SWIM search and refer process

SWIM could be the next generation of Google. Instead of finding a web page, users could find an expert directly to answer their questions using SWIM in the future. I was not alone on this view. SWIM was named as one of the “Most Important Technical Innovations of the Year in the Internet Category” by Technology Research News at the end of 2004 (Patch, 2004). I also got a lot of interest from venture capital seeking to commercialize the system. However, while the original prototype design sounds like a perfect solution, I found that I didn't have a good theoretical answer for the core issue: how to find the right people who have the sought-after expertise. Actually, there are multiple problems, including:

- How can the system automatically spread queries quickly and efficiently in a social network?
- How can we evaluate a user's expertise regarding both subjects and levels?
- How is expertise distributed and shared in people's real lives currently?

These issues will fundamentally affect the design and adoption of the SWIM like systems. They are also key research issues for researchers who work on expertise sharing systems. Based on the literature review, I proposed and designed three studies to answer these questions, which are the Enron email study, the Java Forum study, and the Yahoo Answers study. Next I describe the focus of each these studies and the methodologies they used, as well as how they are related to each other.

Enron Email Study

The Enron Email study targets to solve a problem I faced in designing the SWIM system. When I designed the original SWIM, a big problem was how the system could spread the expertise queries in a large social network efficiently. In other words, what algorithm should a swim agent use to select the next person in the network to pass the query? This is an important problem for almost any social network based information searching systems. There is very limited work in the literature. In my original SWIM design, I used an algorithm called "information scent" (see details in Zhang and Van Alstyne 2004) which is similar to an algorithm reported by Yu et al. (2003). However, I was not sure whether this algorithm would work. Since it was impossible for me to do a large scale lab experiment, I decided to use a simulation to study this problem.

The design of the simulation focused on exploring how different social characteristics of expertise networks can affect the expertise searching process in organizations. People in social networks vary in their connectivity, expertise, status, availability, and sociability. These social characteristics can lead to sizeable differences

in the way expertise is searched and shared. With these factors in consideration, I proposed and tested three families of searching algorithms, each based on the structure of social networks, the strength of individual relationship, and the similarity of expertise. The simulation was conducted on the Enron email dataset, with a carefully designed sensitivity analysis.

As the first step of my thesis research, the Enron simulation study provided me with guidelines for designing searching algorithms for SWIM like systems. More importantly, it led me to realize the importance of social networks in the expertise-sharing problem, and helped me further develop the framework of my thesis.

Java Forum Study

In the Enron email study, I found that social networks indeed have great impact on expertise searching processes. However, there was a new challenge I identified during the study: how can a system identify the right person with the sought expertise? When an expertise searcher sends a query into his social network, he wants to find a person who not only knows about topic, but may also have greater knowledge on this topic than himself. Identifying expertise is a significant ongoing research problem. In the Enron email study, I adopted the keyword indexing and matching method that most previous expertise finders used. A person's expertise is described as a term vector that can be built from his email or other documents, and then used later for matching expertise queries using standard information retrieval techniques. However, while this method may reflect whether a person knows about a topic in general, it is difficult to determine that person's expertise level. Thus, I decided to find new ways of identifying expertise, especially ones that also evaluate people's expertise levels.

The Enron email data used in the first study was not a good data set for the purpose of the second study. Not all the Enron email communications was about

expertise sharing. We wanted to find some dataset that better reflected people's expertise sharing activities. Data in online technical help communities fit this need and were relatively easy to collect. Of course, an online community is very different from an organization. However, I believed many lessons learned from online communities can be applied into organizational environments if one can keep these differences in mind. Furthermore, online communities themselves have also become very important places for people to share expertise, especially in the Web2.0 era. Thus, in the second study, I used the data collected from a technical help community—the Java Forum—where people ask and answer questions about the Java programming language.

The study was divided into two steps. The first step of the study sought to understand what's going on in Java Forum, especially from a network analysis perspective. I analyzed the expertise network constructed from the Java Forum thread structure using advanced network analysis metrics, including bow-tie structure, degree distribution, community structure, motif profiling, and correlation profiling. Furthermore, empirical observations and simulations were also used to explore the relationship between the social settings in these communities (e.g. the expertise distribution among users, people's preferences to ask and answer questions) and the structural properties of these networks. The second step of the study explores opportunities for using the characteristics of expertise networks to develop new algorithms for evaluating expertise levels. Different graph-based ranking algorithms (e.g. PageRank, HITS) were proposed and evaluated. To understand the results, we further simulated the community dynamics and produced networks that not only matched the observed aggregate network characteristics but also allowed us to understand why automated expertise-ranking algorithms perform differently in differently structured networks.

The Java Forum study is the capstone of my dissertation research. Compared to the Enron study, it goes much deeper into the theoretical understanding of relationships among individual interactions, social settings in the community, and the flow of expertise

sharing. The algorithms developed and evaluated in this study can also be directly applied to SWIM and other similar expertise sharing system designs.

Yahoo Answers Study

We had many interesting findings in the study of Java Forum community. However, the Java Forum is a single topic community and is very technical. All the questions and answers are about Java. In most expertise finding systems, there will be very diverse topics and many of them may not be technical. Thus, I decide to extend the study to Yahoo Answers, one of the largest, if not the largest, question answer forum on the Web.

In this study, we harvested one month of questions and answers posted in Yahoo Answers using an automatic crawler. Then, using network and non-network metrics, we examined several aspects of question-answer dynamics in Yahoo Answers. First, we analyzed both content properties and social network interactions across different categories. We identified a set of features that could be used to cluster the categories, and found that thread length and overlap between the set of users who asked and those who replied are the most distinct features that separate different types of categories in Yahoo Answers. Second, we related categories to each other by analyzing users' cross categories posting patterns. For instance, we examined whether if a user is answering questions in one category, they are also likely to answer in another. Third, we examined the range and depth of knowledge that users share across different categories in Yahoo Answers, as well as what factors will affect whether one's answer is rated as "best answer."

As the last piece of my dissertation work, this study revealed what is going on in a large scale general knowledge sharing community. For instance, we found that questions in Yahoo Answers are very diverse. There are not only questions for seeking technical

instructions to repair a computer, but also questions like seeking advice on a medical treatment, gathering opinions on a newborn's name, and satisfying one's curiosity about a celebrity. Using network and non-network metrics, we attempted to identify the different asking and answering patterns among these different types of questions. Furthermore, users' interests and expertise are usually broad. Some users answer questions in many different categories. However, in specialized technical categories, this breadth could come at the detriment of expertise depth. We should expect that the similar diverse topics and activities will also happen in a large scale SWIM like networks. When we design algorithms or mechanism for large scale social network based expertise sharing systems, we should put these diversities into consideration.

Above all, although the studies of Enron email, Java Forum, and Yahoo Answers each focused on different research questions and adopted different methods, we can see that there is an inherent connection among them. They are all conducted around one goal in mind: to gain a better understanding of expertise sharing in social networks, thus helping us design better expertise sharing systems. Together, they helped me understand the expertise networks from different perspectives, as well as providing answers for many expertise system design challenges.

CHAPTER 2

SEARCHING FOR EXPERTISE IN SOCIAL NETWORKS

INTRODUCTION

Imagine you have a question that is blocking your work. For example, you might need help understanding a warning message from a critical application, and you're unable to locate a document explaining the message. Or as another example, you might need to understand how to work around a specific rule for ordering equipment.

In both of these situations, someone knows the answer to the question. Finding that person, however, can be difficult. Ideally, we would like to find a person who knows the correct answer to that specific problem. Additionally, we would like to ask only the appropriate person and to find a person who has enough free time to answer the question.

In reality, of course, answering questions is not so easy. People are busy, they may lack the requisite expertise to answer the question, or they may lack the social graces to answer well. As a first step, you may not know whom to ask.

Systems that help find others with appropriate expertise are called expertise finders or expertise location engines. These have been explored in a series of studies, including Streeter et al. 23 and McDonald and Ackerman 18 as well as the studies in Ackerman et al. 1. Newer systems, that use the social network of an organization to help find people, have also been explored, most notably in Yenta 11, ReferralWeb 15, ER 18,

and MARS 30. These systems attempt to leverage the social network within an organization to help find the appropriate others, thus reducing the need for specialized data. This is a critical requirement for expertise finders, as requiring specialized data for expertise location makes adoption difficult at best.

Because each of these newer social network based expertise finders uses social network data (which may be derived in a number of ways), we can now move away from research emphasis on the systems and towards an examination of the algorithms used to search the social networks.

This chapter surveys three algorithms in the open literature; it also adds several additional algorithms. These new algorithms, as will be seen, have interesting social characteristics. The main contribution of the paper, accordingly, is to examine those algorithms using a simulation testbed in order to evaluate them and understand their relative tradeoffs. We believe this work is critical if progress is to be made on finding methods and mechanisms for expertise location.

The chapter proceeds as follows: First, we survey the related research. Second, we introduce our simulation experiments, including the data set we used, the algorithms we evaluated, and the performance measures we used. Third, we describe our analyses and findings. At last, we discuss the design implications and future work.

SEARCHING IN SOCIAL NETWORKS

In this section, we first review the rich literature about searching for people in social networks. Then we examine the computational approaches used for finding people in social networks, when the person is known in advance and when he is not. The more interesting case for us is the latter, since this is the expertise location problem.

Small World

The classic study on searching in social networks is the “small world” experiment. In late 1960s, Milgram and Travers found that subjects could successfully send a small packet (with a name, the city, and the profession of the recipient on it) from Nebraska to people in Boston 1924. The subjects did so, even though they had only local knowledge of their acquaintances, by passing the packet to an acquaintance that they believed to be closest to the target. Travers and Milgram found the average length of acquaintance chain is roughly six. The result of this experiment indicated that the social network is searchable and that the paths linking people are short, the so-called “six degrees of separation”.

A key question in such an experiment is how people select the next person to whom to forward the packet or message. Potentially each subject has hundreds of acquaintances, but picks one, which ultimately leads to a short chain between the sender and the target. Later similar experiments found that geographic proximity and similarity of profession to the target person were the most frequently used criteria by subjects 1669.

Recently, mathematical models have been proposed to explain why these simple heuristics are good at forming short paths 1726. These models assume that the social network usually has a structure, in which individuals are grouped together by occupation, location, interest, and so on. As well, these groups are grouped together into bigger groups and so forth. The difference in people’s group identities defines their social distance. By choosing individuals who have the shortest social distance to the target at each step, people can gradually reach the target in a short path with only local information about their own immediate acquaintances.

Searching For Expertise in Social Networks

These studies on the small world problem have led to two lines of computationally-based approaches that concern searching for people within social networks. The first is an automatization of the small world approach, where the target is known by name or unique identifier [328]. The second is locating a person with some specific expertise or knowledge. We consider the latter.

In an expertise location or expert finder problem, a suitable person or set of people is not known in advance. One must be found by matching people against a list of attributes.

A number of expertise location systems have been developed. For example, Who Knows [23] found people with appropriate expertise by doing latent semantic indexing of project reports, Yenta [11] found people by searching email archives in a distributed manner, and Expertise Recommender [18] used locally meaningful data to recommend sets of potential answerers for queries. Other work is surveyed in Ackerman, Wulf and Pipek [1].

Yu and Singh's referral system [30] is, as far as we know, the only paper that explicitly argues for a specific expertise-finding algorithm. In their experiment, they use the similarity between a query vector and a neighbor's expertise vector, plus some consideration of its historical referring performance, as the criteria for picking the next agent in a referral graph. The simulation results using a scientific co-authorship network suggest that this strategy can help people find experts in such a network.

Yu and Singh's algorithm is a useful first step, but their approach has limitations. There are several issues. First, the query vector and expertise vector in their experiment are manually coded; and each is a combination of preset topics from taxonomy. This approach is not practical in real world scenarios: Questions are usually extremely detailed, and people cannot be categorized as one specific type of expert. Second, they

had only a rudimentary consideration of the impact of the *social* network structure on the searching process. Finally, and most importantly for this paper, they did not compare the performance of their algorithm with other possible algorithms; thus, the relative benefits and tradeoffs of their algorithm is unknown. Nonetheless, Yu and Singh's algorithm is an important candidate for examination.

In addition to Yu and Singh's algorithm, several other algorithms can be adapted to the expertise location problem. Adamic et al's best connected search (BCS) algorithm 34, which makes use of the skewed degree distribution of many networks¹, can also be used to find experts. By passing the query to highly connected nodes first, BCS can spread a query quickly in the network. However, Adamic et al. also found that the BCS algorithm is not always efficient in all networks. Nonetheless, Adamic et al's algorithm may be valuable in many cases, and we will also include it in our investigation. Breadth First Search (BFS) 21, which broadcasts a query to every person in a social network, has the strength of finding the closest expert available. But it can have a high cost both computationally and socially in that many people can be bothered.

Thus there are three lines of potential algorithms in the open literature that need be examined. To our knowledge, these algorithms have never been evaluated together nor their tradeoffs and social characteristics examined. These social characteristics include standard attributes of social networks:

- Connections among people are not uniformly distributed. Unlike a theoretically constructed graph, the connections among people in a social network are highly meaningful and vary greatly 25.
- The connections between two individuals can have different strengths. There is a strength of association between individuals. This strength of association varies and is

¹ In such a network, many nodes just have one or several links and a few nodes have many links.

not always symmetrical. Usually, in social networks, the strength of association is divided roughly into strong and weak ties 12.

- People in a social network vary in their expertise, status, availability, and sociability. Unlike theoretically constructed graphs and computational agents, a person weighs his use of his social network by considering these additional characteristics.

These social characteristics could lead to sizeable differences in the way information is transferred, affecting the performance of the searching algorithms. For instance, weak ties have been found to be important in helping people get new information 12 and adopt innovations 7. In Dodd et al.'s small world experiment, successful searches were also found to be conducted primarily through intermediate to weak strength ties.

These social characteristics, in addition to computational efficiency, will guide our outcome measures. The following section introduces the outcome measures, but only after introducing the experimental test-bed and the examined algorithms.

SIMULATION

In this section, we firstly discuss the simulation as our experimental apparatus. Second, we describe the data set we used and its limitations. Third, we introduce our new algorithms along with those previously proposed algorithms. Fourth, we describe the simulation process and the data we collected. Finally, we describe the evaluation criteria we used to compare these algorithms.

Simulation as Experiment

It may seem odd, at first glance, that we would wish to examine the social considerations and tradeoffs of these expertise locating (EL) approaches using simulations instead of field or laboratory experiments. However, simulations appear to

be a much more fruitful experimental apparatus or testbed for examining these issues. This is unusual for CSCW investigations, and some explanation is required.

Constructing well-controlled laboratory experiments of the size required to effectively test these algorithms would at best be extremely difficult. On the other hand, while it might be possible to construct Internet-based experiments of a suitable size, these would be uncontrolled. Alternatively, Internet subjects would be required to run special applications (e.g. email monitoring software or email indexing software); this is extremely unlikely. Finally, a real organization (of a sufficiently large size) would provide us with enough users and use. However, we have been unable to convince any large corporation to either provide us with all of the company's email or to introduce experiments into their ongoing communication systems. It is unlikely that we will.

Accordingly, we examined simulations as a potential experimental apparatus for our investigations. We felt that the major problem with using simulations was the threat to the validity of our results.

Simulations are often too artificial. Overly rational agents, a small set of experimental categories and of agent behaviors (so as to be tractable), and severe limitations on methods of choice can lead to problematic social findings because of the restrictions. This is of course not necessarily true - one can look to the insightful simulations of Hutchens or Axelrod 145. While socially limited, their restricted operationalizations have led to insights about cultural production and coordination, respectively.

In the following work, we have tried to avoid artificiality in two ways. First, we constructed our simulations using a data set from a real organization rather than using artificially or theoretically constructed data. The Enron email data set will be discussed below.

We also tried to operationalize our outcomes in a manner that was not overly restrictive. Of course, any operationalization in an experimental situation must be

restricted, if nothing else to be concrete. Our operationalizations, which will also be discussed below, have implied limitations and restrictions. We have tried to account for these limitations both in our discussion and by doing detailed sensitivity analyses on our results, explicitly to look for the effect of such restrictions. We discuss these sensitivity analyses, and where we were unable to do one, below.

Simulation Data

The simulation data set is the well known Enron email dataset 8. After cleaning the data, it contains data from 147 employees, mostly senior management of Enron. There are a total of 517,431 messages in the data set.

To construct a social network, a sub sample of 32766 messages that were exchanged among these 147 employees was used to construct a directed graph. As shown in Figure 2-1, the network is a relative dense internal social network with 147 nodes. The density of the network is 0.096, and the average shortest path is 2.498.

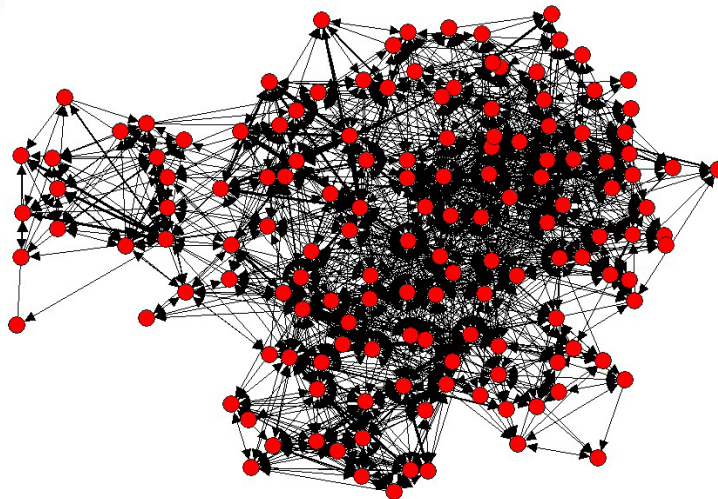


Figure 2-1: Enron Email Network

Figure 2-2 displays the cumulative out-degree² distribution of the network. It is highly skewed with some nodes having high degree in the tail.

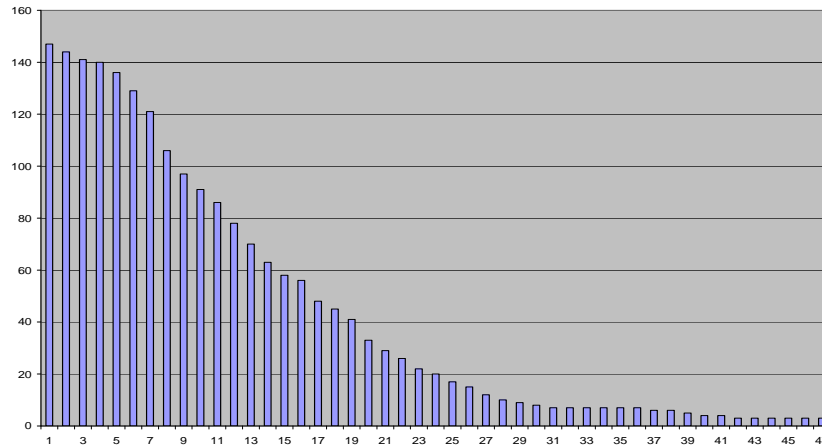


Figure 2-2: Cumulative Out-Degree Distribution of Enron email network

Second, using a standard document indexing method 29, we indexed all the messages that a person sent or received for each of the other 146 users. The indexed result for a single user is a keyword vector, in which a keyword is weighted by its term frequency inverse document (message) frequency. These indexes are used as information profiles for the users.

By doing this, we get a test-bed with a network structure and information profiles derived from a real organization. There are, however, two possible limitations with this data set.

First, the simulation network is not the complete email social network of the Enron organization. It consists of the management level subset of that network. Compared to the complete organization network (which we cannot obtain), this network is likely to be more dense with a smaller average shortest path. This is because general employees usually have a lower probability than managers of knowing people in other groups. Furthermore, managers may have different information profiles (or expertise)

² Here we simply use the number of other users to whom a user had sent emails as his/her out-degree.

than other employees, such as engineers or clerical personnel. To examine whether the data set has different properties than the full social network, we constructed two new networks by removing edges that were weaker than a threshold and by removing high degree nodes. This gave us social networks with different network characteristics. We ran the same simulations on these two additional networks as a sensitivity analysis; we discuss the impact on the findings below.

Second, using this data set to determine expertise may be problematic. A keyword in one's email folder does not necessarily mean that one has expertise concerning that keyword. This is a limitation of our operationalization: we are assuming a perfect match between the information profiles and expertise. Determining an ontology of expertise and determining where it is located in an organization is a significant, ongoing research problem 2, and we believe this operationalization is a good surrogate. Furthermore, for transactive knowledge, communication is likely to be an indicator. Accordingly the information profiles are not only an indicator of this aspect of organizational expertise, they serve as an approximation of the location of other types of organizational expertise.

We believe that despite these limitations, the Enron data set gives us a realistic test bed, reducing the amount of artificiality in the simulation. We will discuss where its limitations affect the results below.

Searching Strategies Evaluated

We evaluated total 8 searching strategies from three families in this simulation; include 3 found in the public literatures and 5 proposed based on related theories. They are shown in Table 2-1.

Table 2-1 Evaluated Algorithms (* indicates algorithms that we proposed based on related works)

	Name	Heuristic	From
General computational	BFS	Breadth first Search	Classic AI
	RWS	Random walk	Adamic et al.
Network structure based	BCS	Best connected	Adamic et al.
	WTS	Weak tie	* Granovetter, Burt
	STS	Strong tie	* Granovetter
	CSS	Cosine similarity	* Wasserman et al.
	HDS	Hamming distance	* Hamming
Similarity based	ISS	Information scent	Extract from Yu and Singh

All these strategies are based on the information that a user can gather or derive locally from their email communications with peers. There may be additional strategies, such as using people’s position in physical space or in an organizational hierarchy 3, but information available in the Enron data set limits us to the ones examined in this study. In any case, the strategies examined here are, we believe, the most important ones to examine first.

Details of these algorithms are:

Breadth First Search (BFS) broadcasts a query to all of one’s neighbors instead of picking a neighbor according to a heuristic. It can find the target closest to the source but with extremely high bandwidth costs (as in p2p file sharing networks).

Random Walk Search (RWS) randomly chooses one of the current query holder’s neighbors to whom to spread the query. RWS will be our baseline to determine whether other heuristics are “better” in spreading a query.

Best Connected Search (BCS) is the algorithm Adamic et al. used. The only difference from their algorithm is that we construct our network as a directed network. Social relations are not symmetrical, and it may affect people's information seeking behavior in a social network. In any case the Enron data set has both outgoing and incoming messages. We use out-degree of connectivity instead of both in-degree and out-degree in evaluating one's neighbors.

Weak Tie Search (WTS) is proposed based on Granovetter's weak tie concept as discussed above. There are different ways of measuring relationship strength ²², here we simply assume that a peer who receives the fewest messages from a user is a weak tie and will be chosen as the next one to forward the query to.

Strong Tie Search (STS) is proposed as a comparison with WTS. It picks the neighbor who has received the most messages from the current user. It may be a reasonable strategy in practice because there is usually lower social cost when one asks for help through strong ties.

Hamming Distance Search (HDS) and *Cosine Similarity Search (CSS)* strategies are two structural dissimilarity strategies based on definitions of structure equivalence in social network studies ²⁵. HDS picks the neighbor who has the most uncommon friends from the current user. The definition of Hamming distance ¹³ favors the nodes with high out-degree. HDS could be viewed as an improved version of BCS. CCS decreases the high degree impact by dividing the Hamming distance by the total number of out-degree relations (friends) a neighbor has.

Information Scent Search (ISS) is extracted from Yu and Singh's algorithm ³⁰, leaving aside the sociability learning part. ISS picks the next person who has the highest match score (which we call information scent) between the query and his profile. Our implementation of the algorithm is slightly different from Yu and Singh, since we needed to adapt their algorithm to the Enron data set. (Remember that Yu and Singh used only 19 categories or keywords.) We use the automatic generated keywords profile instead.

Process and Data Collection

We manually generated 147 questions by picking keywords from one or several messages from the sent folder of each of 147 email users. Each question has three to five keywords that do not include common words, such as “hi” “the”, “ok”, and “enron”. Thus, we assumed each question has at least one expert available in the network.

During each round of the simulation, a question and an asker are selected at random. Each searching strategy is executed simultaneously at each round.

The match between a query and a person is calculated in two steps:

- 1) Message level matching: This is a standard information retrieval matching based on the TF/IDF measure. In general, if a message has the exactly the same combination of keywords as the query, it has the highest score; if it has only several keywords out of many, it has a low score.
- 2) Personal level aggregation: A person might have multiple messages with different matching scores that related to the query. This raises some issues. For instance, how could we compare a person who only has one document that has a very good match score with another person who has hundreds of related messages none of which is a good match? We chose to weight the documents by their ranking in one person's results in the personal level aggregating step.

So, a person's match to the query is measured as $\sum_{i=1}^n MessageScore(i)/(i+1)$. We

use the top 20 messages.

The criterion of a satisfied match is calculated by multiplying the best match score available, which is pre-calculated using a global search, with a satisfaction factor S (S=0.8 here).

The general query propagating process is as follows:

- 1) A user receives a query message (or the asker has a query).

- 2) The simulation engine searches all of the user's directed neighbors' information profiles. If there is a match above a desired threshold, it returns that person to the asker and stops the search. If there is not, the BFS strategy will broadcast the query to all of the -neighbors; other strategies will pick a neighbor according to their definitions. The visited node's ID is appended to the query message so a node would not be visited twice. Except for BFS, the asker starts a new searching path if the previous path reaches a dead end³.
- 3) The query will be continually propagated in the network until no node is not visited (BFS) or no path is left (other strategies).

Note that in step 2, we assume that each user has knowledge of his direct neighbors' knowledge or has access to their profiles. It corresponds to transactive memory 27. It is also the assumption used by Adamic et al in their small world experiments.

The data we collected during each round of the simulation include: asker's information scent on the query, steps (people used) to complete a query, number of paths tried (how many times a query needed to be restarted), number of people used, and the expertise score of the target. Since not all queries are successful because some nodes are not reachable from some other nodes, we record the number of search failures as well. After all rounds (N=30,000) are finished, we summarize overall how many times a user has been queried in each strategy.

We then calculate the out-degree and in-degree of each user. We used these to analyze their influence on the performance of algorithms.

³ It could reach a dead end or the Time-to-live (TTL) of the query message expires. The TTL is set to infinite in this simulation.

Evaluation Criteria

Compared to searching a file in peer-to-peer file networks or searching for a person in a small world experiment, searching for expertise in social networks is a far more complex process. It involves many more social interactions. Speed and computational resource are not the only concerns; psychological and social costs are very important. After a social network based expertise system is adopted into an organization, the searching activities will be embedded into people's daily lives. So, an evaluation should not only consider the computational performance per query, but also needs to consider the social consequences of the strategies.

Based on these considerations and related work, besides analyzing the result from a computational efficiency perspectives, we compare the social cost of the evaluated algorithms using three measures:

- Number of people used per query (how many people were bothered).
- Depth of query chain (i.e., how deep the query went).
- Total labor distribution in all queries.

The number of people used per query is the measure Adamic et al. used in their simulation [3]. It counts how many nodes (people) processed the query during a search. It is a measure of social cost per query as well as the speed of the algorithm. When searching for information in social networks, we usually want to bother as few people as possible. If each used person took one unit of time to process a query and the query is propagated sequentially, we want the search process to be fast and bother the fewest number of people possible.

The depth of query chain measure, in many cases, is equal to the number of people used per query. It becomes different when there is more than one path used for a query. The depth of query chain counts only the number of people involved in the final

successful path. In real life, less distance also means a high probability of getting response from an expert.

Labor distribution measures the overall social cost in an organization related to people's expertise seeking activities. Different from the people used per query, it counts how frequently a person is used by each searching strategy (during an entire simulation).

DATA ANALYSIS

In this section, first, we describe the general computation results. Second, we introduce the general findings related to the social cost measures. Third, we briefly analyze the impact of social characteristics on these algorithms. Finally, we discuss the sensitivity of the results by examining two modified networks.

General Computational Results

Table 2-2 displays the overall success rates of the algorithms. In the table, there are two categories of query failures. The first is when there is no path between the asker and available experts. All the failures in using BFS belong to this category. The second is when the algorithm cannot find available experts even when there are paths. For expertise location, we are primarily concerned with this type of failure. (The adjusted rate in the table shows the successful rate of a query presuming the first type of failure does not occur.)

From the table, we can see that these algorithms are reasonably successful. They can all find a qualified expert for most of the queries in this network. (Note with $N=30,000$, all differences are statistically significant. We omit p-values from our discussion except where important.)

Table 2-2 : The success rate of various algorithms

Algorithm	BFS(b)	RWS(r)	WTS(w)	STS(s)	BCS(h)	ISS(i)	CSS(c)	HDS(d)
Success (%)	97.9	94.7	96.2	95.8	97.1	97.1	97.1	97.1
Adjusted(%)	100	96.8	98.3	97.9	99.2	99.2	99.2	99.2

Figure 2-3 further shows the percentage of successful queries within a given number of search steps using the various strategies. As one can see in the figure, for different search lengths, the rank of these algorithms changes very little. Although HDS and BCS are a little slower than BFS⁴, they are still very fast and successfully finish 80% of the queries within six steps. CCS and ISS can still finish more than 60% queries, WTS can find 55%, but RWS and STS can only find about 40% within six steps.

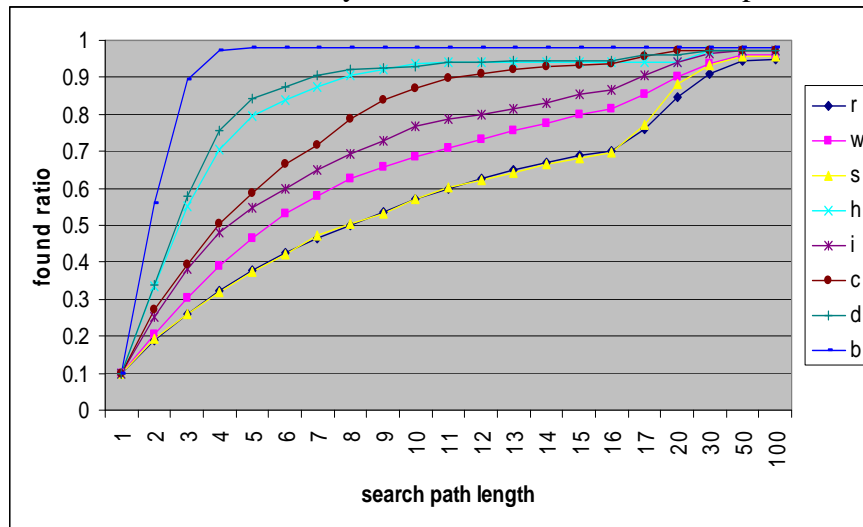


Figure 2-3: Percentage of succeed queries within different search length using various strategies

We can also see that when targets are far away from the askers, there is much less difference among these strategies.

⁴ Note in regarding the speed of BFS, we use the depth of the search instead of the number of people used in the query.

Comparison of Social Costs

Number of People Used Per Query

Figure 2-4 shows the distribution of the number of people used per query using the different algorithms. As the system becomes less completely automatic, this value becomes increasingly important to people’s user experience. Measures for these values are shown in table 2-3. Compared to the BFS broadcasting, HDS, BCS, and CSS strategies bother many fewer people. ISS and WTS are also clearly better.

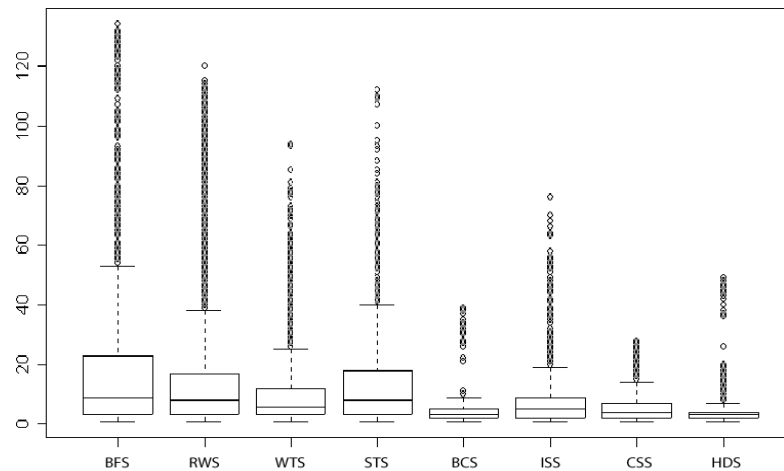


Figure 2-4: Distribution of Number of People used using various strategies

Table 2-3: Number of People used using various strategies

Algorithm	BFS	RWS	WTS	STS	BCS	ISS	CSS	HDS
	(b)	(r)	(w)	(s)	(h)	(i)	(c)	(d)
Median	9	8	6	8	3	5	4	3
Max.	134	117	94	112	39	76	28	49

In Figure 2-4, also note that there are a lot of outliers: Some queries used a lot of people before finding a desired target. Regarding the worst queries, as shown in Table 3, CSS handles them best and BFS handles them worst.

Based on these results, we can see that in this network, HDS, BCS, and CSS clearly have advantages over BFS and RWS regarding the number of people used per query. Also, ISS and WTS are better, but less so. Yu and Singh considered ISS to be very promising, but here we have found it less so than HDS, BCS, and CSS. Considering their performance as shown in Figure 3, HDS, BCS, and CSS could be promising algorithms to replace BFS when the speed and depth of searching chain are not the most important factors while the number of people being bothered is. The other interesting finding is that STS is obviously worse than WTS. The implications of this finding will be discussed later.

Depth of Query Path

Figure 2-5 shows how the depth of the query is distributed for each algorithm. This measures how long a query is in the social network; when designing a system, one would like to minimize these values. Except BFS, which we already knew always found the closest target, the result is not very different from measuring the number of people used per query. We checked the number of paths tried for the various algorithms (except BFS) and found that most successful queries are finished using only one path. This indicates that at least in this social network, there is little need to send queries simultaneously to multiple users to achieve a successful result. This implied in our dataset, therefore, the two measures of depth of query path and number of people used would have the nearly same value.

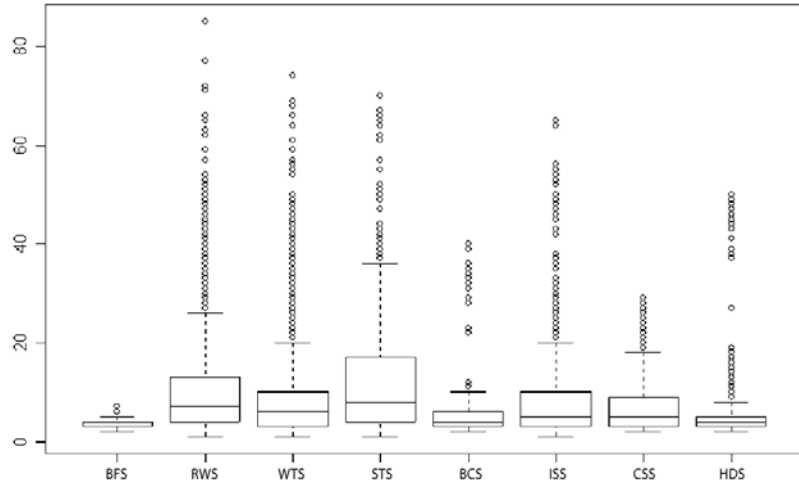


Figure 2-5: Distribution of depth of query chain using various strategies

Labor distribution

Figure 2-6 and Table 2-4 show the labor distribution (i.e., how distributed the search is) for these strategies. We can see that when using BCS, HDS, and CSS, most people are used less frequently, but some users are used extremely frequently. This indicates that referring is mainly loaded on very few members of the network. ISS is a little more balanced than these three algorithms, and BFS bothers people much more frequently than the other strategies. We will further discuss what strategies bother people more in next section.

Table 2-4: Distribution of Labor using various strategies

Algorithm	BFS	RWS	WTS	STS	BCS	ISS	CSS	HDS
	(b)	(r)	(w)	(s)	(h)	(i)	(c)	(d)
Median	19.1	8.2	2.2	6.3	0.6	3.5	0.7	0.5
%								
Max. %	60.1	25.1	48.4	33.9	61.1	23.2	45.2	63.4

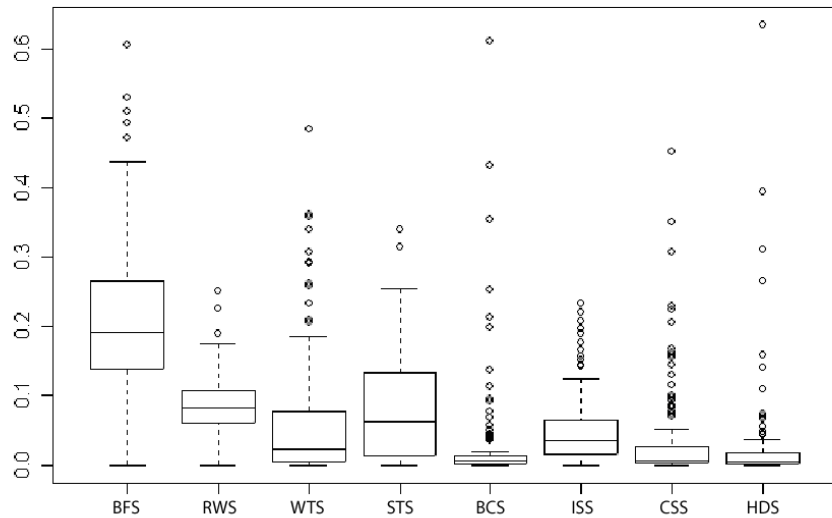


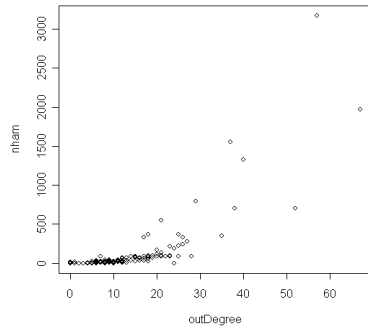
Figure 2-6: Distribution of a user's frequency of being used using various strategies.

Impact of Social Characteristics

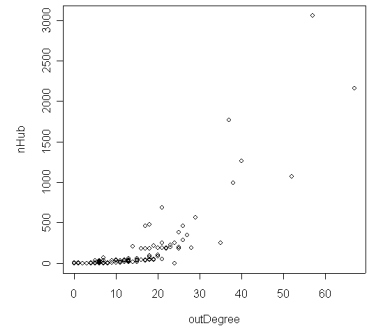
We briefly looked at how different social characteristics influence the performance of these algorithms. Based on the findings from the previous section about social costs, we mainly discuss the impact of two characteristics of the social network: user's out-degree and tie strength.

Impact of User's Out-degree

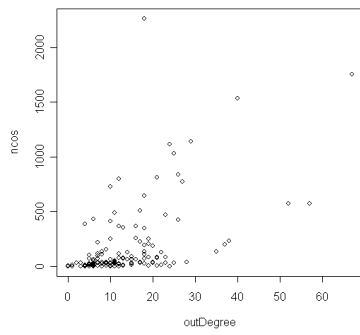
Figure 2-7 displays correlations between a user's out-degree and frequency of being used using various strategies.



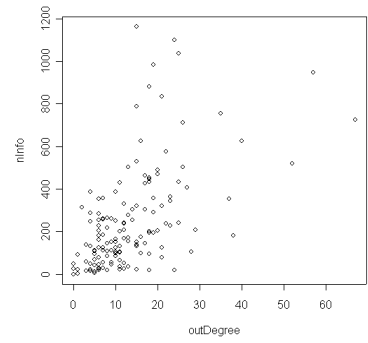
HDS



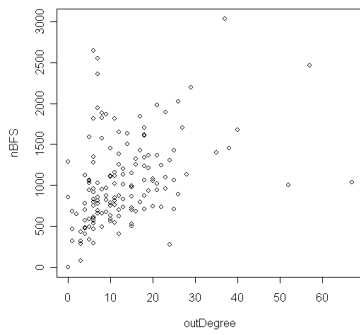
BCS



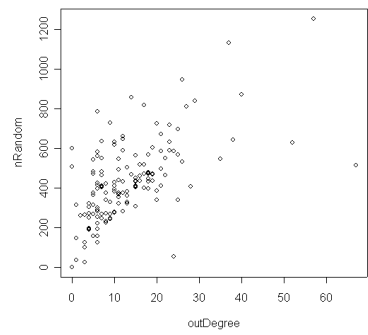
CSS



ISS



BFS



RWS

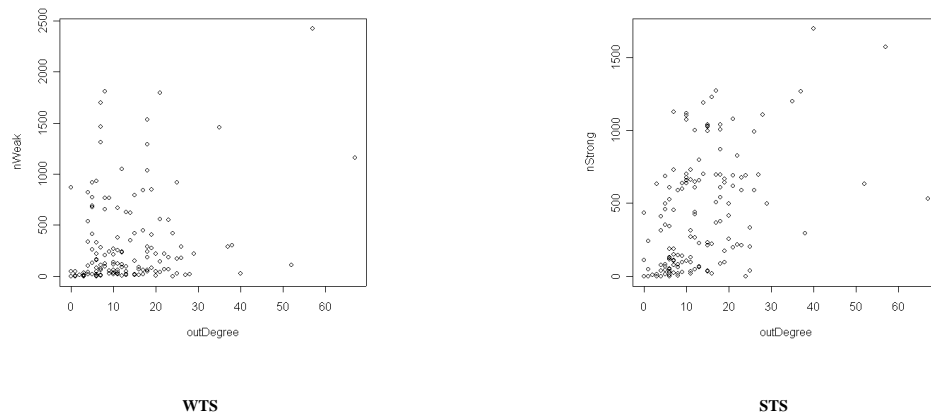


Figure2-7: Correlation between a user's frequency of being used and his out-degree using an algorithm

Surprisingly, out-degree is important even when the algorithms are not explicitly designed with this in mind. This echoes findings regarding the importance of position in networks and node centrality in organizations (e.g., Burt 7). We can see that when HDS and BCS are used, the relation looks close to exponential. People used most frequently are those highly connected people. This is not a surprise, since this is how these two algorithms are defined. More importantly, we checked the social status of those frequently used people and found that the CEO, CIO, and the president of the company are central nodes (or social network hubs). This strongly suggests that if a similar algorithm and system are not totally automatic, they will not be practical in this organization. CSS is designed to decrease the impact of people's out-degree; thus, the correlation in its case is weak. However, it still uses a lot of highly connected users.

As well, there is an intermediate correlation when using RWS. This indicates that random walk is actually not random. As Newman 20 pointed out, nodes with high in-degrees have a high probability of being picked by other nodes in a random walk in a network. We found that there is a correlation between a user's in-degree and out-degree in our network, thus explaining the result here. The case of BFS is similar to RWS. People with high in-degrees also have a high chance being searched during the whole simulation process.

An interesting finding is that there seems some correlation even when IIS is used. The adjusted r-square is 0.31, $p < .001$. This indicates that the IIS strategy is not independent from the network structure. For instance, people have more social connections may have more diverse knowledge. This relationship is worth further investigating in the future.

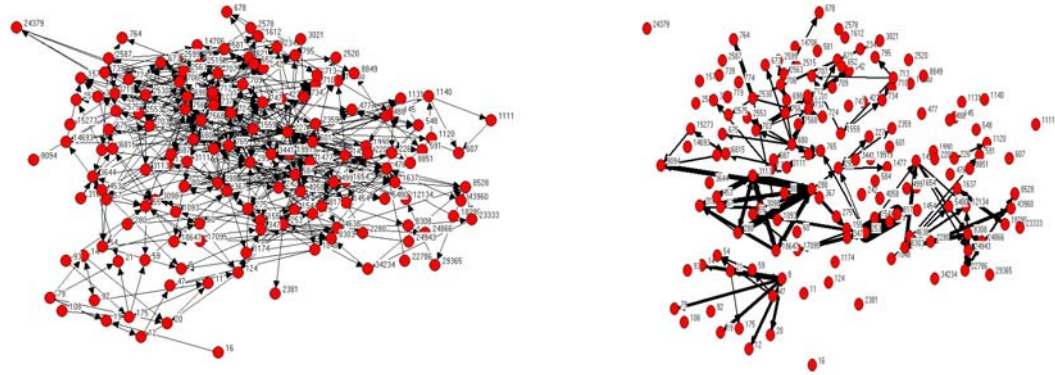
There is no clear correlation when WTS and STS are used.

Impact of Weak Ties

As described in previous findings, WTS seems more effective than STS. It spreads a query faster and bothers fewer people. To explore the reason for this difference, we visualized the distribution of these two types of ties into two network views, as shown in Figure 2-8. From these two views, we can see that weak ties are evenly distributed but strong ties form several local clusters. Thus, it seems the weak tie strategy propagates queries relatively evenly to other parts of the network, and the strong tie strategy usually makes local loops when forwarding the messages.

From this point of view, we can see that strong ties are not useful for seeking new information. However, we noted the motivational advantages of using strong ties. Any algorithms using strong ties, or thresholds for interpersonal association, may need to consider disjoint subgraphs.

The other interesting question is: since the different strengths of ties are not evenly distributed in this social network, what is the impact to other information searching algorithms? We further discuss this issue in section 4.4.1.



a) Tie < 5

b) Tie >= 5

Figure 2-8: Layered network with various tie strength

SENSITIVITY ANALYSIS

As we discussed earlier, the availability of high degree nodes and weak ties are important for searching algorithms we evaluated in this study. However, different networks will have different degree distributions: the Enron data set only presents a single case and it is a very dense social network. Even within other social networks, the availability of weak ties is not stable and changes frequently¹². To evaluate how these algorithms will accommodate to changes of density and tie strength, we carried out two sensitivity analyses using modified networks. The first redefined weak ties and the other removed users with varying out-degrees.

Removing Weak Ties

We first modified the network by removing ties that had less than 5 messages⁵. The result network is the one shown in Figure 8b (density=0.041, average shortest path=3.435).

⁵ We also tried other thresholds for “cuts”. A threshold of 5 was selected because it changed the network enough but still kept the network roughly connected. It is also close to the cut point that Adamic et al. used in their simulation.

We then ran the same simulation on this modified network with the same settings. Because of the changed cut point (or threshold for weak ties), note that the operationalization of “weak” tie here is not the same as in the previous simulation.

Table 2-5 shows the successful rate of this simulation. As can be seen, compared to original network, about 23% more queries cannot be finished because the network became less connected. More interestingly, as can be seen from the adjusted rate, there is a clear performance drop for RWS, WTS, STS, and ISS. This indicates that RWS, WTS, STS, and ISS are sensitive to weak ties and related network structure changes while BCS, HDS, and CSS are less so.

Table 2-5: the success rate of all algorithms in modified network

Algorithm	BFS(b)	RWS(r)	WTS(w)	STS(s)	BCS(h)	ISS(i)	CSS(c)	HDS(d)
Success (%)	76.3	44.0	40.0	45.8	73.2	57.6	73.3	72.8
Adjusted (%)	100	57.6	52.4	60.1	95.9	75.5	96.1	95.5

Furthermore, we can see that performances of HDS, CSS, and BCS are also affected. Figure 2-9 shows the changes of average path length of successful queries in the modified network. Compared to little change in BFS strategy, the changes in HDS, CSS, and BCS are noticeable.

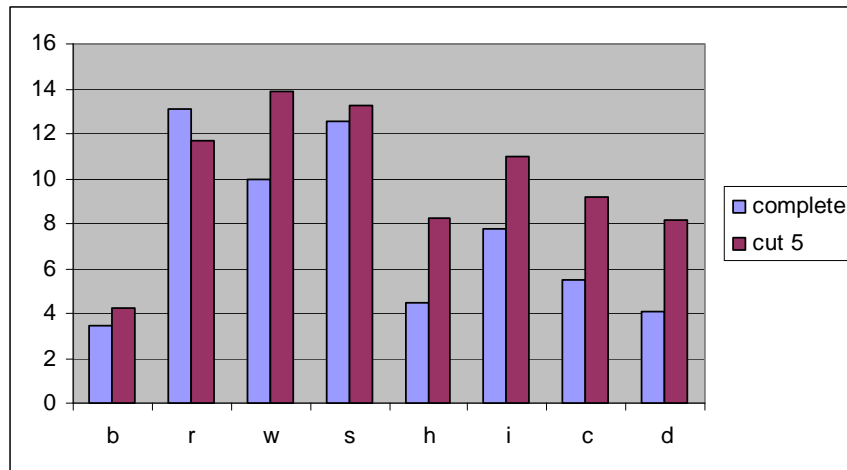


Figure 2-9: Comparison of average path length of successful queries

Above all, results in this modified network suggest that weak ties are really important information channels. They should not be simply ignored in designing social network based systems or in doing email based social network analysis.

Removing Users with High Out-Degree

For the second sensitivity analysis, we modified the original network by removing the 10 users who had the highest out-degrees. Most of them are also the most frequently used users in the original simulation. In this modified network, the average shortest path length became 2.754 and density became 0.076.

Table 2-6 shows the success rates in this simulation. Surprisingly, we find that the performances of BCS, HDS, and CSS are not affected at all. Actually, their relative performances got better with regard to the adjusted success rates.

Table 2-6: the success rate of all algorithms in modified network

Algorithm	BFS(b)	RWS(r)	WTS(w)	STS(s)	BCS(h)	ISS(i)	CSS(c)	HDS(d)
Success (%)	81.9	76.4	79.0	73.1	81.7	81.5	81.8	81.5
Adjusted(%)	100	93.2	96.4	89.3	99.7	99.5	99.9	99.4

However, in Figure 2-10, which shows the changes in average path length of successful queries with this modified network, we can see that BFS is the only one that is not clearly affected. BCS, HDS, and CCS algorithms are affected much more. This suggests their sensitivity to those highly connected nodes. As well, the performance rank of these algorithms did not change.

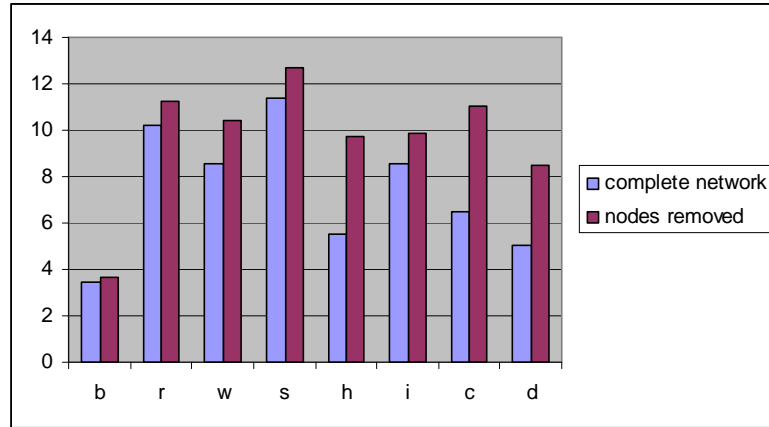


Figure 2-10: Comparison of average path length of successful queries

Interestingly, although designed from different perspectives, these algorithms still are affected by the change of network characteristics. Note these changes result from losing some specific ties or people who are particularly useful for the strategies used. A good example is the ISS strategy: While its design does not consider the effect of weak ties (as in Yu and Singh), the removal of weak ties changes its performance considerably.

SUMMARY

Searching within social networks has gained more theoretical support over the last decade with a better understanding of network dynamics and structure. However, compared to approaches automatizing the small world problem, we know relatively little about searching for expertise in social networks

Searching for expertise is not only affected by the graph characteristics of the network, such as the degree distribution, but also social characteristics of the network, such as people's social interactions and expertise. A human social network is not simply a graph structure, it also includes different social characteristics.

We used a simulation on an organization's email data set, compared three families of searching strategies that utilize both graph and social characteristics of the derived

social network, and then explored the algorithms' tradeoffs and social characteristics. Our results indicate these characteristics can affect the searching process in important ways:

- The relative rank of different algorithms changes little when examining social costs.
- The Information Scent Strategy's advantage (IIS), surprisingly, is not obviously better than out-degree based strategies (BCS and HDS). IIS's performance is close to the Weak Tie Strategy (WTS). Furthermore, we actually found that it also tends to use high out-degree nodes more frequently than low out-degree nodes
- As Granovetter suggested, when compared to the Strong Tie Strategy (STS), the Weak Tie Strategy (WTS) is better. Furthermore, when the weak ties are removed, we also found that performance of IIS also decreased considerably. This indicates weak ties are likely to be critical for automated or augmented expertise finding.
- Our findings confirmed that out-degree based strategies, such as BCS and HDS, in networks like Enron's social network, have a clear advantage over other strategies. However, a very few nodes turn out to be very key in affecting the performance of such social network searching. .
- Simulation, in combination with carefully considered data and analysis, can be very useful in exploring the complex relations among different strategies, social costs, and social characteristics of networks.

As a first-step study, our findings can provide insights for designing future social network based information searching systems. They also open up some interesting avenues for further research. We plan to further look at how the information scent strategy (ISS) really works and its correlation with degree distributions. Then, based on that work, we will try exploring some mixed, dynamic, and learning strategies. We are planning to extend our simulation to examine people's availability and related issues by

using data from people's email exchange patterns. If possible, we are also planning to run the simulations on other data sets.

REFERENCES

1. Ackerman, M. S., Pipek, V., Wulf, V. *Sharing Expertise: Beyond Knowledge Management*, MIT Press, Cambridge MA, 2003.
2. Ackerman, M.S., Boster, J., Lutters, W., McDonald, D. Who's there? The knowledge mapping approximation project, in Ackerman, M. S., Pipek, V., Wulf, V. *Sharing Expertise: Beyond Knowledge Management*, MIT Press, Cambridge MA, 2002.
3. Adamic, L.A., and Adar. E. How to search a social network. *Social Networks*, 27(3), 2005, 187-203.
4. Adamic, L.A., Lukose, R.M., Puniyani, A.R., and Huberman, B.A. Search in power-law networks. *Physics Review E*, 64(46135), 2001.
5. Axelrod, R. Advancing the Art of Simulation in the Social Science, *Simulating Social Phenomena*, 1997.
6. Bernard, H. R., Killworth, P. D., McCarty, C. Index: An informant-defined experiment in social structure. *Social Forces*, 61 (1), 1982, 99-133.
7. Burt, R.S. The network structure of social capital. *Research in Organizational Behavior*. JAI Press, 2000, forthcoming.
8. Cohen, W. Enron Email Dataset, <http://www-2.cs.cmu.edu/~enron/>
9. Dodds, P. S., Muhamad, R., Watts, D. J. An Experimental Study of Search in Global Social Networks. *Science*, 301, 2003, 827-829 .
10. Nardi, BA., Whittaker, S., and Schwarz, H. It's not what you know, it's who you know: work in the information age. *First Monday*, 5, 2000.
11. Foner, L. Yenta: A multi-agent, referral-based matchmaking system. *In Proceedings of the 1st International Conference on Autonomous Agents*, 1997, 301-307.
12. Granovetter, S. The strength of weak ties. *American Journal of Sociology*, 78, 1973, 1360-80.
13. Hamming, R.W. Error-detecting and error-correcting codes, *Bell System Technical Journal*, 29(2), 1950, 147-160.
14. Hutchins, E. *Cognition in the Wild*, MIT Press, 1995.
15. Kautz, H., Selman, B., and Shah, M. The hidden Web. *AI Magazine*, 18(2), 1997, 27-36.

16. Killworth, P., and Bernard, H. Reverse small world experiment. *Social Networks*, 1, 1978, 159-192.
17. Kleinberg, J. Navigation in a small world. *Nature*, 406, 2000, 845.
18. McDonald, D. W. and Ackerman, M.S. Expertise Recommender: A Flexible Recommendation Architecture. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW '00)*, 2000, 231-240.
19. Milgram, S. The small-world problem. *Psychology Today*, 1, 1967, 62-67.
20. Newman, M.E.J. A measure of betweenness centrality based on random walks, Arxiv preprint cond-mat/0309045, 2003.
21. Russell, S., and Norvig, P. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 1995.
22. Whittaker, S., Jones, Q., Terveen, L, Contact Management: Identifying Contacts to Support Long-Term Communication. *Proceedings of the ACM Conference on Computer Supported Cooperative Work.*, 2002, 216-225.
23. Streeter, L.A. and Lochbaum, K.E., Who Knows: A System Based on Automatic Representation of Semantic Structure. RIAO, 1988, 380-388.
24. Travers, J., Milgram, S., 1969. An experimental study of the small world problem. *Sociometry*, 32, 425-443.
25. Wasserman, S., Faust, K., Iacobucci, D, and Granovetter, M. *Social Network Analysis: Methods and Applications*, Cambridge University, 1994, 130-142.
26. Watts, D. J., Dodds, P. S., Newman, M. E. J. Identity and search in social networks. *Science*, 296, 2002, 1302-1305.
27. Wegner, B., Erber, R., and Raymond, P. Transactive Memory in Close Relationships, *Journal of Personality and Social Psychology*, 61 (6), 1991, 923-929.
28. Yang, S. B., and Garcia-Molina, H. Improving search in peer-to-peer networks. *In Proceedings of 22nd International Conference on Distributed Computing Systems*, 2002, 5-14.
29. Yates, R.A, Ribeiro, B. *Modern Information Retrieval*. ACM Press/Addison-Wesley, 1999.
30. Yu, B., and Singh, M.P. Searching Social Networks, *Proceedings of Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*, 2003, 65-72.

31. Yu, B., Venkatraman, M., and Singh, M.P. An Adaptive Social Network for Information Access: Theoretical and Experimental Results, *Journal of the Applied Artificial Intelligence*, 17 (1), 2003, 21-38.

CHAPTER 3

EXPERTISE NETWORKS IN ONLINE COMMUNITIES

INTRODUCTION

Steve is a Java programmer who just started working on a project using Java Speech on a new mobile platform. But he cannot run his first Java Speech program on the new platform and needs some help. Steve is unable to tell whether the problem has arisen because he does not understand how to use the Java Speech package, or because Java Speech does not support the mobile platform well.

It can be difficult to get a satisfactory answer to Steve's problem by searching Google directly. Instead, he may prefer to find and ask someone who has related expertise or experience, and online communities have emerged as one of the most important places for people to seek advice or help. The topics range from advice on medical treatment, programming, software, building a computer from scratch to repairing the kitchen sink. These communities are usually bound by shared professions, interests, or products among their participants. For instance, the Sun Java Forum has thousands of Java developers coming to the site to ask and answer questions related to Java programming every day. The Microsoft TechNet newsgroup is a major place for programmers to seek help for programming questions relating to Microsoft products.

Even though users in these online communities usually do not know each other and are identified using pseudonyms, they are willing to help each other for various reasons, such as altruism, reputation-enhancement benefits, expected reciprocity, and direct learning benefits [16, 18].

This work seeks to enhance online communities with expertise finders. Expertise finders, or expertise location engines, are systems that help find others with the appropriate expertise to answer a question. These systems have been explored in a series of studies, including Streeter and Lochbaum [24], Krulwich and Burkey [17], and McDonald and Ackerman [2] as well as the studies in Ackerman et al. [3]. Newer systems, which use a social network to help find people, have also been explored, most notably in Yenta [12], ReferralWeb [14], and most recently commercial systems from Tacit and Microsoft. These systems attempt to leverage the social network within an organization or community to help find the appropriate others.

Aside from relying on social networks, another interesting characteristic of these systems is that they tend to blur the dichotomy between experts and seekers. They treat one's expertise as a relative concept [3]. In reality, relatively few people will claim themselves as an expert, but many people agree that they have some measure of expertise in some area. These systems allow everyone to contribute as they can.

For these expertise finder systems to be of significant assistance, they must effectively identify people who have expertise in the area desired by the asker. Most current systems use modern information retrieval techniques to discover expertise from implicit or secondary electronic resources. A person's expertise is usually described as a term vector and is used later for matching expertise queries using standard IR techniques. The result usually is a list of related people with no intrinsic ranking order or ranks derived from term frequencies. It may reflect whether a person knows about a topic, but it is difficult to distinguish that person's relative expertise levels. Relying on word and document frequencies has proven to be limited [19].

To ameliorate this, Campbell et al. [8] and Dom et al. [9] used graph-based ranking algorithms in addition to content analysis to rank users' expertise levels. This work, done at IBM Research, applied several graph-based algorithms, including PageRank and HITS, to both a synthetic network set and a small email network to rank correspondents according to their degree of expertise on subjects of interest. They found that using a graph-based algorithm effectively extracts more information than is found in content alone. However, there is a weakness in these studies. The size of their networks is very small and does not reflect the characteristics of realistic social networks.

As a result, we wished to revisit the possibilities of using graph-based algorithms on social networks of users in online communities. In this study, we analyze a large online help seeking community, the Java Forum, using social network analysis methods. We then test a set of network-based algorithms, including PageRank and HITS, on this large size social network. Using a set of simulations, we explore how various network structures affect the performance of these algorithms. We find a small number of structural characteristics in the social networks that we believe lead to differences in the algorithms' performance for online communities. We expect that not only will these characteristics be fruitful for practical algorithm design and implementation, but that they will offer new research insights for others to explore.

The chapter proceeds as follows. In Section 2, we introduce the community expertise network and briefly review related work. In section 3, we describe the network characteristics of our test online community, the Java Forum. In section 4, we describe some expertise ranking algorithms. In section 5, we present an evaluation comparing the rankings produced by human raters and by the algorithms. In section 6, we then explore the network characteristics that affect the performance of these algorithms using a simulation study. And finally, we summarize our findings in Section 7.

EXPERTISE NETWORK IN ONLINE COMMUNITIES

Online communities usually have a discussion thread structure. A user posts a topic or question, and then some other users post replies to either participate in the discussion or to answer a question posed in the original post. Using these posting/replying threads in a community, we can create a post-reply network by viewing each participating user as a node, and linking the ID of a user starting a topic thread to a replier's ID, as shown in Figure 3-1.

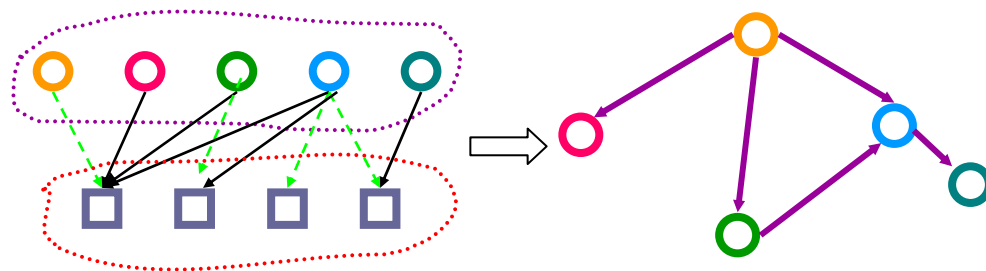


Figure 3-1: We map a replying relationship into a directed graph. On the left we have a bipartite graph of users (circles) and the discussion threads (squares) they participated in. This is transformed to a directed graph where an edge is drawn from the user making the initial post (the dashed edge shown in green) to everyone who replied to it.

This post-reply network has some interesting characteristics. First, it is not intentionally built by its users for the purpose of forming ties. Thus, it is not a network focused on social relationships. Instead, it reflects community members' shared interests. Whether it is a community centered on questions and answers, social support, or discussion, the reason that a user replies to a topic is usually because of an interest in the content of the topic rather than who started the thread. This indirectly reflects a particular shared interest between the original poster and the repliers (although the repliers' sentiment about the topic may differ).

Furthermore, in a question and answer community, the direction of the links carries more information than just shared interest. A user replying to another user's question usually indicates that the replier has superior expertise on the subject than the asker. The distribution of expertise, along with the network of responses, is what we will call the *community expertise network (CEN)*. It indicates what expertise exists within an online community, as well as how it is distributed in practice.

The full dynamic of a CEN may be much complex in some communities. For example, there may be trolls, spammers, etc. An answer thread to a question can be the result of a complex social process and the first few replies may actually not answer the question but try to clarify the problem. The network could be weighted according to the frequency of how often a user helps another. We will discuss these issues in later sections.

Structural Prestige in Social Networks

Expertise is closely related to structural prestige measures and rankings in social network studies. In directed networks, people who receive many positive choices are considered to be prestigious, and prestige becomes salient especially if positive choices are not reciprocated [25].

Researchers in various fields have applied these prestige ideas to different types of networks. Fisher et al. [11] used social network visualization and analysis on the patterns of replies for each author in selected newsgroups to find different types of participants. For instance, they used the indegree (how many people a user replied to) and outdegree (how many people replied to the user) of a user's egocentric network to identify the roles within the group (e.g., general asker or replier). Bollen et al. [5] used a similar ranking measure to evaluate the prestige of academic journals. Liu et al. [20]

used it to evaluate the impact of an individual author in a co-authorship network. And, of course Page et al. [22] used PageRank to rank web pages.

In online help-seeking communities, the social network is an expertise network. Because the way links are constructed, the prestige measure of the network is highly correlated with a user's expertise. Thus, this hints that there are opportunities to make use of such network structures to rank people's expertise in online communities, and build related applications/systems that further improve the expertise sharing in the online world.

Next we turn to the investigation of an expertise network in one online community, the Java Forum.

EMPIRICAL STUDY OF AN ONLINE COMMUNITY

The Java Forum

The Java Developer Forum is an online community where people come to ask questions about Java. It has 87 sub-forums that focus on various topics concerning Java programming. There is a large diversity of users, ranging from students learning Java to the top Java experts. Users usually can get an answer relatively quickly because of the large number of participants. In this study, we used the Java programming sub-forum (called here "Java Forum"), which is a place for people to ask general Java programming questions. The Java Forum had a total of 333,314 messages in 49,888 threads.

We used the network constructed upon these threads to evaluate the usefulness of our expertise-ranking algorithms. The Java Forum network had 13,739 nodes and 55,761 edges.

The next section describes the characteristics of the Java Forum network. This will provide both a test bed for the algorithms and, later in the chapter, will help in

understanding the underlying network characteristics that expertise ranking algorithms operate upon.

Characterizing the Network

The Bow tie structure analysis

Not all users in the Java Forum ask questions, nor do all users answer questions. Using a bow tie structure analysis, we examine the general structure of the Java Forum network.

The bow tie structure, first proposed by researchers at IBM, AltaVista, and Compaq, yields insights into the complex organization of the Web network structure. Its key idea is that the web is a bow tie and has four distinct components: Core, In, Out, and 'Tendrils' and 'Tubes' (see Broder et al. [7]). In our bow tie model, a central core contains users that frequently help each other. It is a strongly connected component (SCC), meaning that one can reach every user from every other by following questioner-answerer links. The 'In' component contains users that usually only ask questions. The 'Out' consists of users that usually only answer questions posted by users in the Core. Other users, the 'Tendrils' and 'Tubes', connect to either the 'In' or 'Out' clusters, or both, but not to the Core. They are users who only answer questions posed by 'In' users or whose questions are only answered by 'Out' users.

Figure 3-2, 3-3 and Table 3-1 compare the bow tie structure of the Java Forum network with that of the Web (as reported in [7]).

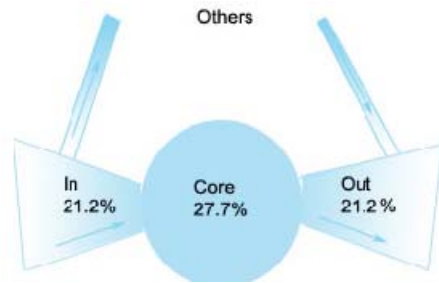


Figure 3-2: The web is a bow tie

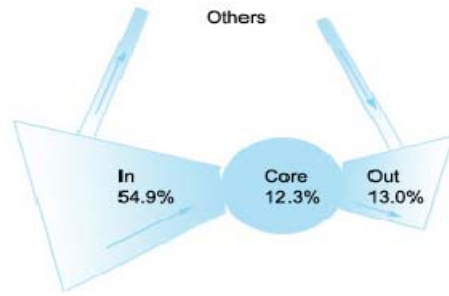


Figure 3-3: The Java Forum network is an uneven bow tie

Table 3-1: Comparison of bow tie analysis between Web and the Java Forum network

	Core	In	Out	Tendrils	Tubes	Disconnect
Web	27.7%	21.2%	21.2%	21.5%	0.4%	8.0%
Forum	12.3%	54.9%	13.0%	17.5%	0.4%	1.9%

These results show the Java Forum network looks much different from the Web. The Java Forum has a much bigger 'In' component and a relatively smaller Core than the Web. This indicates that in this online community, only about 12% of users actively ask and answer questions for each other. More than half of the users usually only ask questions, and about 13% users usually only answer questions. This result also indicates that instead of being a public place where people help each other reciprocally, this online help seeking community is more closely a place where askers come to seek help from volunteer helpers.

Distribution of degree

We can use the bow tie structure to show the role of users in the network, but it does not capture the level of their interaction. Looking at degree distributions is a general way to describe users relative connectedness in a large complex network [21]. The degree distribution is a function describing the number of users in the network with a given degree (number of neighbors). An interesting common feature of many known complex networks is their scale-free nature. In a scale-free network, the majority of nodes are each connected to just a handful of neighbors, but there are a few hub nodes that have a disproportionately large number of neighbors. Figure 3-4 shows the indegree distribution histogram for the Java Forum network. It is highly skewed (and in fact scale-free except for a cutoff at very high degrees), similar to a distribution observed for Web pages and for co-authorship networks. The scale-free degree distribution is a reflection of the highly uneven distribution of participation. Instead of everybody helping each other equally, in the Java Forum, there are some extremely active users who answer a lot of questions while a majority of users answer only a few. Likewise, many users ask only a single question, but some ask a dozen or more.

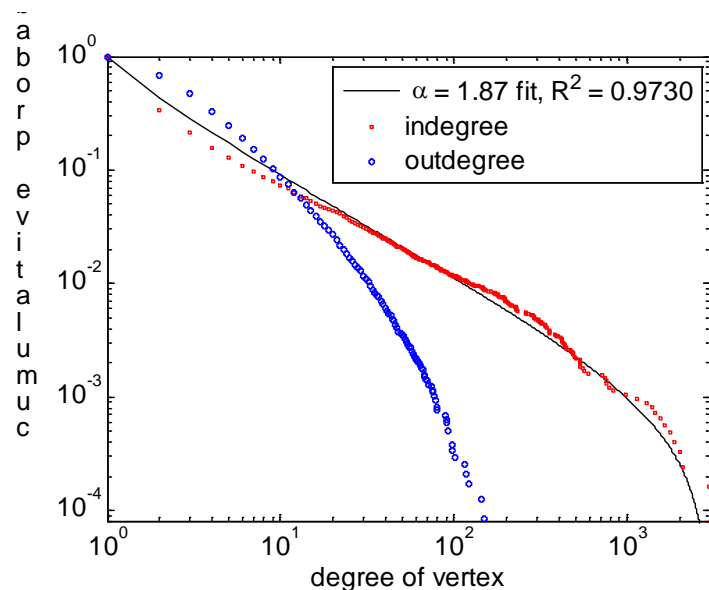


Figure 3-4: Degree distribution of the Java Forum network

Degree correlations

While the indegree distribution shows how many people a given user helps, it gives no information about those users' own tendency to provide help. For example, one might like to know whether high volume repliers only reply to newbies, or if they mostly talk to others similar to themselves. We can answer both of these questions by looking at the correlation profile (see Maslov et al. [23]) Here we consider a simplified correlation profile that for each asker-replier pair counts the indegree of the replier versus the indegree of the asker, as shown in Figure 3-55. We also report a simple correlation coefficient between the askers' and helpers' indegree.

Positive assortativity is common in social networks, where people with many connections tend to know other people with many connections while hermits tend to know other hermits. We find however, that the Java Forum is far from an exclusive club where high volume repliers correspond with other high volume repliers, leaving the newbies to talk to one another. Rather, the Java forum is neither assortative nor disassortative. The correlation coefficient is ever so slightly negative at -0.013 , and the correlation plot shows that the highest degree nodes (usually the experts) tend to answer questions across the board from whoever asks them. As one might expect, low degree users (ones who probably lack the expertise to answer others' questions) typically do not reply to high-degree users.

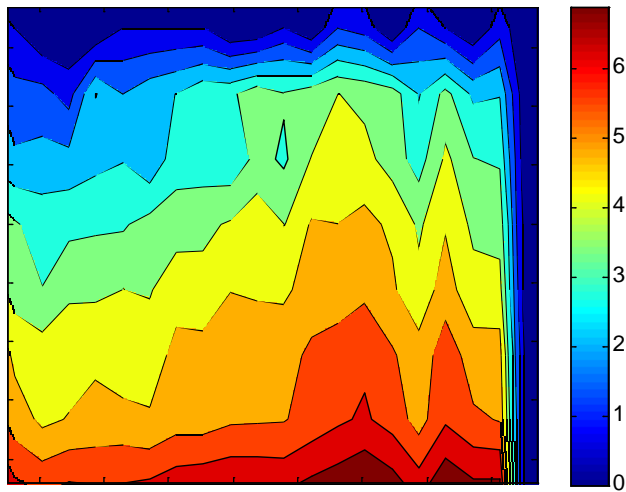


Figure 3-5: The correlation profile of the Java Forum network. The color corresponds to the logarithm of the frequency of such degree pairings.

In summary, from these network analyses, we can see that the Java Forum network has some unique characteristics, including:

- Different groups of users fall into structurally distinct parts of the network: There is a big 'In' group and relatively small Core and 'Out' groups.
- The users' indegree distribution is skewed, with few users answering a large number of questions while the majority of users only answer a few.
- Top repliers answer questions for everyone. However, less expert users tend to answer questions of others with lower expertise level.

Since these characteristics are different from the World Wide Web graph, they can potentially affect the performance of various expertise ranking algorithms, as we will discuss next.

EXPERTISE RANKING ALGORITHMS

After constructing an expertise network from the post-reply patterns in the online community, and having discovered interesting regularities in the structure of the network

which might correlate with a user's expertise, we now present several algorithms designed to automatically infer a user's expertise level. After presenting the algorithms, we will provide the results of their tests.

Simple Statistical Measures

We surmise that if a person answers a lot of questions on a topic, it is often the case that he or she knows the topic well. Exceptions include spammers who may be posting advertisements or trolls who may be making inflammatory or otherwise disruptive posts. We found little trolling or spamming behavior on the Java Forum. However, our observations here would also be applicable to forums where spamming is more prevalent, but can be curbed or identified through users' relevance feedback. Returning to the Java Forum, the simplest method for evaluating a user's expertise may be counting the number of questions answered. We call it the "AnswerNum" measure.

A slightly different measure is counting how many other users a user helped. Some users may have a big AnswerNum but all these replies are answering questions repeatedly from several specific users. On the other hand, a user who posts fewer answers, but in the process helps a greater number of users, could have broader or greater expertise. Thus, counting how many people one helps may be a better indicator than counting the number of replies. In a social network, this could be calculated using the indegree of a node.

Z-score Measures

While replying to many questions implies that one has high expertise, asking a lot of questions is usually an indicator that one lacks expertise on some topics. Thus, we propose the "z-score" as a measure that combines one's asking and replying patterns, as shown in following formula: If a user makes $n=q+a$ posts, q of them questions and a of

them answers, we would like to measure how different this behavior is from a ‘random’ user who posts answers with probability $p = 0.5$ and posts new questions with probability $1-p = 0.5$. We would expect such a random user to post $n \cdot p = n/2$ replies with a standard deviation of $\sqrt{n \cdot p \cdot (1-p)} = \sqrt{n}/2$. The z-score measures how many standard deviations above or below the expected ‘random’ value a user lies:

$$z = \frac{a - n/2}{\sqrt{n}/2} = \frac{a - q}{\sqrt{a + q}}$$

If a user asks and answers about equally often, their z-score will be close to 0. If they answer more than ask, the z score will be positive, otherwise, negative. We calculate the z-score for both the number of questions one asked and answered and the number of users one replied to and received replies from, denoted separately as “Z_number” and “Z_degree”.

ExpertiseRank Algorithm

There is a potential problem in counting the number of answers one posted or the number of people one helped. A user who answers 100 newbies’ questions will be ranked as equally expert as another user who answers 100 advanced users’ questions. Obviously the latter usually has greater expertise than the former.

The well known PageRank algorithm, proposed by Page et al. [22] for ranking web pages, improves this. It provides a kind of peer assessment of the value of a Web page by taking into account not just the number of pages linking to it, but also the number of pages pointing to those pages, and so on. Thus, a link from a popular page is given a higher weighting than one from an unpopular page. Intuitively, the ranking in PageRank corresponds to the fraction of time a random walker would spend ‘visiting’ a page by iteratively following links from page to page. There are various versions of PageRank or similar measures; for an overview, see [4, 6].

We propose using a PageRank-like algorithm to generate a measure that not only considers how many other people one helped, but also whom he/she helped. We call it “ExpertiseRank”. The intuition behind ExpertiseRank is that if B is able to answer A’s question, and C is able to answer B’s question, C’s expertise rank should be boosted not just because they were able to answer a question, but because they were able to answer a question of someone who herself had some expertise. In a sense, ExpertiseRank propagates expertise scores through the question-answer network.

Table 3-2 lists the ExpertiseRank algorithm that is similar to PageRank.

Table 3-2: Basic ExpertiseRank algorithm

Assume User A has answered questions for users $U_1 \dots U_n$, then the ExpertiseRank (ER) of User A is given as follows:

$$ER(A) = (1-d) + d (ER(U_1)/C(U_1) + \dots + ER(U_n)/C(U_n))$$

$C(U_i)$ is defined as the total number of users helping U_i , and the parameter d is a damping factor which can be set between 0 and 1. We set d to 0.85^2 here. The damping factor allows the random walker to ‘escape’ cycles by jumping to a random point in the network rather than following links a fraction $(1-d)$ of the time.

ExpertiseRank or ER (A) can be calculated using a simple iterative algorithm.

Note that an expertise network could be weighted. For instance, we can add values to edges by how frequent one replies another. We can also weight each ask-reply occurrence differently based on how many replies there are in a question thread. It is straightforward to extend the notion of Expertise rank to incorporate the weights of the edges by substituting $ER(U_i)$ with $ER(U_i) * w_{iA}$, where w_{iA} is the number of times i was

² We tried various values (such as 0.95 and 0.70), but it did not make a significant difference.

helped by A and $C(U_i) = \sum w_{ij}$. In our particular study, we found that weighting does not improve the accuracy of our results, so for simplicity we treat the networks as unweighted, although weights can easily be reintroduced for other applications.

HITS Authority

Another ranking algorithm similar to PageRank is HITS (“Hypertext induced topic selection”) [15]. It also uses an iterative approach, but assigns two scores to each node: a hub score and an authority score. In our context, a good hub is a user who is helped by many expert users, and a good authority (an expert) is a user who helps many good hubs. The definition is recursive and converges after a few iterations. In our study, we used the Authority value of HITS to correspond to the expertise rank of the user.

EVALUATION

Since there was no explicit user-supplied expertise ranking data in the Java Forum, we needed to use human raters to generate a “gold standard” for comparison. Because it was not possible for us to rate a large number of these users, we randomly selected 135 users from the network for use as a comparison sample. By omitting those users posting fewer than 10 times, we ensured that the sampled users had generated enough Forum content for a reviewer to evaluate their expertise levels.

While some of the ranking algorithms such as ExpertiseRank and HITS can in principle produce continuous values that can potentially differentiate between all users, it is very difficult for humans to sort 135 users into a ranked list. Raters must read from ten to hundreds of messages posted by a user to evaluate his/her expertise level. It is also difficult to compare two users when they both have posted many messages but have not replied to each other.

Based on our observation of the forum and the results of a pilot rating set, we decided to categorize the users into 5 expertise levels instead of a complete ranked list. Table 3-3 displays details of these categorizations.

Table 3-3: Five levels of expertise rating

Level	Category	Description
5	Top Java expert	Knows the core Java theory and related advanced topics deeply.
4	Java professional	Can answer all or most of Java concept questions. Also knows one or some sub topics very well,
3	Java user	Knows advanced Java concepts. Can program relatively well.
2	Java learner	Knows basic concepts and can program, but is not good at advanced topics of Java.
1	Newbie	Just starting to learn java.

We found two raters who are Java programming experts to rate the 135 users' expertise. (These experts were not part of the research team; they were independent consultants.)

Statistical Metrics

Two of the most frequently used correlation measures between two ranks are Spearman's rho and Kendall's Tau [10, 13].

Both of these metrics have their limitations. The Spearman correlation does not handle weak orderings well (weak ordering means that there are multiple items in the ranking such that neither item is preferred over the other) and our rankings have a lot of weak orderings because multiple users are assigned the same rating. Kendall's Tau, on the other hand, gives equal weight to any interchange of equal distance, no matter where it occurs. For instance, an interchange between rank 1 and 2 will be just as bad as

interchange between rank 100 and rank 101. Kendall's Tau may be a better metric for our purpose. Nevertheless, for the evaluation, we present both Kendall's Tau and Spearman's rho. Furthermore, we have also added a "TopK" metric, which calculates a Kendall's Tau for only the highest 20 ranks.

After each human rater submitted his ratings, we tested the reliability of raters by looking at their inter-rater correlation. The Kendall's Tau distance between the two human raters was 0.736, and the Spearman's rho correlation coefficient was 0.826 ($p < 0.01$), a sufficiently high rate of inter-rater correlation.

Results

To have a conservative measurement of the possible performance for the automatic algorithms, we further removed 10 samples whose ratings have more than 1 level difference between the two raters. The Spearman's rho is 0.832 and Kendall's Tau is 0.796 between the two raters for the 124 users left. (One user was not rated because raters reported that they didn't have enough evidence.) Therefore, we may expect that any automated algorithm would at best achieve around a 0.8 correlation with the human raters. For each of these users, in the data analysis below, we summed the ratings from the two raters together as the standard human rating (HR).

Figure 3-6 shows the statistical correlations between various algorithms and the human ratings of the 124 users. (A sensitivity analysis including all 134 users showed insignificant differences.)

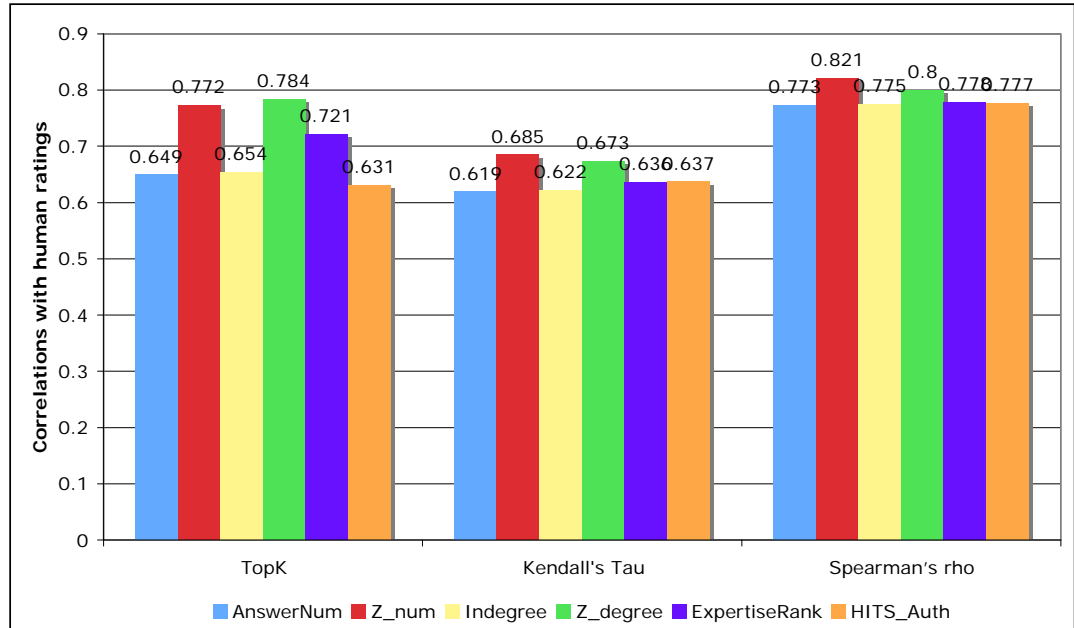


Figure 3-6: The performance of various algorithms in different statistical metrics

From the figure, one can see that all of these ranking algorithms give a relatively high correlation with the human-assigned ratings. This tells us that, indeed, structural information could be used to help evaluate users' expertise in online community networks.

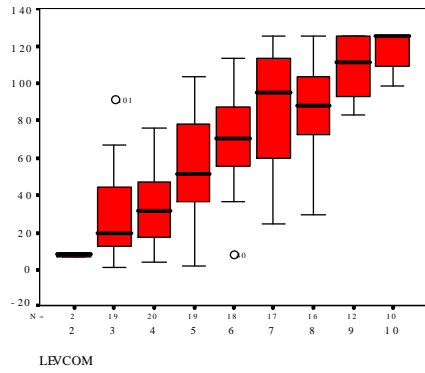
Surprisingly, contrary to what Campbell et al. [8] and Dom et al. [9] found in their simulation studies, we found that, in this real network data set, ExpertiseRank actually does not perform better than other simpler methods. Instead, the z-score-based ranks tend to produce slightly better results than other methods. We will return to this in the subsequent analysis, where we try to find social network features that explain this result.

We can also see that different correlation metrics produce different results when comparing the same data. For instance, while Z_degree shows the highest correlation with the TopK metric, it is the Z_number that shows the highest correlation with the complete Kendall's Tau and Spearman's Rho metrics. In many applications, we may care more about whether the algorithm can identify the top K experts, rather than whether it can rate everyone's relative expertise. Being aware of these differences in metrics can

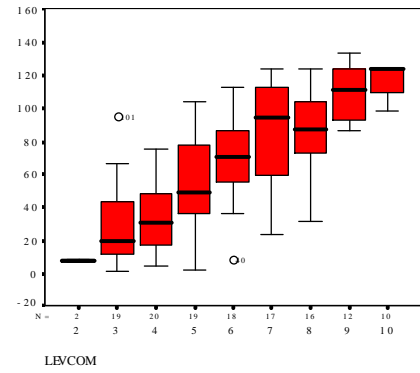
help one choose an appropriate algorithm depending on whether it is the top experts one is after.

We further looked at the distribution of automatic rankings (summarized by the box plots shown in figure 3-7) corresponding to the human rating levels³. From these box plots, we can see the results are consistent with what we found in Figure 6. We can see that the Z_number, Z_degree, and ExpertiseRank all have a slightly smaller inter-quartile range at each human rating level, which indicates that they typically have smaller errors.

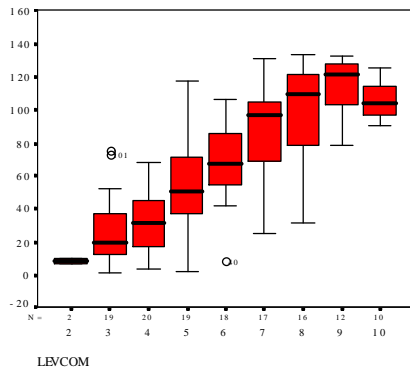
³ We use the rating combination of two raters here, so there is a total of 10 categories.



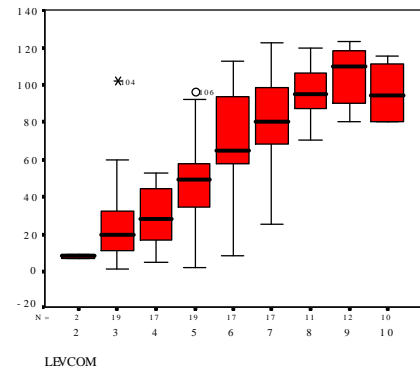
(a). AnswerNum



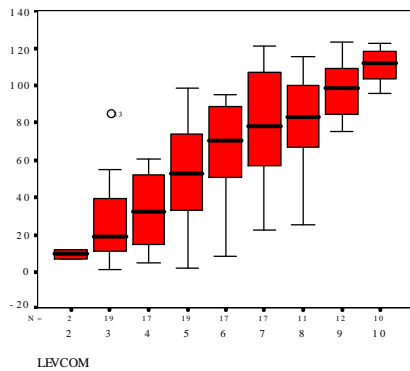
(b). Indegree



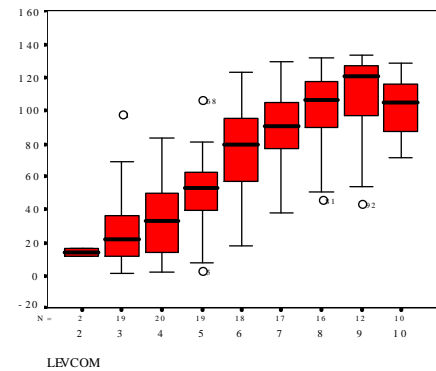
(c). Z_number



(d). Z_degree



(e). HITS_Authority



(f). ExpertiseRank

Figure 3-7: Box plots of algorithm rankings vs. human ratings

While it is interesting to look at the details of these results, it is more important to think about the big picture. We have observed a network structure different from the Web, and we have also seen that some algorithms, such as PageRank and HITS, which excel at ranking Web pages, do not outperform simpler algorithms in this network. The key to understanding the performance of the algorithms is in understanding the human

dynamics that shape an online community. This understanding will then help select algorithms that may be more appropriate for other online communities where the dynamics may be different from the Java Forum. The approach we took was simulation: taking the simplest set of interaction rules that both replicated the observed structure and the relative performance of various algorithms.

We next present the results of those simulations.

SIMULATIONS

Much recent work on modeling of complex networks in social, biological and technological domains has focused on replicating one or more aggregate characteristics of real world networks, such as scale-free degree distributions, clustering, and average path lengths[21]. For instance, the preferential attachment network growth model of Barabasi et al. [1], where new nodes joining preferentially connect to well connected nodes, yields scale-free degree distributions.

Here, we take a different approach. We place an emphasis on studying the various factors that possibly affect the structure of the network. Instead of having a targeted network to generate, we let various factors determine the growth of the network and observe how changes in those factors affect the structure of the network. Figure 3-8 shows a snapshot of the simulator we developed to study how these various network characteristics (the corresponding controls are hidden in the figure) will affect the structure of the network in an online help-seeking community and in turn how they affect the performance of various ranking algorithms (shown in the plots and tables adjacent to the network layout). Details of this simulator can be found in [26].

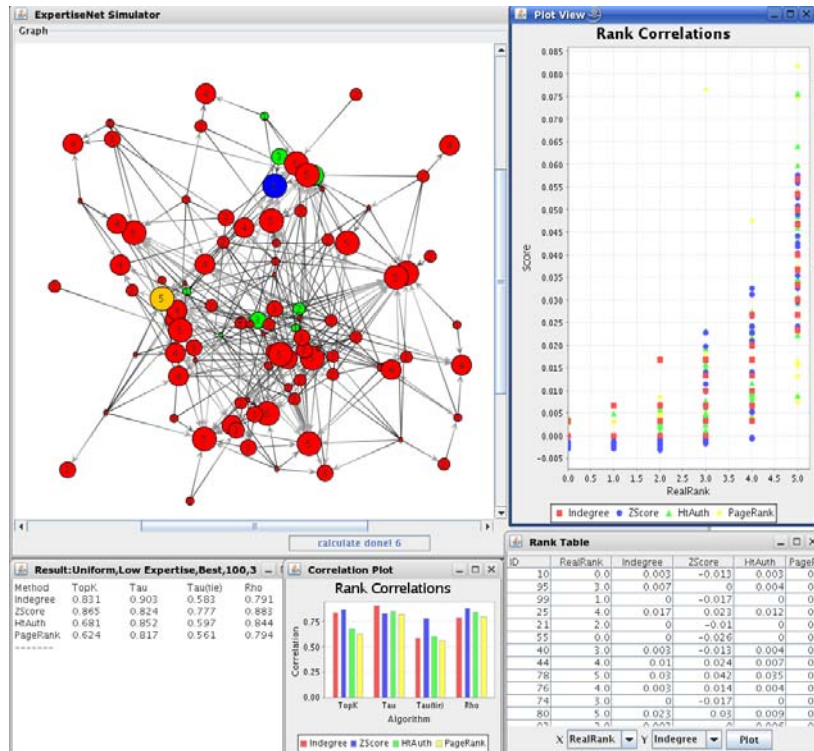


Figure 3-8: Snapshot of the network simulator interface

Modeling Java Forum's Network

From the empirical analysis of the Java Forum, we incorporated the following dynamics governing the forum into our model:

- The majority of users made few posts, either because they were new or had low expertise.
- There were a number of experts who mainly answered others' questions and seldom asked questions themselves.
- Users seemed to answer others' questions according to their own ability corresponding to their level of expertise.

First, we initialized the community with 1,374 users in the community (one-tenth of the observed population of the Java Forum) with a power law distribution for the levels of expertise. There were many level 1 (novice) users and relatively few level 5 (expert) users.

Second, we modeled that low-level users have high probabilities to ask questions. A user u with expertise level $L(u)$ has the probability to ask questions $P_A(u)$ determined by the formula below:

$$P_A(u) = \frac{(L(u) + 1)^{-1}}{\sum_v (L(v) + 1)^{-1}}$$

Third, we modeled which users were most likely to answer a question posed by a user a with expertise level $L(a)$ by using a “best preferred expert” rule, where the probability $P_H(u,a)$ of replying increases exponentially with the expertise level difference between the two users:

$$P_H(u,a) = \frac{\text{Exp}(L(u) - L(a))}{\sum_v \text{Exp}(L(v) - L(a))}$$

Note that according to this formula, even a user with a lower level of expertise than the asker has a small probability of answering the question, just as is the case in the actual Java Forum.

After setting up the model, we ran the simulation to generate networks. At each step, an asker was picked to ask a question and a helper was picked to answer based on the related probabilities.

After we ran the simulation for 5576 steps, we got a network with the same average degree as the Java Forum network. From scaled down versions, shown in

Figure 3-8 and Figure3-13, one can see that in this model, most of links are from low expertise (small nodes in the network visualization) to high expertise (big nodes).

Then, we analyzed the degree distribution of the simulated network to test whether it was similar to the Java Forum network. By comparing Figure 3-9 with Figure 3-4, one can see that while the indegree distribution replicates the heavy skew of the empirical network, the outdegree distribution does not. There are not as many single-post askers with low

outdegree (0, 1, etc) in the simulated network. This is to be expected, since we are not modeling the growth dynamics where newcomers, by virtue of not being in community long enough to ask a large number of questions, contribute to the lower end of the distribution. When we updated our simulation to allow users to join the community with some probability at in each step, we were able to replicate the outdegree distribution (shown in Figure 3-10).

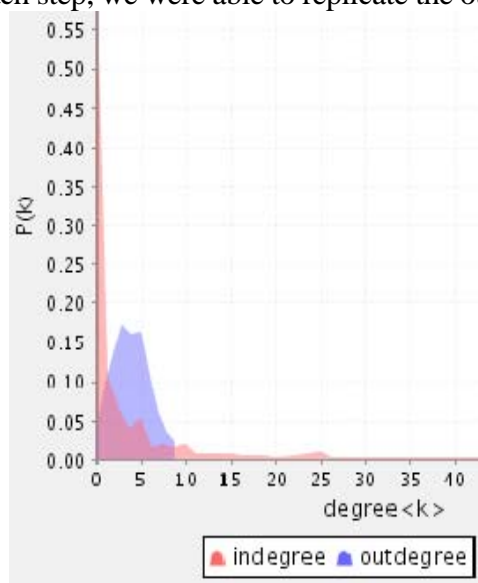


Figure 3-9: Simulated degree distributions with 'best preferred' helpers

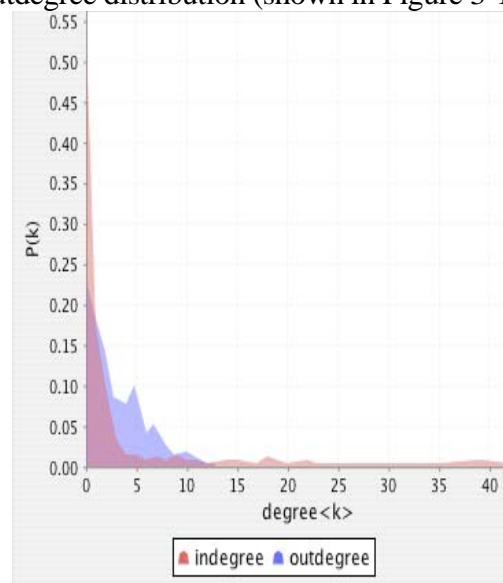


Figure 3-10: Simulated degree distributions with a growing network

We further looked at other characteristics of the network. Table 3-4 shows that the bow tie structure of the simulated network is similar to the Java Forum network. The only significant difference is that we have a relatively larger portion of disconnected users. This is because in the simulation, we built the network based on posting-replying patterns, but in the Java Forum, the lurkers (corresponding to disconnected nodes in our network) do not post in the community and therefore are not part of the empirical network.

Table 3-4: Bow tie structure of the 'best preferred' network

Core	In	Out	Tendrils	Tubes	Disc
13.8%	59.7%	3.6%	5.1%	1.0%	13.7%

Figure 3-11 shows that the indegree correlation profile fits rather closely with that of the Java Forum network. The correlation between asker and helper indegree is indistinguishable from 0 ($\rho = 0.009$, $p = 0.35$)

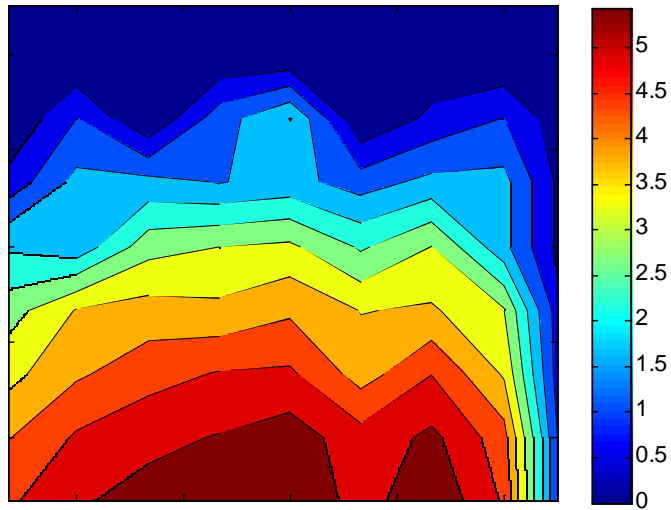


Figure 3-11: Degree correlation profile of the “best preferred” network

We tested various algorithms in this network and compared their ranks with the nodes’ assigned rank in the simulation process. Figure 3-12 displays the result.

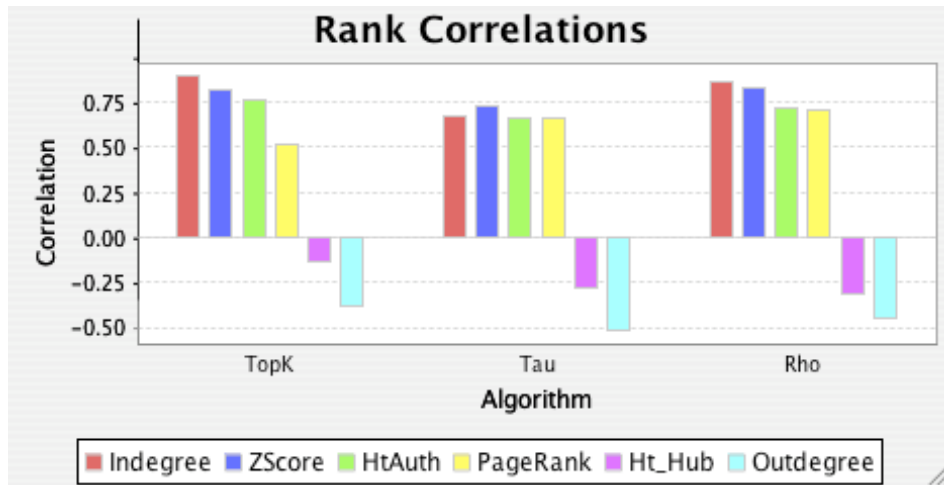


Figure 3-12: Performance of expertise-detection algorithms on the ‘best preferred’ network

From this figure, one can see that algorithms like ExpertiseRank and HITS do not perform better than simpler methods like indegree and z-score, much like what we found

empirically in the Java community. This confirms our intuition that structural differences may be a major reason why complex algorithms like ExpertiseRank do not always work well in various network structures.

An Alternative Network Model

As we saw in the previous section, our simple model dynamics capture both the structural features and expertise ranking algorithm performance of the actual Java Forum. However, not all online expertise communities will follow the same dynamics as the Java Forum. We can glean useful insights by modeling different dynamics and then evaluating the expertise ranking algorithms on the models they create. For example, in other communities, especially ones that may be situated within an organization, experts may be under time constraints and choose to answer only those questions that make best use of their expertise. They would therefore be more likely to answer the questions of those slightly less expert than themselves. It may be the best way for people to make use of one another's time and expertise [2]. Such user behavior was not modeled in our "best preferred" model.

We thus constructed an alternate model, where users who have a slightly better level of expertise than the asker have a higher probability of answering the question, rather than those with a much larger difference in expertise. This model uses a "just better" rule, where a user u 's probability of answering a question posed by user a is decided by the formula below:

$$P_H(u,a) = \frac{\text{Exp}(L(a) - L(u))}{\sum_v \text{Exp}(L(a) - L(v))} \text{ when } L(u) > L(a)$$

Figure 3-14 shows a network generated using this model. In contrast to the "best preferred network" shown in Figure3-13, we can see that the links are not all pointing to

the highest experts. Rather, questions are answered by users with higher, but not highest, expertise.

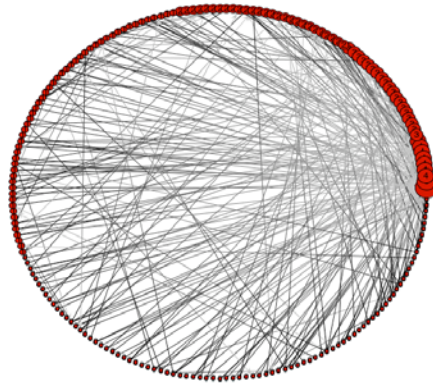


Figure 3-13: 'best preferred' network

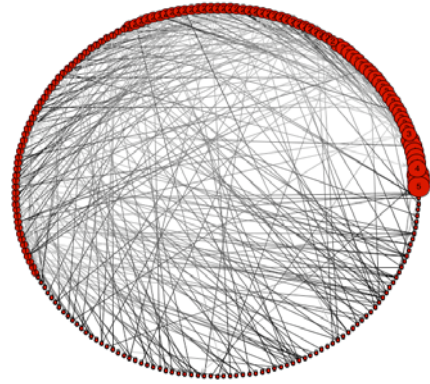


Figure 3-14: 'just better' network

Figure 3-15 shows the degree distribution of the network and Table 3-5 shows the bow tie structure analysis result. They are not very similar to Java Forum (note the very tiny Core in the bow tie structure), but some patterns are close (such as the highly skewed degree distribution and the biggest bow tie part being “In”).

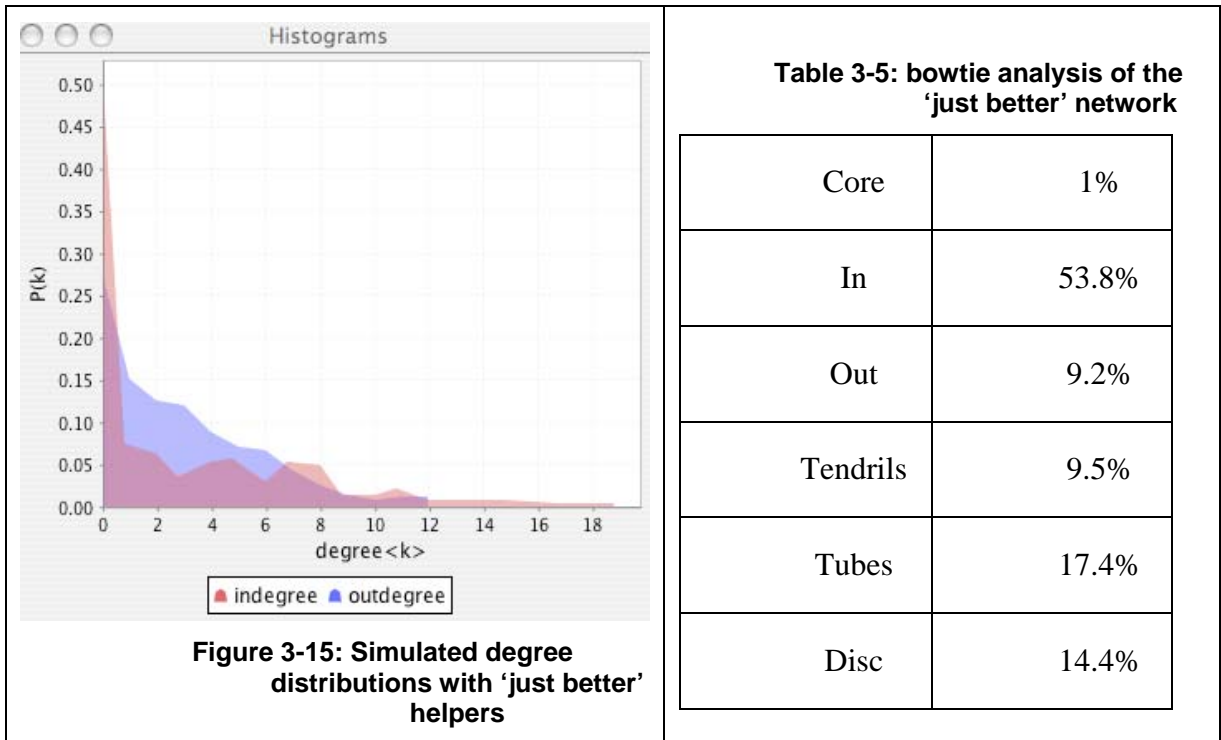


Figure 3-16 shows the degree correlation profile, with an interesting appearance of strong correlation along the diagonal where users are helping those slightly less expert than themselves. At 0.14, the correlation coefficient is positive in contrast to the lack of correlation observed in both the empirical network and the “best preferred” model.

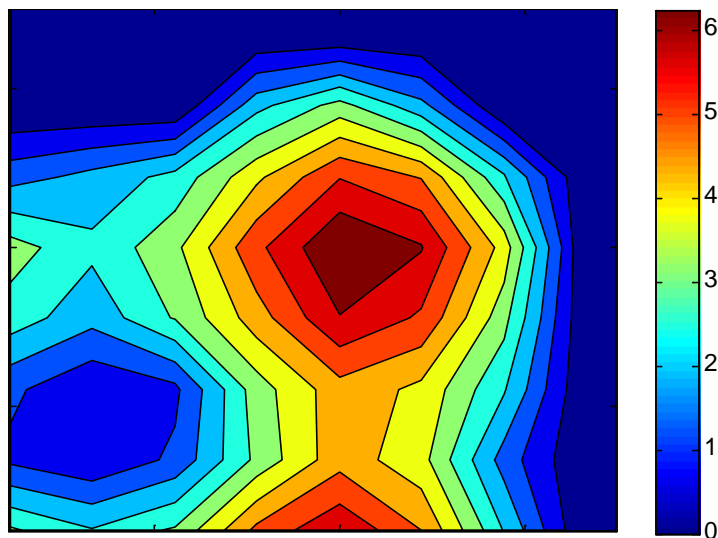


Figure 3-16: Correlation profile of the 'just-better network'

Figure 3-17 displays the performance comparisons of the various ranking algorithms in this new network: ExpertiseRank and Z_score perform the best, and HITS_authority is the worst. Since hubs and authorities reinforce one another in the iterative HITS algorithm, in the ‘best preferred’ network, the newbies who have their questions answered by the best experts reinforce the scores of those experts. However, in the ‘just better’ algorithm, the newbies who are asking the most questions are often helped by users with only slightly higher expertise. Therefore HITS identifies individuals with medium expertise as the highest experts. Similarly Figure 3-18 shows an example of a high expert user who is helping other expert users. Since experts have low HITS hub scores, they thus impart a low HITS authority score to the expert helping them. On the other hand, ExpertiseRank propagates the expertise score from the newbies to the intermediate users who answer their questions and from the intermediate users to the best experts. Thus we expect that PageRank-based algorithms such as ExpertiseRank will in general outperform other algorithms when the askers’ and helpers’ expertise is correlated. The Java Forum did not display this behavior (in fact, it is already very well described by our first model). But, as mentioned, such a scenario is plausible where users make the best use of their time by being more selective in choosing questions that are challenging to them yet they are still capable of answering.

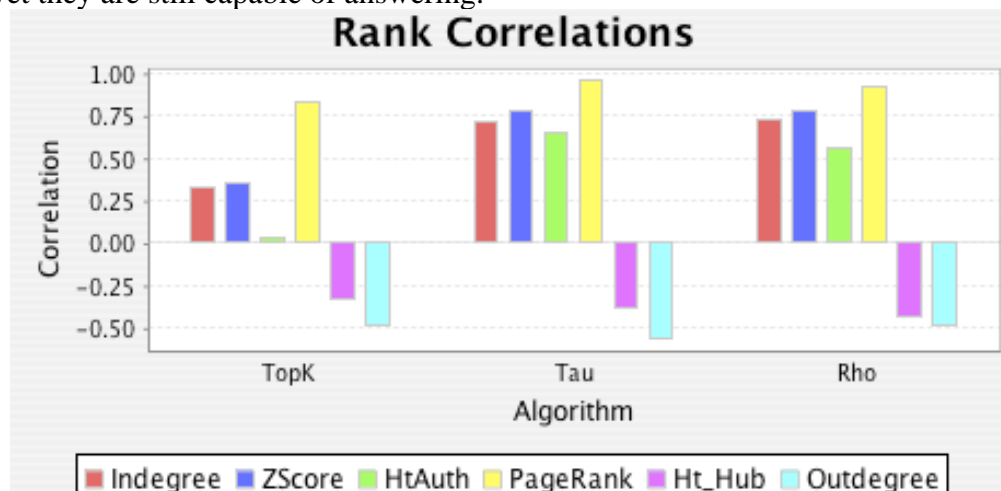


Figure 3-17: Performance of expertise ranking algorithms in the ‘just better’ network

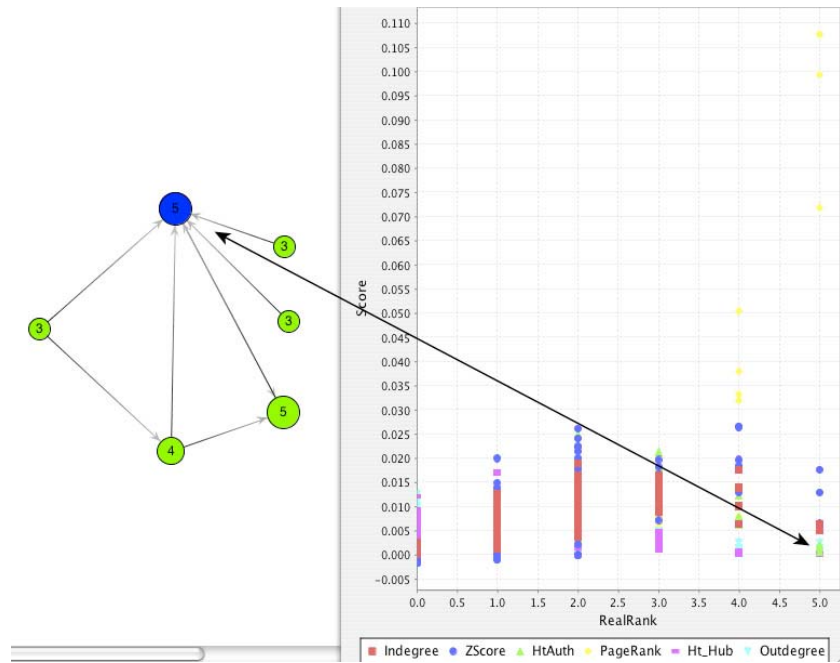


Figure 3-18: A case where a high expertise node has low authority

SUMMARY AND FUTURE WORK

In summary, we wanted to augment how people can help one another in online communities, particularly help-seeking or technical support communities. To do this, we wished to augment what we call the expertise network here – the way that expertise is distributed and deployed in practice.

To do this, we went through three steps. First, we wanted to know what went on socially in a typical help-seeking community. We analyzed the network representing asker-helper interactions in an online community, the Java Forum. Among them were highly skewed degree distributions, much like the graph of the World Wide Web. But unlike the Web, specific dynamics governing this particular forum produce a different bowtie structure and degree correlation profile.

We then ran an evaluation of expertise ranking algorithms – algorithms to analyze the relative expertise of different users – in this community.

To understand the results, we simulated these dynamics and produced networks that not only matched the observed aggregate network characteristics but also allowed us to understand why automated expertise ranking algorithms perform differently in differently structured networks. This understanding should help us weigh the tradeoffs in algorithm design and use for networks we encounter in the future. In fact, it is critical to do so.

In this work, then, we found:

- Structural information *can* be used for evaluating an expertise network in an online setting, and relative expertise can be automatically determined using social network-based algorithms. We also found, however, that the network's structural characteristics matter.
- These algorithms did nearly as well as human raters. However, there were significant tradeoffs among the algorithms. Sometimes a relatively simple measure was as good as more complex algorithms, such as an adaptation of PageRank.
- We believe, and have tested with simulations, that the structural characteristics of the online communities lead to differences in the performance of these algorithms.
- Indirectly, we also determined that simulation is a useful method for the analysis of expertise networks and expertise finding. We were able to tie the performance of the algorithm directly back to the dynamics of the communities. The simulations indicated under what structural conditions, or in what kind of networks, those algorithms will perform best. And we were able to do this without requiring interventions in real organizations, experimental conditions which we cannot obtain.

Work remains to be done. First, we would like to look at several other help-seeking communities (such as an intranet community) and compare it with our results and simulations. This would enable us to gain more insights about the tradeoffs in using these algorithms as well as in modeling online communities. Second, we will explore algorithms that combine content information (to differentiate specific knowledge) and

structural information in order to develop more advanced online community based expertise finders.

REFERENCES

1. Barabasi, A.L. and Albert, R. Emergence of Scaling in Random Networks. *Science*, 286, 509-512.
2. Ackerman, M.S. and McDonald, D.W. Answer Garden 2: merging organizational memory with collaborative help. In *Proceedings of CSCW '96*, Boston, MA, 1996, ACM Press, 97-105
3. Ackerman, M.S., Wulf, V. and Pipek, V. (eds.). *Sharing Expertise: Beyond Knowledge Management*. MIT Press, 2002.
4. Berkhin, P. A Survey on PageRank Computing. *Internet Math*. 2 (1), 2005, 73-120
5. Bollen, J., de Sompel, H., Smith, J. and Luce, R. Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing & Management*, 41 (6). 1419-1440.
6. Borodin, A., Roberts, G.O., Rosenthal, J.S. and Tsaparas, P. Link Analysis Ranking Algorithms Theory And Experiments. *ACM Transactions on Internet Technology*, 5 (1). 231-297.
7. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. Graph structure in the Web. *Computer Networks*, 33 (1-6). 309-320.
8. Campbell, C.S., Maglio, P.P., Cozzi, A. and Dom, B., Expertise identification using email communications. In *the twelfth international conference on Information and knowledge management*, New Orleans, LA, 2003, 528-231
9. Dom, B., Eiron, I., Cozzi, A. and Zhang, Y., Graph-based ranking algorithms for e-mail expertise analysis. In *DMKD*, New York, NY, 2003, ACM Press, 42-48.
10. Fagin, R., Kumar, R. and Sivakumar, D., Comparing top k lists. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, Baltimore, MA, 2003, Society for Industrial and Applied Mathematics, 28-36
11. Fisher, D., Smith, M. and Welsch, H., You Are Who You Talk To. In *HICSS '06*, Hawaii, <http://www.hicss.hawaii.edu/HICSS39/Best%20Papers/DM/03-03-08.pdf>
12. Foner, L.N. Yenta: a multi-agent, referral-based matchmaking system. In *Proceedings of Agents '97*, ACM Press, Marina del Rey, CA, 1997, 301-307
13. Herlocker, J.L., Konstan, J.A., Terveen, L.G. and Riedl, J.T. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22 (1). 5-53

14. Kautz, H., Selman, B. and Shah, M. Referral Web: combining social networks and collaborative filtering. *Commun. ACM*, 40 (3). 63-65
15. Kleinberg, J.M. Hubs, authorities, and communities. *Acm Computing Surveys*, 31. U21-U23
16. Kollock, P. The economies of online cooperation: gifts and public goods in cyberspace. In Smith, M.A. and Kollock, P. eds. *Communities in Cyberspace*, Routledge, London, 1999.
17. Krulwich, B. and Burkey, C., ContactFinder agent: answering bulletin board questions with referrals. In *the 13th National Conference on Artificial Intelligence*, Portland, OR, 1996, 10-15
18. Lakhani, K. and von Hippel, E. How open source software works: "free" user-to-user assistance. *Research Policy*, 32 (6), 923-943
19. Littlepage, G.E. and Mueller, A.L. Recognition and utilization of expertise in problem-solving groups: Expert characteristics and behavior. *Group Dynamics: Theory, Research, and Practice*, 1. 324-328
20. Liu, X., Bollen, J., Nelson, M.L. and Sompel, H.V.D. Co-authorship networks in the digital library research community. *Information Processing and Management*, 41 (6). 1462-1480
21. Newman, M.E.J. The structure and function of complex networks. *Siam Review*, 45 (2). 167-256
22. Page, L., Brin, S., Motwani, R. and Winograd., T. The Pagerank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project, 1998
23. Sergei Maslov, K.S., Alexei Zaliznyak. Pattern Detection in Complex Networks: Correlation Profile of the Internet *eprint arXiv:cond-mat/0205379*, 2002
24. Streeter, L. and Lochbaum, K., Who Knows: A System Based on Automatic Representation of Semantic Structure. In *Proceedings of RIAO*, 1988, 380-388
25. Wasserman, S. and Faust, K. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994
26. Zhang, J., Ackerman, M.S. and Adamic, L. CommunityNetSimulator: Using Simulation to Study Online Community Network Formation and Implications, In *Proceedings of C&T '07*, East Lansing, MI, 2007

CHAPTER 4

EXAMINING KNOWLEDGE SHARING ON YAHOO ANSWERS

INTRODUCTION

Every day, there is an enormous amount of knowledge and expertise sharing occurring online. One of the largest knowledge exchange communities is Yahoo! Answers (YA). Currently, YA has approximately 23 million resolved questions. This makes YA by far the largest English-language site devoted to questions and answers. These questions are answered by other users, without payment. Eckhart Walther of Yahoo Research has claimed that "(YA is) the next generation of search... (it) is a kind of collective brain - a searchable database of everything everyone knows. It's a culture of generosity. The fundamental belief is that everyone knows something." [1]. Indeed, if there is something that someone knows, there is certainly ample opportunity to share it on YA.

Because of the sheer size of the YA community, and its breadth of forums, we wished to conduct a large scale analysis of knowledge sharing within YA. Knowledge sharing has been traditionally difficult to achieve, and yet, YA appeared to have solved the problem, providing a society-wide mechanism by which to bootstrap knowledge and perhaps collective intelligence [2].

In short, we found YA to be an astonishingly active social world with a great diversity of knowledge and opinion being exchanged. The knowledge shown in YA is very broad (in several senses) but generally not very deep.

In this chapter, we examine YA's diversity of questions and answers, the breadth of answering, and the quality of those answers. Accordingly, we analyze the YA categories (or forums), using network and non-network analysis, finding that some resemble a technical expertise sharing forum, while others have a different dynamics (support, advice, or discussion). We then use the concept of entropy to measure knowledge spread, based on a user's answer patterns across categories. We find that having lower entropy, or equivalently, higher focus, correlates with the proportion of best answers given in a particular category. However, this is only true for categories where requests for factual answers dominate. Finally, we examine answer quality and find that we can use replier and answer attributes to predict what answers are more likely to be rated as best.

First, however, we discuss the prior literature and describe YA.

PRIOR WORK

Sharing knowledge has been a research topic for at least 15 years. At first, it was largely studied within organizational settings (e.g., Davenport and Prusak[3]), but now Internet-scale knowledge sharing is of considerable interest. This knowledge sharing includes repositories (including those socially constructed as with Wikipedia[4]) as well as online forums designed for sharing knowledge and expertise. As mentioned, these forums promise – and often deliver – being able to tap other users' expertise to answer all sorts of questions – mundane and everyday questions to complex and expert ones.

In general, there is a large body of literature examining online interaction spaces, especially Usenet. Four perspectives were important for this study. The first attempts to

understand different forums (or newsgroups in Usenet). Whittaker et al. [5] conducted an insightful quantitative data analysis on a large sample of Usenet newsgroups, uncovering the general demographic patterns (i.e. number of users, message length, and thread depth). Interesting findings in their work included the highly unequal levels of participation in newsgroups, cross-posting behaviors across different newsgroups, and a common ground model designed to explore relations between demographics, conversational strategies, and interactivity.

This line of research also used social network analysis to examine forums. For example, Kou and Zhang [6] used network analysis to study the asking-replying network structure in bulletin board systems and found that people's online interactions patterns are highly affected by their personal interest spaces. Fischer et al.[7] and Turner et al. [7] developed visualization techniques to observe various interaction patterns in Usenet groups. These visualization techniques have been very helpful in understanding the big picture of these large online interaction spaces [8].

While the work above mostly focused on the forum level, there have also been studies focusing on the user level. Wenger [9] discussed the importance of different roles in online communities and how they affect community formation and continuation. Nonnecke & Preece [10] studied lurker behavior in different online forums. Donath [11] explored techniques to mine users' virtual identities and detect deception in online communities. Recently, Welser et al. [12] argued that one can use users' ego- networks as "structural signatures" to identify "discussion persons" and "answer persons" in online forums. This work described role differences in online communities and provided insights on how to analyze user level data. However, the work lacks a strong quantitative basis.

There has also been work focusing on the thread and message level. For example, Sack [13] used visualization to show that there are various conversation patterns in discussion threads. Using message level content analysis, Joyce and Kraut [14, 15]

studied whether the formulation of a newcomer's post and related responses influenced the extent to which they continue to participate.

Besides studying the conversation patterns in online communities, researchers have also focused on understanding why people participate in and contribute to online communities. This work has been usually based on small scale data collection and surveys (e.g., Lakhani and von Hippel [16] and Butler et al. [17]). These studies have informed us in this study by delineating possible reasons why users engage in different activities in YA.

In the previous chapter, we have been studying one kind of online forum, online expertise sharing communities – those spaces devoted to answering one another's technical questions. We analyzed a technical question answering community (Java Forum) and explored algorithms using network structure to evaluate expertise levels in [18]. Using simulations, we explored possible social settings and dynamics that may affect the interaction patterns and network structures in online communities [19]. The goal of these studies was to design better systems and online spaces to support people in sharing knowledge and expertise in the Internet age.

During the course of our studies, we realized that relatively little is known about extremely large scale knowledge sharing and expertise distribution through online communities. YA presents an excellent place to study this problem because of its breadth of topic and high level of participation. More importantly, YA is a space that was designed for the sole purpose of knowledge sharing, although as we will see, it is used for much more. To our knowledge, there have been only two studies examining YA to date. Su et al. [20] used YA's answer ratings to test the quality of human reviewed data on the Internet. Kim et al. [21] studied the selection criteria for best answers in YA using content analysis and human coding. This has left open both the need for a large scale systematic analysis of YA, and the opportunity to study the depth and breadth of direct knowledge sharing from several perspectives that are only visible in such a large space.

YAHOO ANSWERS AND DATASET

The format of interaction on YA is entirely through questions and answers. A user posts a question, and other users reply directly to that question with their answers.

On YA, questions and their answers are posted within categories. YA has 25 top-level and 1002 (continually expanding) lower level categories. The categories range from software to celebrities to riddles to physics to politics. There are some "fact"-based threads, such as the following from the Programming & Design (Programming) category. In this thread, user asks for information on how to read a file using the C programming language⁶.

```
Q: How to read a binary file in C ?
```

```
I want to know what function from which header I must use to read a binary file. I will need to know how big a file is in byte. Then I want to move N byte into a char * variable.
```

She garners two responses. One is:

```
use the function fopen() with the last parameter as "rb" (read, binary).
```

The other, selected as the "best answer" by the asker, is more detailed:

```
#include <stdio.h>

FILE *fp;

fp = fopen("Data.txt", "rb");

fseek(fp, 0, SEEK_END);

filesize = ftell(fp);
```

⁶ We have anonymized any identifying data for publication and reworded the messages slightly for publication.

```
rewind(fp);  
  
fread(DataChar, 5000, 1, fp);  
  
fclose(fp);
```

This is a typical level of depth and complexity of the questions and answers for the Programming category. Indeed, many questions and their answers on YA are relatively simple. For example, math and science categories appear to be dominated by high school students trying to find easy solutions to their homework.

Not all categories are strictly focused around expertise seeking, however. The following question is from the Cancer category and appears to be soliciting both help and support:

```
My uncle was recently diagnosed with some rare cancer and does not have medical insurance. He has tried to apply for medical but has been denied. He does not have much money because he had to quit his job because he is getting too weak. Who can help him?"
```

This question received 10 answers, including a pointer to the local cancer society office. On average, a question in the Cancer category receives 5.2 replies, and only 6% go unanswered.

What is surprising is just how much of the interaction in YA is in fact just pure discussion, in spite of the question-answer format. There are many categories where questions are asking for neither expertise nor support, but rather opinion and conversation. For example, in the celebrity category, one finds the following question:

```
Who is the better actress, Angelina Jolie or Jennifer Aniston?
```

This question has appeared at least twice, once garnering 33 answers and once garnering 50. While one might expect that a large question-answer forum may show a

more diverse range of behavior than a narrowly focused software forum, we were nevertheless surprised to see the full range of topic and user types previously seen in general online newsgroups [8].

It is important to note that these discussions are constrained by the question-answer format of YA. Threads must still start with a question. YA users discuss by answering the question, not by addressing one another. Furthermore one cannot answer more than once nor can one answer oneself, making Usenet-type discussions difficult. This clearly changes the thread interactions relative to other online systems; the YA system was specifically set up for technical expertise sharing with a strict question and answer format.

In order to study the characteristics and dynamics of YA in a systematic manner, we harvested, using a automated crawler, one month of YA activity. The dataset includes 8,452,337 answers to 1,178,983 questions, with 433,402 unique repliers and 495,414 unique askers. Of those users, 211,372 both asked and replied. These numbers are already a hint to the diversity of user behavior in YA. Many users make very few posts. Even those who actively post will sometimes reply without asking much, while others do the opposite. These behaviors will vary by YA category, so we will briefly describe our analysis of those categories first.

CHARACTERIZING YA CATEGORIES

Basic characteristics

Based on an initial examination of YA, we expected that every category would have some mix of requests for factual information, advice seeking, and social conversation or discussion. While it would be difficult to determine the precise mix for each category without reading the individual posts, we can indirectly infer the category type by observing characteristics such as thread length (the number of replies per post)

and post length (how verbose the answers are). Figure 4-1 shows just such a scatter plot, with several categories highlighted.

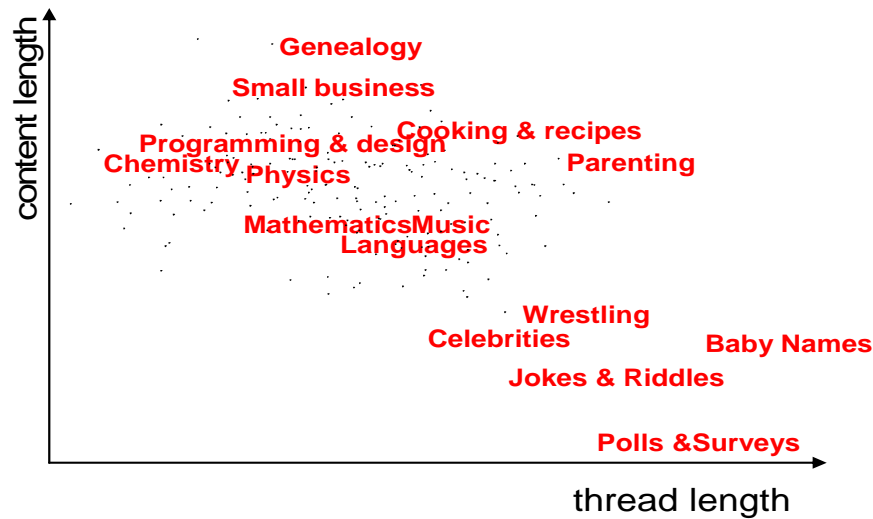


Figure 4-1 Post length vs. thread length

We observe that factual answers on technical subjects such as Programming , Chemistry, and Physics will tend to attract few replies, but those replies will be relatively lengthy. In fact, all of the math and science subcategories have a relatively low answer-to-question ratio, from 2 answers per question in chemistry to 4 answers per math question.

Astronomy has a higher question-answer ratio at 7, due to occasional questions about extraterrestrial travel and life that garner many replies (e.g. "What will you think if NASA comes clean about UFOs?" attracted 21 answers in 3 hours). In fact, the one science subcategory that stands out starkly is Alternative Science with 12 replies on average per question. These questions deal with the paranormal and by their very nature can lead to long discussions. (A typical question might be: "Can you use a RMS Multimeter for Ghost Hunting?")

On the other extreme are categories with many short replies. The Jokes and Riddles category contains many jokes whose implicit question is "Is this funny?" Most of the replies are short, "hahaha. that's funny" or "I've heard that one before". Also in this corner of the figure is the category Baby Names, where threads center around

brainstorming and suggestions of names, and many users chime in (24 people per question on average).

We can recognize discussion categories, those attracting many replies of moderate length: sports categories like Wrestling, as well as other categories such as Philosophy, Religion, and Politics. Also among those categories attracting many replies of moderate length are topics where many individuals have some experience and advice is sought. These include Marriage & Divorce and Parenting (including the subcategories for newborns, toddlers, grade schoolers, and the especially lengthy threads about adolescents). The Cats and Dogs categories generate fairly long threads of moderate reply lengths as well.

Another distinguishing characteristic for categories is the asker/replier overlap, whether the people who pose questions are also the ones who reply. In a forum where users share technical expertise, but the majority of askers are novices, one might expect that the population of askers and repliers is rather distinct [18]. Those who have expertise will primarily answer, while those who do not have it will be posing the majority of the questions. In a forum centered on advice and support, users may seek and offer both, becoming both askers and repliers. In a discussion forum, both asking and replying are ways of continuing the conversation. It is therefore unsurprising that the technical categories have a lower overlap in users who are both askers and repliers, while the discussion forums have the highest overlap. We will revisit this question in section 0.

Cluster analysis of categories

We calculated the three metrics, thread length, content length, and asker/replier overlap, for each of the 189 most active categories (categories that have at least 1000 questions) and ran a k-means clustering algorithm on them.

We find that clustering the categories into 3 groups gives us a result we find the most intuitively meaningful. Figure 4-2 shows how these three clusters are distributed regarding to thread length and the overlap in the population of askers and repliers.

Categories with high participation by the same users in both asking and replying tend to be discussion forums (yellow diamonds in Figure 4-2) – where users are discussing who is likely to win tomorrow’s game in a sports category, why Democrats and/or Republicans are a bunch of crooks in the Politics category, or debating the true nature of a god in the Religion category.

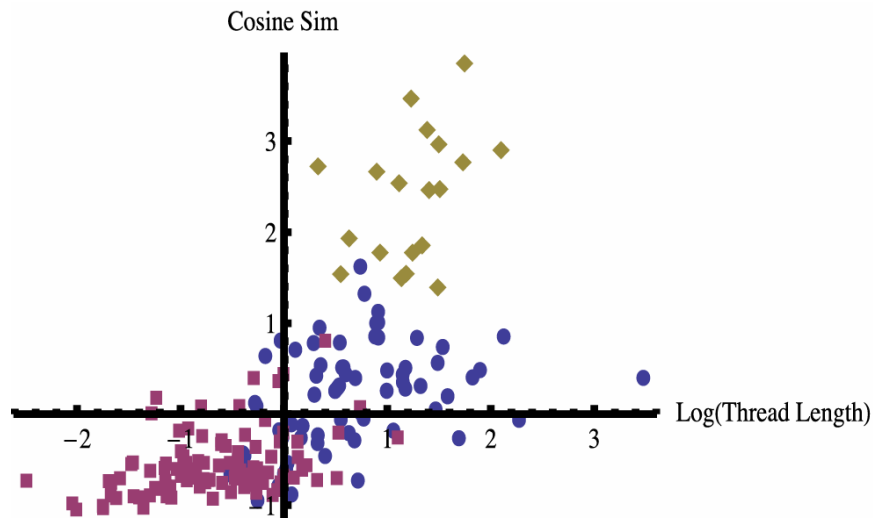


Figure 4-2 Clustering of categories by thread length and overlap between askers and repliers.

These kinds of stimulating questions tend to attract long thread lengths. Cluster two (blue circles) consists of categories in which people both seek and provide advice on questions where there may be several legitimate answers or no single factual answer. Perhaps because there is rarely a definitive answer, and at the same time many feel qualified to give advice, the threads tend to be long. This cluster includes the categories Fashion, Baby Names, Fast Food, Cancer, Cats, and Dogs. In Cluster 3 (purple squares), we observe categories where many questions have factual answers, e.g. identifying a spider

based on markings. People tend to either ask or reply, and thread lengths tend to be shorter. These categories include Botany, Zoology, and Programming.

In next section, we examine the question-answer dynamics further by analyzing how network structure differs in representative categories for each of these clusters. This more carefully considers how expertise and knowledge is arranged and structured in YA.

Network structure analysis

By connecting users who ask questions to users who answer them, we can create an asking-replying network; we call these QA networks. Analyzing the network structure of these QA networks reveals some interaction patterns that the non-network metrics cannot uncover.

Ego network analysis

Welser et al. [12] suggested that one can distinguish an “answer person” from a “discussion person” in online forums by looking at users’ ego networks. Thus, we examined what types of users appeared in different categories. Figure 4-5 shows the ego networks of randomly sampled 100 users from categories that are at the center of our 3 clusters. These categories are Programming, Cancer, and Wrestling.

From this figure, we can see that the neighbors of some of the highly active users in Wrestling are themselves highly connected, which indicates that they are more likely to be “discussion persons”. On the contrary, in the Programming and Cancer categories, the most active users are “answer people” because most of their neighbors, the people they are helping, are not connected [12].

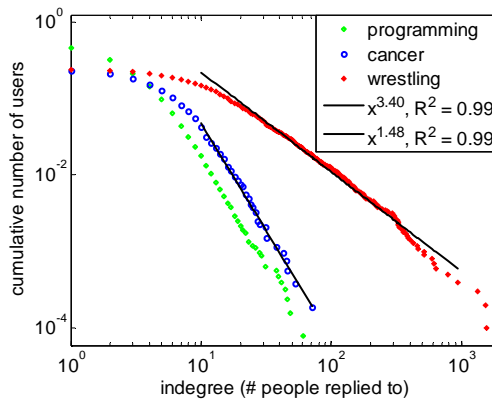


Figure 4-3 Indegree distributions for different categories.

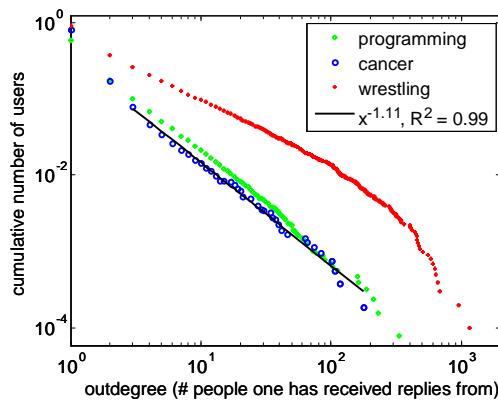
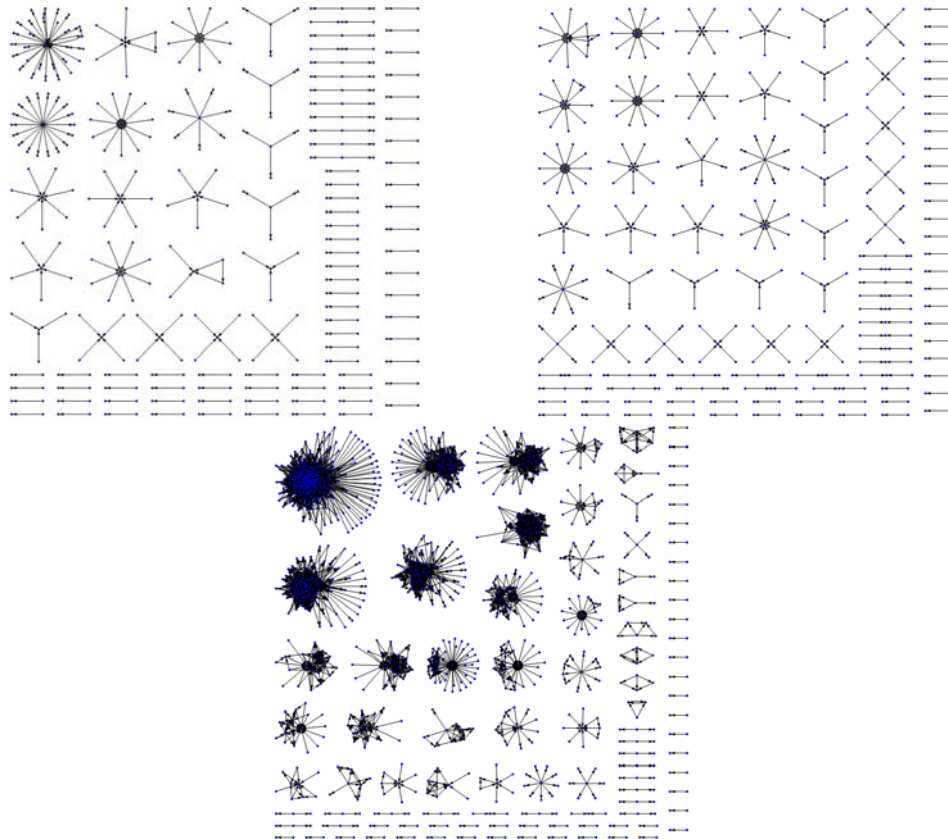


Figure 4-4 Outdegree distributions.

Figure 4-3 shows the indegree distribution (number of people one answered) and Figure 4-4 shows the output degree distribution (number of people one has received replies from) for categories that are central in our clusters. From these figures, first, we can see that the users differ in their activity level in all three categories. Some answer many questions, others merely stop by to ask or answer a question or two. For instance, in Programming, about 57% of the users who asked questions did not answer any during this time period, and similarly 51% who answered questions did not ask. On the other extreme there were users who asked or answered dozens of questions. Second, we can also see that there are differences among these three categories. Although all three categories display heavy tailed distributions, Wrestling distinguishes itself as a topic from



**Figure 4-5: Sampled ego network of three selected categories:
*Programming, Cancer, Wrestling***

Programming and Cancer, by having a much broader distribution of indegrees (with a couple of people replying to thousands of others in just the one month sample). In contrast, the most active repliers in Programming and Cancer replied to a few dozen others. A similar pattern applies for outdegree. Some users posing wrestling questions attracted answers from hundreds of users, while the most successful or active askers in Programming or Cancer attracted replies from only a hundred.

Strongly connected components

Given that some people reply almost exclusively, and others ask almost exclusively, it is unclear whether these categories contain strongly connected components (SCCs). These strongly connected components are those sets of users, such that one user

can be reached from any other, following directed edges from asker to replier. Table 4-1 summarizes general statistics of the networks of these three selected categories.

Table 4-1 Summary statistics for selected QA networks

Categories	Nodes	Edges	Avg degree	Mutual Edges	SCC %
Wrestling	9960	54961	5.51	1898	13.5%
Programming	12539	18311	1.46	0	0.01%
Cancer	5237	7575	1.45	4	0.04%

From this table, we can see, consistent with the degree distributions shown in the previous section, that the Wrestling category is more connected. More importantly, it has a strongly connected component and a relatively large number of mutual edges (two users who have answered to each other's questions), which indicates that there may be a core social group forming in this category. There is almost no strongly connected component in either Programming or Cancer (even a random network of this size and density should have a modestly sized SCC). We believe that is due to the separation of roles of "helper" and "askers" in these two categories. We delve into this further in the next section.

Motif analysis

Motif analysis allows one to discover small local patterns of interaction that are indicative of particular social dynamics. Here we focus on all possible directed interactions between three connected users within a forum. Figure 4-6 displays the motif profiles of the three selected categories, showing for example, how often interactions are reciprocal (the asker becomes the replier for another question) and how often the triads

are complete (a user interacting with two others often corresponds to those two users interacting as well). These profiles are constructed by counting the actual frequency of each triad in the QA network for that category, and then comparing that frequency against the expected frequency for randomized versions of the same network [22-24].

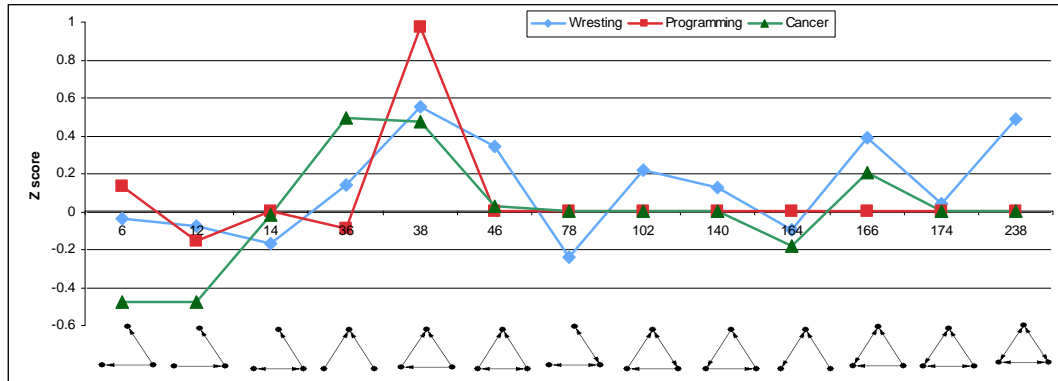


Figure 4-6 Motif profiles of selected categories

From Figure 4-6, first, we can see that all three categories have a significantly expressed feed forward loop (see triad 38 in the figure) compared to random networks. In this motif, a user is helped by two others, but one of the helpers has helped the other helper. The motif, most pronounced in the Programming category, indicates a common characteristic in help-seeking online communities, where people with high levels of expertise are willing to help people of all levels, while people of lower expertise help those with even less expertise than their own [19].

As well, we can see that the Wrestling category has a high number of fully reciprocal triads, indicating symmetric interaction. Another triad that is significant in Wrestling involves two users who have replied to one another (who may be regulars in the forum) and have also replied to a third user, perhaps someone who is just briefly joining the discussion to ask a question. This triad is also significant for the Cancer category. Interestingly, the triad of two users who have replied to one another, and have also both received replies from a third user, is not significant for Programming or Cancer (it would imply that the regulars are drawing answers from less active users), but it is

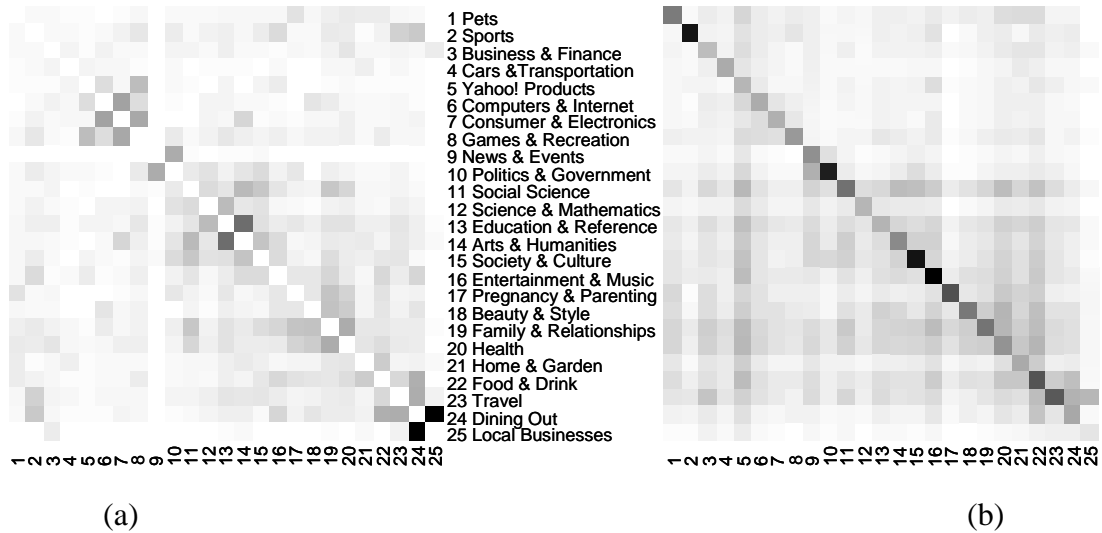


Figure 4-7 Similarities between categories: a) overlap in users who replied in both categories, b) overlap in users who answered in one category (rows) and asked in another (columns). A cosine similarity was used in both, but the shades correspond to different scales.

significant for Wrestling, where even questions posed by regulars are of an inviting nature.

Expertise depth

Is expertise being shared? We have already alluded to the relative simplicity of many questions. It often seems as though users are sharing the answers to one another's homework questions. To determine the depth of the questions asked in YA, we rated 100 randomly selected questions from the Programming category. We rated these questions into 5 levels of expertise (as discussed in [18]). In this rating scheme, level 3 expertise is that of a student with a year's experience in a programming topic, for example, someone who could pull details from an API specification. A level 4 expert, on the other hand, would be a professional programmer, someone with experience in implementation or deployment issues and their effects on design (such as compiled Java applications and their speed). We found only one question (1%) in the Programming category that required above level 3 expertise. In short, the questions are very shallow. This is not a

definitive test, of course, but it indicates that YA is very broad but not very deep. We explore that breadth in the next section.

EXPERTISE AND KNOWLEDGE ACROSS CATEGORIES

Given the wide variety of behavior and interests in the different forums, we saw an opportunity to describe how knowledge and expertise are spread across different domains. In this section, we describe the breadth of YA from two perspectives. The first considers the relationships between the categories, where users who are actively answering questions in one category are also likely to do so in another. The second measures the users *entropy*, namely the breadth of topics their answers fall in.

Relationships between categories

By tracking answer patterns, it is easy to discern related categories, shown in Figure 4-7(a). In short, people who answer questions in one category are likely to answer questions in related categories. Computer-centric categories, including Computers & Internet, Consumer Electronics, Yahoo! Products, and Games & Recreation (dominated by questions about video and online games), are all clustered together. Similarly, Politics and Government is linked to News and Events, while the Home and Garden category is linked to Food and Drink, which is in turn linked to Dining Out, which is in turn linked to the topic of Local Businesses. The above cross-category correlations suggest a focus of interest on the part of the users.

Reply patterns only reveal the topics that a user feels comfortable discussing. The overlap of asking and replying patterns, on the other hand, reveal whether people who reply in one topic are likely to ask questions in the same topic or another. In Figure 4-7(b), we can observe that users are likely to post both questions and replies in the same forum, if that forum deals with topics that are prone to discussions: Sports, Politics, and

Society & Culture (including Religion). On the other hand, topics dominated by factual, straightforward questions, such as those found in the Education & Reference and Science & Math subcategories, have a smaller percentage of users who both seek and offer help. Most users are either asking for help (as mentioned, many apparently looking for easy answers to their homework questions), while others almost exclusively provide that help, without posing questions of their own.

Other interesting patterns emerge when one looks at question answer patterns across categories. As a silly hypothetical example, consider users who answer many car repair questions, but may need lots of advice about beauty and style. As amusing as it would be to find this connection, we find that those posting answers about cars and transportation tended to not ask for help in other categories, as much as people answering in other categories asked for help with cars. In fact, sports and politics were the only other large categories from which the helpers were less likely to be the ones asking questions about beauty and style.

No matter the category that users post answers in, they almost uniformly also ask about Yahoo products, including YA itself. Health was a category that many users asked questions in, no matter where else they answered. But it was also a category that many answered in, no matter where most of their questions were posed. The latter was also true of Family & Relationships (with users apparently willingly chiming in with their advice), but asking questions about relationships typically did not correlate with answering in other categories. There was again an asymmetry between technical and support categories: people who answered in Relationships, Health, or Parenting tended to ask in the Computers & Internet category, while the opposite was not true: those answering in Computers & Internet did not have a high proportion of questions in Health, Relationships, or Parenting.

The above connections between categories are apparent because at least some users are not replying in all categories at random, they have a certain degree of focus. So

while YA gives the opportunity to individuals to seek and share knowledge on a myriad of different topics, any *individual* user is likely to only do so for a limited range. In the next section, we will turn to studying YA on the individual level in order to pinpoint just how broad users' participation is.

User entropy

We have already observed in previous sections that users differ in their activity level. Due to this wide variation in activity levels, we decided to focus just on those users who had posted at least 40 replies, so that by observing a sufficient number of replies, we could discern whether they were truly focused on just a few topics, or simply had not been active long enough to reveal their full range of interests. We sought a measure that would capture the degree of randomness in a person's reply patterns. Entropy is just such a measure – the more evenly distributed a person's activity is among the categories, the higher the entropy. We also wanted our entropy measure to capture the hierarchical organization of the categories, such that a user who answers in a variety of subcategories of the same top level category would have a lower entropy than someone who answered in the same number of subcategories, but with each falling into different top level categories.

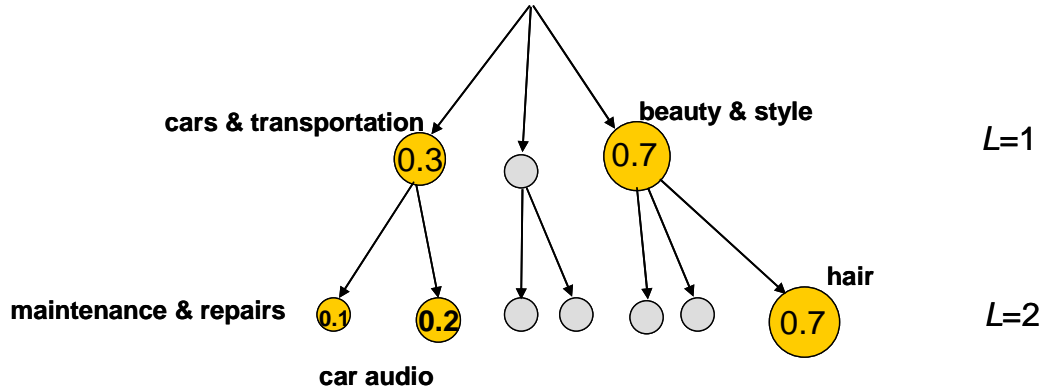


Figure 4-8 Illustration of the hierarchical entropy calculation,

$$H_1 = -0.3 * \log(0.3) - 0.7 * \log(0.7) = 0.61,$$

$$H_1 = -0.2 * \log(0.2) - 0.1 * \log(0.1) - 0.7 * \log(0.7) = 0.81 \text{ and}$$

$$H_T = H_1 + H_2 = 1.42$$

Figure 4-8 illustrates a hypothetical user's distribution of questions. To obtain the total entropy for a user, we first calculate the entropy H_L for each level separately.

$$H_L = -\sum_i p_{L,i} \log(p_{L,i})$$

where $p_{L,i}$ is the proportion of answers by the user in category i at level L . We then sum the entropies for the different levels together:

$$H_T = \sum_L H_L .$$

If we look at several users who have answered, for example, exactly 40 questions, we can observe a range of entropies. For one user, who describes herself as a dog trainer who shows shelties at dog shows, we find that all her answers are in the Dog subcategory. Therefore her entropy is 0. On the other end of the spectrum is a user whose 40 questions are scattered among 17 of the 25 top-level categories and 26 subcategories. He posted no more than 4 answers in any one category and his combined 2-level entropy is 5.75.

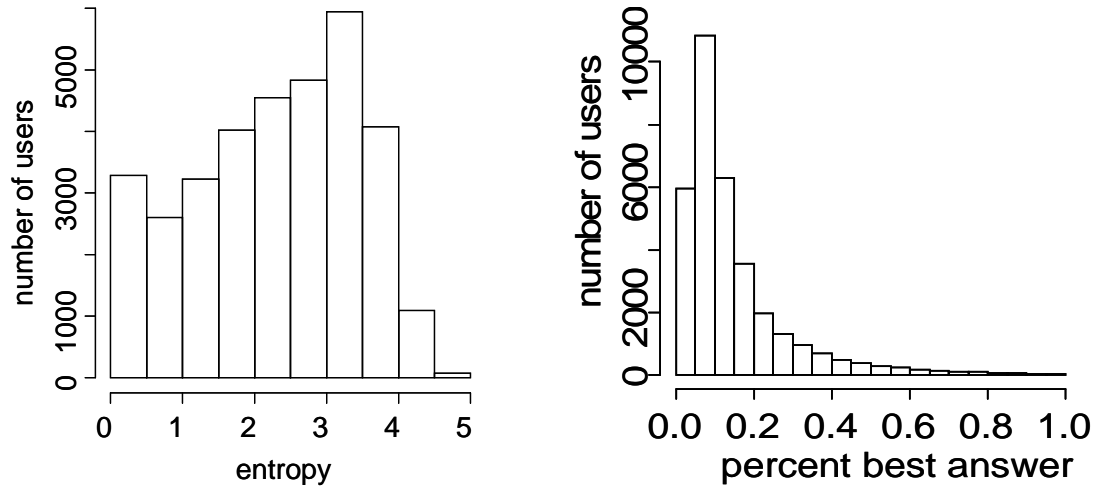


Figure 4-9 The distribution of entropy and percent best answer across users who had answered at least 40 questions

Figure 4-9 shows the entropy distribution of all users who posted 40 or more questions. The distribution is surprisingly flat. It is not the case that only a few users are very diverse. Rather, some users have a very low entropy, being focused on a subcategory or two, but higher entropies are relatively common, until one encounters a limit in terms of the number of possible categories that are specified by the YA hierarchy.

We also examined the proportion of best answers by users. (Again, best answers are those answers rated as such by the asker or voted as such by YA users.) This distribution is skewed, with a mode around 6-8% best answers. Some users obtain much higher percentages of best answers. In the next section we will correlate the two metrics applied to users in order to determine whether being focused corresponds to greater success in having one's answers rated as best.

Correlating focus to best answers

Intuitively, one might expect that users who are focused to a limited range of topics tend to have their answers selected as best more frequently. For example, a dog trainer/breeder who answers questions about dogs may be expected to have a higher

proportion of best answers because all of her answers are focused on her specialty. Interestingly, we found no correlation between total entropy of a user across all categories and their overall percentage of best answers ($\rho = -0.02$, $p < 10^{-3}$). Users do not provide better answers (at least according to their best answer count) when they specialize. The value of the correlation has the correct sign (more scattered users have a lower proportion of best answers), but is only significant because of the large number of users ($n=33,720$).

While it may well be the case that posting answers in several discussion forums does not correlate with whether others like those answers, we still expected to see a correlation in some cases. From our earlier examination of the different categories, we know that only some topics reflect requesting and sharing factual information. This brings to question what the criteria for best answer selection are in other forums. In support forums, the best answer may be the one with the most empathy or most caring advice. In a discussion forum, the best answer may be the one that agrees with the askers' opinions, while for entertainment categories, the wittiest reply may win. A previous study that sampled users comments upon selecting a best answer to their question found that content value (such as accuracy and detail) was used in selecting the best answer in just 17% of the cases, compared to 33% for socio-emotional value, including agreement, affect, and emotional support[21].

Another idiosyncrasy of selecting just one best answer, instead of rating individual ones, is that there may be several good answers, but only one is selected. We randomly sampled 100 questions each from categories of Programming, Cancer and Celebrity and coded them according to how well they answered the question. We found that replies selected as best answers were mostly indeed best answers for the question. For those best answers not rated as the best answer by us, we found that they could still be second or third best answers. This beneficial glut of good answers means that even if a user always provides good answers, we may not be able to discern this, because their

good answer will not always be selected as best, depending on how many other replies were posted. This means that users focusing on categories with a high answer-to-question ratio will on average have a lower best answer percentage, and any correlation between user attributes and this percentage will be weakened by the noise introduced through answers being pitted against one another for first place.

Table 4-2 Correlating category entropy and % best answers

Level 1 category(ies)	Pearson (entropy,score)	p-value
computers & internet science and math	-0.22	10^{-7}
Family & relationships	-0.13	10^{-13}
sports	-0.01	0.65

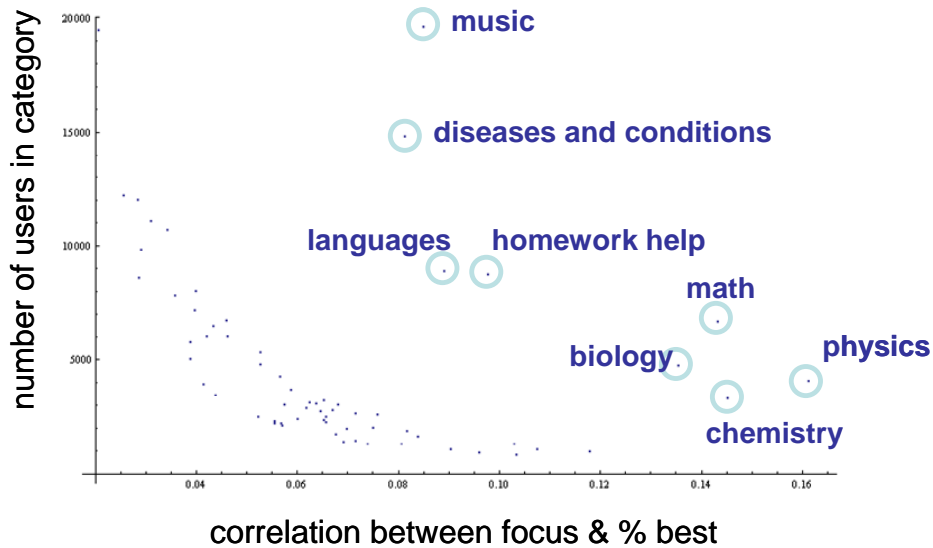


Figure 4-10 Relationships between focus and best. Categories with many users tend to have a weaker correlation between focus and score on the user level, with the exception of technical categories that stand out as having high correlation beyond the overall trend.

Despite these caveats, we still expected lower entropy to be correlated with performance

for categories where many questions were of a technical or factual nature. To verify this claim we computed separate second level entropies for several first level categories. Indeed, for the technical categories of Computers & Internet and Science & Math, we find a significant correlation between the users' entropy within those top level categories and their scores. The correlation is weaker, but still present for the advice-laden category of Family & Relationships. It is absent in the discussion category of Sports. Table 4-2 the above.

Finally, we used a very simple measure, the proportion of a user's answers in the category, and correlated it with a user's proportion of best answers in that category across all of YA. We found that for technical categories, focus tended to correlate with better scores. For categories that still required some domain knowledge to answer questions, there was a weaker, but significant correlation. And finally, in discussion categories, there was no relationship between focus and score within that category. A listing of typical categories for each level of correlation is shown in Table 4-3. Note the predominance of a single cluster corresponding to low asker-replier overlap and short thread length for the categories where correlation between focus and score is highest.

Table 4-3 Correlation between focus and score*.

moderate correlation $\rho > 10, p < 10^{-5}$	low correlation $0.05 < \rho < 0.10, p < 0.01$	no correlation $\rho < 0.05$
physics chemistry math biology Y! products	garden & landscape genealogy dogs hobbies & crafts cooking & recipes	marriage & divorce American football alternative medicine religion & spirituality baby names

PREDICTING BEST ANSWERS

So far, we have observed distinct question-answer dynamics in different forums. We have also observed a range of interests among users – some focusing quite narrowly on a particular topic, while many participate in several forums at once. Furthermore, focusing on a particular category (having low entropy) only correlated with obtaining “best” ratings for one’s answers in categories where questions centered on factual or technical content. Here, we test our ability to predict whether an answer will be selected as the best answer, as a function of several variables, some of which will correspond closely with our previous observations.

Table 4-4 Predicting the best answer

	Programming	Cancer	Wrestling
reply length	+ ***	+ ***	+ ***
reply position	- **		
user # replies	+ ***		+ ***
user # questions			+ *
prediction accuracy	0.732	0.723	0.709
+ (positive coefficient), - (negative coefficient)			
*(p<0.05), ** (p < 0.01), *** (p < 0.001)			

We constructed randomly selected balanced sets of answers that were and were not chosen as best answers. We excluded those instances where the answer was the only answer, which would make it very likely to be selected as best. We then ran a logistic regression on a number of variables. We normalized the variables so that they summed to 1. For example, the reply length of each answer is divided by the sum of the lengths of all

replies in that category. This both takes into account the relative lengths of answers to the same question, and also lowers the probability that an answer is selected when many other answers are supplied. We omitted entropy and focus measures because the majority of users had posted too few replies to produce meaningful entropy values. We ran a logistic regression to predict the best answer, and performed a ten-fold cross-validation to obtain a prediction accuracy, with a baseline of 0.5 for random guesses.

Table 4-4 summarizes the prediction results for the three categories from the category clusters. For all categories, the length of the reply is most significant, and is in fact the only significant feature for the Cancer category. We can achieve about 70% prediction accuracy across all three categories based on this feature alone – showing a strong preference for the asker to receive a lengthier reply. Figure 4-11 shows the difference in length distribution for best answers and non-best answers in the Programming category.

The interesting differences arise in the other features found to be predictive. The total number of wrestling questions answered by a user is only very mildly predictive of whether their answer is selected as best. Since Wrestling is a discussion forum, with a strong correlation ($\rho = 0.55$) between the number of questions posted and answered, the number of questions asked is also marginally significant in predicting best answers, but can be dropped from the model without loss of accuracy. In contrast, for Programming, the number of questions asked is irrelevant, but the number of questions answered, though trailing behind answer length by a long margin, is also fairly significant. That there is a correlation for frequent repliers, but there is an absence of correlation for frequent askers, is also true, for example, for the Small Business category, which we saw

grouped with Programming in the factual question cluster in early section.

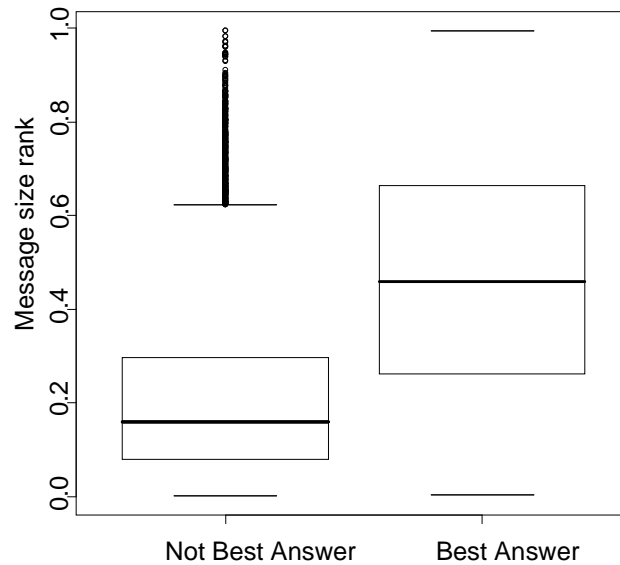


Figure 4-11 Answer length and best answer selection

Our results for Yahoo Answers stand in stark contrast to our previous analysis of Sun's Java Forum, where the number of previous replies strongly correlated with the expertise level as judged by independent human raters. Here, we see that when the raters are the askers themselves, there is a preference for longer answers, but not always by more active repliers. It would of course be interesting to pit the askers' choice of best answer against best answers selected by experts in the subject. It would also be interesting to examine whether frequency of replies correlates with expertise level, and even whether there is as much of a differentiation in expertise level on a general community such as Yahoo Answers, as opposed to a specialized community such as the Java Forum. We leave these and other questions for future work.

CONCLUSIONS

Yahoo Answers is a diverse and broad question answer community, acting not only as a medium for knowledge sharing, but as a place to seek advice, gather opinions, and satisfy one's curiosity about things which may not have a single best answer. One may dispute the validity of the knowledge in Alternative Science and even the degree of knowledge in Celebrities. However, the YA participants believe this is knowledge, and they are certainly exchanging it.

We took advantage of the range of user behavior in YA to inquire into several aspects of question-answer dynamics. First, we contrasted content properties and social network interactions across different YA categories (or topics). We found that we could cluster the categories according to thread length and overlap between the set of users who asked and those who replied. Discussion topics or topics that did not focus on factual answers tended to have longer threads, broader distributions of activity levels, and their users tended to participate by both posing and replying to questions. On the other hand, YA categories favoring factual questions (what are usually called question-answer forums) had shorter thread lengths on average and users typically did not occupy both a helper and asker role in the same forum. We found differing interaction motifs in the question-answer networks corresponding to these distinct dynamics. Consistent with prior work on online forums, we found that the ego-networks easily revealed YA categories where discussion threads, even in this constrained question-answer format, tended to dominate.

Second, we related the categories to one another, both in terms of relating knowledge, by identifying pairs of topics such that if a user is answering questions in one, they are also likely to answer in another. We found many expected relationships between the categories, but also interesting asymmetries when linking asking questions in one category with answering questions in another. Many users answered questions about

familiar topics such as Family & Relationships, no matter where they tended to ask their questions. On the other hand, users who answered in specialized, technical categories, such as Car Maintenance & Repair or Computers & Internet, asked fewer questions in other categories, where the users they were helping predominantly supplied answers.

This lead us to examine the range of knowledge that users share across the many categories of YA. We found that while many users are quite broad, answering questions in many different categories, this was of a mild detriment for specialized, technical categories. In those categories, users who focused the most (had a lower entropy and a higher proportion of answers just in that category) tended to have their answers selected as best more often.

Finally, we attempted to predict best answers based on attributes of the question and the replier. Our results showed that just the very basic metric of reply length was most predictive of whether the answer would be selected. The number of replies by a user, indirectly reflecting focus, correlated with best answer ratings, but most significantly so for technical categories.

In future work we would like to examine further the level of expertise being shared on YA. By democratizing knowledge sharing, YA has accomplished a large feat – everyone knows something, and through our analysis, we know that many know even several things and can share them on YA. But it remains unclear whether depth was sacrificed for breadth. We would like to know whether different incentive mechanisms could encourage YA participation by top level experts – who may currently still prefer more specialized, boutique forums – while at the same time allowing the rest of us to get our everyday, simple questions answered.

REFERENCES

1. Y. Noguchi, "Web Searches Go Low-Tech: You Ask, a Person Answers," in Washington Post, 2006, pp. A01.
2. D. Engelbart and J. Ruilifson, "Bootstrapping our collective intelligence," ACM Computing Surveys (CSUR), vol. 31, 1999.
3. T. H. Davenport and L. Prusak, "Working knowledge: how organizations manage what they know," Ubiquity, vol. 1, 2000.
4. T. Holloway, M. Bozicevic, and K. Börner, "Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors.", Complexity, Vol.12, No. 3, 2007.s
5. S. Whittaker, L. Terveen, W. Hill, and L. Cherny, "The dynamics of mass interaction," Proceedings of the 1998 ACM conference on Computer supported cooperative work, pp. 257-264, 1998.
6. Z. Kou and C. Zhang, "Reply networks on a bulletin board system," Physical Review E, vol. 67, pp. 36117, 2003.
7. T. C. Turner, M. A. Smith, D. Fisher, and H. T. Welser, "Picturing Usenet: Mapping Computer-Mediated Collective Action," Journal of Computer-Mediated Communication, vol. 10, 2005.
8. D. Fisher, M. Smith, and H. T. Welser, "You Are Who You Talk To: Detecting Roles in Usenet Newsgroups," HICSS'06, 2006.
9. E. Wenger, "Communities of Practice: learning, meaning, and identity," 1999.
10. J. Preece, B. Nonnecke, and D. Andrews, "The top 5 reasons for lurking: Improving community experiences for everyone," Computers in Human Behavior, vol. 2, pp. 42, 2004.
11. J. S. Donath, "Identity and Deception in the Virtual Community," Communities in Cyberspace, pp. 29-59, 1999.
12. H. T. Welser, E. Gleave, D. Fisher, and M. Smith, "Visualizing the Signatures of Social Roles in Online Discussion Groups," Journal of Social Structure, vol. 8, 2007.
13. W. Sack, "Conversation map: a content-based Usenet newsgroup browser," Proceedings of the 5th international conference on Intelligent user interfaces, pp. 233-240, 2000.
14. J. Arguello, B. S. Butler, L. Joyce, R. Kraut, K. S. Ling, and X. Wang, "Talk to me: foundations for successful individual-group interactions in online communities,"

- Proceedings of the SIGCHI conference on Human Factors in computing systems, pp. 959-968, 2006.
15. E. Joyce and R. E. Kraut, "Predicting Continued Participation in Newsgroups," *Journal of Computer-Mediated Communication*, vol. 11, pp. 723-747, 2006.
 16. K. R. Lakhani and E. von Hippel, "How open source software works:" free" user-to-user assistance," *Research Policy*, vol. 32, pp. 923-943, 2003.
 17. B. S. Butler, "Membership Size, Communication Activity, and Sustainability: A Resource-Based Model of Online Social Structures," *Information Systems Research*, vol. 12, pp. 346-362, 2001.
 18. J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," *Proceedings of the 16th international conference on World Wide Web*, pp. 221-230, 2007.
 19. J. Zhang, M. S. Ackerman, and L. A. Adamic, "CommunityNetSimulator: Using simulations to study online community networks " presented at *Communities and Technology 2007*, Lansing, MI, 2007.
 20. Q. Su, D. Pavlov, J. H. Chow, and W. C. Baker, "Internet-scale collection of human-reviewed data," *Proceedings of the 16th international conference on World Wide Web*, pp. 231-240, 2007.
 21. S. Kim, J. S. Oh, and S. Oh, "Best-Answer Selection Criteria in a Social Q&A site from the User-Oriented Relevance Perspective," presented at *ASIST*, 2007.
 22. S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," *Bioinformatics*, vol. 22, pp. 1152, 2006.
 23. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network Motifs: Simple Building Blocks of Complex Networks," vol. 298, 2002, pp. 824-827.
 24. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of Evolved and Designed Networks," *Science*, vol. 303, pp. 1538-1542, 2004..

CHAPTER 5

CONCLUSION

In this chapter, I firstly summarize the related findings in my three dissertation studies. Then, I discuss their design implications, as well as showing a prototype system we developed based on these findings. At last, I briefly discuss the contributions of my dissertation work and the future work.

SUMMARY

As I mentioned in chapter 1, the goal of my dissertation research was to gain a better understanding of expertise sharing in social networks to help us design better expertise sharing systems. To fulfill this goal, I conducted three studies based on data set collected from the Enron Corporation, the Java Forum, and Yahoo Answers. Each of studies provided insights into expertise networks from a different perspective.

The Java Forum study is the capstone of this thesis. In this study, I firstly analyzed the social network characteristics of the Java Forum expertise network using a set of social network metrics. We found that networks in these communities typically differ in their topology from other online networks such as the World Wide Web. People vary in their participations of online expertise sharing in different ways. There are different groups of users: users who mostly only answer questions, users who mostly only

ask questions, and users who both ask and answer questions. The users' contributions to the online community also vary dramatically, with few users answering a large number of questions while the majority of users only answer a few. Furthermore, top repliers answer questions for everyone, and less expert users tend to answer questions of others with lower expertise level.

We then tested a set of network-based ranking algorithms, including Z-score, PageRank, and HITS, on the Java Forum expertise network in order to identify users with a high level of expertise. We found that structural information can be used for evaluating an expertise network in an online setting, and relative expertise can be automatically determined using these social network-based algorithms. These algorithms did nearly as well as human raters. However, surprisingly, we found that relatively simple measures like Z-score are as good as more complex algorithms. We then used simulation to explore the reasons of such results. We identified a small number of simple simulation rules governing the question-answer dynamic in the network using simulations. These simple rules not only replicate the structural characteristics and algorithm performance on the empirically observed Java Forum, but also allow us to evaluate how these algorithms may perform in other communities with different characteristics.

The Yahoo Answers study extends the Java Forum study into a more general community setting and covers much more diverse perspectives of knowledge sharing dynamics. In this study, we firstly analyzed this large scale knowledge sharing community using both network and non-network methods. We found that Yahoo Answers is indeed a very diverse and broad question answer community. It is actually not only a place for knowledge sharing, but also a place for seeking advice, gathering opinions, and satisfying one's curiosity. We found that there are many expected relationships among different knowledge categories (i.e. users who answer in category "Baby Names" also answer in category "Family"), but also interesting asymmetries regarding users' asking and answering places (i.e. users who answered in "Computer"

categories did not ask many questions in other categories). At last, we found that many users are very broad, answering questions in many different categories. However, in specialized technical categories, this breadth leads to the detriment of expertise depth, which is reflected on the portion of one's answers being selected as "best answers".

The Enron study investigates how social network structure could affect the expertise searching process in organizational communication networks using simulations and social network analysis. With the analysis of the Enron email network, I showed that a social network has specific social characteristics and these social characteristics lead to sizeable differences in the way expertise is searched and shared. In detail:

- Network degree based searching strategies in networks like Enron's email network have a clear advantage over other strategies. A very few nodes turn out to be key in affecting the performance of such social network searching.
- As Granovetter suggested, when compared to the strong tie based strategy, the weak tie based strategy is faster. Furthermore, when the weak ties are removed, the performance of information similarity based strategies also decreased considerably. This indicates that weak ties are likely to be critical for automated or augmented expertise finding.
- The information similarity based strategy, surprisingly, is not as fast as network degree based strategies. However, its social cost is more evenly distributed.

With these three studies, I gained the insight into theoretical understanding of factors that one should consider when developing social network based expertise sharing systems. The next sections discuss the design implications of these findings.

DESIGN IMPLICATIONS

The findings in these three dissertation studies have a set of design implications for social network based information systems, especially for expertise finding systems in organizations or online communities. In this section, I firstly discuss the two most important design implications from my studies. Then I briefly describe a prototype system developed based on these findings.

Designing for different communities and users

One of the most important lessons I learned from these three studies is that social networks are not random graphs, whether they are organizational email networks or replying networks in online communities. They have specific social structures which are outcomes of the interaction among their users guided by specific social dynamics and rules. Users in these communities are very different regarding their backgrounds, motivations, expertise depths and breadths, and activities. These differences in users and community structure can make significant impact on how the system being designed and deployed.

In the Enron simulation study, we found that not all users were queried equally during the expertise searching processes. There are some people with much higher social connections than other people in the organization, and these people were referred significantly more frequently than other users and have a much higher workload no matter which query spreading algorithms we used. Thus, if we design a SWIM like peer-to-peer based expertise searching system in an organization, we need design specific functions to support these highly connected people to relieve some workload for them. We might also need design a workload balance system to make sure that no users are overloaded.

In the Java Forum study, we found that instead of being a public place where people help each other reciprocally; this online community is more closely a place where

askers come to seek help from volunteer helpers. The current forum interface was designed to support the former situation, in which each user has the same interface and capabilities. With the understanding of the separation of asking and answering roles in such online forum, we may think to provide different users with different interfaces and functions. For instance, we may provide a better question browsing interface for the general helpers so they can find the interesting and un-answered questions more easily. At the same time, for the askers (especially the new ones), we should consider building functions to help them formulate the question for others to read and interpret. Furthermore, we should also consider design functions or mechanisms to encourage askers to answer questions.

In the Yahoo Answers study, we found that people have different expertise depth and breadth, and they usually focus on one or several topic categories while lightly involved in some others. In different topic categories, the activeness and dynamics are different based on the nature of the topic. For instance, in topics about entertainment, a question usually gets many answers but most of them are very short; in topics about health, an answer to a question can be long and very personal. These differences show the diversities of human's question asking and answering behaviors. Thus, when we design an expertise sharing system or a community, we need keep these differences in mind. For instance, when we design a reputation system for a multiple topic community like Yahoo Answers, we may consider putting some topic based labeling and weighting into the ranking system, thus an active user will be ranked as "Top expert in Physics" instead of "Top expert", which could be a top expert in "Movies and Music".

Above all, social networks based expertise sharing systems were largely designed to let everybody contribute as much as they can. Based on this assumption, most early system designs have treated every user the same and provided the same functions for everyone. However, in reality, people have different backgrounds, expertise, needs, and

resources, these differences will influence their roles and behaviors in expertise sharing related activities. Thus, in design, we should keep these differences in mind and think about how to design functions to support different user groups.

Matching Expertise Levels

Current expertise finders, both commercial and research, cannot infer expertise levels automatically. Traditionally, expertise finders have relied on the standard information similarity measures (such as term vector comparisons) and the approach has proved to be very limited. In the Java Forum study, we found that we can use network structure based metrics to infer users' expertise levels based on their previous question asking and answering histories. The ability to add the level of expertise would be a major step forward for expertise finders, and would likely open up a range of new application possibilities.

For example, in Java Forums, one problem we found is that high expertise users' questions need significantly longer time to get an answer than questions asked by low expertise users. While the low availability of high expertise users in the forum is the main reason for this situation, the current interface makes the problem worse. The high expertise users' questions are often lost in the flood of newbie questions. In online forums, the cost of a question and answerer match-up falls upon the potential helpers, and the helpers accomplish the match-up task by reading or scanning questions posted in the forum. The current forum interface does not support these match-up tasks well. With the availability of expertise ranking algorithms to infer both askers' and helpers' expertise levels, we may address this problem by matching high expertise users' questions to other high expertise users who are capable of answering them.

Such expertise matching approach will also have great potentials in organizations. In organizations, top experts are usually very busy and expensive. As we found in the literature review, reaching a top expert may also have high social and psychological cost. However, in many expertise finding situations, we don't need to find the best expert available but someone who has enough knowledge to help us solve the problem. Thus, when we design an expertise finder, we should provide such expertise level information to the end users, so that they can use it to decide whom they want to contact.

QuME Prototype

I have developed a prototype system based on findings in these three studies, called Question Matching Engine (QuME), shown in Figure 5-1. The QuME system is a combination of web forum with expertise finder techniques. Its core technique is a question-user matching engine that uses both the expertise ranking and topic match information.

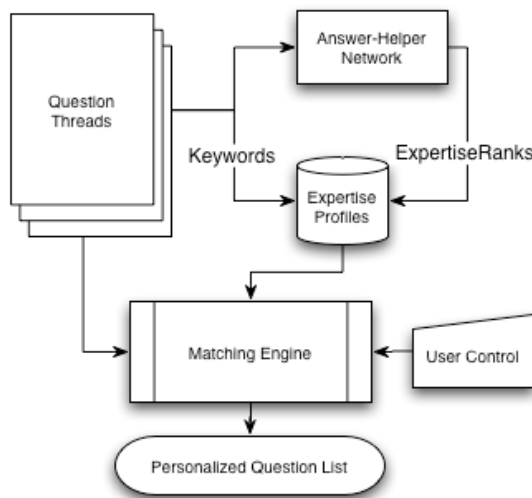







Figure 5-1 The system structure of QuME

Figure 5-2 shows two screenshots of a QuME system interface (as published in Zhang et al. 2007). In this interface, the order of the questions is customized for each viewer according to their expertise profiles. The screenshot in the front is what a high

expertise level user would see, and the one in the back is what a low expertise level user would see. Note the bar next to author's name that indicates his Expertise Rank score. You can see that questions are listed in a roughly descending order according to the asker's expertise ranks. Thus, a user will first see the questions that are slightly below his expertise level. These questions have a higher probability of allowing the answerer to gain new experience while still being capable of providing answers. The expertise-level bar will also help users know whom they are helping.

Topics	Author	Replies	View
.java recovery problem	el3orian 	0	17
Time Updater - clocks now set forward an hour	crackhead 	0	17
reading data out to a txt file	cake 	1	18
How to do this?	aditya15 	3	54
Java applet problem	SilverSurfer 	0	5







Topics	Author	Repli
Obtaining container shutdown	kilyas 	0
Java time Memory Problems	TuringPest 	2
More pri Design recommendation	CSAngel 	0
Help! whe unchecked conversion	CSAngel 	2
BST: Find Reading Annotations Elements	lphazygma 	1
Interactive Text	Chris.G 	1

Figure 5-2 Screenshots of QuME interface

Figure 5-3 shows another interface that is designed specifically for new users who arrive the site to ask questions. After a new user posts a question, the system prompts him to answer some questions. These questions are picked according to their askers' rankings in an ascending order. If a user answers these questions, he will get a high initial rank which will increase the probabilities of his questions being shown to high expertise users. If he does not answer any question, he will be assigned to the lowest rank.

Your question is posted in the forum.

While you are waiting, can you see if you can answer some questions for other users?

This will not only help our community get better, but also will increase the chance of your questions being viewed and answered by an advanced helper. This is especially useful if your question is a difficult one.

Here are some questions you may be able to answer:

Topics	Author
Need help about classpath and packages	ap7926
Newbie - Can't get method to trigger...	NocturnalManNz

Figure 5-3 an interface for new users

With interfaces such as QuME, we believe that one can allocate questions more efficiently to various users. An advanced user's question will have a higher probability of being viewed by more advanced users, thus increasing its chance to be answered faster. For general users, they will receive questions that they can answer, and this will encourage them to answer more questions, thus improving their expertise ranking. This will eventually benefit both themselves and the community.

QuME can be either deployed in an organization as an expertise finding system or added into online communities to augment their expertise sharing functions. QuME has not been tested yet, the users' satisfaction in using such new interfaces, the closeness of the expertise match, and the distribution of reply times for questions of varying difficulty are key things to be tested in future work.

FUTURE WORK

There are still many open questions about expertise networks left in these three studies (e.g. people's motivations to answer others' questions online, the formation process of an expertise sharing community). Furthermore, there are many new intriguing questions raised during the course of these studies (e.g. ranking expertise in multiple topic communities, the impact of user feedbacks on the ranking systems). I am planning

to research on these questions in the future studies. In particular, I want to focus on exploring new ranking algorithms in Yahoo Answers community, studying people's motivations for online expertise sharing, and analyzing the dynamics of community expertise networks.

Expertise Ranking in Yahoo Answers Community

Ranking expertise using people's online interaction histories is a major research topic in my thesis. In the Java Forum study, I have successfully used the network structural information to rank people's expertise. However, in the Yahoo Answer studies, I found that the similar approach did not work very well. Based on the Yahoo Answers study reported in chapter 5, I think that it is very likely because expertise exchanged in Yahoo Answers does not have much depth but has a very broad breadth. Furthermore, most of the questions asked in Yahoo Answers are low expertise questions, thus, they are not interesting enough to solicit answers from high expertise users. However, I still think that there are ways to rank users' expertise in Yahoo Answers. Actually, a good expertise ranking and question-user matching mechanism (like QuME) may help the Yahoo Answers community to solicit high expertise questions and attract high expertise users. I plan to continually explore the ranking algorithms in Yahoo Answers based on the findings reported in chapter 5. In particular, I will focus on following directions:

- Different categories in Yahoo Answers show different social characteristics and reflect users' different expertise needs. Thus, their expertise ranking mechanisms should have different purposes and adopt different algorithms. For instance, a user with highest "best answer" count in the "Celebrity" category may play a very different role from a user with highest "best answer" count in the "Cancer" health category. The former may be more likely to be a social hub while the latter may be more likely to be an active helper with expertise. I want to explore new ways

to automatically clarify different categories into “social expertise sharing” and “technical expertise sharing”, and develop different ranking algorithms for them.

- An important feature of Yahoo Answers is its user feedback ratings. For every question that is being answered, there is a “best answer” selected by the asker or voted by other community users. Currently we use the number of best answers as a direct indicator of one’s expertise levels. This approach is limited and problematic. First, a “best answer” vote from a question with only 1 answer is not as meaningful as a “best answer” vote from a question with 10 answers. Second, a lot of “best answer” votes gained by beating many low expertise users may not be a better expertise indicator than a vote gained by beating one high expertise user. We need find ways to weight these ratings.
- During the Java Forum study, I explored different ways to construct weighted expertise networks. I found that these different ways of constructing expertise networks could not improve the expertise ranking results in Java Forum. However, since Yahoo Answers is very different and has the user feedback information, it will be interesting to re-explore these different network constructing techniques in the Yahoo Answers community.

Motivations of Online Expertise Sharing

People’s motivation to provide help for other people in the online communities has been an important ongoing research topic. It is also one of the design challenges for my early development of SWIM system. Most of previous work relied on interviewing top contributors or surveying general users. Although I have not studied the motivation problem directly in my thesis research, I gained a better understanding of this problem during the Java Forum study and the Yahoo Answers study. An important lesson I learned is that different people have different motivations, especially for people who

played different roles in the expertise sharing process, as well as people who are active in different topic forums. Thus, we should combine interview and survey methods with the analysis of individual users' contributions or activity histories in the communities. Furthermore, it is very likely that people's motivations may change with their experience in the community (i.e. from a newbie to a regular member). Thus, it may also be interesting to conduct longitudinal surveys with the same group of users in a relative long period of time.

Dynamics of Expertise Networks

How does an expertise network evolve? While this problem is difficult to study in organizations, online communities provides us the data that could be traced back to the very beginning of the community launching. By doing longitude analysis on such data set, we could explore how users' roles changes during the community formation, how community network structure changes, and how these characteristics can affect the community sustention and user contributions. I am planning to conduct this study on the Java Forum data set I collected.

Besides the problems above, as a system builder, I always want to know whether social network based expertise sharing systems will work in real organizations or communities, and how people will adopt and react to such new systems. I hope to introduce the QuME system into a real organization to study this issue.

SUMMARY AND CONTRIBUTION

Searching for expertise in one's social networks is one of the most important ways for people to get help when they face intelligence challenges. This thesis investigates large scale knowledge searching and sharing processes in online

communities and organizations. It focuses on understanding the relationship between social networks and expertise sharing activities. The work explores design opportunities of these social networks to bootstrap knowledge sharing, by using the specific social characteristics of social networks which can lead to sizeable differences in the way expertise is searched and shared. The potential impact of this approach was examined in three related studies using data from Java Forum, Yahoo Answers, and Enron.

The Java Forum study investigated how people asked and answered questions in this online community using advanced social network analysis metrics. Furthermore, it explored algorithms that made use of the network structure to evaluate expertise levels. It also used simulations to explore possible social structures and dynamics that would affect the interaction patterns and network structure in online communities. The Yahoo Answers study extended the Java Forum study into a more general community setting and covered much more diverse knowledge sharing dynamics. It analyzed both content properties and social network interactions across sub-forums with different types of knowledge, as well as examined the range and depth of knowledge that users share across these sub-forums. The Enron study, on the other hand, investigated how social network structure could affect the expertise searching process in organizational communication networks using simulations and social network analysis. Based on findings in these studies, a novel expertise sharing system, QuME, was proposed and developed.

This thesis provides a network theoretical foundation for the analysis and design of knowledge sharing communities. It explores new opportunities and challenges that arise in online social interaction environments, which are becoming increasingly ubiquitous and important. This work also has direct implications for practitioners. The ability to add the level of expertise would be a major step forward for expertise finding systems, and would likely open up a range of new application possibilities.

APPENDICES

APPENDIX A

LITERATURE REVIEW

In this literature review, I first survey related studies in the field of expertise sharing because it studies this topic from both social and technical perspectives. Researchers working on this topic come from different disciplines, which include CSCW, AI, and KM, etc. A shared characteristic of these studies is that they are trying to find new ways to help organizations manage knowledge and they have agreed expertise sharing is a direction in which to go.

Then, I survey related work in the field of social network studies. Social networks are the infrastructure for interpersonal interactions. Their structure and dynamics heavily influence people's expertise seeking processes. Studies in social networks provide us a lot of insights and methods to understand and use network structure. Thus, it is important to include them in this review.

This chapter is organized as follows: Section 1 surveys empirical works that try to understand how people search for expertise in practice. Section 2 surveys available technical solutions. Section 3 is my brief survey of related social network studies. Section 4 summarizes previous findings and a concept map is provided and discussed.

UNDERSTANDING EXPERTISE SEARCHING IN PRACTICE

Expertise sharing systems target to augment how people can search and use expertise in organizations. Thus, we must understand the related personal, organizational, and social processes in expertise sharing before we design systems to augment them or before we evaluate available systems. In this section, I am going to survey the empirical work that has studied the actual practices of expertise sharing in various types of organizations. The findings from these studies can provide us a better understanding of the inherent complexity of expertise sharing.

Expertise as an information seeking source compared to others

Information seekers usually need to choose which information sources (human, library, internet, etc) to use before they start the searching process. An important question is what are the benefits and limitations of human as an information seeking source compared to others, such as library and World Wide Web.

When do people search for expertise

Yimam-seid and Kobsa (2003) divided the needs of people searching for expertise into two categories: looking for another person as a source of information and as someone who can perform a given organizational or social function, such as giving a speech. I focus on the first one: finding people as information sources. Yimam-seid and Kobsa suggested that there are different reasons for people choosing a person over other sources. The major ones include:

- Accessing undocumented or nonpublic information. Not all information is accessible because of different cognitive, economic, social, or political reasons (Kautz, Selman et al. 1997) (Volkmar, Hinrichs; et al. 2003).

- Solving problems that are situated. For instance, Orr (1986) showed how informal interpersonal interactions in the form of narratives lead individuals to new understandings of work related problems.
- Leveraging others' expertise to minimize the time and effort in information seeking. For many information seeking tasks, it may take a lot of work for novices but only a little work for experts, especially when people search for information in areas in which they are not familiar (Bhavnani 2005). Experts can help users quickly formulate their information needs into query terms and point them to the valuable information sources available without spending much time (Taylor 1962).

Furthermore, Penuel and Cohen (2003) pointed out that the need of expertise is also related to individual experience. They found that there are two different types of knowledge learning needs in organizations: the learning of newcomers or novices on the job, and the learning of experts. They have different backgrounds and need different supporting strategies. For a newcomer, the most important thing is to find out where expertise is distributed and how they can access it. For experts, they may already know these things, and their needs may be more related to interaction with other experts or people to update and expand their knowledge or solve new problems.

In summary, these analyses indicate that people's expertise needs are diverse and situated. One should consider these various needs in system designs.

Social psychological cost for information seekers

Although there are a lot of unique benefits from seeking information from people directly, in reality, people are not always the first choice for information seekers. Research has found that there are various barriers for people seeking expertise from their colleagues, including social costs and logistical costs (i.e. easy access to the source).

ignorance. This social psychological cost seems to outweigh the benefits of consulting people directly. For instance, Allen found that even when they needed to consult their colleagues, engineers tend to go to the literature first to improve their background in the area so they will not appear ignorant.

Similar findings can be found in later studies in the field of social psychology. Lee (Lee 2002) found that in an organization, fewer than one-third of participants who actually needed help to solve a problem proactively asked other people for help, even though help was available. Lee found that this is because the social cost, including admitting incompetence, inferiority, and dependence, is expensive for a help seeker as it hurts self-esteem and public impression. Furthermore, DePaulo and Fisher (1980) found that a person deciding whether to ask for help not only takes account his own costs, but also the “anticipated cost-reward contingencies” of the helper. An excellent review of various factors that affect people help-seeking behavior can be found in (Gall 1985).

We should note that the social psychological costs for asking for informational help are changeable and vary in different circumstances.

Allen [1977] found that developing social relationships is an effective strategy to decrease the concerns of social psychological cost. When information seekers have good social relationships with available helpers, they tend to worry less about the social cost and can communicate more effectively. The benefit of using social relationships to seek help can also be found in the social network literature (Haythornthwaite 2002) (Shapiro 1980).

Furthermore, Lee (2002) found that the social cost of help seeking is lower for peripheral tasks than central tasks. This implies that when expertise sought is not related to people’s competence evaluation (such as programmers sharing how to cook dinner), the social cost may not be as important to them.

How people Search for Expertise

We need to understand how people find expertise in practice before designing systems to augment it. To my best knowledge, the field study conducted by McDonald and Ackerman in 1998 is the only work that systemically studied how people search for expertise in organizations. McDonald and Ackerman (1998) suggested that the process of finding expertise includes three steps: “expertise identification”, “expertise selection”, and “escalation processes”. In following sub-sections, I used this framework, combined with other related studies, to discuss how people search for expertise in organizations.

Identifying expertise

Expertise identification is about “knowing what information or special skills other individuals have” (McDonald and Ackerman 1998). It is the first crucial step in the process of expertise searching. Understanding how people identify expertise in real life can help us understand how to augment this process in the system designs. McDonald and Ackerman found that there are three ways for people in their site to identify expertise: everyday expertise, historical artifacts, and expertise concierges.

- “Everyday expertise” is about knowing who knows what by everyday “experience”. Similar findings can be found in the studies of “transactive memory” (Moreland, 1999, Wegner 1995). The key idea is that people get to know their colleagues’ expertise based on their daily interactions. “Everyday expertise” is affected by people’s professional experience, organizational tenure, and geographical proximity.
- “Historical artifacts” are historical or archival data that are available, such as programming code changing history, which can indicate one’s previous works and related expertise.

- “Expertise concierges” is about using some specific people who know others well to refer information-seekers to the possible helpers. This concept is similar to “technological gatekeepers” described in Allen (Allen 1977) and “contact brokers” described in Paepcke (Paepcke 1996). In organizations, these are people who usually have strong social networks. They maintain “a sophisticated map of the individuals in the organization and what they know” (MacDonald and Ackerman, 1998). They play the role that mediates information-seeking requests to those who are most likely to have the information. In their study, MacDonald and Ackerman noted that these people are usually managers, who have a “high level of technical competence” and “relatively long tenure with the organization” and “high-status positions”.

Another interesting work on how people get to know one another’s expertise is Fitzpatrick’s case study in a new community. Fitzpatrick (Fitzpatrick 2003) summarized how people get to know others by “finding out in the large” and “finding out in the small”. Information “in the large” is that information “of relatively coarse grain and likely to be easy to find out. ... People are more likely to self-report or that is more amenable to being recorded in some form or to being publicly available” (page 92). Such information includes who worked on what and who knew whom. Fitzpatrick found that people are likely to gain such information through previous experience or from general conversation. Information “in the small” is that “information which is at a much finer level of granularity that people would rarely think to self-report because they would not deem it relevant or important at the time” (page 93), such as small tricks to do a specific task. Such information is usually discovered and shared “by accident in the course of casual conversation”, such as “finding out accidentally, finding out by snooping, finding out incidentally, finding out incrementally, and finding out the real story” (page 94).

In summary, although the task of searching for expertise takes place in only a short time, the process of knowing where expertise is actually is situated complexly in people's everyday activities, including their experience, social interactions, and artifacts.

Selecting Expertise

Expertise seekers usually are faced with choosing from among several possible people, who all possibly have the wanted expertise. In reality, people find it easy to choose. However, to augment this process, we need to understand what criteria are important.

As mentioned, similar social costs (i.e. loss of status), expected reciprocity (i.e. can I return the favor later) and social equity (i.e. how well do they know each other socially) are the key factors that affect decisions on whom to ask for help [Allen 1977]. Lee (2002) found that people prefer to seek help from peers instead of higher or lower levels of their organization's hierarchy because of such social cost considerations.

McDonald and Ackerman (McDonald and Ackerman 1998) further explored the expertise selection problem in detail. They identified three general expertise selection mechanisms: organizational criteria, the load on the source, and performance. Their findings include that people tend to go to local experts first, they compare expert candidates' workload (both regular and over time) before going to them, and they consider an expert's ability for problem comprehension and providing a suitable explanation, as well as their attitude.

In summary, we can see the social and psychological complexity of the expertise selection problem. As MacDonald and Ackerman summarized, "expertise selection is achieved through combination of many, slightly different, behaviors each adding to an individual's judgment about the appropriateness of one or more expertise" (page 6).

Escalation process

Finally, MacDonald and Ackerman also indicated that Expertise finding often involves escalation processes. Escalation is “the way in which people repair failures in identification and selection” (MacDonald and Ackerman, 1998, page 8). Expertise identification can fail in three ways: over-identification (the set of candidates provided is too large), under-identification (the set of candidates provided is too small), or misidentification (none of the candidates provided has the required expertise at a sufficient level). Expertise selection can fail when the selected expert is too busy to respond or does not really understand the problem. MacDonald and Ackerman pointed out that escalation provides a way to either adjust the set of candidates previously identified or to reselect from among those candidates utilizing information gained in the previous attempts. They suggest that expertise searching systems should support such escalation process, such as having some feedback and modification techniques to support users’ previous histories or personal preferences.

Summary and Design Implications

Compared to seeking information from a library or the web, searching expertise from people has many unique benefits. However, it also raises many issues socially, such as various expertise needs and the social psychological cost for askers. Although the expertise searching task seems to take place in only a short time, from the analysis of people’s search for expertise in organizations, we can see that it is actually situated complexly in people’s everyday lives. It is tightly related to an individual’s social experience, the organizational structures, and the surrounding organizational cultures.

Based on the literature, I suggested that these issues are important for designing systems:

- 1) To consider and support various ways to identify different types of expertise (or information):
 - a) Using historical artifacts is a practical way of identifying expertise. With the increasing volume of electronic records (codes, documents, emails, etc) and the development of information retrieval technology, we can easily mine documents to find out what people created or accessed, which hints at what people know or are good at. This is a common method used in most current automatic expertise locating systems. However, relying only on historical artifacts has its limitations. For instance, it is difficult to automatically evaluate one's relative expertise levels.
 - b) Individual experience is an important factor affecting how people identify expertise. However, it is difficult to automatically code personal experience into a computer system. Many available systems provide a person's role and their position in the organizational hierarchical structure, which indirectly reflect one's experience. But such information is not always available. There is also a lack of study on how this "extra" information affects people's usage of the systems.
 - c) In real life, what people know about others is an incremental process, and it is embedded in everyday activities. Lack of support for such incremental processes is one important reason that information in many systems becomes outdated and less valuable. New systems should find more flexible ways to support such processes. Another interesting idea is adding a "new expertise gained by my peers" function into expertise systems, which may foster better expertise awareness.
 - d) We should explore the possibilities in supporting "expertise concierges". This is an extremely important method for people to find possible helpers outside of their immediate social environment or daily experience. There may be two ways to augment this: to provide needed help to these concierges/brokers by decreasing

their workload and increasing their accessibility⁸; or to build automatic broker systems to replace the human concierge. Most of available systems follow the second way, and future studies should direct more attention to the first way.

- 2) To consider various factors that affect people's decisions on expertise selection.
 - Identifying expertise is not the end of the expertise searching process. Simply giving people the best expert available may not work. A more preferred way is providing information seekers candidates who have a satisfying (instead of best) expertise but a low social cost to access.
 - It is difficult (if not impossible) to implement one "identifying and selecting algorithm" for expertise selection (Zhang and Ackerman, 2005). As McDonald and Ackerman pointed out, systems should not automatically select an expert for information seekers. Instead, it should provide a list of candidates with related information to support people's decision making. Such important information includes: availability, social status, and workload, previous interaction histories, etc.
 - Good social relationships can decrease the social cost of expertise searching. The process of asking and answering questions is also a process of using and building social relationships. A system should make use of social networks in expertise selection, as well as helping people maintain and build social networks.
 - Expertise searching may include multiple rounds of expertise identifying and selecting processes. We need consider this in system designs.
- 3) To decrease the negative impact of related social psychological cost. As mentioned earlier, information seekers are hesitating to seek helps from peers because of the

⁸ In many situations, these "brokers" are usually high status people, so they may mean higher social cost for general seekers.

concerns of related social cost. While this issue is more related to organization culture and people's social relationship. It is still interesting to explore ways to decrease its impact in the system design.

EXPERTISE SHARING SYSTEMS

In this section, I will survey available expertise sharing systems. First, I will survey systems that were specifically designed to help information seekers find experts. These applications are usually developed in the field of artificial intelligence (which usually focus on technical perspectives) and CSCW (which usually emphasize both technical and social problems). Second, I will discuss online communities as expertise sharing systems. Lastly, I will summarize available system models and key techniques.

The screenshots of several typical systems can be found in the appendix.

Systems Designed for Finding Experts

Expert Databases

Early systems were usually called expertise databases, knowledge directories, yellow pages, or knowledge maps. Typical systems include Microsoft SPUD and the NASA expertseeker [Davenport et al., 1997]. These systems were usually designed for identifying experts to help solve technical problems or to match employee competencies with positions within the company.

A key challenge for the success of these systems is feeding systems with expertise data. Common approaches include assessment interviews, skill inventories, and extensive surveys (Hoffman, 1995). These methods have several limitations (Dawit Yiman-Seid and Kobsa 2003; Volkmar, Hinrichs; et al. 2003, Lutters et al, 2000). Firstly, they are usually labor intensive and time consuming. Secondly, they tend to collect only fairly flat, one-dimensional assessments of expertise and expertise topics. A lot of expertise information is not inputted into the databases for various reasons. For instance, owners may think some information will not be important for others. People may also input only

expertise that they are really good at. At last, because of the dynamic nature of knowledge practice in organizations, collected data becomes obsolete very quickly, and it can be difficult to update such a system in a timely manner to reflect new and changing expertise.

Furthermore, these systems usually need a taxonomy to describe and catalog people's knowledge. Some use a standard (i.e. Library of Congress) taxonomy, and some design their own. Developing and maintaining taxonomies is not very user friendly in practice. As well, expertise-related queries are usually very fine-grained and context specific. It is hard to find a match between such a query and abstract and general taxonomy description (Kautz, Selman et al. 1997).

Automatic Expert Locators

With the development of the information retrieval technology and the availability of large electronic records of organizations and individuals, researchers developed more helpful systems, which are usually called expertise finders or expertise locators. The key characteristic of these systems is trying to automatically discover expertise profiles from implicit or secondary electronic resources using information retrieval techniques (e.g. indexing). A person's expertise is usually described as a term vector and is used later for matching expertise queries using standard IR techniques.

Typical systems in this category include Who-Knows (Streeter and Lochbaum 1988), ContactFinder (Krulwich and Burkey 1996), and MITRE MII Expert Finder (Maybury, Ray D'Amore et al. 2003). Who-Knows identified experts across an organization by using Latent Semantic Indexing (LSI) techniques on the project documents people produced. ContactFinder monitored discussion groups and extracted indications of expertise from messages, and then answered askers' questions by referring them to people who might have expertise on those topics. Expert Finder not only used

electronic historical artifacts (i.e. documents) people produced to represent their expertise, it provided some experience related information as well, such as people's basic employment information and projects in which they participated.

These systems are more automatic and have lower updating cost. They solve or reduce many problems in the previous generation of systems. However, there are still many problems remaining unsolved. From the technical perspective, we still need to improve ways of selecting and integrating different sources and types of data to better reflect people's expertise. We also need to improve the ways of matching information seekers' fine-grained information needs with the large and amorphous expertise profiles.

However, more importantly, these systems largely do not consider the social perspectives of expertise sharing; for instance, their results are usually ranked purely based on the computed information similarity between the query and profiles⁹. As we have discussed in the previous chapter, this is not how people select experts in real life.

Expertise systems that address social perspectives

Rooted in the field of CSCW, Ackerman and other researchers developed a series of systems that address both social and technical issues.

Answer Garden (AG) (Ackerman 1998) is a system designed to help in situations like technical support, where there is a continuing stream of questions, many of which occur repeatedly, but some of which have never been seen before. It has a branching network of diagnostic questions that helps users find the answers. If there is no available answer, it automatically routes the question to the appropriate expert, then, the expert can answer the user as well as inserting it into the branching network. The design of AG addresses two important social issues in expertise finding. First, askers are anonymous to

⁹ Some are a little bit advanced, such as considering time of the expertise updated.

the experts, thus decreasing the asker's social psychological cost related to status implications and need for reciprocity. Second, by continually adding questions and answers into the corpus, it decreases the expert's workload in answering the same questions repeatedly as well as it grows an organizational memory incrementally. In the field study of AG, experts were manually selected, and there is not much direct interaction support between askers and experts because of the anonymity. In a field study (Ackerman 1998), Ackerman found these designs to be helpful. A number of users reported that it is beneficial to be able to ask questions anonymously. The other interesting finding is that "a large proportion of the users did not get answers that were at the right level or length of explanation." This indicates that expertise systems should route organizational members more effectively to the right level of expertise instead of to the experts with the highest level of expertise¹⁰. Furthermore, Pipek and Wulf (2003) applied the Answer Garden approach into different organizational setting. They found that the incomplete of data, continually changing classification schemes, and domain-specific needs for technically mediation communications made adoption of Answer Garden like system difficult. More importantly, they found that the Answer Garden approach is subject to the impact of the given division of labor and organizational micro-politics.

In Answer Garden 2 (AG2) (Ackerman and McDonald 1996), an expertise location engine is provided. Various computer-mediated communication mechanisms are also added. AG2 also prefers to "stay local" when selecting expertise to allow contextualization and it supports an escalation process. Another interesting change of

¹⁰ This corresponds to the findings on psychological difficulty in communications between experts and novices.

AG2 is that the system tends to blur the dichotomy between experts and seekers¹¹.

MacDonald and Ackerman explained the reason as follow:

“While there was nothing in the underlying technology to force this dichotomy [in AG], it was a simplifying assumption in the field study to have separate user and expert groups. Real collectivities do not function this way. Most people range in their expertise among many different skills and fields of knowledge.. We would like to allow everyone to contribute as they can, promoting both individual and collective learning.” (Page 2)

Expertise Recommender (ER)¹² is another system developed by MacDonald and Ackerman (McDonald and Ackerman 2000). Its design is guided by their findings in their field study that I described in chapter 2. It has a profiling supervisor and an identification supervisor that focus on the expertise identification process. What makes ER different is that it has a selection supervisor that provides various modules to the preference database that maintain personally and organizationally relevant data, such as social networks, workload, etc, thus supporting different ways of expertise selection. The evaluation of ER in his later study (McDonald 2001) suggests that “ER performs well when compared to human performance.” (Page 221) McDonald further addressed that the prior work designed to find the expert or the small set of experts for the whole organization or the whole community “discounts the importance of local knowledge (context) and the inherently social aspects of expertise locating” (Page 221) and suggested the usefulness of using various social factors for expertise selection.

In summary, these systems started to consider various social factors in the support of the expertise searching and sharing process. Their usability studies indicated that these considerations are useful. However, there are two perspectives that I think can be further improved. First, although these systems started to make use of the underlying social network, the implementation of related functions is still preliminary. Second, many

¹¹ AG2 and ER actually start to address this issue.

important techniques, such as identifying expertise, can be further improved. For instance, these systems usually either pick the top level experts or just give a rough estimation of relatedness of one's knowledge. There is no solution that automatically evaluates people's expertise levels.

Referral systems

There is another type of expertise search systems that uses a different approach to find experts. Their designs are based on observations of "referral processes" of how people find experts in their social life, whereby seekers find needed experts through referral by colleagues or friends. Besides providing functions to automatically generate people's expertise profiles, they utilize people's social network information to support information searching or sharing.

ReferralWeb (Kautz, Selman et al. 1997) was the first well known system that utilized social network information for expertise finding. In ReferralWeb, people's expertise is indexed from their publications and other published documents. Social network information is extracted from the co-authorships or co-appearances in their web pages.

Yenta¹³ (Foner 1997) used people's social networks directly as interaction channels and has a distributed system structure. The way Yenta works is really close to how people share and search for information with their social connections. It is basically like a personal agent. It creates people's personal interest profiles by mining documents in their local machines. The profile is stored locally and uses inter-agent communication to find people who have information similar to the query. Yenta also clusters people based on their shared interests to built social coalitions. Another similar system is the

¹³ Yenta actually is not designed specifically for expertise sharing purpose, but it can be used as an expertise sharing system very easily.

SmallBlue system developed by Ehrlich et al. (2007). SmallBlue used people's email and chat logs to infer content and dynamic social networks. A user can first see profile information of the potential experts and get information about the social distance to them, then he can decide whether and how to initiate contact.

Other similar systems can be found in Vivacqua (1999), Maybury (2003), Yu and Munindar (2003), and Lukose et al. (2003). Recently, with the advancement of social network theory research, there are increasing number of peer-to-peer applications designed to share information and resources (files, contacts) using social networks, as well as commercial social network systems (spoke, visiblepath, etc) that are designed to help people share contact information.

We can see that these social network based systems have their advantages. They work in the same way how people find information through their social contacts in real life. They can give people flexible control on what to share and with whom. The technical development of these systems is relatively easy with available peer-to-peer and information retrieval techniques. Their design still needs more consideration of social issues, such as privacy and trust.

Social network based systems could be one of the most promising solutions for expertise sharing. However, current systems are usually based only on the basic idea of using social connections as an interaction channel but lack a deep understanding of the social network characteristics and their relationships with expertise sharing processes. In some sense, they are more like peer-to-peer based systems instead of social network based systems. For instance, as we discussed in Chapter 3, people do not go to every peer equally; they rely more on expertise concierges or information brokers. I will further discuss social networks and expertise sharing in Chapter 4.

Online Community as an Expertise Sharing System

Different from previous expert finder systems, an online community, instead of users seeking experts, has experts come and select whom to help. There are many online communities of this type that are designed for different purposes. Here we address two types: local “knowledge communities” and online learning/technical support communities. They both support expertise sharing in virtual public spaces.

Local Knowledge communities

The idea of “knowledge communities”, which aims to “use computer networks to provide a medium in which individual can come together to share knowledge and expertise” (Ramo 1961, Page 9) dates back to early 1960s. These systems are designed to augment expertise sharing in an organizational community of practice (Lave and Wenger 1991). Available solutions usually include: providing a virtual place (include various communication media) for people to interact and share information, such as a MUD (Nichols 1993), bulletin board systems, or online chatting-like systems for question posting and answering (Ackerman and Palen, 1996; Ackerman and Halverson, 2003).

In these systems, there are usually no assigned roles of experts and users. In theory, everybody can be an information seeker and a helper at the same time¹⁴. People seek help for what they do not know and help others on problems they know. Participation is often voluntary. Ackerman and Palen studied various technical and social issues in “The Zephyr Help Instance” (Ackerman and Palen, 1996). They found that although Zephyr is a very simple technical system, it provides great flexibilities to people

¹⁴ The idea is like what MacDonald and Ackerman argued in their AG2 design.

to negotiate their roles and status (i.e., questioners and answers), as well as social norms for acceptable and preferred behavior.

Internet Online Community

Internet online communities are different from local knowledge communities. They are not bounded by organizational boundaries and as that supported by organizational management or culture. Instead, they are usually bounded by shared professions, interests, or products among their participants. These communities usually have a very large number of participants. People use pseudonyms and usually do not know each other offline. Seeking information to solve a problem is usually one of major reasons for people to come to an online community, especially for technical support or online learning communities, such as Javaforum (forum.java.com), Aximsite (aximsite.com), and Apache support forum (Apache UseNet).

A significant difference between an online learning community and the previous discussed expertise systems is how they develop. As Preece (2000) pointed out, online communities are neither built nor do they just emerge. They evolve organically and change over time. Developers cannot control online community development but they can influence it. I found this point interesting because it may raise some interesting research questions. For instance, how can we influence the development of online community by providing new techniques? Can we link the online community with other expertise systems more tightly to make use of all of their advantages?

Why People Help in Online Community

There is a rich literature on online communities. A thorough and deep analysis of online communities can be found in Wenger (Wenger, 1999). Regarding expertise searching, we are especially interested in the motivation problem.

One study is Constant et al. (Constant et al., 1996), which examined the practices of distant employees exchanging technical advice through a large organizational computer network. They found that employees were willing to share information even they did not know one another personally. They found that there are various reasons and the impact of corporate ties is more important than personal benefits.

However, participating in an online community is obviously different from participating in a corporate one. Other proposed motives include altruism, incentives to support one's community, reputation-enhancement benefits, and expected reciprocity (including specific and generalized), and contributors' sense of efficacy (Kollock, 1999, Ekeh, 1974, Bandura, 1995). Recently, Lakhani and von Hippel proposed a new explanation which is very useful for explaining some findings in my studies. From their study of an Apache (an open source project) helping forum, Lakhani and van Hippel (2003) suggested that the major motivation of information providers is the direct learning benefits plus the low cost of the help process. They found that when they partition the help process into component tasks, 98% of the effort expended by information providers in fact returns direct learning benefits to those providers.

Lastly, there is a new type of motivating method on the Internet –paying people directly. These systems are called “electronic markets for expertise/human competencies” (Lang and Pigneur 1997). A well known example is “Google Answer”. The idea of such system is that people can pay for other people to answer questions or search information for them.

Summary

Many other systems can be found in the survey of (Ackerman and Halverson 2003). But I think the systems discussed above are pretty typical and they represent the various models, structures, and techniques used in expertise sharing system development.

Table A-1 lists my summary of the major aspects a system design should consider and the available technical solutions.

Table A-1 Key perspectives of expertise searching and available solutions

- Types of Expertise searching supported:
 - Domain knowledge: (expertise database systems)
 - Detail problems/keywords: (ContactFinder, Who-Knows)
 - Mixed: (AG)
- System structure
 - Repository (early systems)
 - Repository + matcher (AG, AG2 partly)
 - Peer-to-peer agent based (Yenta, AG2 partly)
 - Public place (Java Forum, Google Answer)
- Life cycle
 - Build once, and maintained slowly (early expertise database systems)
 - Continually growing content in the repository (AG)
 - Continually updating experts' profiles (expertise locators)
 - Continually growing participants and connections (Yenta)
 - Community life cycle (online community).
- Expertise identification
 - Explicit input: self-declaration, professional position, or peer evaluation (early experts database systems)
 - Implicit identifying:
 - Documents/papers authored, projects participated, web pages, emails. (ER)
 - Using brokers (Referral systems)

- Self-identifying (online communities)
- Expertise selection
 - Only based on expertise
 - Selection based on ontology (early expert database systems)
 - Similarity matching using techniques like LSI. (expertise locators)
 - Consider both expertise level and other social factors (ER)
 - Experts select themselves (online community)
- Handling of social cost related issues
 - Rely on organizational policy and culture (most systems)
 - Trying to decrease status implications and need for reciprocity (AG).
 - Emphasize the use of social relationships and reciprocity (Referral systems, AG2, ER, and Yenta)
 - Rely on various self-participating motivations (online community)
 - Rely on direct money incentives (Google Answer)

From this table, we can see that there are multiple solutions for each dimension of the expertise searching process. Each of them has its benefits and limitations. They should be selected based on the target of the system as well as the context in which the system will be deployed. In many situations, it may be a good idea to combine multiple methods and structures together. Actually, after I wrote the first version of this literature review, Reichling et al. (2007) finished a nice case study of designing a real expertise sharing system for a large and complex organization. In their findings, they argued that the research area of knowledge management is not yet mature enough to come up with general concepts. Thus, they suggested that expertise-sharing systems should be flexible

and provides modules for different matching strategies and their parameterization, and design decisions should be grounded in the specific requirements of the organization.

EXPERTISE SEARCHING AND SOCIAL NETWORK

A social network is the infrastructure for interpersonal information interactions. Its structure and dynamics heavily influence people's expertise seeking processes. Researchers in expertise sharing have noted the importance of the social networks very early and built systems using social networks as channels for expertise sharing. However, as I mentioned earlier, most current systems, like ReferralWeb and Yenta, focus only on using social ties as a referral path or interaction channel. They have not considered much the impact of the social structure of organizations or communities, as well as various characteristics of social networks. To design a better system, we need to learn more from social network studies.

Overview of social network studies

Currently, there are basically two lines of social network research: research in the field of sociology and research in the field of statistical physics. Each field has a different research focus and uses different methods.

In the field of sociology, social network analysis (SNA) focuses on relationships between actors rather than attributes of actors (Wasserman 1994). Based on the mathematical foundations of graph theory, statistical and probability theory, and algebraic models, SNA provides a set of metrics to study network properties, including:

- Individual actor level: connectedness, reachability, prominence, betweenness, isolation, and centrality.
- Dyads, triads, and group levels: reciprocity, symmetry, transitivity, clustering coefficient, and cohesions.

- Global level: network density, connectivity, heterogeneity.

In the field of statistical physics, research has focused on common properties of many different kinds networks¹⁵, including social and non-social networks (i.e. Internet, World Wide Web, and biological networks). The research topics include topology, evolution, and complex processes occurring in networks (Dorogovtsev and Mendes 2002; Newman 2003). Compared to focusing on various metrics that measure the individual or network attributes in the field of sociology, these researches usually focus on the general scaling properties of the network, such as the so-called “scale free network” and “small world effect”. Findings in this area have given computer science researchers great help in designing better searching algorithms in various information networks (i.e. web, p2p file sharing, and blogs)(Adamic 1999; Brin 2000; Adamic, Lukose et al. 2001; Menczer 2002; Adar and Adamic 2005).

For the purpose of this paper, I will focus only on several topics I feel important for expertise searching research. In next two sub sections, I will first survey related work on the searchability of social networks, as well as how can we search them efficiently. Then, I will look at some social network characteristics that are important for information searching.

¹⁵ These networks usually are very big compared to networks studied in social science, which are usually within an organization or a community.

Searching¹⁶ in Social Networks

Small world

The classic study on searching in social networks is the “small world” experiment. In the late 1960s, Milgram and Travers found that subjects could successfully send a small packet (with a name, the city, and the profession of the recipient on it) from Nebraska to people in Boston (Travers and Milgram 1969). The subjects did so, even though they had only local knowledge of their acquaintances, by passing the packet to an acquaintance that they believed to be closest to the target. Travers and Milgram found the average length of acquaintance chain is roughly six. The result of this experiment indicated that the social network is searchable¹⁷ and that the paths linking people are short, the so-called the “six degrees of separation.”

A key question in such experiments is how people select the next person to forward the packet or message from among their hundreds of acquaintances, which ultimately leads to a short chain between the sender and the target. Later experiments found that geographic proximity and similarity of profession to the target are the most frequently used criteria by participants [Killworth and Bernard, 1978; Bernard et al., 1982; Dodds et al., 2003]. For instance, in Dodds et al.’s global level small world experiment that involved 60,000 email users and 18 target persons in 13 countries, they found that the geography proximity of the acquaintance to the target dominated the early stage of the chain, because senders are geographically distant. Occupational proximity was used more frequently after the third step. Other related findings in Dodd et al.’s

¹⁶ Actually, we should use “navigating” instead of “searching” because small world local searching process is really a navigating process. But most of early publications used “searching”.

¹⁷ We should be aware that the successful ratio is not very high; originally only 5% of the letters successfully reached the targets.

experiment is that successful searches were conducted primarily through intermediate to weak strength ties, and that the success of the search did not rely on a small minority of exceptional individuals (i.e. social hubs).

Recently, mathematical models have been proposed to explain why these simple heuristics are good at forming short paths (Kleinberg 2000; Watts, Dodds et al. 2002). In general, I prefer the hierarchical network model of Watts et al to Kleinberg's. It assumes that the social network usually has a structure, in which individuals are grouped together by occupation, location, interest, and so on. As well, these groups are grouped together into bigger groups and so forth. The difference in people's group identities defines their social distance. By choosing individuals who have the shortest social distance to the target at each step, people can gradually reach the target in a short path with only local information about their own immediate acquaintances.

This small world model can be easily adopted into a social network based expertise searching processes. Table A-2 is my attempt to understand this model in an expertise searching perspective by comparing the "six contentions" of the small world model and searching approaches.

Table A-2 "Expertise searching" view of small world model, adopted from Watts, Dodds et al. 2002

Original contentions and approaches (find a named person)	To find a person with some type of expertise
1. Individuals in social networks are endowed not only with network ties, but identities	Identities can be viewed as belonging to different expertise groups
2. Individuals break down, or cluster,	This might corresponds to a taxonomic

<p>the world hierarchically into a series of layers, where the top layer accounts for the entire world and each successively deeper layer represents a cognitive division into a greater number of increasingly specific groups.</p>	<p>view of knowledge. Such as Science->Computer Science->AI->referral system.¹⁸</p>
<p>3. Group membership, in addition to defining individual identity, is a primary basis for social interaction, and therefore acquaintanceship.</p>	<p>We interact with people with similar interests and knowledge. People who have similar knowledge often belong to either the same departments or the same professional associations.</p>
<p>4. Individuals hierarchically cluster the social world in more than one way (for example, by geography and by occupation).</p>	<p>Similar things happen in the clustering of the “knowledge world”. George is viewed as a SI professor, as well as a researcher in the field of IR.</p>
<p>5. Based on their perceived similarity with other nodes, individuals construct a measure of “social distance”, which we define as the minimum ultra metric distance over all dimensions between two nodes. This minimum metric captures the</p>	<p>We can define a similar measure for “knowledge distance”.</p>

¹⁸ These clusters and taxonomic category are not necessary mutually exclusive and crisp. The overlap of clusters will only decrease the social distance, thus shorten the search distance.

intuitive notion that closeness in only a single dimension is sufficient to connote affiliation.	
6. Individuals forward a message to a single neighbor given only local information about the network.	In real life, people don't want to spam their colleagues. And they prefer seek expertise locally.
<p style="text-align: center;">Searching approach:</p> <p>Each member of a message chain forwards the message to his or her neighbor who is perceived to be closer to the target in terms of social distance</p>	<p style="text-align: center;">Searching approach:</p> <p>Each member of a query chain forwards the query to his or her neighbor who is perceived to be closer to the target in terms of "<i>knowledge distance</i>".</p>

The analysis above is preliminary. However, we can see that there are many similarities between searching a named person and searching any person that carries wanted expertise. Building a similar small world model for expertise searching would be a very interesting research topic.

Automatization of network searching

In those small world experiments, it is a person who decided to whom the messages were forwarded. Since participants knew the target's location or profession as well as their own local neighbors' related attributes, with the help of their own understanding of the relations and similarities between the target's and their neighbor's identifiable characteristics, they could pick the next person in the searching chain effectively.

Adamic and her colleagues did several simulation studies to explore strategies that could be used in the automatization of the network searching (Adamic, Lukose et al. 2001; Adamic and Adar 2005). They found that the best-connected searching algorithm that makes use of the skewed degree distribution of many networks is an efficient algorithm in power law networks. By passing the query to highly collected nodes first, the query can be spread broadly in the network and find the desired results quickly.¹⁹ Similar algorithms were later adopted in peer-to-peer file sharing networks, such as Gnutella, to replace the traditional broadcast strategies. Compared to the classical breadth-first-search algorithm, which can find the target quickly but with extremely high cost in terms of bandwidth, searching utilizing these high degree nodes proved to be relatively fast and used much less resources.

In another computer simulation study on the HP email network, Adamic and Adar (2005) found that some simple strategies are more effective than best-connected strategies in automatically finding a named person with some known identities, such as using a contact's position in physical space or an organizational hierarchy. Adamic and Adar suggested that "this was due in large part to the agreement with theoretical predictions by Watts et al. and Kleinberg about optimal linking probabilities relative to separation in physical space or in the organizational hierarchy" (Page18).

In summary, Adamic's studies suggest we can find efficient ways to automatically navigate to a person in social networks. Then, is it possible to use similar approaches to automatically search for expertise in social networks?

¹⁹ This strategy did not perform well in the HP email network search because the degree distribution of HP network is not power law.

Automatization of expertise searching in social networks

Recently, some work has been done on automating expertise searching in social networks (Yu and Munindar 2003; Zhang and Ackerman 2005). It is different from the work of Adamic and her colleagues or other small world experiments in which the desired person is known by name or unique identifier. In the expertise searching problem, a suitable person or set of people is not known in advance. One must be found by matching people against a list of attributes.

In their work on “MARS” referral system, Yu and Singh (2003) proposed a distributed expertise searching algorithm and studied related dynamics using simulation. They used the similarity between a query vector and a neighbor’s expertise vector, plus some consideration of one’s historical referring performance, as the criteria for picking the next agent in a referral graph. The simulation results using a scientific co-authorship network indicate using “information scent”²⁰ can help people find experts in such a network.

Following Adamic et al and Yu and Singh’s work, Zhang and Ackerman (2005) compared various strategies that could be used in searching expertise in social networks. We found that using highly connected person or using weak ties is more efficient regarding the searching speed and per-query cost than other strategies. More importantly, we found that “information scent” strategy is not as efficient as Yu and Singh claimed. There could be many reasons for these different results. First, Yu and Singh never compared their searching strategies with other possible strategies. Second, Yu and Singh’s simulation was conducted in a co-authorship network while our simulation was on an email network. Information distribution on these two types of networks may be

²⁰ “Information Scent” is a word I borrowed from Furnas (1997) and Pirolli (1997). The key idea here is that a seeker will follow the information scent (which nodes have the highest information similarity between current node’s profile and the query) in a network search.

different. These results and discussions suggest that we should further look at how information is distributed in social networks.

Important network characteristics that affect network searching

We have discussed the searchability of social networks in previous sections. But to design better searching strategies, we need to understand what characteristics of social networks are important. In this section, we will look at three of these characteristics, including: structural properties of social networks, various centrality measures, and impact of ties.

General structural characteristic of social networks

A social network is usually represented as a graph. However, different from a random graph or other non-social networks, the structure of social networks is highly meaningful and has its special characteristics.

The small world network model suggests a general characteristic of many large scale social networks. The key idea of the small world network model is that most people have a relatively small circle of friends who generally all know each other, but the shortest-path length from one person to any other in the whole world is possible very short (Newman 2003).

Newman and Park (2003) further proposed two important properties that differ between social networks and non-social networks:

- Different patterns of correlation between the degrees of adjacent vertices:
Degrees are usually positively related in most social networks while negatively correlated in most non-social networks. In other words, in social networks, a person who has a lot of social connections tends to connect to other persons who also have a lot of social connections.

- Level of clustering or transitivity: Social networks usually show a high level of clustering while non-social networks do not.

Centralities of actors

The studies on the structural properties of networks have mostly been concerned about an actor's position in a network, which can affect his role in the information dissemination and access. The key idea is that people in different positions in a network will have different access to information, resources, and social support. The most commonly used measures of people's network position are centralities. There are many different types of centralities (Freeman 1979; Bonacich 1987; Newman 2005). Following are several widely used ones:

- The simplest one is *degree centrality*, which simply counts the number of direct connections an actor has. In general, a person with high degree centrality is viewed as socially popular and is like a social hub. The best connect strategy used in Adamic's simulation used this type of centrality. Furthermore, there are in-degree centrality and out-degree centrality that consider the direction of social ties. A person with high in-degree is good at collecting information, while a person with high out-degree is good at spreading information. The weakness of degree centrality is that it takes into account only the immediate ties that an actor has, rather than indirect ties to all others.
- To address the weakness of degree centrality, *closeness centrality* approaches consider the distance of an actor to all others in the network by focusing on the distance from each actor to all others instead of only to local ones. Depending on the definition of "close", there are several slightly different measures for closeness centrality, such as the ones based on the Eigenvector of geodesic distance or based on reachability. People with high closeness centrality are in an

excellent position to monitor the information flow in the network, and they usually have the best visibility into what is happening in the network

- *Betweenness centrality* is another important centrality measure of information flows in the network. It examines “the extent to which an actor is situated among others in the network, the extent to which information must pass through them to get to others, and consequently, the extent to which they are exposed to information circulation within the network” (Freeman 1979, page 215). If a person has high betweenness centrality, he frequently acts as a local bridge that connects the individual to other people outside a group. The technological gatekeepers mentioned in Allen’s study probably had high betweenness centrality. There are also multiple variants of betweenness centrality, such as ones based on information flow or based on random walk.

These measures provide us methods to quantitatively describe the network structure as well as to compare individuals’ differences. More importantly, by comparing these different measures and noting how sociologists explain them, we can better understand that connections among people are not uniformly distributed in the social network. Unlike a theoretically constructed graph, the connections among people in a social network are highly meaningful and vary greatly (Newman and Park 2003; Newman 2003). People with various degrees in social networks also vary on their information access abilities as well as social status (Wasserman 1994). People in different network positions need to be supported differently in designing peer-to-peer based expertise sharing systems because of different accessibility and workload concerns.

The impact of ties

An individual’s network position affects his overall ability to access and diffuse information. However, for each individual’s information seeking behavior in social networks, the strength of his social ties may have an important impact.

The connections between two individuals can have different strengths. The strength of association varies and is not always symmetrical. Usually, in social networks, the strength of association is divided roughly into strong and weak ties²¹. The term of weak tie is firstly used by Granovetter (1973) to represent the ties in a social network that are not strong, such as loose acquaintances that people met at a party. By contrast, strong ties usually mean those who are kin relations or close personal friends. These different tie strengths have different benefits and tradeoffs in searching for information. Weak ties display an important bridging function, allowing information travel from one subgroup to another subgroup in a social network. they can help people get new information and adopt innovations (Granovetter 1973; Brown and Reingen, 1987; Haythornthwaite 2002; Burt 2004). Strong ties have found been more likely activated for the flow of referral information. They are usually perceived to be as bearing lower social psychological cost in the searching process (corresponding to my early discussion in chapter 2) (Granovetter 1973, Allen 1977, Brown and Reingen). When designing local searching algorithms, one needs to consider tie strength.

Summary

The findings in social networks research we discussed above can provide many aids to an expertise sharing study. They provide a deep understanding of the structure that underlies expertise sharing activities. They also provide us methods and tools to analyze this structure. More importantly, they may provide us a new ways of designing expertise sharing system searching expertise in social networks. Different from previous peer-to-peer based referral systems, such new systems should emphasize the understanding of the

²¹ How to measure the tie strength is very vague.

human social network, and small world network searching problem, as well as consider the impact of various network structure properties and the characteristics of social ties.

SUMMARY AND DISCUSSION

The past few years have seen a burst of interest within the CSCW and related communities in building social network based information sharing systems. In this paper, I have targeted related social and technical issues in building a new generation expertise sharing system; I therefore surveyed the previous work from areas of expertise sharing, social networks, and closely related fields. From this survey, we can see that although the development of information technology has provided us various possibilities to develop systems, there are many social issues that still need to be addressed and understood, such as the social psychological cost for askers, processes involved in the expertise seeking tasks, and the impact of social structure.

Figure 2 shows a simple concept map of the expertise sharing problem in my view. From this figure, we can see that expertise sharing is really about finding better ways to tie people, information, and social networks together with the help of technology. To build systems to augment the expertise sharing, we need consider each component and their intersections.

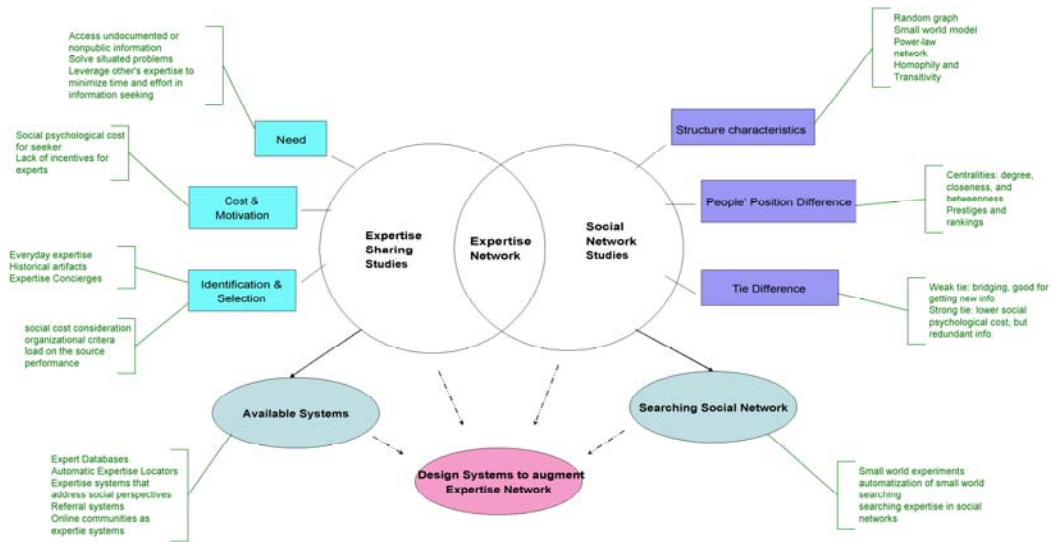


Figure A-2: the concept map of Expertise Network

More importantly, these two figures reveal two gaps which could be our research opportunities.

First, previous systems have not put much attention into the consideration of the social network structure that underlies the expertise searching and accessing process. If we look at the concept map, they basically focus on the intersection between the domain of “individual” and the domain of “expertise”. Newer systems, like ReferralWeb and ER, have started to look at using social networks as a searching and accessing sources. However, their designs lack the consideration of various social network characteristics and their implications. Developing systems with the consideration of social network characteristics should be an interesting research direction.

Second, the intersection between the domain of “social networks” and the domain of “expertise (information)” is under-explored. I think this “blank” area is about the relationship between expertise (information) distribution and social networks. Ackerman and Halverson (2003) suggested that “Expertise is socially arranged and organized”. The idea is that people in a social network vary in their expertise, status, availability, and sociability. The organization structures like roles, responsibilities, and departments “provide orientation and also a cultural background on how people proceed when in need

of expertise” (Ackerman and Halverson 2003). I think there are two questions hidden in these suggestions. The first is how expertise is distributed in social networks. The second is how individuals know and use such distributions of the expertise in social networks. These two problems are fundamental problems for designing social network based expertise searching systems.

REFERENCES

1. Ackerman, M. S. (1998). "Augmenting organizational memory: A field study of answer garden." *ACM Transactions On Information Systems* **16**(3): 203-224.
2. Ackerman, M. S. (2000). "The intellectual challenge of CSCW: The gap between social requirements and technical feasibility." *Human-Computer Interaction* **15**(2-3): 179-203.
3. Ackerman, M. S. and C. A. Halverson (2003). "Sharing Expertise: The Next Step for Knowledge Management." *Social Capital and Information Technology*. MIT Press
4. Ackerman, M. S. and D. W. McDonald (1996). Answer Garden 2: merging organizational memory with collaborative help. *Proceedings of the 1996 ACM conference on Computer supported cooperative work*. Boston, Massachusetts, United States, ACM Press. 97-105
5. Ackerman, M. S. and B. Starr (1996). "Social activity indicators for groupware." *Computer* **29**(6): 37-42.
6. Ackerman, M. S., V. Wulf, et al. (2002). *Sharing Expertise: Beyond Knowledge Management*, MIT Press.
7. Ackerman, M. and L. Palen (1996). "The Zephyr Help Instance: Promoting Ongoing Activity in a CSCW System." *Proceedings of CHI'96*: 268-275
8. Adamic, L. and E. Adar (2005). "How to search a social network." *SOCIAL NETWORKS* **27**(3).
9. Adamic, L. A. (1999). *The Small World Web*. ECDL, Springer.
10. Adamic, L. A., R. M. Lukose, et al. (2001). "Search in power-law networks." *Physical Review E* **64**04(4): art. no.-046135.
11. Adar, E. and L. A. Adamic (2005). Tracking information epidemics in Blogspace. *Web Intelligence*. Compiegne, France.
12. Allen, T. J. (1977). *Managing the Flow of Technology*. Cambridge, MIT Press.
13. Anderson, J. (1999). *Cognitive Psychology and Its Implications*, Worth Publishers.
14. Argote, L. (1999). *Organizational Learning: Creating, Retaining & Transferring Knowledge*, Springer.

15. Bhavnani, S. K. (2005). "Why is it difficult to find comprehensive information? Implications of information scatter for search and design: Research Articles." *J. Am. Soc. Inf. Sci. Technol.* **56**(9): 989-1003.
16. Bill Penuel and A. Cohen (2003). Coming to the Crossroads of Knowledge, learning, and technology: Integrating knowledge management and workplace learning.
17. Bonacich, P. (1987). "Power and Centrality: A Family of Measures." *The American Journal of Sociology* **92**(5): 1170-1182.
18. Brin, S. (2000). The Anatomy of a Large-Scale Hypertextual Web Search Engine.
19. Burt, R. S. (2004). "Structural holes and good ideas." *American Journal Of Sociology* **110**(2): 349-399.
20. Bush, V. (1945). As We May Think. *The Atlantic Monthly*.
21. Carley, K. M. (1999). "ON THE EVOLUTION OF SOCIAL AND ORGANIZATIONAL NETWORKS." *Research in the Sociology of Organizations*.
22. Castells, M. (1996). *The Rise of the Network Society (Castells, Manuel. Information Age, 1.)*, Blackwell Pub.
23. Constant, D., Sproull, L., & Kiesler, S. (1996). "The kindness of strangers: On the usefulness of weak ties for technical advice." *Organization Science*, 7, 119-135. (download PDF).
24. Contractor, N., K. M. Carley, et al. (2000). Co-evolution of knowledge networks and 21st Century Organizational Forms: Computational Modeling and Empirical Testing, UIUC.
25. Contractor, N., D. Zink, et al. (1998). IKNOW: A tool to assist and study the creation, maintenance, and dissolution of knowledge networks. *Community Computing and Support Systems, Lecture Notes in Computer Science*. T. Ishida. Berlin, Springer-Verlag. **1519**: 201-217.
26. Dawit Yiman-Seid and A. Kobsa (2003). Expert-Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach.
27. DePaulo, B. M. and J. D. Fisher (1980). "The Costs of Asking for Help." *Basic & Applied Social Psychology* **1**(1): 23.
28. Dorogovtsev, S. N. and J. F. F. Mendes (2002). "Evolution of networks." *Advances in Physics* **51**: 1079.

29. Ehrlich, K., Lin, C., and Griffiths-Fisher, V. 2007. Searching for experts in the enterprise: combining text and social network analysis. In Proceedings of the 2007 international ACM Conference on Supporting Group Work (Sanibel Island, Florida, USA, November 04 - 07, 2007). GROUP '07. ACM, New York, NY, 117-126.
30. Englebart, D. (1962). Augmenting Human Intellect: A Conceptual Framework. *AFOSSR*. Menlo Park, Calif, Stanford Research Institute.
31. Fitzpatrick, G. (2003). Emergent Expertise Sharing in a New Community.
32. Foner, L. N. (1997). Yenta: a multi-agent, referral-based matchmaking system. Proceedings of the first international conference on Autonomous agents. Marina del Rey, California, United States, ACM Press.
33. Freeman, L. (1979). "Centrality in social networks: Conceptual clarification." *Social Networks* **1**: 215-239.
34. Friedkin, N. E. (1983). "Horizons of Observability and Limits of Informal Control in Organizations." *Social Forces* **62**(1): 54-77.
35. Furnas, G. W. (1997). Effective view navigation. CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM Press.
36. Gall, S. N.-L. (1985). "Help-Seeking Behavior in Learning." *Review of Research in Education* **12**: 55-90.
37. Granovetter, M. S. (1973). "Strength Of Weak Ties." *American Journal Of Sociology* **78**(6): 1360-1380.
38. Grant, R. M. (1996). "Toward a Knowledge-Based Theory of the Firm."
39. Grudin, J. (1989). "Why groupware applications fail: Problems in design and evaluation." *Office: Technology and People* **4**(3): 245-264.
40. Haythornthwaite, C. (2002). "Strong, weak, and latent ties and the impact of new media." *Information Society* **18**(5): 385-401.
41. Hinds, P. J. and J. Pfeffer, Eds. (2003). Why Organizations Don't "Know What They Know": Cognitive and Motivational Factors Affecting the Transfer of Expertise. *Sharing Expertise: Beyond Knowledge Management*. Cambridge, Massachusetts, MIT Press.
42. Kautz, H., B. Selman, et al. (1997). "The hidden web." *Ai Magazine* **18**(2): 27-36.
43. Kautz, H., B. Selman, et al. (1997). "Referral Web: combining social networks and collaborative filtering." *Commun. ACM* **40**(3): 63-65.

44. Kleinberg, J. M. (2000). "Navigation in a small world - It is easier to find short chains between points in some networks than others." *Nature* **406**(6798): 845-845.
45. Krulwich, B. and C. Burkey (1996). *ContactFinder agent: answering bulletin board questions with referrals*. The 1996 13th National Conference on Artificial Intelligence, Portland, OR; USA.
46. Lakhani, K. and E. von Hippel (2003). "How open source software works: "free" user-to-user assistance." *RESEARCH POLICY* **32**(6).
47. Lave, J. and E. Wenger (1991). *Situated Learning: Legitimate Peripheral Participation (Learning in Doing: Social, Cognitive & Computational Perspectives)*, Cambridge University Press.
48. Lee, F. (2002). "The Social Costs of Seeking Help." **38**(1): 17-35.
49. Lukose, R. M., Adar, E., Tyler, J. R., and Sengupta, C. 2003. SHOCK: communicating with computational messages and automatic private profiles. In *Proceedings of the 12th international Conference on World Wide Web (Budapest, Hungary, May 20 - 24, 2003)*. WWW '03. ACM, New York, NY, 291-300.
50. Lutters, W. G., M. S. Ackerman, et al. *Mapping Knowledge Networks in Organizations: Creating a Knowledge Mapping Instrument*. In *proceedings of Americas Conference on Information Systems, Long Beach, CA, 2000*.
51. Malone, T. W. (2004). *The future of work: How the new order of business will shape your organization, your management style, and your life*. MIT press
52. Maybury, M., R. D'Amore, et al. (2000). "Automating the finding of experts." *Research-Technology Management* **43**(6): 12-15.
53. Maybury, M., Ray D'Amore, et al. (2003). *Automated Discovery and Mapping of Expertise*.
54. McDonald, D. W. (2001). *Evaluating expertise recommendations*. *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*. Boulder, Colorado, USA, ACM Press.
55. McDonald, D. W. and M. S. Ackerman (1998). *Just talk to me: a field study of expertise location*. *Proceedings of the 1998 ACM conference on Computer supported cooperative work*. Seattle, Washington, United States, ACM Press.
56. McDonald, D. W. and M. S. Ackerman (2000). *Expertise recommender: a flexible recommendation system and architecture*. *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. Philadelphia, Pennsylvania, United States, ACM Press.

57. Menczer, F. (2002). "Growing and navigating the small world Web by local content." PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA **99**(22).
58. Nardi, B. A., S. Whittaker, et al. (2000). It's Not What You Know, It's Who You Know: Work in the Information Age, *First Monday* 5 (May 1).
59. Newman, M. (2005). "A measure of betweenness centrality based on random walks." *SOCIAL NETWORKS* **27**(1).
60. Newman, M. and J. Park (2003). "Why social networks are different from other types of networks." *PHYSICAL REVIEW E* **68**(3).
61. Newman, M. E. J. (2003). "The structure and function of complex networks." *Siam Review* **45**(2): 167-256.
62. Nichols, P. C. a. D. A. (1993). "MUDs Grow Up: Social Virtual Reality in the Real World."
63. Paepcke, A. (1996). "Information Needs in Technical Work Settings and Their Implications for the Design of Computer Tools." *CSCW*(5): 63-92.
64. Penuel, B., & Cohen, A. (2003). Coming to the crossroads of knowledge, learning, and technology: Integrating knowledge management and workplace learning. *Sharing expertise: Beyond knowledge management*. V. P. M. Ackerman, & V. Wulf. Cambridge, MA, MIT Press: 57-76.
65. Resnick, P., N. Iacovou, et al. (1994). GroupLens: an open architecture for collaborative filtering of netnews. Proceedings of the 1994 ACM conference on Computer supported cooperative work. Chapel Hill, North Carolina, United States, ACM Press.
66. Pipek, V., Wulf V. (2003) Pruning the Answer Garden: Knowledge Sharing in Maintenance Engineering. In proceedings of the Eighth European Conference on CSCW, Helsinki, Finland, 2003, 1-20
67. Reichling, T., Veith, M., and Wulf, V. 2007. Expert Recommender: Designing for a Network Organization. *Comput. Supported Coop. Work* 16, 4-5 (Oct. 2007), 431-465.
68. Shapiro, E. G. (1980). "Is Seeking Help from a Friend Like Seeking Help from a Stranger?" *Social Psychology Quarterly* **43**(2): 259-263.
69. Stanley Wasserman, K. F., Dawn Iacobucci, Mark Granovetter (1994). "Social network analysis: Methods and applications."
70. Streeter, L. and K. Lochbaum (1988). Who Knows: A System Based on Automatic Representation of Semantic Structure. Proceedings of RIAO.

71. Taylor, R. (1962). "The process of asking questions." *American Documentation*.
72. Travers, J. and S. Milgram (1969). "Experimental Study Of Small World Problem." *Sociometry* **32**(4): 425-443.
73. Volkmar, P., J. Hinrichs, et al. (2003). Sharing expertises: challenge for technical support. working paper
74. Watts, D. J., P. S. Dodds, et al. (2002). "Identity and search in social networks." *Science* **296**(5571): 1302-1305.
75. Wegner, D. (1995). "A computer network model of human transactive memory." *SOCIAL COGNITION* **13**(3).
76. Wegner, D. M. (1987). "Transactive memory: A contemporary analysis of the group mind." *Theories of group behavior*, 185-208.
77. Wegner, D. M. (1991). "Transactive Memory in Close Relationships." *Journal of Personality and Social Psychology* **61**(6): 923-929.
78. 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. Denver, Colorado, United States, ACM Press.
79. Wellman, B. (1998). "A computer network is a social network." *SIGGROUP Bull.* **19**(3): 41-45.
80. Wenger, E. (1999). *Communities of Practice: Learning, Meaning, and Identity*, Cambridge University Press.
81. Yu, B. and P. S. Munindar (2003). Searching social networks. Proceedings of the second international joint conference on Autonomous agents and multiagent systems. Melbourne, Australia, ACM Press.
82. Zhang, J. and M. V. Alstyn (2004). SWIM: fostering social network based information search. CHI '04: CHI '04 extended abstracts on Human factors in computing systems, New York, NY, USA, ACM Press.
83. Moreland, R. L. (1999). Transactive memory: Learning who knows what in work groups and organizations. L. L. Thompson, J. M. Levin, D. M. Messick, eds. *Shared Cognition in Organizations: The Management of Knowledge*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 3-31.

APPENDIX B

COMMUNITY NETWORK SIMULATOR

INTRODUCTION

Help-seeking communities have been playing an increasingly critical role the way people seek and share information online, forming the basis for knowledge dissemination and accumulation. Consider:

- About.com, a popular help site (<http://about.com>), boasts 30 million distinct users each month
- Knowledge-iN, a Korean site (<http://kin.naver.com/>), has accumulated 1.5 million question and answers.

Many additional sites exist from online stock trading discussions to medical advice communities. These range from simple text-based newsgroups to intricate immersive virtual reality multi-user worlds.

Unfortunately, the very size of these communities may impede an individual's ability to find relevant answers or advice. Which replies were written by experts and which by novices? As these help-seeking communities are also often primitive technically, they often cannot help the user distinguish between e.g. expert and novice advice. We would therefore like to find mechanisms to augment their functionality and social life. Research is proceeding to make use of the available structure in online

communities to design new systems and algorithms (e.g., [4], [10]). These are largely focused on social network characteristics of these communities.

However, differing network structures and dynamics will affect possible algorithms that attempt to make use of these networks, but little is known of these impacts.

Accordingly, we developed a CommunityNetSimulator (CNS), a simulator that combines various network models, as well as various new social network analysis techniques that are useful to study online community (or virtual organization) network formation and dynamics.

The paper is organized as follows: First, in the next section, we discuss social networks in online communities and their implications, as well as review related work. Second, we describe our CommunityNetSimulator (CNS) and its functionality. Third, using the example of a real-world question and answer forum, we show why simulation is a powerful method to study online community networks. Finally, we discuss CNS' limitations and our future work.

SOCIAL NETWORKS IN ONLINE COMMUNITIES

The Community Expertise Network

There are many forms of social networks. As Wasserman and Faust point out,

“In the network analytic framework, the ties may be any relationship existing between units; for example, kinship, material transactions, flow of resources or support, behavioral interaction, group co-membership, or the affective evaluation of one person by another”. ([26], p. 8)

The main goal of social network analysis is detecting and interpreting patterns of these connections and their implications [20].

Accordingly, while usually the term "social network" implies affinity networks, there are different types of social networks and the meanings attached to them are

different. Some of them are obvious and easy to interpret. For instance, a network generated from the email archives of an organization reflects the communication network of the organization. This can help analysts understand how the information flows [16]. A network generated by co-authorship histories reflects which scientist collaborated together. It helps people understand scientists' collaboration patterns and their shared research interests [18].

But some networks are not obvious. For instance, Amazon generates a co-buying network from customers' transaction histories and uses it to recommend products bought by people with similar purchase histories. People in such a network usually do not know one another even though there is a link between them. The meaning of a link in such a network reflects people's shared interest instead of a direct relationship between two individuals. Sometimes, these "co-interests" can be compared to direct ties, for example, in blogs of different political leanings preferentially linking to one another [3].

Another social network is the flow of expertise and knowledge in online communities (such as newsgroups or web forums). Online communities usually have a thread structure like what is shown in figure 1(a). A user posts a topic or question, and then some other users post replies to either participate in the discussion or to answer a question posed in the original post. Using these threads in a community, we can create a post-reply network by viewing each participating user as a node, and linking the ID of a user starting a topic thread to a replier's ID, as shown in Figure B-1.²²

²² Note there could be multiple ways to convert a topic thread into a network; this is only one of them.

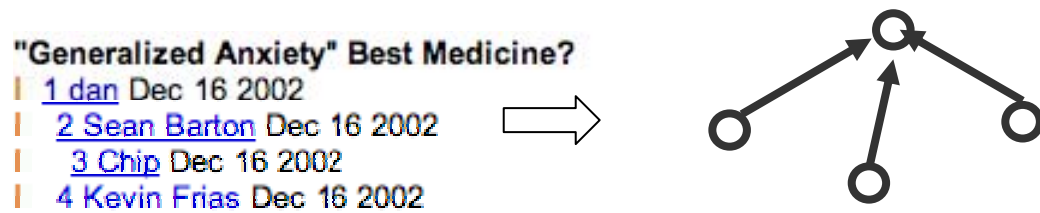


Figure B-1. Method for converting a topic thread into a network

This post-reply network reflects community members' shared interests. Whether it is a community centered around questions and answers, social support, or discussion, the reason that a user usually replies to a topic is because of an interest in the topic. This indirectly reflects that shared interest between the original poster and the repliers²³ (although the repliers' sentiment about the topic may differ).

Furthermore, in some types of communities, the direction of the links may carry more information than just shared interest. For instance, in a question and answer community, a user's replying to another user's question usually indicates that the replier has superior expertise on the subject than the asker. The distribution of expertise, along with the network of responses, is what we will call the *community expertise network* (*CEN*). It indicates what expertise exists within an online community, as well as how it is distributed in practice.

All organizations and communities have their own community expertise network. We might imagine, however, that CENs have differing characteristics among organizations, communities of practice, communities of interest, and the corresponding online communities; that is, they may differ more between types of collectivities than within. Understanding CENs and their differences is critical for knowing how to provide better technical support through online communities, facilitate the flow of technical or knowledge transfer within organizations, and construct effective online communities of practice.

²³ The full dynamic may be much complex in some communities. For example, there may be trolls, spammer, etc.

Studying these community expertise networks, especially with post-reply data, is non-trivial. The next section surveys the work on studying these networks, particularly from a network-analytic perspective.

Research on online community networks

Researchers in various fields have tried to analyze and make use of community expertise networks in different ways.

The first line of study mainly uses network techniques to gain an understanding of the interaction patterns in online communities. Garton et al. [11] describe how online networks could be constructed and analyzed like an offline network, such as measuring the size of the network, individual roles, or using partition techniques to find the formation of groups. But since the network in many online communities is very large, dynamic, and not socially bounded, the methods developed for studying relatively small offline social networks are of limited use.

Many other studies focus on the visualization of the network. Sack[22] used network visualization to display ties between users who either responded to or quoted from one another. Similar work could be found in Donath et al. [9] and Tuener et al. [25].

These visualizations are usually used as an interface to browse and understand patterns of the online community. While these visualizations are interesting and helpful to show various patterns of network structure, these studies focus on building visualization tools instead of further using them to research on various community network structures and the meanings behind them.

To our knowledge, Fisher, et al. [10] was the first to use network structure visualization and analysis to compare and identify different types of online communities, in their case to post-reply networks in newsgroups on Usenet. They found, for example, a correlation between a newsgroup thread's length and time duration and the thread's

content type: question-answer, discussion, flame war, and posting of binaries. They also found that these networks have different ego-centric network patterns and degree distributions, which in turn could be used to categorize different types of participation and to analyze and identify different types of communities.

A second line of study tries to utilize the underlying network structure to develop new applications or algorithms for online communities. For instance, Campbell et al [4, 8] demonstrated, using a synthetic data set, that graph based ranking algorithms, such as PageRank [19], may be applied to conversation networks to rank participants' expertise levels. However, we found that, when applied to a real online question and answer forum, the performance of PageRank was not significantly better than just counting how many other users a user helped [28]. Without simulating a network, it is difficult to pinpoint what factors can account for differences in performance, and moreover, which algorithms are best suited to different online conversation structures. In this case, Campbell's dataset was based on a randomly generated network; but the online community network we studied showed interesting patterns that were actually very different from a random network. These studies indicate that a better understanding of community networks will be required before designing or evaluating new applications.

Above all, these studies indicate that post-reply patterns in online community networks do not follow random patterns. Rather it is the ways in which these networks deviate from random graphs that are important to factor into the design of new systems targeting the use of such underlying networks. Because these communities are self-organizing systems [13], their network structure is an outcome of community users' collective activities that are supported and shaped by various community settings and user preferences and behaviors. How these various factors affect the formation of the community network is an important research question; and the next section discusses how one could use a simulation tool to address it.

Simulation as a Method to Study Community Expertise Networks

Techniques like visualization are useful in providing an overview of the network, as well as helping researchers to find patterns in the network structure. Combined with some careful empirical analysis of the community, researchers may be able to explain why a network has some specific patterns, such as those found by Fisher et al.[10]

However, such an approach has two limitations. First, the size of the online community network is usually very large and dynamic. It can be very difficult to find the meaningful patterns of the network by just looking at the visualization of the network or limited number of available metrics. More importantly, while a visualization may help in identifying some patterns, it does not reveal the underlying factors that influence people's interaction patterns, such as the proportion of various types of users.

Instead in this work, we attempt to borrow theories and methods from organizational studies and complex networks to explore these topics.

Scholars in organizational research have proposed many theoretical mechanisms to explain the emergence and dynamics of communication networks in organizations [15]. These theories, including social capital, mutual self-interest, collective action, social support, and evolution, can help us to gain an understanding of community expertise networks and their emergence. However, we found it was difficult to directly apply these theories and methods to community expertise networks. Most of these theories are constructed based on empirical studies in formal organizations which differ widely from community expertise networks, in which people are less bounded by organizational settings and culture.

Therefore, we used a simulation methodology to examine these theories against observed online interaction patterns. In fact, social network simulations have been used to do this, albeit in a limited manner. For instance, Zeggelink et al [27] used simulation to model and study the subgroup formation in the evolution of friendship

networks. However, these simulations are limited in a small scale and the network metrics used are limited in scope. These simulations are also usually not combined and re-tested with the studies of real networks. In comparison, our work allows for the exploration of a wide variety of network formation algorithms relevant to online communities, and a range of metrics to probe their structure.

Researchers in complex systems have been focused on large scale networks. They developed various models and use simulations to study the formation of some widely observed real-world network characteristics, such as scale-free degree distributions, clustering, and average path lengths[17]. For instance, the preferential attachment network growth model of Barabasi et al. [1] yields scale-free networks just by having new nodes joining the network by linking to existing nodes in proportion to the number of connections they already have. These scale-free networks have a few vertices that become highly-connected hubs, while most vertices have very few connections. Watts and Strogatz' [6] small world model replicates the small-world phenomenon of high clustering and short average path length, by randomly rewiring links in a regular lattice. The regular lattice contributes to clustering – friends of friends are more likely to know one another, and the random links shorten the distance between any two individuals in the network. These models are rather simple, but they proved to be very powerful for understanding the formation of many network structures.

Given that these simple models have been extremely insightful for understanding networks in general, the question remains whether one can apply these models directly to the study of the formation of an online community network. One of their drawbacks is that these models do not consider the social factors that affect the individual interactions. Rather, they usually have a specific network structure in mind as target, and focus on finding simple rules to generate a network that is not in contradiction to real world situations. To do so without a basis in an empirical analysis of the online community, however, would not lead to meaningful models. Indeed, we have tried these models

directly without modification, and found that they did not fit well to observed communities.

For example, in the preferential attachment model applied to the web, a page with many hyperlinks leading to it is more likely to be discovered by a user browsing by following hyperlinks or using a search engine. That user may subsequently include a link to the discovered page on a new page he/she creates. Many models, however, can create scale-free distributions, and may have entirely different underlying dynamics, which are then reflected in very different network characteristics using other measures. And finally, models such as preferential attachment may not make sense in an online community. If we define an edge to exist between someone who starts a thread and everyone who replies to that initial post, then there may or may not be intuitive rationale for preferential attachment.

Thus, we believe that simulations of the online community networks should combine the approaches in both social science and complex system studies. First, we should place an emphasis on studying various factors that possibly affect the structure of the network. Instead of having a targeted network to generate, we should let various factors determine the growth of the network and observe how changing those factors affects the structure of the network. The candidates for these factors should come from empirical studies of online communities. Second, we should have a set of metrics that are very useful for characterizing and comparing the simulated networks against each other and against real world networks. Thus, we could then use such simulations to study how various factors will affect the formation of the network and ultimately the suitability of algorithms that can be applied to the network.

The power of interdisciplinary study is that we can borrow ideas and knowledge from various fields like organizational studies, online community studies and complex network studies. The empirical analysis of the online communities can help us gain some understanding of the important factors that affect people's interaction patterns and how

the network is developed. The simulation models and various network metrics in social sciences and complex system studies provide us tools to further explore their relationship and consequences.

This approach has some additional benefits. Our goal, as mentioned, is to look for the underlying structural characteristics that help determine the community expertise networks for various online activities. One cannot hope to do only empirical examination of these online activities, it would be impossible to intervene sufficiently in real community expertise networks or communication networks. For example, it would be impossible to find companies that would allow us to change their communication patterns. Instead, we can use simulations – bootstrapped from empirically derived data – to investigate changes in underlying structural characteristics.

In the next section, we demonstrate how our CNS simulator provides a powerful and fruitful way to explore the formation of online community networks and their implications.

THE CNS SIMULATOR

Originally, the motivation for us to build the CNS came from our desire to construct network-based algorithms. The goal of these algorithms was to augment an online community by identifying a forum participant's expertise level from the question-answer patterns of his/her posts. We spent a lot of time trying to understand our preliminary results (especially as compared to the literature). While it was clear that the major reason for the different results was that an online community has a very different network structure from a random or web graph, we did not know how and why they were different, as well as what the implications of these differences might be. We decided to try using simulation to explore this issue since there was no other possible way.

Based on our analysis of the question and answer communities we have studied, we found that there were three factors to model for help-seeking communities:

- **Who is more likely to ask questions or initialize topics?**

People have different likelihoods of initiating a question in online communities. For instance, in some communities, it may be that most of the questions are posted by newcomers. But in some internal organization online forums, perhaps all users have an equal likelihood of asking questions.

- **What are users' preferences in replying to a topic?**

People have different motivations for and preferences about replying to a topic. For instance, Lakhani [14] suggested that learning by answering questions is a major reason that people help in an online technical community. In this case, it is very possible that users may prefer to answer questions that are closer to their level of expertise. On the other hand, some researchers argue that altruism or organizational ties are the major reason for answering [5, 12]. In this case, users may just randomly answer the questions that they are capable of answering.

- **What is the distribution of the users with various levels of expertise?**

Users in an online community have various levels of expertise. The distribution of users' expertise (and experience) has a big impact on the formation of the network in an online help seeking community. For instance, if a majority of the users are users new to the products or the domain, then they must rely on a few available experts to help them. If the level of expertise is more evenly distributed, then it is more possible for a greater proportion of users to help one another.

Of course there are many other potential factors. For instance, an incentive system in the community could change users' helping behavior. The diversity of the topics in the community will affect users' chances to have opportunities to use their specific expertise to help others. But the three factors above are most obvious ones, and they were

relatively easy to model as a starting point. As we will show shortly, these three factors create a rich landscape which allows us not only to explain the differences in algorithm performance between our test community and a random graph, but also to explore network structures that may plausibly exist in other contexts.

As mentioned, CNS was mainly developed to examine how these three structural properties affect the formation of the network in a help-seeking community (and in turn how they affect the performance of various ranking algorithms). It is closest in spirit to NetLogo [24]. However, because of the intended use, CNS has two additional capabilities. It provides a set of advanced network analysis methods that can help researchers compare the structural characteristics of the network. As well, CNS provides flexible visualizations and related layout algorithms that were specifically designed to help look for related patterns.

Below we will detail the features of CNS, primarily focused on examining the community expertise network of an online community. The goal is to understand the structural characteristics in order to construct technical mechanisms to support the community. We will give an example of a different use of CNS, understanding an empirical study of an online community, in section 5.

Overview

Figure B-2 shows a snapshot of our CommunityNetSimulator. This snapshot shows the formation of a network.

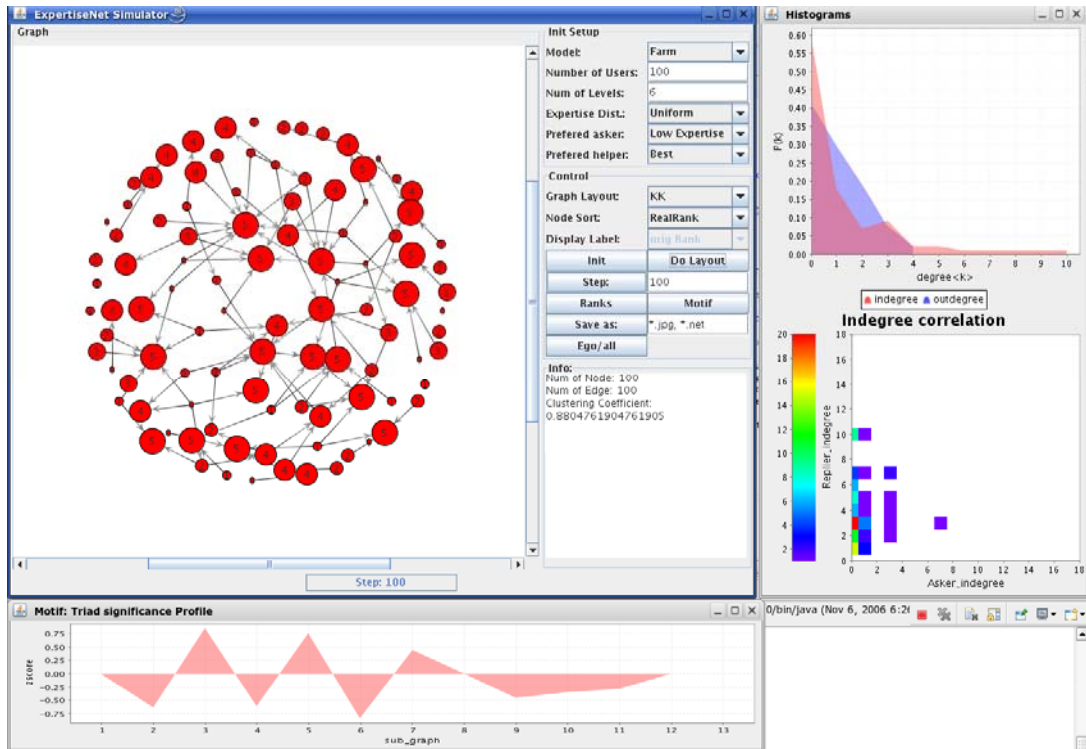


Figure B-2. An overview of CNS

As shown in the figure, there are three types of components in this interface:

- The simulation parameters setup and process controls, through which users can set up the parameters of the simulation and control the process of the simulation.
- The network visualization, which allows users to directly examine the visual patterns of the network being created.
- Network analysis result displays, which include a general network statistic measure report, an in- and out-degree histogram, a degree correlation plot, and a motif profiling analysis plot. These results are automatically calculated and visualized when the network is changed. It gives the user the summary characteristics of generated networks instantly. We will describe these analyses in detail later.

Next we describe the details of several components by walking through the simulator.

Generating Networks

Figure B-3 shows the parameters that we need to set up to create a network like an online community network.



Model:	Farm
Number of Users:	100
Num of Levels:	6
Expertise Dist.:	Uniform
Prefered asker:	Low Expertise
Prefered helper:	Best
<input checked="" type="checkbox"/> Preferential Attach?	

Figure B-3. The simulation parameters

The first step of the simulation is to initialize the parameters of the community to be simulated. There are four parameters that need to be setup: the model, number of users, number of levels, and expertise distribution.

The model parameter determines the basic model of the network. There are two types of network models: “Farm” and “Grow”. In a “Farm” model, the number of users is fixed in the network; and only the links indicating communication or relations are added or altered. In a “Grow” model, a node can be added or removed during the simulation process. The number of users specifies the total number of users in a “Farm” model and the starting number of users in a “Grow” model.

One must also set up the expertise distribution of users in the community. Currently, we assume that there is only one type of expertise in the community and users have different levels. This simulates forums on topics such as “apache server development” or “Sony digital cameras.” One also sets the levels of expertise. For instance, “6” in the “number of levels” creates 6 levels of expertise among the community users. These different levels of expertise can also have different distributions, including Uniform, Normal, and Power Law distributions. Other distributions can be easily added.

After this step, we will have an initial “blank” community that is ready to be developed. Figure B-4 shows two such initialized communities. The first community has 100 users with 6 different levels of expertise that are uniformly distributed. The other has 100 users with 6 levels of expertise but with a power law distribution. Note that the size of the node represents the user’s expertise level.

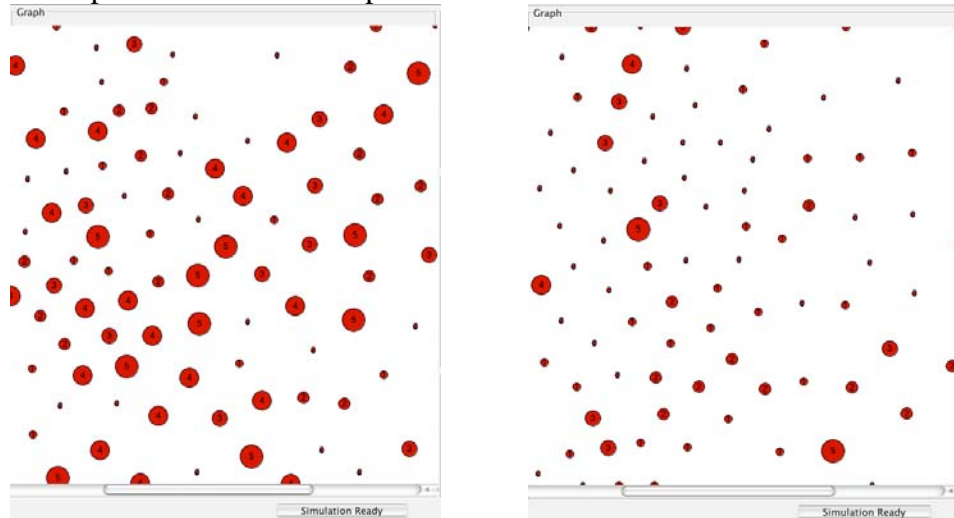


Figure B-4. Two initialized communities. The community on the left has expertise levels uniformly distributed. The community on the right has an uneven power-law distribution: most users have very little expertise, but a few users have high levels.

After we configure the initial condition of the community, we must still set up how the community is going to develop. This is decided by the three parameters controlling the network growth process: “preferred asker”, “preferred helper”, and “preferential attachment”.

The “preferred asker” parameter decides who is more likely to ask questions. We have implemented two “preferred asker” choices in CNS: “Anybody” and “Low expertise”. In the “low expertise” case, a user’s probability to ask questions is determined by the formula below:

$$\text{PossibilityToAskScore}(U_i) = 1 / (EL(U_i) + 1) \quad (1)$$

$$\text{PossibilityToAsk}(U_i) = \text{PossibilityToAskScore}(U_i) / \text{SUM}(\text{PossibilityToAskScore}(U)) \quad (2)$$

Here “EL” stands for “Expertise Level, “Ui” stands for a “user i”, and “U” stands for all users.

Thus, low expertise level users tend to ask more questions. In the case of “Anybody”, everybody has an equal likelihood to ask questions. The former pattern is frequently observed in online forms, where many newbies are seeking help, while the latter may occur within an organization.

The “preferred helper” parameter decides who is more likely to answer the question. There are four basic choices in “Preferred Helpers”: “Best”, “Best better”, “Just better”, “Any better”. We describe only the two typical ones here.

When the “Best” is selected, a user’s probability of answering a question is decided by the formula below:

$$\text{PossibilityToHelpScore}(U_i) = \text{Exp}(EL(U_i) - EL(U_{\text{asker}})) \quad (3)$$

$$\text{PossibilityToHelp}(U_i) = \text{PossibilityToHelpScore}(U_i) / \text{SUM}(\text{PossibilityToHelpScore}(U)) \quad (4)$$

Thus, users who have highest levels of expertise have a higher probability of answering a question. Note that according to this formula, even a user with a lower level of expertise than the asker has a small probability of answering the question. This is natural in many online help seeking communities.

In the case of “Just Better”:

$$\text{PossibilityToHelpScore}(U_i) = \text{Exp}(EL(U_{\text{asker}}) - EL(U_i)) \text{ when } EL(U_i) > EL(U_{\text{asker}}) \quad (5)$$

Thus, users who have slightly better level of expertise than the asker have a higher probability of answering the question, rather than those with a much larger difference in expertise. This may be the case in organizations or communities where

experts' time is limited: It may be the best way for people to make use of each other's time and expertise [2].

The "preferential attachment" selection is used to decide whether a user's previous helping behavior will affect whether he has a high possibility to help more[1]. If it is selected, a user's likelihood to answer a question is not only decided by the expertise level difference between the user and the asker, but also the previous in-degree of the users. The idea is that the more askers a user has helped, the higher the probability that he may help again.

After setting up these parameters, we can run the simulation to generate networks. At each step, an asker is randomly picked based on the "preferred asker" policy. Then a helper is picked to answer the question based on how the "preferred helper" was set up. A directed link is added starting from the asker to the helpers. Figure B-5 shows a growing process of a network when the preferring asker is "low expertise" and preferred helper is "best." Note that while most of the links are from lower level nodes to high-level nodes, there are still some links between high-level nodes because it is still possible for a high level user to ask a question even though this probability is lower than that for low level nodes.

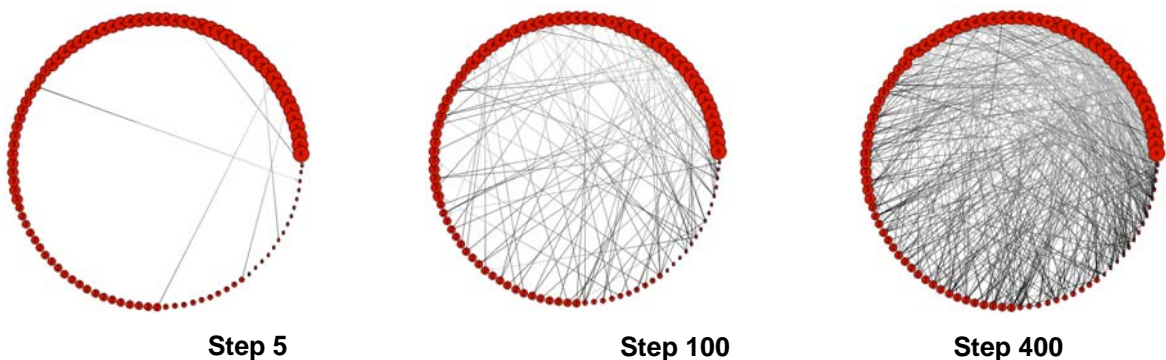


Fig. B-5. The growth of a network. The nodes representing users are arranged on a ring and sized according to their expertise level. Links are drawn between each asker-helper pair, with the direction indicated by the color gradient.

Analyzing Networks

Network Visualization as an Analysis Tool

Network visualization is almost always the first method used to analyze social networks. CNS has a very flexible visualization interface to support visually examining the network. For instance, CNS has various layout algorithms and many filters to highlight or select specific nodes or edges for detailed analysis.

Figure B-6 shows two networks generated by CNS using slightly different parameters. Each network is displayed using two layouts, the top is “Kamada-Kawai” (KK) and the bottom is “circle” [7]. They both are using the farm model, 100 users, 6 levels, normal distribution, and a preferred asker set to “low expertise”. The only difference is the preferred helper. The first one uses “best” while the second uses “just better”. From the visualizations of these two networks, we can see that the network visualization, with the help of different layouts, indeed can help us to observe some patterns that are different between the networks. For instance, from the KK layout, we can see that most high level expertise nodes have a high in-degree in network 1 but not in network 2. From the circle layout, we can see that most of links are connected from low level nodes to high level nodes in network 1 but not in network 2.

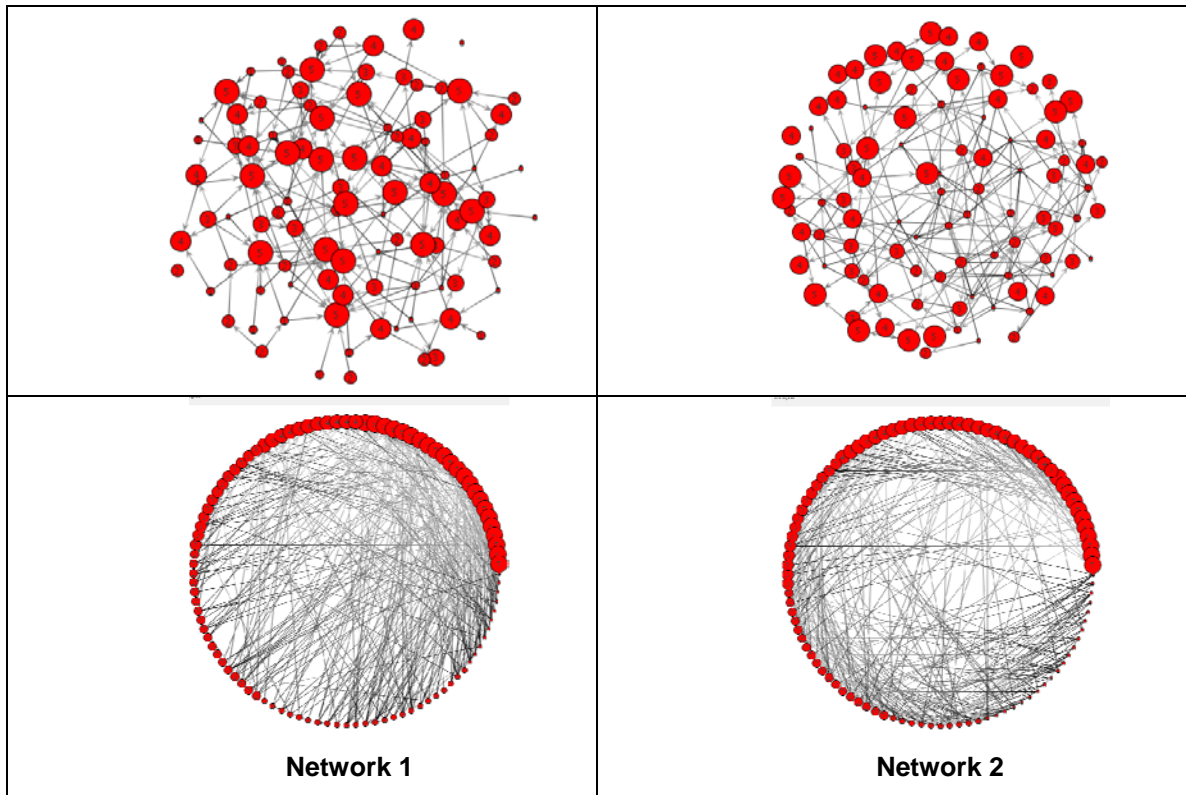


Fig. B-6. Two generated networks

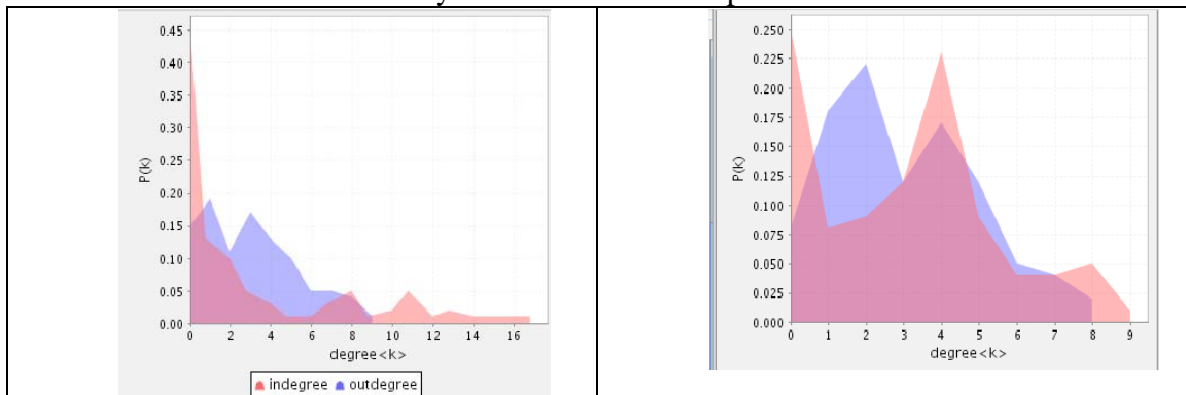
However, besides these findings, the patterns that could be observed from the network visualization are limited. Furthermore, when the network becomes very big or highly connected, it is hard to use visualization to analyze the networks. Below we describe some advanced measures to further compare the various network characteristics.

Advanced Network Analysis Methods

Social network analysis has developed many, by now well established, metrics, such as the average degrees of nodes, density of the network, and the average shortest path. These metrics reveal some overall features of the community and CNS shows them in the general network information panel. However, some more recently developed features lead to three innovative visualizations that CNS can display that we will discuss below. We will use the two networks we visualized in figure 6 to demonstrate the usefulness of these methods.

Degree Histogram

Degree histograms are one of the most frequently used methods to examine large-scale complex networks. A histogram basically characterizes how nodes vary in the number of connections they have. In the context of community expertise networks, it tells us whether some nodes have very different connection patterns from others.



Network 1

Network 2

Fig. B-7. Degree histograms of two networks

Figure B-7 shows the degree histogram of the two example networks. In each histogram, the X-axis represents the degree, and the Y-axis represents what fraction of the total nodes have that many connections. Note that two separate degree distributions are shown, the in-degree corresponding to the number of users the particular user had replied to, and the out-degree corresponding to the number of users who have replied to this particular user.

From these two histograms, we can see that the most significant difference between the two networks is their in-degree distribution. In network 1, the distribution is highly skewed, with a small portion of the nodes having a very high in-degree, while others have a few. In network 2, the in-degree is much more balanced. This tells us that there are some “star” repliers in this network who answered a lot of questions in network 1, while the work of “answering” in network 2 is relative evenly distributed among all community users.

Correlation Histogram

While the in-degree distribution shows how many people a given user helps, it gives no information about the identity of that user's neighbors. For instance, do high volume repliers mainly reply to those who haven't posted many replies, or do they mostly talk to others who are similar to themselves? Correlation histograms are often used in studying network assortativity (characteristics of a node's neighbors) in complex network studies [23], and they are useful in answering such questions.

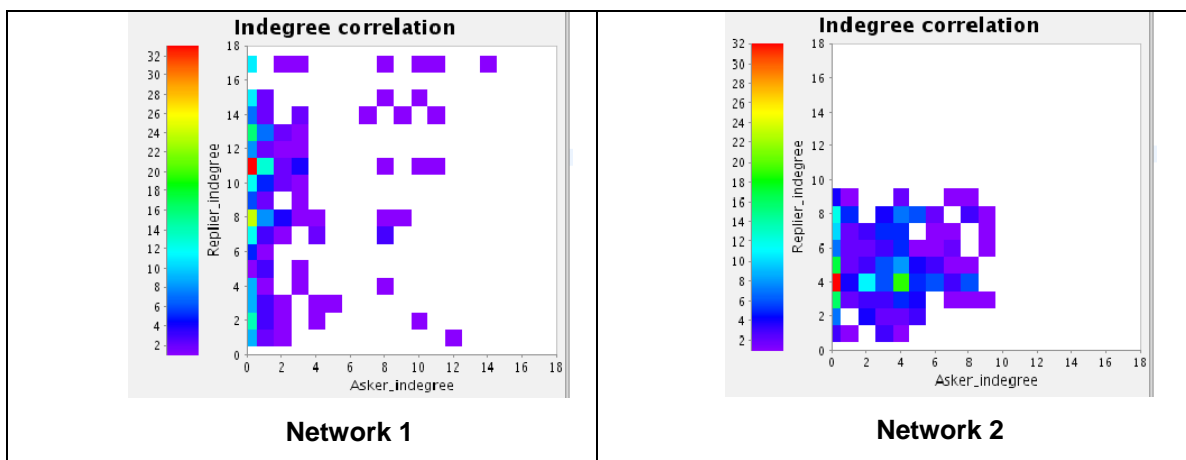


Fig. B-8. Correlation histograms of two networks

Figure B-8 shows the in-degree correlation histograms of the two example networks. In each histogram, the X-axis represents the in-degree of askers, and the Y-axis represents the in-degree for helpers. The color represents the number of pairs of askers and helpers who have the corresponding in-degree.

From these two histograms, we can see that these two networks show very different patterns. In network 1, most of the connections are between high in-degree users and low in-degree users, and there are a few links among high in-degree nodes. In this case, there is a sharp distinction between askers and answerers. In network 2, there are still a lot of links between high in-degree users and low in-degree users, but there are also

a lot of links between medium in-degree users. There is more overlap between askers and answerers in network 2

Motif Profiling Analysis

Are there dyads (two interacting nodes) that indicate reciprocities in the network (i.e., does asking someone a question mean that that user will answer later)? Are there sequential triads that indicate indirect reciprocities in the network, e.g. A helps B who in turn helps C who in turn helps A? The motif profiling analysis, first developed for analyzing biological networks, could be very helpful in answering such questions [21].

There are triad and dyad motif profiles. Figure B-9 shows the triad motif profile of two example networks. The X-axis demarks the different triad subgraphs that are possible (numbered and listed below the motif profile plots). Each graph's Y-axis shows the difference, for each possible subgraph, between the analyzed network and a random network with same connectivity. In the randomized network, each node has the same number of people they helped and received help from as in the original network, but who exactly those other users are is randomized.

From these two diagrams, we can see that the “best” and “just better” helper preferences produce networks with very different triad profiles. For example, network 1, where the ‘best’ helper has a higher likelihood of answering, has many more instances of subgraph 4 than a random network but much fewer of subgraph 5. In subgraph 4, two users help one another, and one of those users also helps a third user. This could correspond to two experts. In subgraph 5, two users are helping one another, and one of those users is also being helped by a third. If the pattern is that of a very good expert typically answering questions, then motifs 4 and 9 might correspond to two experts helping one another and also helping a third user. Motif 5 is unlikely in this scenario because, two people helping each other are much more likely to have a high level of expertise and are therefore unlikely to be helped by others. However, network 2 has a

totally different profile. For example, it has many instances of profile 3, which means that A helps B who helps C. This is possible because questions are answered by someone who is “just better”, meaning that A could have a slightly higher expertise than B and B a slightly higher expertise than C. Such a chain is not particularly likely in network 1, which would prefer to have A answer both B’s and C’s question. The above motif analysis pointed out interesting structure corresponding to two different user behaviors. In this instance, we observe most reciprocity occurring among high-expertise nodes in network 1 but among lower expertise nodes in network 2.

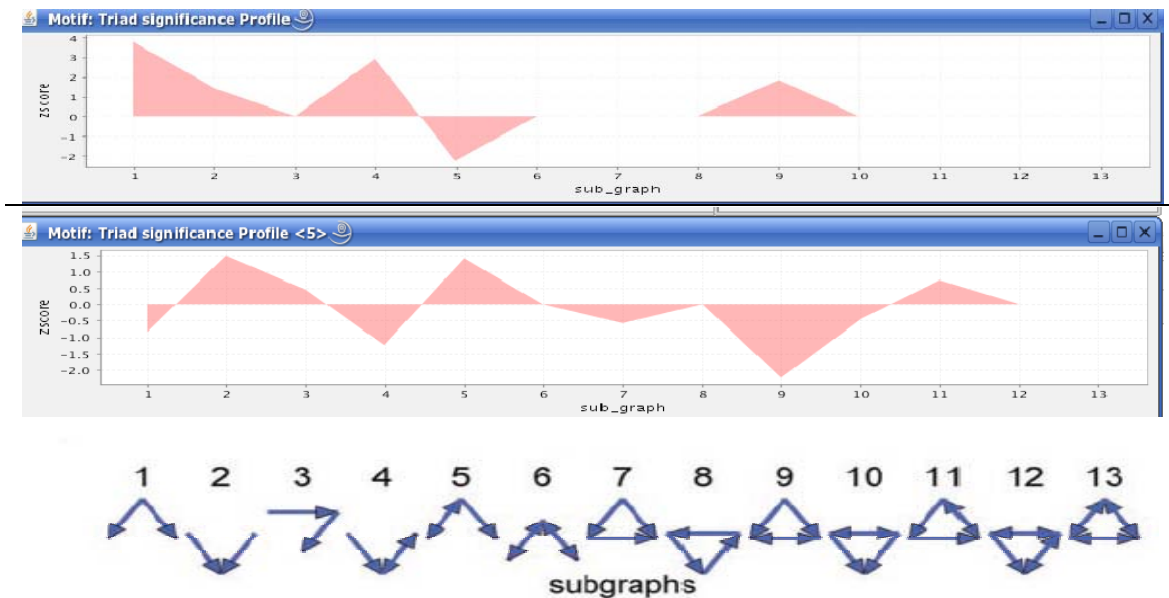


Fig. B-9. The motif analysis plots of two networks

Algorithm Analysis Interface

Concomitant with the original research goals of this project, CNS has a very powerful analysis interface for exploring the performance of various expertise ranking algorithms.

Figure B-10 displays a snapshot of CNS used for analyzing various centrality measures and rankings.

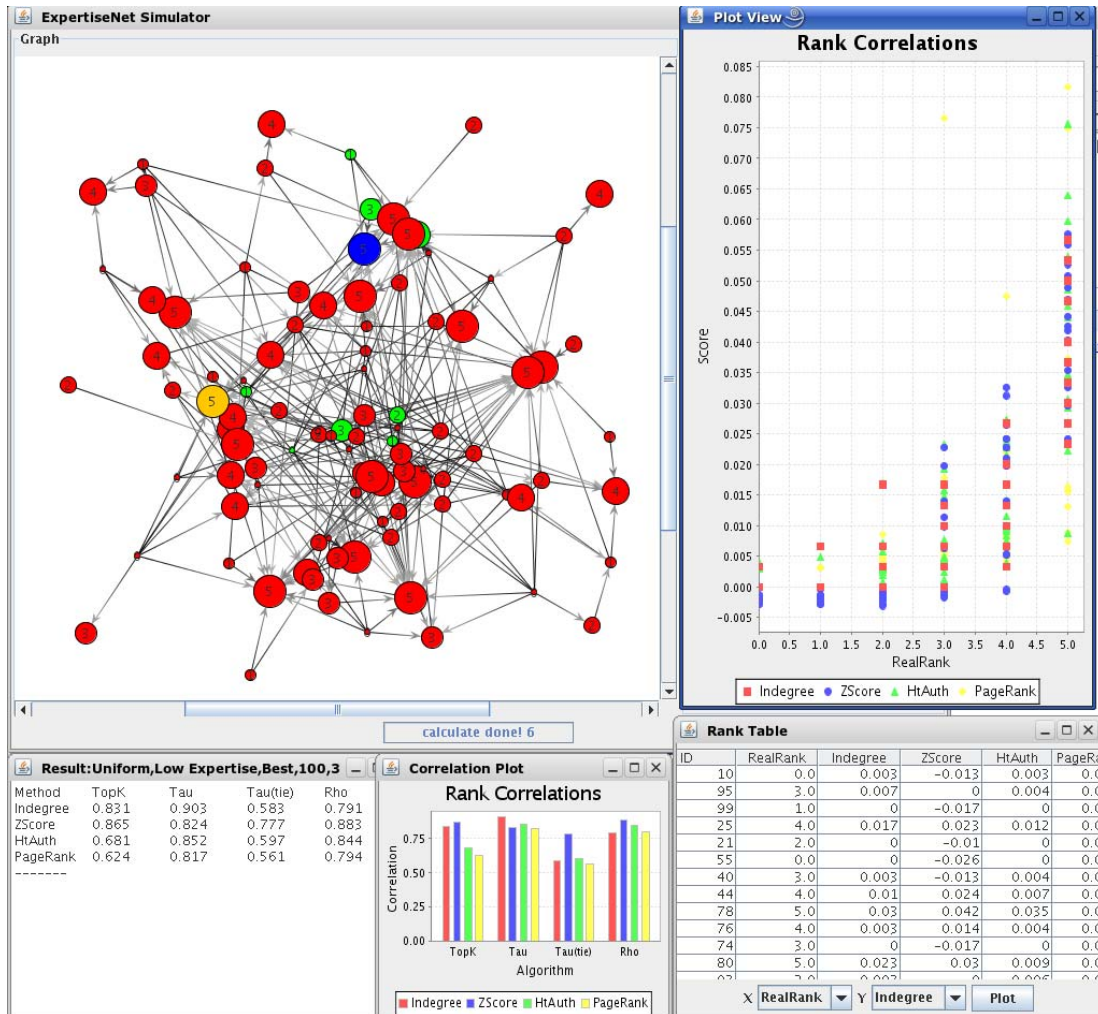


Figure B-10. The algorithm analysis interface

As shown in the figure, the algorithm analysis interface includes five windows: network visualization, a plot of ranks, a table of the ranks, the statistical correlation results for the algorithms, and a chart visualizing the results. The plot of ranks plots the expertise level assigned by the simulation setup on the X-axis, and the expertise level ‘surmised’ through use of the algorithms on the Y-axis. The rank correlation plot shows various rank correlation coefficients between these two variables. From the correlation results window and the chart, one can easily see which algorithm generates ranks that are more correlated to users’ expertise levels assigned by the simulator in the initialization of the community, according to different statistic techniques. Using rank plots and tables, we can examine the individual users and why they are ranked higher or lower than

expected. The rank plot and table are tightly coupled with the network visualization, so clicking on a point in the rank plot or table will highlight the corresponding users in the graph. To further unclutter the view, nodes not in the immediate neighborhood of the node that was clicked on may be temporarily hidden. These visualizations allow one to quickly and easily discover the patterns of interaction between a user and the users they are interacting with that lead to particular outcomes when using ranking algorithms.

While these ranking tools and the algorithm analysis interface are designed for comparing various expertise ranking algorithms, they can be easily modified to study other network-based algorithms (such as those for spreading queries in organizations), as well as issues related to individual prestige in community networks.

CNS AND EMPIRICAL STUDIES

In previous sections, we introduced CNS and its functionality. In this section, we describe how we used CNS to help explain the result we found in an empirical examination of an online community study. We hope this can further demonstrate the utility of our simulator.

In our empirical study, we examined JavaHelpers (not its real name), a place where people come to post questions about Java and get answers from other programmers. We used the "Java Programming" forum in JavaHelpers to examine who asked and who answered questions. At the time of our analysis, the forum had 2,320,345 messages, and the total number of posters, including askers and helpers, was 196,191.

Our goal was to see whether expertise-ranking algorithms worked as reported with a large empirical dataset. The results were a surprise to us. We suspected that the network structure might be the reason and set about using CNS to simulate JavaHelpers. After two rounds of simulation, we were able to find some basic structural characteristics that appear to explain most of the behavior on JavaHelpers.

Initially, based on our empirical analysis of the community, we believed that there were three patterns there.

- There were a number of experts in this online community who mainly answered questions and seldom asked questions.
- The majority the users were either new or had low expertise.
- The experts seemed to answer everyone’s questions.

In the first round of simulation, then, the majority of the askers had low expertise, and high expertise users played the role of helpers. The simulation's results showed a distinction between those who asked and those who answered, as depicted in Figure B-11.

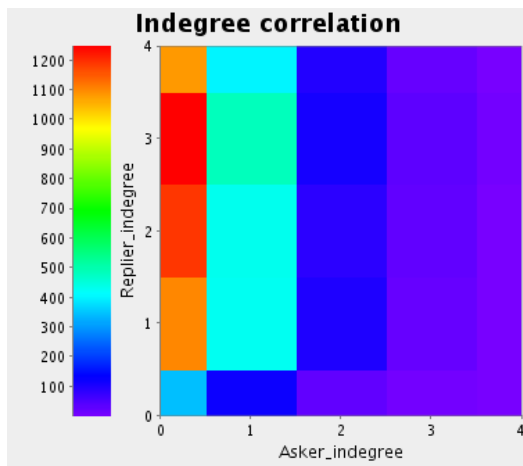


Fig. B-11. The network characteristics of first simulation

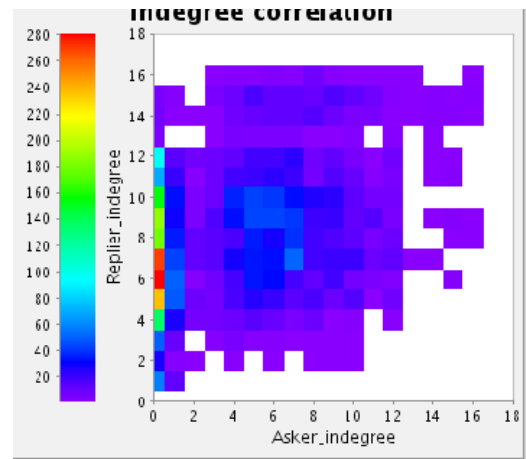


Fig. B-12. The network characteristics of second simulation

However, this simulation did not correspond completely with the empirical dataset. The correlation profile is a bit different from what we found in the empirical study. While most experts in JavaHelpers helped anyone, the other users tended to help people who had a similar level (or just lower) of expertise. Thus, instead of askers always

being helped by the “best” experts available, there were instances where askers being helped by “just better” users, as shown in Figure B-12. (Figure 12 is clearer in color.)

We believe the community is the combination of two subpopulations: the “best” and “just better” groups, each with different response characteristics. Algorithms and other mechanisms (technical or social) must consider both, as should research designs.

Simulations using CNS, then, helped to answer our questions about why algorithms do or do not perform as expected in the communities. When running the algorithms on the real and simulated networks, when the degree distributions and correlations coincide between the real and simulated networks, the algorithms perform similarly as well. Since we know what kind of conditions led to the formation of the simulated network (since we created it), we can tie the performance of the algorithm directly back to the dynamics of the communities. They indicate under what structural conditions, or in what kind of networks, those algorithms will perform best. (And we can do this without requiring interventions in real organizations, experimental conditions which we cannot obtain.) In addition, the simulations can tell us what structural conditions best fit empirical data and help us understand how to better model real communities. So far no other method can accomplish this task.

DISCUSSION AND FUTURE WORK

CNS is a powerful tool for examining online community networks, as well as exploring network-based algorithms. However, CNS, as it currently stands, has some limitations. It does not consider multiple types of expertise, as is the case in real help-seeking communities. In most help-seeking communities, there will be different topics, and individuals will have different levels of expertise for each topic. CNS also does not model learning effects from continued involvement on either individuals or on the community as a whole. Most importantly, we do not yet model tie strengths (types of

relationships) among users. These are all things we would like to add in the future, to better model help-seeking and question-and-answer communities.

Furthermore, the simulations are themselves limited. We have tried, where possible, to tie our simulations to empirically-determined data. However, any simulation is necessarily a simplification of actual practice and social structures. There are important effects, for example, from organizational reward systems, turnover in community participation, conflict over goals, and the like. Nonetheless, we believe we have found important structural characteristics through these simulations that explain a great deal of questioner and answerer behavior. More empirical work will further refine the empirical bases for these models and provide us with a greater understanding of the important factors to model.

It should be noted that CNS can be easily modified through the addition of new capabilities. For example, we can add different probability functions to how people answer questions, and we can add additional visualizations as required. In addition, CNS can be easily modified to study other community network related issues. For instance, we can simulate how hierarchical structures are formed in an online game world by modeling who defeats whom in an adversarial encounter and who talks with whom. Or, we could look at whether the centralities in an organization email network really reflect the importance of a person in the network.

SUMMARY

Simulations are a powerful technique for understanding online communities, especially help-seeking communities. Since we are unable to directly modify a community's expertise network or communication network, we need alternative ways of studying the underlying characteristics that influence how the community functions. Simulations allow us to understand the important characteristics and provide us with data that may not

be obtainable otherwise. (Of course, empirically-based examinations of actual online communities will provide us with the data that we need to bootstrap and to doublecheck simulations.) Coming to an understanding of these help-seeking communities would allow us to better create new ways (technical or social) to augment these communities.

In this paper we have presented the CommunityNetSimulator (CNS), a simulator that combines various network models as well as various new social network analysis techniques that are very useful to study online community networks. CNS' visualizations include degree histograms, correlation histograms, and motif analysis profiles. We have also tried to argue for CNS' utility in community studies. CNS provides substantial capabilities to understand the expertise networks of communities and to consider new augmentations for those networks. This paper has attempted to demonstrate those capabilities.

We believe that simulations, especially combined with empirically based examinations, will be a very fruitful path through which to explore online communities.

REFERENCES

1. Barabasi, A.L. and Albert, R., Emergence of Scaling in Random Networks. *Science*, Vol 286, 1999, 509-512
2. Ackerman, M.S. and McDonald, D.W., Answer Garden 2: merging organizational memory with collaborative help. In *Proceedings of CSCW'96*, ACM Press, Boston, MA, 1996, 97-105
3. Adamic, L.A. and Glance, N., The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In *LinkKDD'05*, Chicago, IL, 2005
4. Campbell, C.S., Maglio, P.P., Cozzi, A. and Dom, B., Expertise identification using email communications. In *the 12th international conference on Information and knowledge management*, New Orleans, LA, 2003, 528-531
5. Constant, D., Sproull, L. and Kiesler, S., The kindness of strangers: the usefulness of electronic weak ties for technical advice. *Organization Science* 7(2). 1996, 119-135
6. Watts, D.J., and Strogatz, S.H., Collective dynamics of 'small-world' networks. *Nature* (393), 1998, 440-442.
7. Díaz, J., Petit, J. and Serna, M., A survey of graph layout problems. *ACM Computing Surveys*, 34 (3). 2002, 313-356.
8. Dom, B., Eiron, I., Cozzi, A. and Zhang, Y., Graph-based ranking algorithms for e-mail expertise analysis. in *DMKD*, New York, NY, ACM Press, 2003, 42-48.
9. Donath, J., Karahalios, K. and Viegas, F. Visualizing Conversations. *Journal of Computer Mediated Communication*, 4 (4), 1999, p.2023
10. Fisher, D., Smith, M. and Welser, H., You Are Who You Talk To. In *HICSS*, Hawaii, 2006, <http://www.hicss.hawaii.edu/HICSS39/Best%20Papers/DM/03-03-08.pdf>
11. Garton, L., Haythornthwaite, C. and Wellman, B., Studying online social networks. *Journal of Computer-Mediated Communication*, 3 (1), 1997,
12. Kollock, P., The economies of online cooperation: gifts and public goods in cyberspace. In Smith, M.A. and Kollock, P. eds. *Communities in Cyberspace*, Routledge, London, 1999, 220-239
13. Krikorian, D. and Kiyomiya, T., Bona fide groups as self-organizing systems: Applications to electronic newsgroups. In Frey, L.R. ed. *Group communication in context: Studies of bona fide groups*, Lawrence Erlbaum, New York, 2002.

14. Lakhani, K. and Hippel, E.v., How open source software works: "free" user-to-user assistance. *Research Policy*, 32 (6). 2003, 923-943
15. Monge, P.R. and Contractor, N.S., Emergence of communication networks. In F. Jablin and Putnam, L. eds. *Handbook of organizational communication*, Sage, Thousand Oaks, CA, 1999.
16. Muir, H. Email traffic patterns can reveal ringleaders. *New Science*, 2003, <http://www.newscientist.com/article.ns?id=dn3550>
17. Newman, M.E.J., The structure and function of complex networks. *Siam Review*, 45 (2). 2003, 167-256.
18. Newman, M.E.J., Who is the best connected scientist? A study of scientific coauthorship networks. *Phys.Rev.*, E64 (016131), 2000
19. Page, L., Brin, S., Motwani, R. and Winograd., T., The Pagerank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project, 1998.
20. Nooy, W.D., Mrvar, A., and Batagelj, V., *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2005.
21. Milo, S.S.-O., Itzkovitz, S., Kashtan, N., Chklovskii, D, and Alon, U., Network Motifs: Simple Building Blocks of Complex Networks *Science*, 298. 2002, 824-827.
22. Sack, W., Discourse Diagrams: Interface Design for Very Large Scale Conversations. In *HICSS 2000*, p.3034.
23. Maslov, S., Sneppen, K., Zaliznyak, A., Pattern Detection in Complex Networks: Correlation Profile of the Internet *eprint arXiv:cond-mat/0205379*, 2002.
24. Tisue, S. and Wilensky, U., NetLogo: A Simple Environment for Modeling Complexity. In *International Conference on Complex Systems*, Boston, MA, 2004
25. Turner, T.C., Smith, M.A., Fisher, D. and Welser, H.T., Picturing Usenet: Mapping computer-mediated collective action. *Journal of Computer Mediated Communication*, 10 (4). 7, 2005 <http://jcmc.indiana.edu/vol10/issue4/turner.html>
26. Wasserman, S. and Faust, K., *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994
27. Zeggelink, E.P.H., Stokman, F.N. and van de Bunt, G.G., The emergence of groups in the evolution of friendship networks. *Journal of Mathematical Sociology*, 21. 1996, 29-55
28. Zhang, J. and Mark, A.S., Adamic, L., Using ExpertiseRank to evaluate expertise in online communities, Technical Report, University of Michigan, 2006