CommunityNetSimulator: Using Simulations to Study Online Community Networks

Jun Zhang[1], Mark S. Ackerman[1,2], Lada Adamic[1,2]

[1]School of Information, University of Michigan

[2]Dept of Electrical Engineering and Computer Science, University of Michigan

{junzh, ackerm, ladamic@umich.edu}

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 1(b).[1]

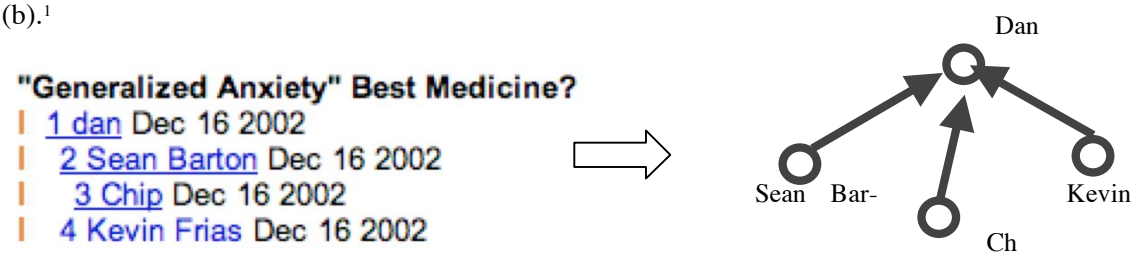

Fig. 1. Method for converting a topic thread into a network

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

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

shared interest between the original poster and the repliers[2] (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.

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,

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

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

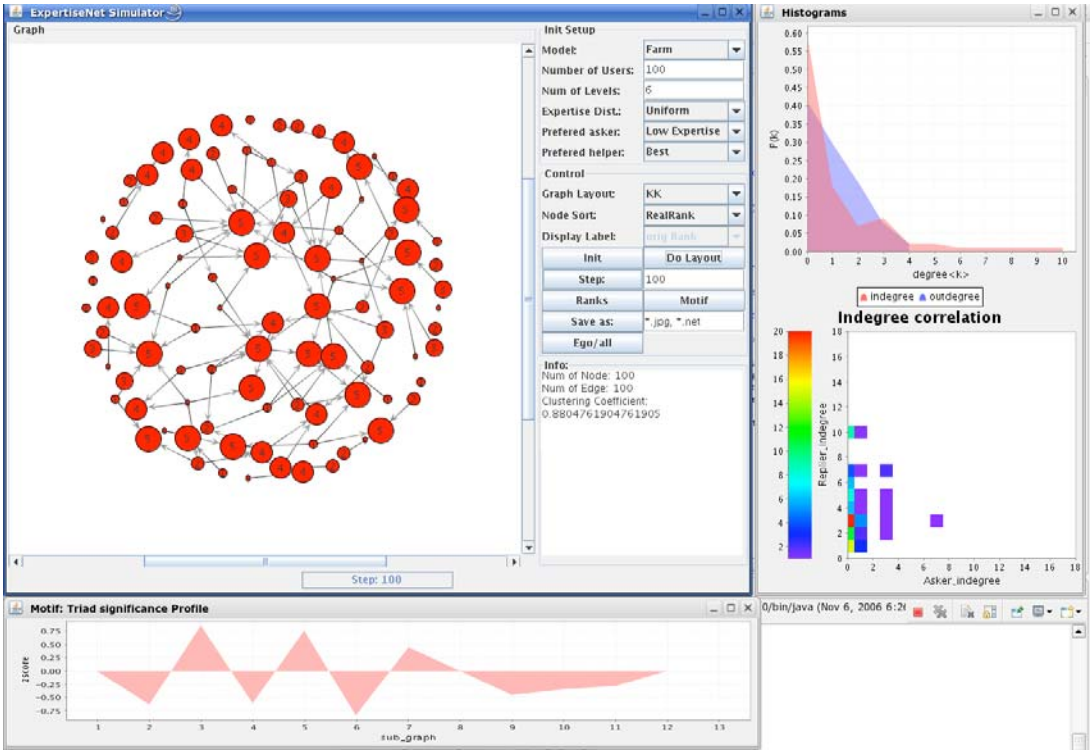

Fig. 2. An overview of CNS

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

Fig. 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 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.

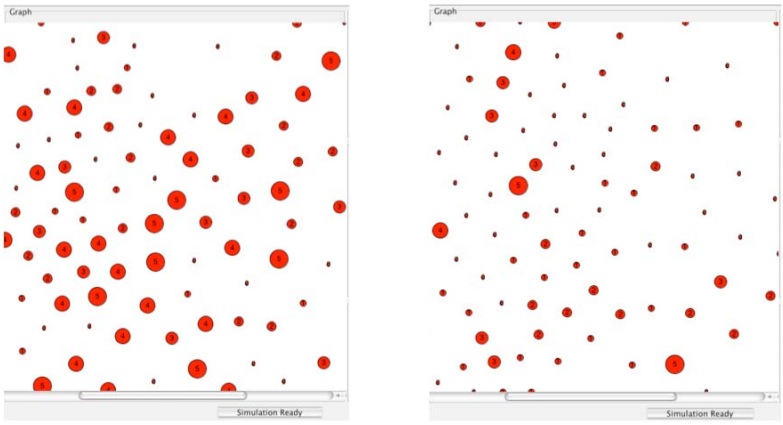

Fig. 4. Two initialized communities. The community on the left has expertise levels uniformly distributed. The community on the right has an uneven power-law distribution: most users have very little expertise, but a few users have high levels.

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:

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

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

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:

$$\text{PossibilityToHelpScore}(Ui) = \text{Exp}(\text{EL}(Ui) - \text{EL}(\text{Uasker})) \quad (3)$$

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

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":

$$\text{PossibilityToHelpScore}(Ui) = \text{Exp}(\text{EL}(\text{Uasker}) - \text{EL}(Ui)) \text{ when } \text{EL}(Ui)>\text{EL}(\text{Uasker}) \quad (5)$$

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 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.

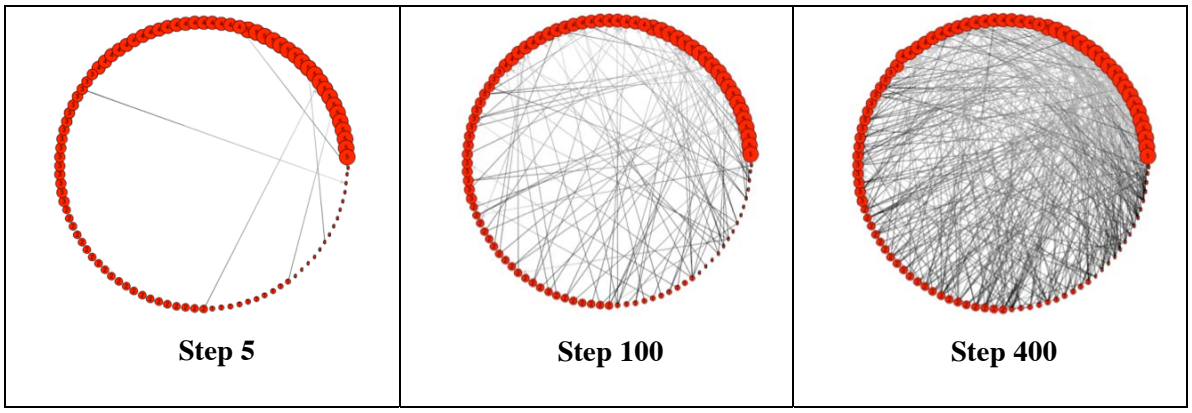

Fig. 5. The growth of a network. The nodes representing users are arranged on a ring and sized according to their expertise level. Links are drawn between each asker-helper pair, with the direction indicated by the color gradient.

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 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.

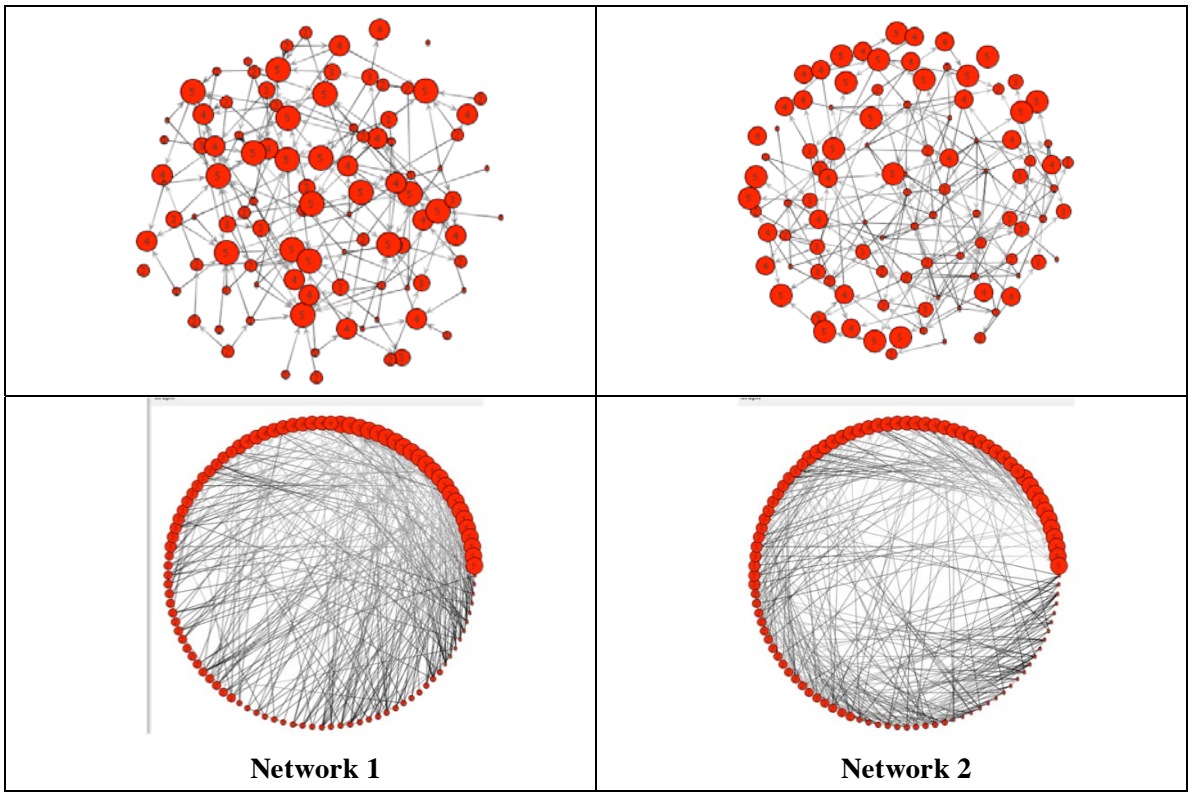

Fig. 6. Two generated networks

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.

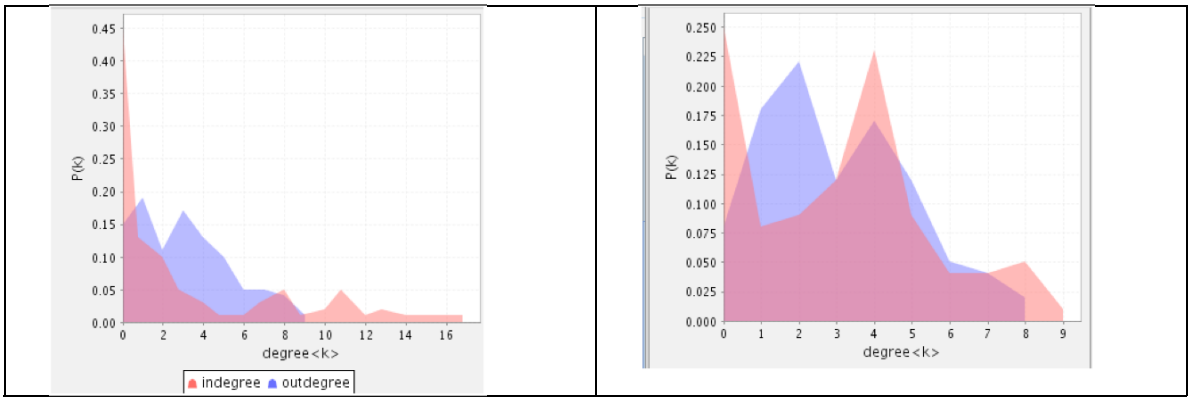

Fig. 7. Degree histograms of two networks

Figure 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.

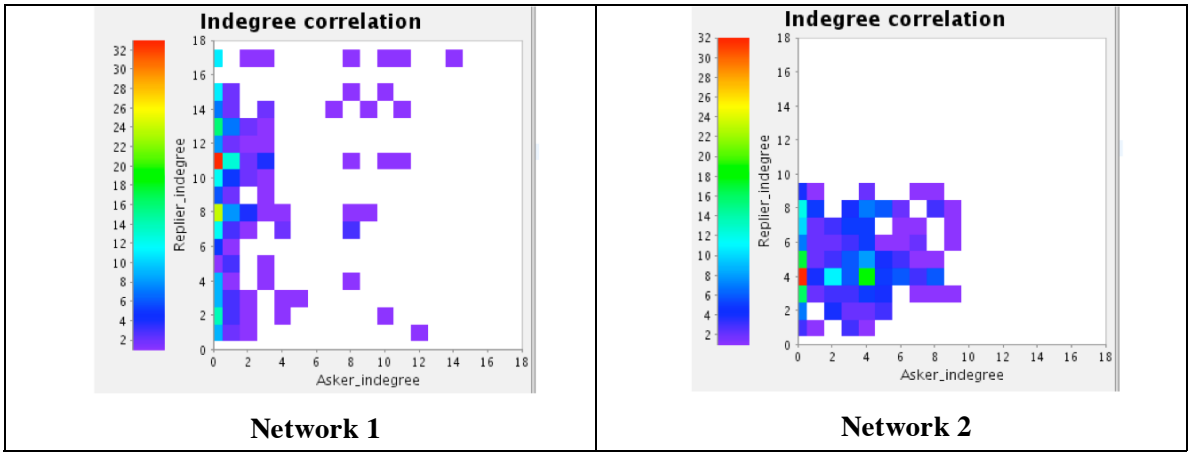

Fig. 8. Correlation histograms of two networks

Figure 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 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.

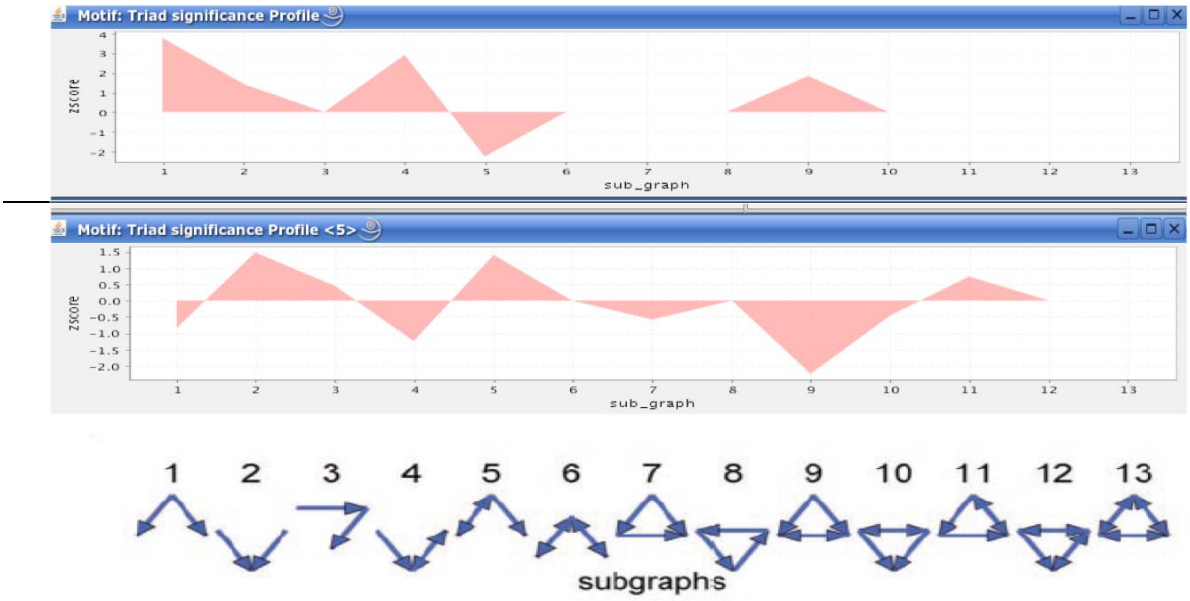

Fig. 9. The motif analysis plots of two networks

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

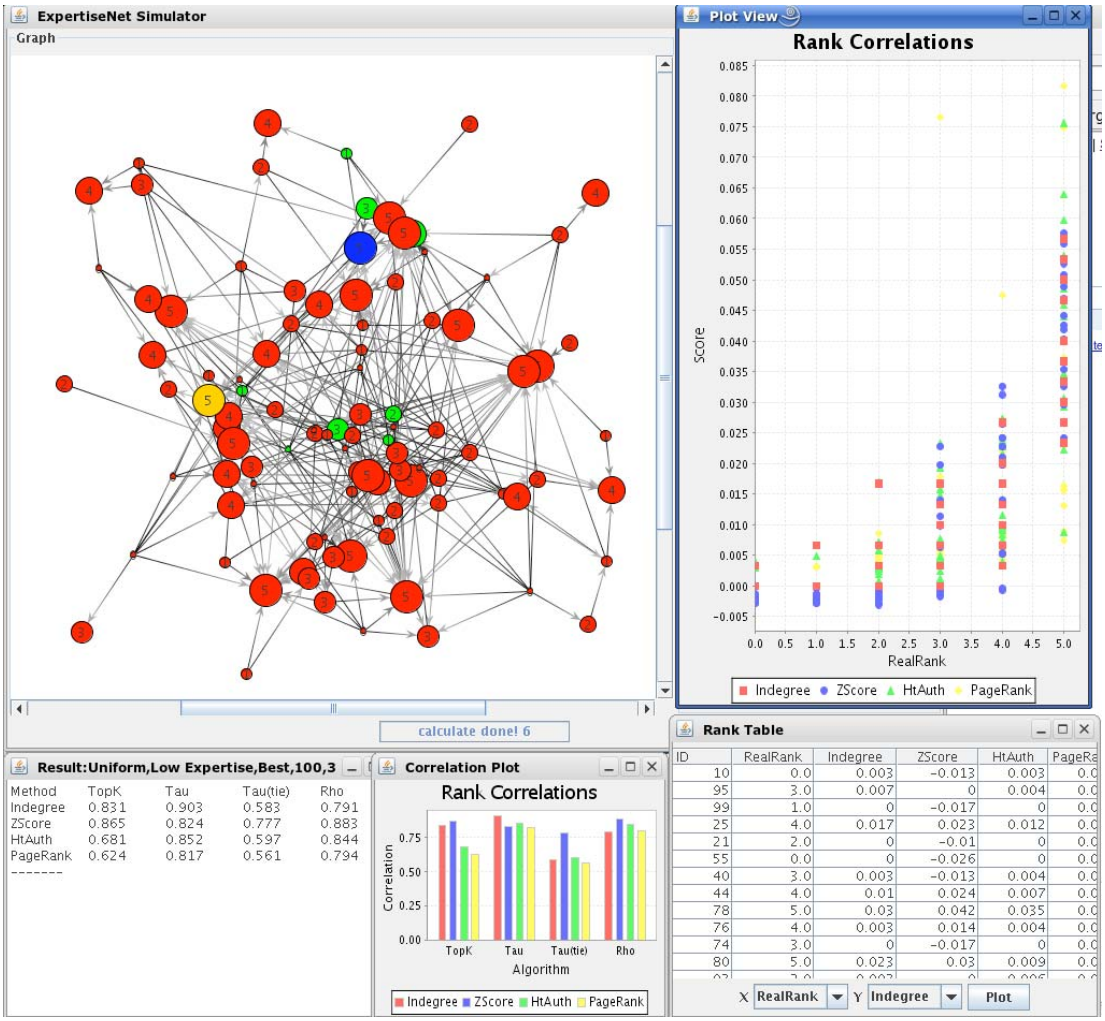

Fig. 10. The algorithm analysis interface

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 11.

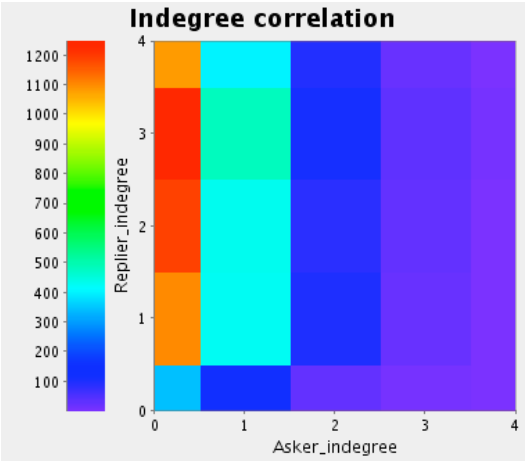

Fig. 11. The network characteristics of first simulation

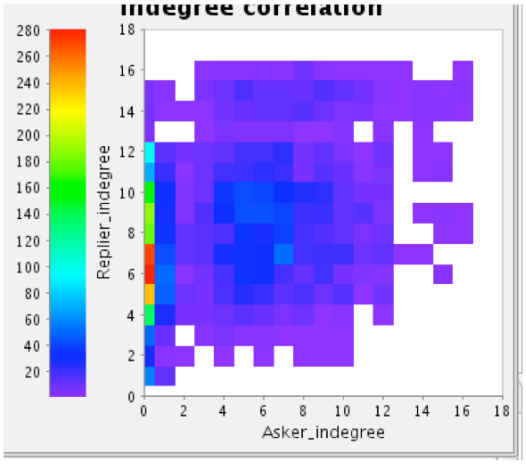

Fig. 12. The network characteristics of second simulation

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 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.

Acknowledgements

This work was supported in part by the National Science Foundation (IIS-0325347). The authors would also like to thank George Furnas, Michael Cohen, the participants in the UM NetSeminar, our research group colleagues, and the anonymous reviewers.

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