Visualizing Threaded Conversation Networks: Mining Message Boards and Email Lists for Actionable Insights*

Derek L. Hansen[1], Ben Shneiderman[2], and Marc Smith[3]

[1] College of Information Studies & Center for the Advanced Study of Communities and Information, University of Maryland, College Park, Maryland, USA
dlhansen@umd.edu
[2] Dept. Of Computer Science & Human-Computer Interaction Lab, University of Maryland, College Park, Maryland, USA
ben@cs.umd.edu
[3] Connected Action Consulting Group, Silicon Valley, California, USA
marc@connectedaction.net

Abstract. Analyzing complex online relationships is a difficult job, but new information visualization tools are enabling a wider range of users to make actionable insights from the growing volume of online data. This paper describes the challenges and methods for conducting analyses of threaded conversations such as found in enterprise message boards, email lists, and forums. After defining threaded conversation, we characterize the types of networks that can be extracted from them. We then provide 3 mini case studies to illustrate how actionable insights for community managers can be gained by applying the network analysis metrics and visualizations available in the free, open source NodeXL tool, which is a powerful, yet easy-to-use tool embedded in Excel 2007/2010.

1 Introduction

Threads are the things that hold the net together. Since the inception of the Internet most virtual communities have relied on asynchronous threaded conversation platforms as a main channel of communication. Usenet newsgroups, email lists, web boards, and discussion forums all contain collections of messages in reply to one another. The natural conversation style supported by the basic post-and-reply threaded message structure has proven enormously versatile, serving communities ranging widely in focus and goals. Cancer survivors and those seeking technical support or religious guidance are as likely to use a threaded discussion as a corporate workgroup. Modern incarnations of threaded conversation are embedded in social networking site wall posts, blog comments, Google Wave threads, YouTube or Flickr comments, and Twitter 'reply to' (RT) tweets. Traditional forums now include profile pages, participation statistics, reputation systems, and private messaging.

* This paper is a revised version of a chapter from "Analyzing Social Media Networks with NodeXL: Insights from a Connected World" by Hansen, Shneiderman, and Smith to be published by Morgan Kaufmann Publishers in Fall 2010.

A. An et al. (Eds.): AMT 2010, LNCS 6335, pp. 47–62, 2010.

Despite the differences in types of threaded conversation, the common structure lends itself well to network analysis, due to its easily identifiable reply structure that captures communication patterns between people. Unfortunately, most threaded conversation systems do not make this networked data easily accessible. The majority of threaded message content is not easily accessible due to the number of different software platforms used and the fact that many groups only make content accessible to subscribed members. Many threaded message systems do report participation statistics and ratings (e.g., top 10 contributors), which are important metrics but fail to capture the social connections between members – a critical component of virtual communities and corporate communities of practice.

This paper considers how to analyze threaded conversations from a network perspective. We begin by defining threaded conversation and characterizing some of the most important networks that can be created from threaded conversation. We then include several brief case studies that demonstrate the value of taking a network approach. The major contribution is to demonstrate novel analysis and visualization approaches that provide users with powerful methods for extracting actionable insights. We rely upon a novel, open source network analysis tool called NodeXL (www.codeplex.com/nodexl), which enables a wider range of analysts to make discoveries and visual presentations that previously required a higher degree of technical skills. These analysts can apply their rich domain knowledge and understanding of social and organizational structures to handle larger datasets and make appropriate business decisions.

2 Definition and Structure of Threaded Conversation

Threaded conversation is a commonly used design theme that enables online discussion between multiple participants using the ubiquitous post-reply-reply structure. It shows up in many forms from email lists to web discussion forums to photo sharing and customer review sites. The key properties of threaded conversation were enumerated in Resnick, et al. [1] and are listed here with some modification:

- Topics. A set of topics, groups, or spaces, sometimes hierarchically organized to aid users in discovering interesting groups to "join." Topics or groups are persistent, though their contents may change over time. Fig. 1 includes two topics: TOPIC 1: Social Media and TOPIC 2: NodeXL.
- Threads. Within each topic or group, there are top-level messages and responses to those messages. Sometimes further nesting – responses to responses – is permitted. The top-level message and the entire tree of responses to it are called a thread. In Fig. 1, there are 5 unique threads. Thread A includes only 2 messages, while Thread B includes 6 messages. Thread D includes only a single message.
- Single Authored. Each message contributed to a thread is authored by a single user. Typically, the person's username or email address is shown alongside the post so people know who is talking. In Fig. 1, the author of each message and the time of their post are indicated. Users may post to multiple threads (e.g., Beth) or multiple times within a thread (e.g., Cathy).
- Permanence. In many threaded conversations including email lists and Usenet, once a message has been posted it cannot be re-written or edited. A new message

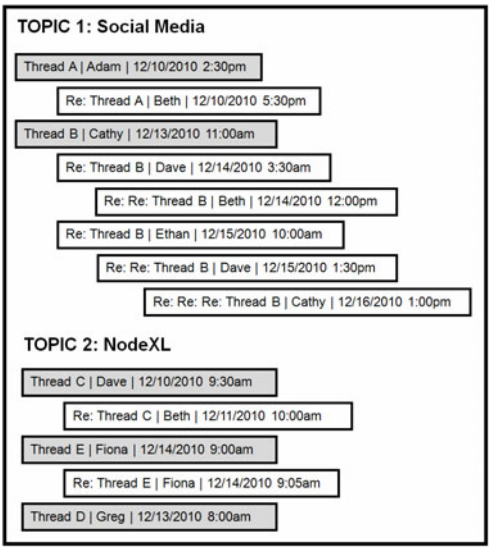

Fig. 1. Threaded Conversation Diagram showing 5 Threads that are part of two different Topics. Each post includes a subject (e.g., Thread A), a single author (e.g., Adam), and a timestamp (e.g., 12/10/2010 2:30pm). Indenting indicates placement in the reply structure. Darker posts initiate new threads (i.e., they are top-level threads), while lighter posts reply to earlier messages in the same thread.

may be posted, but no matter how much someone may wish it, an original post often cannot be retracted. In some discussion boards and newer systems like Google Wave, original posts can be modified after initial contribution.

- Homogeneous View. The partitioning of messages into topics is a feature shared by many discussion interfaces. Moreover, in most systems users all see the same view of the messages in a topic, either in chronological or reverse chronological order. Messages are often sorted into threads (e.g., Fig 1). In some cases, the system will keep track of which messages a user has previously viewed, so that it can highlight unread messages, but that is the only personalization of how people view the messages.

3 Threaded Conversation Research

Research on communities that use threaded conversation began in the early days of Bulletin Board Systems (BBS) and Usenet. Many of the same themes continue to be explored today. For example, Kollock and Smith's book "Communities in Cyberspace" [2] included chapters on identity online, deviant behavior and conflict management, social order and control, community structure and dynamics, visualization, and collective action. All of these topics are still being explored in new contexts and with new technologies such as social networking sites, blogs, microblogging, and wikis. Early books by Preece [3], Kim [4], and Powazek [5] provided some enduring, practical advice and inspiration for those managing online communities. One persistent finding

is the skewed pattern of participation in threaded conversations wherein a few core members contribute the majority of content, many peripheral members contribute infrequently, and a large number of lurkers [6] benefit by overhearing the conversations of others [7].

While most early research on threaded conversations used content analysis, counts of participation patterns, and interviews, a few early researchers applied social network analysis to examine online interactions [e.g., 8-9]. Network analysis approaches are now common, particularly at technical conferences such as the International AAAI Conference on Weblogs and Social Media (ICWSM) that work with large datasets. However, analysis of large-scale networks by academics differs significantly from analysis of bounded networks by community administrators and corporate managers trying to gain insights relevant to their day-to-day actions. In the past couple of years network analysis tools such as NodeXL have made it possible for those without advanced degrees or specialized training to collect, analyze, and visualize networked data from social media sources [10-11]. This has prompted a great need for applied research that clarifies how network analysis techniques can be used to gain actionable insights – the focus of this article.

4 What Questions Can Be Answered?

There are many reasons to explore networks that form within large collections of conversations. New employees or community members need to rapidly catch up with the "story so far" to get to a point that they can make useful contributions. Community managers need tools to help them serve as metaphorical fire rangers and game wardens for huge populations of discussion contributors and the mass of content they produce. When outsiders such as researchers or competitors peer into a set of relationships, social network analysis can point out people, documents, and events that are most notable. A few of the specific questions that can be addressed with network analysis of community conversations are described below:

- Individuals. Who are important individuals within the community? Who are the question answerers, discussion starters, and administrators? Who are the topic experts? Who would be a good replacement for an outgoing administrator? Who fills a unique niche?
- Groups. Who makes up the core members of the community? How interconnected are the core group members? Are there subgroups within the larger community? If so, how are the subgroups interconnected? How do they differ?
- Temporal Comparisons. How have participation patterns and overall structural characteristics of the community changed over time? What does the progression of an individual from peripheral participant to core participant look like and who has made that transition well? How is the community structure affected by a major event like a new administrative team, the leaving of a prominent member, or an initiative to bring in new members?
- Structural Patterns. What network properties are related to community sustainability? What are the common social roles that reoccur among community members (e.g., answer person, discussion starter, questioner, administrator)?

5 Threaded Conversation Networks

Two primary types of networks can be created from threaded conversations: reply networks and affiliation networks, each of which is discussed here and illustrated later in the article with examples.

5.1 Reply Networks

Each time someone replies to another person's message, she creates a directed tie to that other person. If she replies to the same person multiple times, a stronger weighted tie is created. A reply graph treats the message authors as the graph vertices and the reply connections as the graph edges. There are two types of reply networks, depending on how you determine what constitutes a reply. The direct reply network connects a replier to the person they are immediately replying to in the course of a thread (see Fig. 2). In contrast, a top level reply network connects all repliers within a thread to the original thread author.

NodeXL is a free and open source plugin for Microsoft Excel [10]. Network data about edges and vertices are stored in the spreadsheet, while network visualizations are displayed in the graph pane. The spreadsheet portion includes separate worksheets for the Edges (shown in Fig. 2), the Vertices (which includes a unique list of each vertex in the network and visual properties associated with them), and other data of interest such as clusters and overall graph metrics. Different visual properties such as edge width, color, and opacity can be mapped to data properties such as edge weight (i.e., number of messages exchanged) or edge type. Similarly, vertex size, color, opacity, and shape can be mapped to graph metrics (e.g., degree, betweenness centrality) or other attribute data (e.g., demographics). Vertex and edge labels can be displayed in multiple ways. Advanced features allow analysts to import data from social media tools (e.g., email, Twitter, YouTube, Flickr), automatically identify vertex clusters, layout the vertices according to different algorithms, calculate sub-graph images, and dynamically filter out edges and vertices using sliders.

A *top level reply network* emphasizes those who start threads (i.e., post the top-level message), while de-emphasizing conversations that occur midway through a thread. In some communities with short threads where all replies are typically directed at the original poster, such as email based Q&A communities, this network can better reflect the underlying dynamics. However, in discussion communities or forums with longer threads, the *direct reply network* is typically preferred since people later in the thread are often replying to each other. A top level reply network based on data in Fig. 1 would have Dave, Beth, and Ethan all pointing to Cathy who started the longest thread, thus emphasizing her importance. It would also include a self-loop from Cathy to Cathy, which are more common in these types of networks since people like Cathy reply to those who have replied to them.

5.2 Affiliation Networks

Affiliation networks are bi-modal networks that connect people to a set of groups, events, or places. For example, a traditional affiliation network may connect a group of executives to companies for whom they serve on the board of directors. Vertices

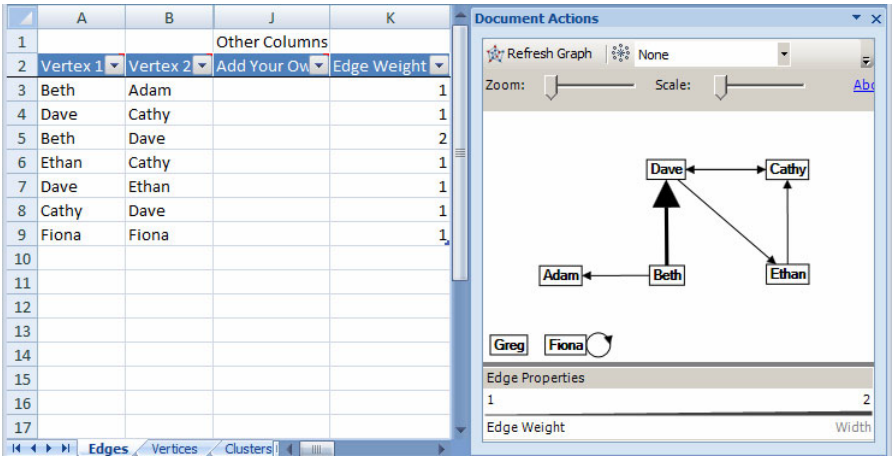

Fig. 2. Direct reply network graph based on data in Fig. 1. The network is constructed by creating an edge pointing from each replier to the person they replied to, and then merging duplicate edges to create an Edge Weight column. Notice that Beth has replied directly to Dave twice, so the edge connecting them is thicker. Fiona replied to her own message so there is a self-loop. Greg started a thread but was not replied to so he is not connected to anyone else.

represent both people and companies (which is why it is a bi-modal network), while edges represent affiliations between them. Affiliation networks for threaded conversations typically connect authors to Topics or Threads. The edges are undirected since there is only one possible direction (a person can post to a thread, but a thread can't post to a person). They are weighted based on the number of times a person posted to a Topic or Thread. For example, an edge would connect Cathy to Thread B with a weight of 2, since she posted to that thread twice. Beth would be connected with a weight of 1 to Thread A, Thread B, and Thread C since she posted to each of them once. This network is ideal for identifying boundary spanners and Forums or Threads that share authors.

Other affiliation networks connect authors to items that conversations are associated with (e.g., YouTube videos, Flickr Photos, blogs). These networks are related to recommender systems, in this case identifying "people who commented on this also comment on that" relationships.

Each affiliation networks can be transformed into 2 additional unimodal, weighted networks: a user-to-user network connecting people based on the number of threads (or forums) they both contribute to, and a thread-to-thread (or forum-to-forum) network connecting threads together based on the number of contributors they share. Or in the case of videos they show connections between videos based on the number of shared authors. These networks are good for creating overview graphs of large communities with many threads or forums. They help to identify content clusters that share many of the same authors, as well as clusters of users that hang out together in similar threads or forums.

6 Analyzing a Technical Support Email List: CCS-D

There are a host of technical support groups that use email lists, Usenet newsgroups, or web discussion forums to help individuals solve problems and make sense of a specific technology like JAVA, a product such as the iPhone, or a topic such as web design. Many companies host these forums to learn about problems with existing products, resolve customer concerns, generate new ideas on future improvements, and build a loyal customer community. To meet these goals it is often important to understand which individuals play important roles within the community, something that can be challenging when managing multiple, active communities. This section describes how to identify key members of the CSS-D email list devoted to the effective use of Cascading Style Sheets (CSS) in web design. It is a highly active list with around 50 messages sent each day. There are a handful of administrators who keep the conversation friendly and encourage contributors to follow the guidelines. See [12] for a complete description of the community and some of the strategies they use that make them so effective.

6.1 Preparing Email List Network Data

Creating network data from email lists such as CSS-D poses a few challenges. Email lists often have people registered with multiple email addresses, making it necessary to combine duplicate addresses for the same person. This process is called de-duplication and is an active area of research [13]. Another problem is that inferring who is replying to whom is not always obvious. By definition, all messages sent to an email list are sent to a single email list address (e.g., css-d@lists.css-discuss.org). The result is a star network connecting all contributors' email addresses to the list email address. Messages that begin a new thread (i.e., initial posts) will be sent to the list address and rarely will Cc other individuals. Replies to initial posts are handled differently depending on email list configuration choices.

Some lists, like css-d, set the default Reply-To address to that of the original sender. Users who click "Reply" to the initial email will send directly to the person, whereas users who click "Reply to All" will send to the initial person in the To field and the email list in the Cc field. This configuration is good for network analysis because it can use the information in the Cc field to identify who is replying to who. It does encourage more private messages however, which are missed by the email list and are thus absent in the network analysis. Other email lists set the default so that when users click on "Reply To" it sends to the list and they must choose "Reply To All" to explicitly Cc the initial sender. This configuration makes it more likely that people just send to the list and don't copy in the person they directly reply to. The result is that analysts may need to look at subject lines and email header information to reconstruct who is replying to whom.

The NodeXL tool includes an email import tool where analysts can generate email-based networks based on the To, Cc, and Bcc fields of an email corpus stored on a Windows indexed machine. It allows users to filter based on a time range, an email folder of interest, text in the subject line or body of the message, email features such as size or containing an attachment, and individual email addresses. It generates standard *direct reply networks*. The analysis of CSS-D is based on data from Jan-Feb of 2007.

6.2 Identifying Important People and Social Roles at CSS-D

In an online community, users contribute in different patterns and styles. In other words, community members fill different social roles. Understanding the composition of social roles within a community can provide many insights that make for more effective community managers. Unfortunately, simple activity and participation metrics are unable to capture the different types of contributions in discussion forums. In contrast, social network analysis provides metrics that can be used to automatically identify those who fill unique social roles and track their prevalence over time. This can help community managers:

- Identify high-value contributors of different types: Which community members are the most important question answers or question starters? Who connects many other users together? Answering these questions can help community managers to know who to thank (and for what) and how to support individuals' needs.
- Determine if a community has the right mix of people: Is this community attracting enough Question Answerers? Are there enough Connectors to hold the community together? Is discussion crowding out Q&A? Is a discussion space dissolving into Q&A? Knowing the answers to these questions can help community managers know who to recruit or encourage more, as well as what policies may be needed.
- Recognize changes and vulnerabilities in the social space: How has the community composition changed as it has grown? What is the effect of a certain prominent member leaving the community going to have? Which members are currently irreplaceable in the type of work they do? What is the effect of a policy change or change in settings on the community dynamics (e.g., changing the default Reply To behavior to send to the Sender versus the entire list)? Answering these questions can help community managers prepare for change, understand the effects of prior decisions and events, and cultivate important relationships.

This section shows how to identify important individuals and social roles within the CSS-D community by using NodeXL's subgraph images (i.e., egonetworks of CSS-D members) and creating a composite metric that helps identify the 2 most important social roles within Q&A communities like CSS-D: Answer People and Discussion People. This metric makes possible visualizations that show the relationships between these individuals as will be shown.

The first step in identifying important contributors to the CSS-D email list is to remove the overwhelming effects of the email list address by removing it from the graph. In NodeXL this is accomplished easily by choosing "Skip" in the Visibility column, which assures that the list email address will not be included in future analysis, such as the calculation of graph metrics, or visualizations where it would just clutter up the graph.

The next step is to create ego-networks of each contributor, which are called 1.5 Subgraph Images in NodeXL. In the examples provided in this section we use the Harel-Koren Fast Multiscale layout to automatically position the vertices in a meaningful organization [14]. NodeXL stores subgraph images of desired size in the spreadsheet itself or in a separate folder where they can be browsed. Once created,

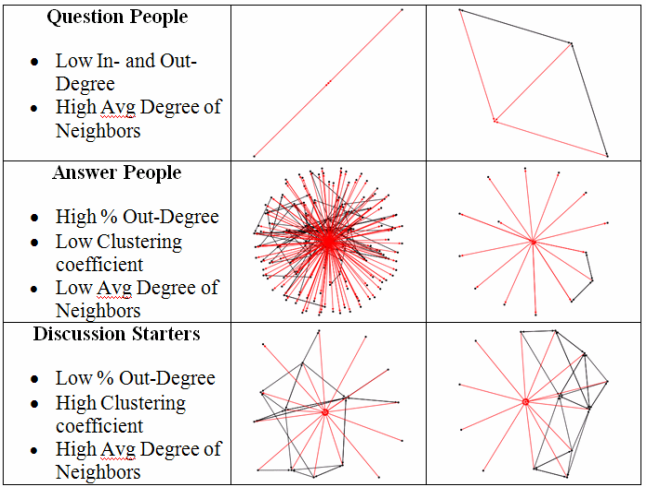

Fig. 3. NodeXL Subgraph Images (1.5 Degree; vertex and incident edges are red/lighter) for 6 CSS-D contributors that fill 3 different social roles within the CSS-D community

analysts can use Excel's built-in features to sort vertices based on graph metrics such as In-Degree (who receives messages from the most people) and Out-Degree (who sends messages to the most people) to bring differently connected individuals to the top. Sorting by centrality measures like Eigenvector reveal the core members of the community because they are active participants and talk to other active participants.

Scanning through the Subgraph Images of CSS-D contributors shows the different social roles that exist within the email list community. Fig. 3 shows examples of 3 types of contributors (Question People, Answer People, and Discussion Starters) along with some of the metrics that could be used to identify them. Question people post a question and receive a reply by one or two individuals who are likely to be Answer People. Answer People mostly send messages (arrows point toward other vertices) to individuals who are not well connected themselves [15]. Discussion Starters mostly receive messages (arrows point toward them), often from people who are well-connected to each other.

While the Fig. 3 graphs help identify the different types of social roles, metrics can also be used to classify individuals automatically. Question People are easy to detect because of their low degree. To identify people along the Answer Person / Discussion Starter spectrum we create a single Answer Person score by multiplying the percent out-degree by the inverse of the clustering coefficient (defined as the percent of neighbors who are connected). Those who score high are Answer People because they reply to others more than they are replied to and those they reply to are primarily isolates (i.e., question people). Those who score low are Discussion Starters since they are replied to often and by others who are well-connected. We only apply this metric to those with an out-degree + in-degree of 15 or higher to focus on active members. In Fig. 4 those with high Answer Person scores are darker disks, those with low scores are lighter disks, and those with a low degree (mostly Question People) are circles that are not filled in.

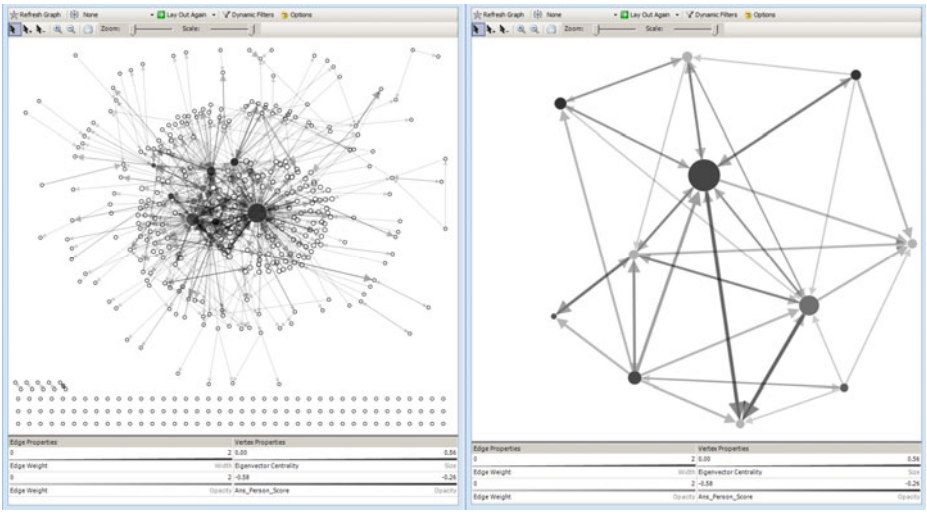

Fig. 4. Two NodeXL graphs of the CSS-D email list network for Jan-Feb of 2007. Answer People (darker) and Discussion Starters (lighter) are identified by the calculated Answer Person Score. Circle vertices (filtered out of the graph on the right) have a total degree of fewer than 15 and mostly consist of Question People. Vertex size is mapped to Eigenvector Centrality. Edge weight (i.e., number of messages sent) is mapped to both edge Size and Opacity, applying a logarithmic scale and ignoring outliers.

The specific social roles and their prevalence within a particular community will depend on the nature of that community. Since the CSS-D community is primarily a Q&A community, it consists of mostly Question Askers, a handful of prominent Answer People, and a small number of Discussion Starters. Other more discussion-based communities would have many more Discussion Starters as well as other social roles such as Flame Warriors, Commentators, and Connectors. Tracking the ratio of people that play different social roles can be a good way to assure that a community is healthy. For example, if the CSS-D community had too few Answer People or an influx of many Question People it could not function as effectively.

Viewing the entire reply network for the CSS-D email list (left graph in Fig. 4) provides some general insights about the composition of its population, although the size of the network makes it challenging to interpret without filtering. Larger nodes have a higher Eigenvector Centrality suggesting they are connected to many people and others who are well-connected. The binned layout is used to identify isolates along the bottom, of which there are many since the email list address itself was removed from the network. Isolates represent those who posted to the list and didn't receive a response (e.g., they posted an announcement) or in some cases those who replied to the list without copying in the address of the person who they were replying to. Overall the entire reply network shows many individuals connected primarily through a handful of central question answerers and a small, but stable core group of members that interact with one another regularly.

To better focus in on the core members of the community and their relationship to one another, analysts can filter out vertices with a total degree of less than 15. The

graph on the right side of Fig. 4 shows the resulting network after manually positioning the vertices. The edge weights, represented in the edge width and opacity provide a good sense of who interacted with whom during the 2 month time period and is thus likely to know each other and perhaps have similar interests. Note that even among these core members, Discussion Starters (light vertices) rarely reply to other Discussion Starters. Also notice that the largest vertex, while categorized as an Answer Person, receives many messages from the core members. This suggests that he plays multiple important roles within the community. In fact, if he were removed from the network there would be considerably fewer connections between the core members. Community administrators should make sure this individual is adequately appreciated and encouraged to remain in the community since his removal would seriously disrupt the community.

7 Finding a New Community Admin for the ABC-D Email List

Administrators of online conversations play pivotal roles in maintaining social order, encouraging participation, and making communities feel like home [3]. They are typically among the most active members of a community [16] and can function better when they are known and respected by the members of the community. Because of the importance of administrators, when one leaves or steps down it has the potential to disrupt the community. In this section we look at how network analysis can help in identifying a potential replacement for an administrator that is going to step down. Data for this analysis comes from an email list we will call ABC-D, based around a specific profession. It is a classic example of a community of practice that spans multiple institutions. Unlike CSS-D, ABC-D encourages in-depth discussion about the community's domain and is not primarily about questions and answers.

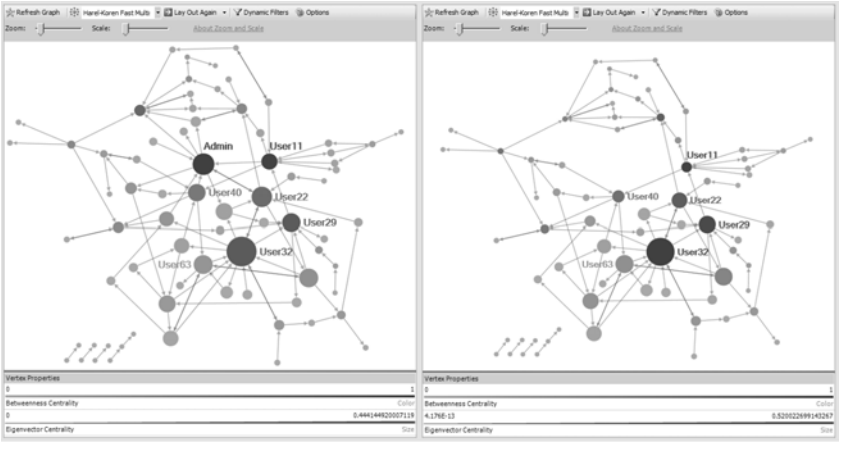

Fig. 5. NodeXL maps of ABC-D's email list direct reply network, with the current Admin (left) and without the current Admin (right). The most central members are labeled including Admin in the left side image. Larger vertices have a higher eigenvector centrality and darker vertices have a higher betweenness centrality.

The network is a *direct reply network*. An arrow pointing from person A to person B indicates that person A replied to a message of person B. Data from ABC-D was collected for a two-week period by Chad Doran, a graduate student at the University of Maryland College of Information Studies, who also came up with the administrator replacement scenario. A more complete analysis would include a longer time-period (e.g., 2 months) and include edge weights, but the current dataset is sufficient to illustrate the key idea. All data, including the name of the community, has been anonymized to respect the privacy of the group members.

The graph on the left side of Fig. 5 shows the entire reply network with a few key individuals (as identified by graph metrics) labeled. The graph on the right is the same graph after removing the Admin and recalculating the graph metrics. The networks show that individuals are almost all connected in one large component, but the degree of any one individual is relatively low (e.g., the maximum total degree is 14 and the average total degree for an individual is about 3). The result is a fairly spread out network. Graph metrics were calculated and used to identify the most central individuals, who presumably are in the best position to serve as an administrator replacement. Darker vertices have a high betweenness centrality, suggesting that they are important at connecting different vertices and integrate the network as a whole. Larger vertices have a higher eigenvector centrality, which in this case suggests the person is well connected to others who are themselves well connected.

As expected, the current administrator (labeled Admin in the left-side graph) has the highest betweenness centrality and a high eigenvector centrality. Interestingly, the individual with the highest eigenvector centrality (User32) is not directly connected to the Admin; in the time period of our data collection neither of them replied to the other. Another important individual is User11 who has a high betweenness centrality because he was the only link to several vertices, but a relatively low eigenvector centrality since most of his connections were with individuals who rarely posted. All of the labeled individuals scored high on the metrics and may be good candidates to replace the administrator. Of course other characteristics not captured in the network structure, such as their willingness to serve, their friendliness, and their experience would also be key determinants.

A key question is: how the community would change if the administrator were removed from the network. The key network metrics, betweenness centrality and eigenvector centrality, will change for the remaining individuals because they are dependent on the network properties of other vertices. Thus, looking at the graph without the admin (on the right-hand side of Fig. 5) can help more accurately assess individual's potential as a replacement. It also helps analysts to see how the network as a whole may be impacted. For example, removing the admin changes the average Closeness centrality from 3.2 to 3.5 suggesting that people will not be as directly connected with others once the admin is gone. Analysts may also notice certain subgroups within the network that lose an important connection to other subgroups, such as the large group at the top of the graph. These differences can be more easily noticed when the location of the vertices has been fixed in both graphs as in Fig. 5.

Looking at the right-hand graph in Fig. 5 confirms that the initial individuals identified as possible replacements are good candidates. It also suggests that if certain candidates were chosen, such as User32, there may be subgroups of the community that would not be as well connected (e.g., the group of nodes at the top of the graph).

The fear is that these individuals may feel alienated by a new administrator they either don't know or don't converse with often. The graph also points out individuals who may be able to keep them involved: User11 and User22. Armed with this information the outgoing administrator may be wise to recommend that User32 and User22 jointly serve the role of administrator, or that whoever is chosen should foster a relationship with these individuals to link to those who may feel alienated.

8 Understanding Groups at Ravelry

Ravelry (http://www.ravelry.com) is a thriving online community for anyone passionate about yarn. As of January 2010, there were over 600,000 knitters and crocheters registered on the site. Users organize their projects, yarn stashes, and needles; share and discover designs, ideas, and techniques; and form friendships through discussions and exploration of shared interests. In this section, the Ravelry community administrator works with data on the top 20 posters to 3 discussion forums created for different groups. The data and initial network analysis for this section was developed by Rachel Collins, a graduate student at Maryland's iSchool.

Imagine a community manager is assigned 3 group discussion forums to monitor and help develop. They are highly active groups, making it hard to keep up with all the messages and get a better sense of how the most important community members relate to one another, as well as how the groups differ. This understanding helps the community manager to recommend the best group for a newcomer to join, as well as identify individuals with certain expertise or social relations. The 3 groups (whose names have been changed for privacy reasons) include one common-interest group (Apathetic, Funloving Crafters [AFC]), one Meet-Up (Chicago Fiber Arts), and one Knit-Along (Project Needy). They are 3 of hundreds of similar groups. Discussion forums for each group serve as their central hubs. Individuals can participate in as many forum groups as they desire. The data includes project output, discussion board usage, blog activity, and community roles for the top 20 posters in each group. This lets the community admin relate many different activities together in a single analysis, focusing attention on the most active members who are typically the most important.

Fig. 6 shows a bi-modal affiliation network of the 3 forums/groups (shown in text boxes) connected to individuals who have posted to them. Edge thickness is based on the number of forum posts (using a logarithmic mapping). The thinnest lines connect users to groups that they are members of, but have not yet posted to. Other visual properties are used to convey individuals' level of activity in other parts of the community as described in the Fig. 6 caption. The graph identifies important individuals, such as those who post to multiple groups or have certain color/size/shape combinations. It also enables comparison of the three groups. For example, the graph makes clear that the AFC forum is very active, includes many bloggers, and includes relatively few people who complete a large number of projects (perhaps explaining the "Apathetic" in their title). In contrast, the Project Needy group includes many highly productive members, many of whom are both administrators and bloggers. In contrast, the Chicago Fiber Arts group has fewer bloggers and less project activity.

Administrators could use a graph like Fig. 6 to identify potential candidates for Volunteer Editors or identify clusters of boundary spanners with which to form new

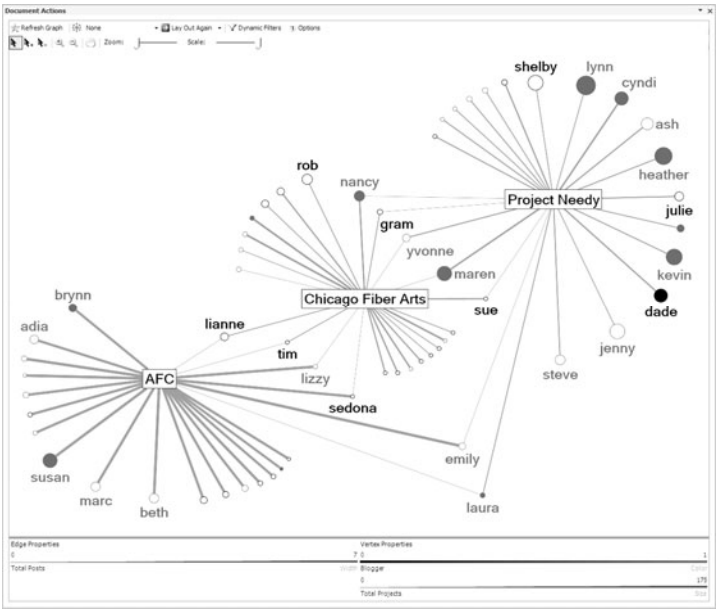

Fig. 6. Bi-modal affiliation network connecting 3 Ravelry groups (i.e., forums AFC, Chicago fiber Arts & Project Needy) to contributors represented as circles. Edge width is based on number of posts (with logarithmic mapping). Vertex size is based on number of completed Ravelry projects. Maroon/lighter vertices have a blog and solid circles are either Community Moderators or Volunteer Editors. The network helps identify important boundary spanners (e.g., those connected to multiple groups), as well as compare groups.

groups because of shared interests. Providing graphs like this one to the groups themselves can also prompt self-reflection and potentially foster new connections. They can also be used to better understand how the activities on the site relate to one another, although use of statistics may be needed to more systematically validate initial claims. For example, Fig. 6 shows that location-based groups have a lower percentage of active members who blog and people who complete many projects seem to cluster into project groups. Finally, simplified versions of this graph may help newcomers to Ravelry get a sense of which group(s) they may want to join, as well as identify some of the prominent members they may want to follow or meet.

9 Conclusion and Future Work

Network analysis and visual presentations of online communities that use threaded conversations can produce valuable insights. In this article we have defined threaded conversation and characterized the different types of networks that are created by them: the directed, weighted *direct reply network* and *top level reply* networks; the undirected, weighted *affiliation network* connecting threads (or forums) to the individuals that posted to them; and the undirected, weighted unimodal networks derived from the affiliation network including user-to-user network and thread-to-thread networks.

We have also demonstrated how new analysis tools such as NodeXL can be used by community administrators to gain actionable insights about the communities they serve. The analysis of the CSS-D technical support community showed how to identify important social roles and individuals who fill those roles including Answer People, Discussion Starters, and Questioners. The analysis of ABC-D discussion-based email list showed how to identify good candidates to replace a community administrator based on network metrics such as Betweeness and Eigenvector Centrality. And the analysis of Ravelry showed how to use a bi-modal affiliation network to understand how forum-based groups are connected, identify important boundary spanners, and relate non-discussion network metrics (e.g., blog activity; project activity) to group discussion activity. We hope these mini case studies provide inspiration for other focused network analyses aimed at gaining actionable insights about online interaction.

Research on threaded conversation communities has a long history as outlined in Section 3, yet there remain many interesting research questions to explore. As threaded conversations become embedded within more complex social spaces with multiple interaction technologies, it is increasingly important to understand how they all interact. For example, Hansen found that technical and patient support groups benefit from combining a threaded conversation (i.e., email list) with a more permanent wiki repository [12]. The Ravelry example showed strategies that have not yet been widely used by the research community to understand how network position relates to use of other tools (i.e., blogs) or activities (i.e., projects). Network-based research is also needed to better understand the determinants of successful online communities. For example, we don't know what proportion of mixtures of Answer People, Discussion Starters, and Questioners lead to better outcomes or what overall network statistics (e.g., clustering coefficient) are correlated to success. From a design perspective, there are many fascinating opportunities to enhance the threaded conversation model as evidenced by Google Wave and other prototype systems. Many opportunities remain to advance techniques to visualize online conversation spaces [17] and threaded conversation networks as demonstrated in this article.

References

1. Resnick, P., Hansen, D., Riedl, J., Terveen, L., Ackerman, M.: Beyond Threaded Conversation. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, pp. 2138–2139. ACM, New York (2005)
2. Smith, M., Kollock, P. (eds.): Communities in Cyberspace. Routeledge, London (1999)
3. Preece, J.: Online Communities: Designing Usability and Supporting Sociability. John Wiley & Sons, Inc., New York (2000)
4. Kim, A.J.: Community Building on the Web: Secret Strategies for Successful Online Communities. Peachpit Press, Berkeley (2000)
5. Powazek, D.: Designing for Community. Waite Group Press, Corte Madera (2001)
6. Nonnecke, B., Preece, J.: Lurker Demographics: Counting the Silent. In: Proceedings of the SIG-CHI Conference on Human Factors in Computing Systems, pp. 73–80. ACM, New York (2000)

7. Hansen, D.L.: Overhearing the Crowd: an Empirical Examination of Conversation Reuse in a Technical Support Community. In: Proceedings of the Fourth International Conference on Communities and Technologies, pp. 155–164. ACM, New York (2009)
8. Garton, L., Haythornthwaite, C., Wellman, B.: Studying Online Social Networks. J. Comput-Mediat Comm. 3(1) (1997)
9. Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., Haythornthwaite, C.: Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community. Annual Review of Sociology 22, 213–238 (1996)
10. Smith, M.A., Shneiderman, B., Milic-Frayling, N., Rodrigues, E.M., Barash, V., Dunne, C., Capone, T., Perer, A., Gleave, E.: Analyzing (social media) networks with NodeXL. In: Proceedings of the Fourth International Conference on Communities and Technologies, pp. 255–264. ACM, New York (2009)
11. Hansen, D.L., Rotman, D., Bonsignore, E., Milic-Frayling, N., Mendes Rodrigues, E., Smith, M., Shneiderman, B., Capone, T.: Do You Know the Way to SNA? A Process Model for Analyzing and Visualizing Social Media Data. HCIL-2009-17 Tech Report (2009)
12. Hansen, D.: Knowledge Sharing, Maintenance, and Use in Online Support Communities. Unpublished Dissertation, University of Michigan (2007)
13. Bilgic, M., Licamele, L., Getoor, L., Shneiderman, B.: D-Dupe: An Interactive Tool for Entity Resolution in Social Networks. In: Proceedings of IEEE Symposium on Visual Analytics Science and Technology. IEEE, Los Alamitos (2006)
14. Harel, D., Koren, Y.: A Fast Multi-scale Method for Drawing Large Graphs. In: Marks, J. (ed.) GD 2000. LNCS, vol. 1984, pp. 183–196. Springer, Heidelberg (2001)
15. Welser, H.T., Gleave, H., Fisher, D., Smith, M.: Visualizing the signatures of social roles in online discussion groups. Journal of Social Structure 8(2) (2007)
16. Butler, B., Sproull, L., Kiesler, S., Kraut, R.E.: Community Effort in Online Groups: Who Does the Work and Why? In: Weisband, S., Atwater, L. (eds.) Leadership at a Distance. Lawrence Erlbaum Associates Inc., Mahwah (2005)
17. Turner, T.C., Smith, M.A., Fisher, D., Welser, H.T.: Picturing Usenet: Mapping Computer-Mediated Collective Action. J. Comput-Mediat Comm. 10(4) (2005)