You Didn't Tell Me That! Visualizing the Hidden Attributes of Online Health Communities

Diana MacLean
Stanford University
Gates Computer Science Building, Rm 268
353 Serra Mall, Stanford, CA 94303
malcdi@stanford.edu

Sudheendra Hangal
Stanford Computer Science
Gates Computer Science Building, Rm 400
353 Serra Mall, Stanford, CA 94303
hangal@stanford.edu

ABSTRACT

Online health communities (OHCs) are a growing source of public medical knowledge; they facilitate several health-related tasks including searching for and acquiring new medical information, and seeking emotional support. Little is known, however, about how knowledge exchange patterns and community structures differ across OHCs. Prior work demonstrates that visualization constitutes an effective technique for discovering and exploring online community attributes that are not readily apparent during simple forum browsing, such as forum type (debate vs. Q&A oriented), or proportion of expert members. Quick access to such knowledge might prove useful to community leaders, who try to maintain (and enhance) the overall well-being of their respective OHCs, as well as to "newbies" who may want to window shop a number of forums before deciding which community to invest in. This paper presents an exploratory study of discovering, analyzing and summarizing OHC attributes using visualization. We find that not only do community dynamics vary across several dimensions between disease forums, but also that many of these variations can be intuitively visualized.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.

General terms: Design, Human Factors, Management, Experimentation

Keywords: Online Health Communities, Visualization, Online Community Management

Introduction

As people rely increasingly on the Internet as a source of medical knowledge, OHCs are becoming more and more prevalent. This shift is attributed mostly to changes in the health care system (lower access to healthcare professionals and higher costs of health care) and increased technological literacy in the general population [8]. Although the question of whether OHCs provide members with *actual* health ben-

efits remains open [4], OHCs have several advantages over traditional health information systems. These include cost effectiveness [8]; always-available, unrestricted access [7]; empowering patients through democratization of medical information [13]; providing comfortable venues for discussing sensitive issues [13]; and enhanced social support stemming from interactions with people suffering from the same illness [9, 13, 8].

Despite this, OHCs also carry several disadvantages. Sites such as Cure Together [1], MedHelp [2] and others, cater to the growing population of OHC participants primarily by providing interactive discussion forums. However, participants have no means of evaluating the quality of knowledge found in these forums [13, 7], or the trustworthiness and personal agenda of participants [7]. Moreover, the vast quantity of peer-to-peer contributions make OHCs difficult to navigate, as users must sift through archives of posts in a search for relevant information [7, 13]. Several of these limitations can be attributed to the lack of a good interface [7] - the typical bulletin-board health forum interface has changed little over the past 10 years.

Prior work indicates that visualizing online communities can be a highly effective technique for uncovering a range of subtextual community attributes, such as member demographics [11], conflict patterns [2], and types of social roles played by community members [11]. In this paper, we present an exploratory analysis of 5 MedHelp OHCs, in which we use visualizations to examine basic hypotheses about community structures and user behavior patterns. Our goal is twofold: to discover a set of OHC attributes easily synthesized from community forum data, and to explore their variation across different communities. We posit that easy access to subtextual community attributes could greatly enhance OHC usability for both members and moderators.

Related Work

For a thorough overview of prior work on online communities, we direct the reader to Iriberri and Leroy's work on online community life cycles and evolution [6].

Visualizing Online Communities Our work is based primarily on Viegas and Smith's canonical paper on community visualization, in which they present two visualization tech-
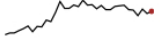
---

[1] http://www.curetogether.org
[2] http://www.medhelp.org

Table 1: Complete Forum Data

Forum	Start Date	# Posts	# Threads	Avg. Posts per Thread	# Members	Post Frequency Trends
Asthma	10/2007	3, 903	1, 369	2.85	1606	
Breast Cancer	9/2006	33, 640	9, 181	3.66	9, 049	
Depression	11/2006	19, 849	4, 563	4.35	5, 125	
Lupus	9/2007	3, 603	836	4.31	862	
Lyme Disease	9/2007	9, 762	1, 630	6.0	879	

niques designed to reveal non-obvious attributes of Usenet newsgroups and newsgroup members [11]. Their visualizations effectively convey community types (for example discussion oriented, or Q&A) as well as member demographics (such as proportion of new users, proportion of members participating in other newsgroups, and proportion of consistent contributors). They were also able to convey several user role patterns within newsgroups, including those of answer person, spammer, and conversationalist.

Welser et al. present visual methods for categorizing "role signatures" in online communities [12]. Although they focus on identifying "answer people" in particular, their work suggests that visualization techniques combined with regression analyses can be used to build successful models for role prediction. Other related work includes Smith and Fiore's research on visualizing online discussions for improving user navigation through online community spaces [10]. The primary focus of this work is on content analysis and categorization, which we do not yet attempt in our research.

Knowledge Sharing in Online Communities In a study of knowledge sharing in Yahoo Answers forums, Adamic et al. investigate participation and interaction patterns of users across an array of online forums [1]. Based on user interaction attributes such as post length and cross-category posting, they were able to classify forums into "discussion" and "Q&A" types. They also found that user role demographics were likely to be different based on community types. Along similar veins, Fiore et al. present a method of employing behavioral descriptors (posts, replies etc.) to estimate compatibility between forum users, essentially performing automatic collaborative filtering to find information relevant to unique members [5].

Data Acquisition
Medhelp [3] is a free, online health community website designed to aid users in the discovery, exploration, and management of personal illnesses. The site boasts a wide array of tools and services, including over 200 "Medical Support Communities", where users discuss medical conditions amongst themselves. Forums are structured in a typical bulletin-board style: users reply to an initial post. Responses are not inlined, so detecting conversations between repliers in longer threads is not possible without text analysis. The

site also provides trackers which members may use to record health-related measurements (such as weight, or mood severity etc.) and chart trends over time. Each user has a profile page linking to her participation on the site as well as other personal information.

We use the BeautifulSoup [4] library to crawl the MedHelp site, using the site's robots.txt file as a guide to which pages are acceptable to crawl [5]. While crawling community forums and personal user pages is permitted, crawling health trackers is not. We crawl 5 full medical support communities: Asthma, Breast Cancer, Depression, Lupus and Lyme Disease, as well as user profiles for all contributors to the Asthma and Lyme forums. Although several community choices exist, we selected these 5 for their representative diversity. Table 1 summarizes the forum data.

Visualizing Online Health Communities
In order to explore the subtextual space of OCHs, we posed several simple hypotheses relating to community structure and communication patterns. We then designed visualizations to answer our hypotheses. We discuss subset of these, as well as additional knowledge gleaned from the process, below.

H1: Communities have different social hierarchies We confirm this hypothesis using simple, node-link diagrams, illustrated by Figure 1. Nodes represent users, and edges are drawn, for each post, between the thread *responder* and the thread *initiator*. Node placement is determined by the Fruchterman-Reingold, force-directed layout algorithm. Red nodes designate official community leaders[6], yellow nodes, users who have received "barn stars" for excellent forum posts, and green nodes, doctors sponsored by MedHelp. Isolates occur when users initiate a thread that never receives a response.

From Figure 1(a), we see that the Asthma community has a large number of isolates, and a clearly hierarchical structure whereby community leaders respond to users' questions. The "bunch of balloons" clumps are indicative of "answer" people dispatching responses to a variety of users who have one-time questions. If long debates were typical in this com-

---

[3]www.medhelp.org

[4]http://www.crummy.com/software/BeautifulSoup
[5]http://www.medhelp.org/robots.txt
[6]Official community leaders have special icons, and are responsible for moderating forums.

munity, we would expect to see a denser network with fewer distinct clumps: a trait that is evident in Figure 1(b). Not only is hierarchical structure much less prevalent in the Lyme Disease community - there are not even any official community leaders. There are also proportionally fewer isolates.
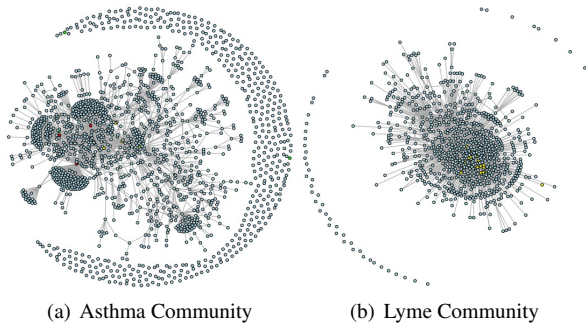

(a) Asthma Community   (b) Lyme Community

Figure 1: Node Link Community Diagrams

H2: Community role demographics differ across OHCs   Prior work indicates that user role demographics differ across online communities [1]: for example, certain communities may have a larger proportion of leaders than usual. Figure 2 shows our "community chromatograph" visualization, designed to facilitate comparative summarization of community demographics. Each user is represented by a bubble; bubble size is proportion to the number of contributions that user made to the forum, and y-axis placement indicates the number of days since the user's first and last post. Thus, users are visually separated according to how long they have been community participants, and the volume of their contributions. Large dots near the top are indicative of community leaders, or answer people, while large dots at the bottom depict new users who (perhaps anxiously) make several posts in a short span of time.

In Figure 2 we see that although the Lyme community has no officially designated community leaders, it has a number of unofficial leaders who clearly make significant forum contributions. The "official" Asthma community leaders (depicted by the 3 larger middle dots) do not contribute nearly as much. "Smudges" of tiny dots at the bottom of each column are indicative of either "Q&A" behavior, in which people do not re-post once they receive a satisfying response, or users' questions being ignored. Similarly large-sized dots closer to the top of each column suggest discursive communication activity, as is shown for both the Breast Cancer and the Lyme Disease forums. Thus, community chromatographs not only depict role demographics, but also suggest what user behaviors particular demographic subgroups engage in.

H3: Question/Answer contribution patterns differ across communities   A community chromatograph does not differentiate between question and answer contributions. Figure 3 depicts simple scatter plots, in which the x and y-axes represent number of questions asked and answered, respectively, and dot size represents number of days since first and last post. The x and y-axes are log-scaled so as to retain a more "core" representation of the communities: those users who both ask and answer questions.
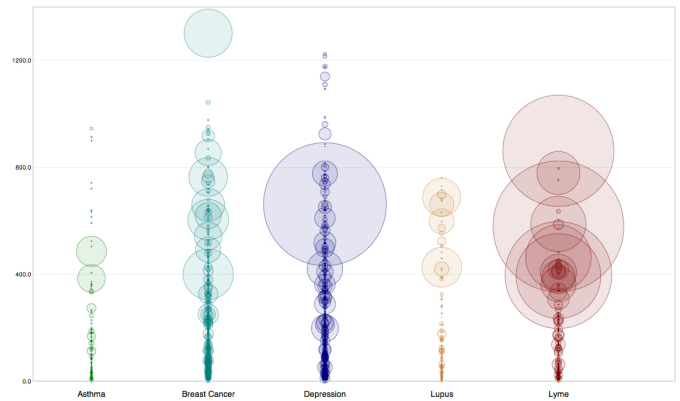

Figure 2: A community chromatograph shows a community's role composition.

In Figure 3 we see strong diagonal trends in the Breast Cancer, Depression and Lyme Disease forums, indicating that long-term members tend to both ask and answer questions. We do not see this trend, which may indicate discussion or support-based forums, as strongly in the Asthma or Lupus data. The forum scatterplots also depict clear horizontal and vertical lines close to the x and y axes. Consider the strong blue line on the y-axis in the Depression forum scatterplot: this line represents people who have asked only one question. Similarly, horizontal lines encode the trend of question asking for people who answered x questions, and asked y. With further analysis, this information could prove useful for understanding community member retention and other structural dynamics.

Figure 3: Forum scatterplots summarize question and answer contributions.

H4: Cross-community postings may confer co-occurrence patterns   Finally, we consider that multiple community membership may be indicative of relationships between particular diseases. Figure 4 presents an arc diagram based on the Asthma community. Nodes represent different communities; the size of a node is proportional to the number of *members* from the primary community (i.e. Asthma) who have also

posted in that community. The width of the arcs encodes the number of *posts* that people in the primary community have posted in the secondary community.

Figure 4 shows two well-known co-occurrences: Asthma and Thyroid disorders, and Asthma and Anxiety. More interesting is the link between Asthma and Fertility. While no proven connection exists between the two, a recent analysis of OHC data from CureTogether suggested a very strong correlation between infertility and asthma [3]. While making medical inferences from co-occurrence data is impractical, understanding community overlap is nevertheless useful information for OHC interface design.
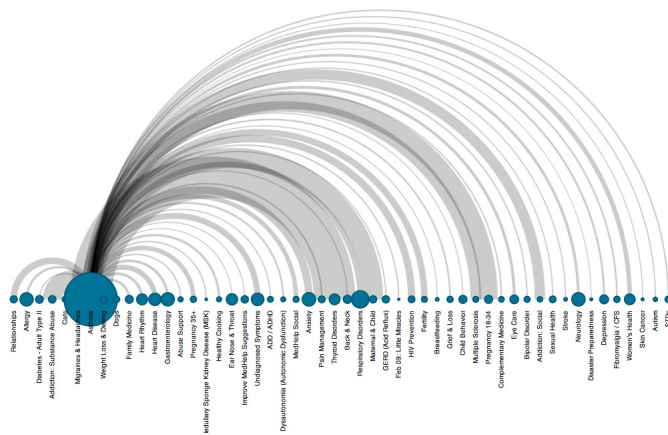

Figure 4: "Forum Panels" summarize communication contributions.

Conclusions

We have presented several visualizations of online health community data, with the goal of eliciting community attributes that while useful for participants and administrators, are not obvious from typical forum-browsing behavior. While our work clearly requires both quantitative analysis and user studies before proceeding much further, we demonstrate that visualization is an effective medium both for discovering subtextual community attributes and for exploring their variation across different communities. We found that not only do OHCs have significant, almost unique traits, but also that these traits do vary strongly across communities.

Our findings suggest several avenues for future work. One next step is to analyze the *social* data that we acquired from our crawls. In fact, the MedHelp data is incredibly rich in social information: members have friends and send each other public messages, providing a unique and uncommon social data set. Another goal is to acquire data from a broader range of communities, with the goal of finding specific OHC categories. While we analyzed communities from MedHelp, several other large, popular OHCs exist, and a cross-site analysis would be valuable for our long term study. Finally, while our visual analysis has afforded us much insight into the workings our our communities, it would be useful to structure our data within a more rigid framework (along the lines of knowledge sharing, or skillset transfer, for example) in order to experiment with mathematical models of cause and prediction.

REFERENCES

1. L.A. Adamic, J. Zhang, E. Bakshy, and M.S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceeding of the 17th international conference on World Wide Web*, pages 665–674. ACM, 2008.

2. U. Brandes and J. Lerner. Visualization of conflict networks. *Building and Using Datasets on Armed Conflicts*, 36.

3. Alexandra Carmichael. Crowdsourced Health Confirms Infertility-Asthma Finding, September 2009.

4. G. Eysenbach, J. Powell, M. Englesakis, C. Rizo, and A. Stern. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *British Medical Journal*, 328(7449):1166, 2004.

5. A.T. Fiore, S.L. Tiernan, and M.A. Smith. Observed behavior and perceived value of authors in usenet newsgroups: bridging the gap. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, page 330. ACM, 2002.

6. A. Iriberri and G. Leroy. A life-cycle perspective on online community success. 2009.

7. D. Maloney-Krichmar and J. Preece. A multilevel analysis of sociability, usability, and community dynamics in an online health community. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):232, 2005.

8. L. Neal, G. Lindgaard, K. Oakley, D. Hansen, S. Kogan, J.M. Leimeister, and T. Selker. Online health communities. In *CHI'06 extended abstracts on Human factors in computing systems*, page 447. ACM, 2006.

9. S. Rodgers and Q. Chen. Internet community group participation: psychosocial benefits for women with breast cancer. *Journal of Computer-Mediated Communication*, 10(4):5, 2005.

10. M.A. Smith and A.T. Fiore. Visualization components for persistent conversations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 136–143. ACM, 2001.

11. FB Viegas and M. Smith. Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspaces. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, page 10, 2004.

12. H.T. Welser, E. Gleave, D. Fisher, and M. Smith. Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure*, 8(2), 2007.

13. M. White and S.M. Dorman. Receiving social support online: implications for health education. *Health Education Research*, 16(6):693, 2001.