Spring 2017

Engagement of users in online health communities - a social support perspective

Xi Wang
University of Iowa

ENGAGEMENT OF USERS IN ONLINE HEALTH COMMUNITIES
- A SOCIAL SUPPORT PERSPECTIVE

by

Xi Wang

A thesis submitted in partial fulfillment
of the requirements for the Doctor of Philosophy
degree in Informatics in the
Graduate College of
The University of Iowa

May 2017

Thesis Supervisor: Assistant Professor Kang Zhao

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

____

PH.D. THESIS

____

This is to certify that the Ph.D. thesis of

Xi Wang

has been approved by the Examining Committee for
the thesis requirement for the Doctor of Philosophy degree
in Informatics at the May 2017 graduation.

Thesis
Committee: ___________________________________
Kang Zhao, Thesis Supervisor

____
Nick Street

____
Gautam Pant

____
Padmini Srinivasan

____
Shelly Campo

To My Mom

ACKNOWLEDGEMENTS

I would like to express my sincere thankfulness to all of you, who have helped, supported and loved me during the past 5 years. Without any of you, I would not be able to finish this thesis and my Ph.D.

Foremost, I will be forever grateful to my advisor, Dr. Kang Zhao, for his patience, guidance and support on both of my research and life in Iowa City. He has set up an excellent example as a researcher and mentor, guided me patiently on my research projects and language skills, and provided immense assistance on my career development. Professor, it is always my greatest honor and fortune of being your first Ph.D. student.

I am thankful to Dr. Nick Street, who introduced Informatics to me at the beginning. I really appreciate his insightful comments to my research, and encouragement when I encountered different kinds of obstacles. Nick, thank you for replying to a stranger's email 6 years ago, I would never have such a wonderful experience in Informatics studies without you.

In addition, I want to thank the other members of my thesis committee, Dr. Gautam Pant, Dr. Padmini Srinivasan, and Dr. Shelly Campo. Thanks for all the constructive suggestions and feedbacks. Your expert guidance from multiple perspectives motivated me to work as an interdisciplinary researcher.

I also need to acknowledge my friends.

Xing Tong, the one who made me a key to United States and my Ph.D. study, thank you for always being my strongest backup here, whenever I need you.

Yan Jiang, the most handsome guy I have ever met, thank you for your romance and concern, your bad temper and sweet smile.

Yuanyuan Jiang, the girl stepped in Iowa City with me from the first day, thank you for still being here as my listener and friend.

Jennifer Catherine Brooke and Jacob Austin Rodgers, two American friends, thank you for your efforts in improving my poor oral English.

Zhong Zhang and Yiwen Cai, two of my best friends during my first two years in Iowa City, thank you for keeping me accompany on my amazing journeys.

Zhiya Zuo, Yang Zhang and Yuanyang Liu, three of my co-authors, thank you for your brilliant ideas and hard work.

My deepest appreciation goes to my families- my grandpa, my dad, especially my mom, Binlin Wang. I am tremendously fortunate to have her guiding, supporting and loving me all the time. Mom, thank you for showing me how big the world is, thank you for teaching me to keep smile in facing all challenges, and thank you for supporting all my dreams. I always love you.

ABSTRACT

Online Health Communities (OHCs) have become an important source of sharing and receiving information and support for people with health-related concerns. These communities provide important benefits to users including enhance medical knowledge, emotional comfort, personal empowerment and the ability to create offline social connections. High levels of user engagement are beneficial to both users and the OHC, so it is important to understand what motivate users' participation, encourage them to contribute and influence their churning behaviors.

This thesis covers why, when, and how users are actively engaged within an OHC. It is based on descriptive and predictive analytics of OHC users' online interactions with text mining techniques. I built explanatory models to reveal how users' motivations and roles evolve over time, the types of social support activities that encourage users' continuous participation, and the forms of social capital that drive users' continued contributions to the community. In addition, I developed predictive models to help an OHC forecast whether and when a user will churn.

The findings of this study have implications for managing and sustaining successful OHCs, and can provide OHC managers with suggestions on how to motivate user contributions and retain users through interventions.

PUBLIC ABSTRACT

The ubiquity of the Internet access brings us an epoch that people can access the information or support they need instantaneously. Online Health Communities (OHCs), as a product of modern Internet, are convenient sources of health-related information that allow users to interact with peers having similar concerns. However, as many other virtual communities, OHCs face many challenges such as low user activity levels and high rates of turnover. Thus, it is important to understand the factors that influence users' engagement in OHCs.

This study analyzed data from a public OHC that deals with breast cancer. Various computational methods made it possible to determine the type(s) of social support existing in each post. By summarizing and analyzing users' seeking and receiving behaviors, I was able to understand how users interact and involve within this OHC. For example, users often joined the community because something sparked their interest, maybe an online article or the diagnosis of a family member. As they sought and received different types of social support, connected with various people or became embedded within different groups, their interests changed over time. Some of users used the site for a very short period, then left, while others were more active and engaged over longer periods.

This study aims at detecting factors that impact users' engagements in OHCs. Its findings have implications for managing and sustaining stable OHCs, such as providing OHC managers with suggestions in improving website design and adopting interventions to motivate user contributions and retain users in the community.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Healthcare is a major challenge for modern society and has attracted the attention of stakeholders well beyond the healthcare industry. The ubiquitousness of the Internet has made it easier for individuals to obtain, process, and understand information related to health. According to a report from the Pew Research Center (Fox, 2014), 72% of United States adults have used the Internet for information about medical conditions. In addition simply to seeking information through web portals, such as Wikipedia and WebMD, Internet users also interact with others online to obtain knowledge and support. People communicate through the Internet for common concerns by forming online communities, such as discussion forums and bulletin boards. The online communities designed specifically for people with a health interest are referred to as Online Health Communities (OHCs).

The most widely accepted definition of an OHC is "a group of individuals with a common interest or a shared purpose, whose interactions are governed by policies in the form of rules, rituals, or protocols; who have ongoing and persistent interactions; and who use computer-mediated communication as the primary form of interaction to support and mediate social interaction and facilitate a sense of togetherness" (Rodgers & Chen, 2005, p.00). As a special type of online community, OHCs share similarities with other online communities, but at the same time feature some unique characteristics. OHCs also have pros and cons compared with traditional offline support groups.

OHCs as Online Communities

Online communities are one example of how Internet technologies can help individuals interact with information contributed by one another in cyberspace. The formats of online communities are diverse, including chat rooms, email lists, discussion boards, and forums. Individuals join online communities for diverse personal preferences:

share content (e.g., Flickr), collaborate (e.g., Stack Overflow and Wikipedia), or have fun together (e.g., online multiplayer games). Preece (2000) summarized that an online community is a group of people who are connected through the Internet and interact over time around a shared purpose, interest, or need. Therefore, online communities span geographical boundaries and allow individuals to exchange knowledge and feelings at any time.

Compared with offline communities, online communities have both strengths and weaknesses. As one computer-mediated communication (CMC) tool, some asynchronous online community systems allow users to spend more time to create optimally desirable messages, with which users can selectively present information to one another. The hyperpersonal model of CMC (Walther, 1996) points out that impressions and relational states in CMC exceed what is expected to occur in face-to-face (FTF) settings. For example, text-based CMC allows senders to select polished sentences for self-presentation, and receivers adopt idealization in filling in the missing information. In addition, the lack of geographic, time, and even social barriers also contribute to the prevalence of online communities. For instance, CMC may hide factors that can lead to members' unequal participation rates, such as age or gender (Sproull & Kiesler, 1986). However, even if possessing many strengths, the drawbacks of an online community are apparent as well. Specifically, the relative lack of nonverbal cues of online community interactions would reduce the socio-emotional quality of communication (Walther, Van Der Heide, Ramirez, Burgoon, & Peña, 2015), which might require more time for a user to understand the message fully.

Similar to other online communities, OHCs allow users to seek, receive and interpret information on a 24/7 basis. Without barriers in location and time, OHCs help users to acquire information about their health concerns, which might not be adequately acquired from traditional FTF networks. For example, rare-disease patients sometimes cannot find people who face similar problems in the same city, and as a result, they spend time surfing online to seek help. In addition, news of novel treatments and effective medicines must spread through the Internet first. According to the Pew Research Center (Fox, 2014), one-in-four adult Internet users stated that they had read someone else's experience about medical issues in the past year. Communicating with peers who have similar health issues is beneficial for users facing common problems.

However, compared with other online communities, OHCs are more likely to emphasize similar experiences or emotions related to health problems. Specifically, users involved in OHCs are usually individuals with health concerns, such as patients and caregivers. In addition to simply sharing health-related information or knowledge, users also provide and receive social support. A quick response or encouragement in OHCs may help a user overcome his or her anxiety. Moreover, users do not care about the monetary value of their contributions. Users post content in OHCs to help others or seek information. Therefore, consistent with all online communities, OHCs serve as an interesting case deserving researcher's attention.

OHCs vs. Offline Support Groups

Besides the broad reach and 24-hour availability mentioned above, other advantages of OHCs have also surfaced compared with offline support groups. First, all the previous posts are warehoused on the website, which means new users can retrieve past related information any time. Although medical knowledge may update rapidly over time, the stored data are still a good resource to provide possible solutions or support to newcomers of the community. Meanwhile, compared with FTF support group, labor and time costs are efficiently saved. In addition, Wright (2015) points out that OHCs are beneficial in reducing users' embarrassment. Compared to face to face setting, online users may be more likely to express themselves in a straightforward and honest way about their concerns. In other words, OHCs can successfully mask physical appearance or disabilities that are a result of the health condition, which may be valuable for some users. Therefore, OHCs become important platforms for users sharing private or sensitive talk.

Nevertheless, OHCs have also been found to suffer from some problems. Risks such as sporadic membership of active users, deception, and insincerity of strangers are all possible reasons that may disappoint a user (Caplan & Turner, 2007; Wright, 2002; Wright & Bell, 2003). In contrast, an offline support group may be more able to mitigate these risks, or at least lower the probabilities of occurrences. Therefore, many users are involved in both online and offline support groups. Table 1 below provides a comparison of OHCs and offline health support groups. The valuable features enable OHCs to become an indispensable social media product in our life.

	Online Health Community	**Offline Support Group**
Features	***1)*** No limitation for time and location ***2)*** Weak ties among more users ***3)*** More confidentiality, free to talk about deeply personal issues ***4)*** Includes sensitive topics ***5)*** Directly seeking information without building relationships ***6)*** Prevents members from being judged by gender, age, or race ***7)*** Hostile behavior ***8)*** More self-experience description ***9)*** All messages are stored	***1)*** Group members meet regularly ***2)*** Strong ties among fewer users ***3)*** Less confidentiality ***4)*** Regular topics ***5)*** Exchange information after establishing personal relationship ***6)*** More respect ***7)*** Explicit suggestions to team members

Table 1 Online health community vs. offline support group (Pfeil, 2009).

Motivation of the Study

OHCs as a product of modern technology integrate scientific achievements from many different disciplines. As a result, studying OHCs is valuable in multiple aspects, such as recognizing communication patterns, improving public health systems, providing managerial insights, and understanding user behaviors.

Scientific contributions of studying OHCs

First, from the perspective of communication, studying OHCs is beneficial for summarizing users' participation with others in the community. Although physicians play a significant role in providing information, the majority of OHC users are patients and their caregivers (e.g., family members or friends). Participation in OHCs is helpful in obtaining therapeutic information, enhancing the understanding of medical knowledge, heightening emotional comfort and personal empowerment, and strengthening offline social connections (Barak, Boniel-Nissim, & Suler, 2008; Lieberman, 2007). Wicks et al., (2010) found that around 57% of users from "Patientslikeme.com" expressed that the OHC is helpful for understanding side effects of their treatments. Moreover, the common experiences promote users' mutual understanding, which may generate amounts of personalized and private discussions (Ferguson, 1996; Pfeil & Zaphiris, 2007). In addition, proper involvement in the social environment is beneficial in shaping individual behaviors as well, such as raising their self-satisfaction and confidence (McLeroy, Bibeau, Steckler, & Glanz, 1988). Last but not least, users can build friendships with other members online and even form social networks offline.

Second, from the perspective of public health, studying OHCs can help to inform and predict critical health events and establish effective interventions. Specifically, assessment (such as monitoring and diagnosing health problems), policy development (such as informing, educating the public, and developing related policies), and assurance (such as enforcing laws, providing care to needed patients, and assuring a competent public healthcare workforce) are three essential services of public health (Turnock, 2015). Timely and effective analysis of topics discussed in OHCs may assist early detection of health events and help to develop appropriate countermeasures. For example, an unexpected hot topic in a daily-used OHC may indicate a pandemic. Quick detection and preparedness can minimize human loss. During the 2009 global H1N1 influenza pandemic, emergency preparedness and a coordinated response of public health departments prevented an estimated 5 to 10 million additional cases, 30,000 hospitalizations, and 1,500 deaths in the United States. Therefore, studying OHCs is useful for improving the public health system.

Third, from the perspective of management sciences, studying OHCs can benefit other stakeholders who expect monetary values from the communities, such as website operators and medical marketers. Studying OHCs could benefit website operators in website management and benefit medical marketers by identifying potential customers. Specifically, the owners of OHCs operate the website to help people as well as earn profits. In the age of big data, large numbers of participants equal huge revenue. A clear understanding of user-generated content and user behavior can help operators to improve the website design. For example, operators can devise the website to attract newcomers as well as maintain senior members to run a sustainable OHC (Young, 2013). In addition, as discussed above, most users of OHCs are patients and caregivers, who are potential consumers of medical products, such as new treatments, drugs, and insurances. Understanding user-consumers in OHCs can benefit medical or insurance representatives to be aware of the extent of user-consumers' medical concerns, and find a way for profit maximization (Jayanti & Singh, 2010).

Finally, studying OHCs is beneficial for comprehending users' behaviors in the community. In particular, with observing activities and profiling information of users, researchers are able to figure out the trajectories of their involvement. For example, users join an OHC driven by different motivations (Hsu & Lin, 2008; Oh, 2012; Prasarnphanich & Wagner, 2009; Wasko & Faraj, 2005). Some of them are interested in meeting new friends, while others may only need a specific piece of information. Some of them become active leaders after several months, while others always behave as readers (Jennifer Preece & Shneiderman, 2009). In addition, users may provide different types of information or support to one another (Mi Zhang & Yang, 2015), and their continued participation in the OHC may be also impacted by receiving different types of support (Wang, Kraut, & Levine, 2012, 2015). Studying users' trajectories of involvement provide researchers a view to understanding how users are supported by each other in OHCs and how OHCs help to relieve stress related to health. The outcome of the studies can guide us in the direction of better serving users.

The study of OHCs in information science

Information science is an emerging discipline in the information age. Borko (1968) summarized information science as an interdisciplinary field that focuses on analyzing information, including its properties, behavior, usage, and transmission. Kling (2007) provided a more formal definition of informatics research: "the interdisciplinary study of the design, uses and consequences of information technologies that takes into account their interaction with institutional and cultural contexts" (p. 205).

Information scientists are interested in studying the relationship between information, technology, and people. In other words, information science not only studies information alone (e.g., the structure, form, and organization of information) but also studies how individuals interact with information (e.g., seeking, receiving, and interpreting information) and how technologies can facilitate such interactions. For example, information scientists study Twitter, with working on algorithms for data storage and information retrieval to improve users' experiences. At the same time, they also focus on the social network of users to see how users connect with each other and how information diffuses in the network. All these findings can be transferred to human knowledge in the end.

Healthcare research in information science focuses on how community members retrieve, comprehend, and use health-related information, as well as how stakeholders adopt and use information technologies in the healthcare system (Eysenbach, Powell, Englesakis, Rizo, & Stern, 2004). Turnock (2015) summarized that the integration of information is necessary to characterize the health status and needs of individuals. Thus, to better serve people who have health concerns, information scientists devote themselves to healthcare-related studies.

In the study of OHCs, an information science approach would not only extract knowledge from user-generated contents but also give an understanding of how users behave and interact with each other in the OHCs, as well as how to sustain a successful OHC. Studying OHCs provides great opportunities to inform science research. On the one hand, the outcome of these studies can make a difference and benefit OHC members. On the other hand, an interdisciplinary approach is needed to understand people's complex behaviors and interactions in OHCs better. Information scientists will need to integrate knowledge and methods from sociology, public health, computer science, and management science to address these problems.

The remainder of this thesis is organized as follows. Chapter Two covers prior literatures related to OHCs. With summarizing some research gaps from previous studies, I use machine learning methods to automatically detect social support from user-generated content (Chapter Three), analyze users' motivations and roles (Chapter Four), users' participation, and predict their churn (Chapter Five) in the OHC. Then, I investigate factors that impact users' future contributions (Chapter Six). Chapter Seven concludes all my findings, summarizes some limitations and offers some future research directions.

CHAPTER TWO

LITERATURE REVIEW

During the past two decades, many studies related to OHCs have been published. Multiple research questions have been addressed using various study designs including big data Internet studies using computational methods (Wang et al., 2015, 2012; M. Zhang, Yang, & Gong, 2013; K. Zhao et al., 2014), traditional social science studies summarizing findings from surveys or interviews (Coulson & Shaw, 2013; Matzat & Rooks, 2014; Setoyama, Yamazaki, & Namayama, 2011; Wicks et al., 2010). Irrespective of the methods used, the topics of most studies can be divided into two categories: community-level analysis and individual-level analysis.

Data Sources and Methodology

Depending on various data sources, prior studies related to OHCs set up different research goals and adopted different methodologies. The health issue of a target OHC is related to the design of the study at first. An OHC may focus on one specific health issue, such as Breastcancer.org for breast cancer patients and survivors and QuitNET for tobacco product quitters. An OHC may also serve audiences with varying health issues, such as PatientsLikeMe. Table 2 summarizes health issues the objective OHC cares about in related literature.

Basically, most OHCs are designed for chronic health conditions, such as cancer and diabetes, or health promotion, such as losing weight. The reason is people who have such chronic health conditions or addictions need continuous support during their long-term battles with these health issues. Furthermore, different types of OHCs offer researchers various formats and volumes of data. For example, users' temporal active spans in an OHC for an acute disease, such as the flu, are usually much shorter than those in an OHC for a chronic disease. In the OHCs related to acute diseases or infectious diseases (e.g., the discussion board for the flu on Yahoo! Answers), it emphasizes the effectiveness and quality of information, and as a result, researchers can collect informative posts from a large number of users during a relatively short time period. In contrast, chronic disease or health promotion OHCs focus more on users' long-term support. Some off-topic content may also be discussed online. Therefore, the nature of a target OHC sometimes resolves possible research questions and designs.

In addition, prior research has varied by discipline and by whether or not the team was multidisciplinary. Researchers with a background in computer science or information science usually built a web crawler to gather large volumes of public data from OHCs and analyze data automatically by means of computers, such as using a Support Vector Machine (Lu, 2013; M. Wen & Rose, 2012), EM clustering (Lu, Zhang, & Deng, 2013), and association rules (Yang, Yang, Jiang, & Zhang, 2012). The machine learning methods can save lots of human resources to finish the task, but the accuracy of content analysis remains open to question. By contrast, researchers with a sociology or communication background typically conducted more quantitative or secondary data analysis, such as manual coding a small scale of data online or summarizing findings from surveys and interviews with OHC users. These rule-based methods can reach higher reliability, but at the same time, the outcome may be influenced by subjective factors of coding experts or interviewees to an even greater extent. Hence, whichever the computational methodology or rule-based methodology adopted by researchers, it requires the methodology to fit the study target.

Reference	Health Issues	Website	Participants/Posts	Data Collection
(Rodgers & Chen, 2005)	Breast Cancer	NA	3-year posts from 100 females	Collaboration
(Coulson, 2005)	Irritable Bowel Syndrome	NA	572 messages from 132 users	Browsing online
(JoAnn Coleman et al., 2005)	Pancreatic Cancer	pathology.jhu.edu	600 posts	Browsing online
(Pfeil & Zaphiris, 2007)	Senior People	Seniornet.org	400 messages from 47 users	Browsing online
(Lieberman, 2007)	Breast Cancer	6 Bulletin Boards	77 users	Survey
(Eichhorn, 2008)	Eating Disorder	Yahoo! OHC	490 posts	Browsing online
(Ginossar, 2008)	Multiple Disease	www.acor.org	1,424 messages	Browsing online
(Idriss, Kvedar, & Watson, 2009)	Psoriasis	5 OHCs	260 users	Survey
(H.-J. Chang, 2009)	Emotional Distress	PTT.CC	689 posts from 438 users	Browsing online
(Mo & Coulson, 2010)	HIV/AIDS	NA	340 users	Survey
(Wicks et al., 2010)	Multiple Disease	PatientsLikeMe	1,323 users	Survey
(Cobb, Graham, & Abrams, 2010)	Smoking Cessation	QuitNET	7,569 users	Web Crawler
(K.-Y. Wen, McTavish, Kreps, Wise, & Gustafson, 2011)	Breast Cancer	Comprehensive Health Enhancement Support System (CHESS)	202 posts from 1 user	Browsing online
(Bender, Jimenez-Marroquin, & Jadad, 2011)	Breast Cancer	Facebook	620 discussion groups	Browsing online
(Setoyama et al., 2011)	Breast Cancer	Breast cancer OHC	253 users	Survey
(Weitzman, Cole, Kaci, & Mandl, 2011)	Diabetes	20 OHCs	2-year data	Browsing online
(Greene, Choudhry, Kilabuk, & Shrank, 2011)	Diabetes	Facebook	690 messages from 180 users	Browsing online

Table 2 A summary of data resources of literature related to OHCs studies.

Reference	Health Issues	Website	Participants/Posts	Data Collection
(Hambly, 2011)	Anterior Cruciate Ligament Reconstruction	KNEEguru OHC	201 users	Survey
(Qiu et al., 2011)	Multiple Disease	Cancer Society Cancer Survivors Network (CSN)	468,000 posts from 27,173 users	Collaboration
(Castleton et al., 2011)	Multiple Disease	NA	500 users	Survey
(X. Tang, Zhang, & Yang, 2012)	NA	ISI- KDD 2012 Challenge	27,968 threads, 129,425 comments from 2,803 users	Collaboration
(Chuang & Yang, 2012)	Alcoholism	MedHelp.org	493 forums, 423 journals and 1180 notes	Web Crawler
(Wang et al., 2012)	Breast Cancer	Breastcancer.org	1.5 million posts	Web Crawler
(M. Wen & Rose, 2012)	Breast Cancer	Breastcancer.org	10-year data from 31,307 users	Web Crawler
(A. T. Chen, 2012)	Multiple Disease	DailyStrength	2,852 posts for BC 2,806 posts for Type 1 Diabetes 6,100 for Fibromyalgia	Web Crawler
(X. Tang & Yang, 2012)	Alcoholism Smoking Cessation	MedHelp.org	352 users, 554 threads and 2339 responses; 446 users, 737 threads, 5,307 responses	Web Crawler
(Yang et al., 2012)	Multiple Disease	Medhelp.org	NA	Web Crawler
(Lu, 2013)	Breast Cancer	Komen.org	4,041 posts	Web Crawler
(Lu et al., 2013)	Multiple Disease	Medhelp.org	96,093 posts from 26,197 users	Web Crawler
(Huh, Yetisgen-Yildiz, & Pratt, 2013)	Diabetes	WebMD.com	8,239 posts from 2,902 users	Web Crawler
(Y. Zhang, He, & Sang, 2013)	Diabetes	Facebook	154 posts and 1,198 responses	Browsing online

Table 2 - continued

Reference	Health Issues	Website	Participants/Posts	Data Collection
(M. Zhang et al., 2013)	Smoking Cessation	QuitNET	228 posts and 1,672 comments	Web Crawler
(van der Eijk et al., 2013)	Parkinson	NA	NA	NA
(Loane & D'Alessandro, 2013)	Amyotrophic Lateral Sclerosis (ALS)	NA	61 threads and 499 posts from 133 users	Browsing online
(Jiang & Yang, 2013)	Multiple Disease	MedHelp.org	16,339 threads and 120,393 messages	Web Crawler
(Wentzer & Bygholm, 2013)	Multiple Disease	Sundhed.dk	4,301 posts from 630 users	Browsing online
(Coulson & Shaw, 2013)	NA	24 OHCS	33 OHC moderators	Survey
(Petrovčič & Petrič, 2014)	NA	2 OHCs in Slovenia	616 users	Survey
(Patki et al., 2014)	Multiple Disease	DailyStrength	20,486 comments	Web Crawler
(Ho, O'Connor, & Mulvaney, 2014)	Diabetes	18 adolescents' Type 1 Diabetes OHCs	NA	Browsing online
(Matzat & Rooks, 2014)	Multiple Disease	Yahoo! OHC	1,050 users	Survey
(Rupert et al., 2014)	Multiple Disease	77 different OHCs	89 users	Collaboration
(K. Zhao, Greer, Yen, Mitra, & Portier, 2015)	Multiple Disease	Cancer Society Cancer Survivors Network (CSN)	27,173 users	Collaboration
(MacLean, Gupta, Lembke, Manning, & Heer, 2015)	Drug Abuse	MedHelp.org	2,848 users	Collaboration
(Goh, Gao, & Agarwal, 2016)	Rare Disease	NA	638 patients	NA
(Kirk & Milnes, 2016)	Cystic Fibrosis	A CF charity website	151 threads from 279 participants	Browsing online
(K. Zhao et al., 2016)	Smoking Cessation	BecomAnEX	1,337 users	Collaboration
(Pearson et al., 2017)	Smoking Cessation	BecomAnEX	Posts between 2008-2015	Collaboration

Table 2 - continued

Community-level Analysis

Posts, as user-generated content in OHCs, are the most applicable resources to help researchers understand OHC users at the community level. Prior studies at the community level mainly focused on several aspects: social support analysis, topic of discussions, medical information retrieval, and sentiment analysis.

Social support analysis

Social support refers to the "exchange of resources between at least two individuals perceived by the provider or the recipient to be intended to enhance the well-being of the recipient" (Shumaker & Brownell, 1984, p.11). Based on the nature of exchanged "resources," community psychology researchers have identified different types of social support. For example, House (1981) defined four types of social support: emotional, instrumental, informational, and appraisal. Another set of researchers created a more fine-grained typology of social support with six categories: material aid, behavioral assistance, intimate interaction, guidance, feedback, and positive social interactions (Barrera & Ainlay, 1983).

The literature on social support suggests that OHCs mainly feature four types of social support: informational support, emotional support, companionship (a.k.a., network support), and instrumental support (Bambina, 2007; Keating, 2013). Informational support is the transmission of information, suggestion, or guidance to the community users (Krause, 1986). The content of such a post in an OHC is usually related to advice, referrals, education, and personal experience with the disease or health problem. Example topics include side effects of a drug, ways to deal with a symptom, experience with a physician, or medical insurance problems. Emotional support, as its name suggests, contains the expression of understanding, encouragement, empathy, affection, affirming, validation, sympathy, caring and concern, etc. Companionship, also known as network support, consists of chatting, humor, teasing, and discussions of offline activities and daily life that are not necessarily related to one’s health problems. Thus, they are sometimes referred to as off-topic discussions. Examples include sharing jokes, birthday wishes, holiday plans, or online Scrabble games. Instrumental support, or tangible support, refers to offline support activities in the physical world, such as transporting others to hospitals and assistance in grocery shopping.

Social support is frequently offered and received in OHCs, the interacting platform for users who have common health concerns. Users connect with one another by replying or commenting, thereby forming a social network. According to a conceptual model proposed by a book Glanz, Rimer, and Viswanath (2008) in Figure 1, building social networks and being involved in different types of social support can directly impact the individual’s coping resources and community resources, such as enhancing one’s ability to solve problems. Meanwhile, it can impact regularity of an individual’s exposure to stress and help the person recovering from the health issues. While people may have different reasons to participate in an OHC, obtaining social support is among the most important needs (Bouma et al., 2015; E. Kim et al., 2012; Rodgers & Chen, 2005). Social support can help them adjusting to the stress of living with and fighting against their diseases (Dunkel-Schetter, 1984; Qiu et al., 2011; K. Zhao et al., 2014) and it is a consistent risk factor for survival (McClellan, Stanwyck, & Anson, 1993).

Users involved in OHCs designed for different health issues may need different types of social support. As mentioned earlier, operators design OHCs for multiple diseases, such as acute disease, chronic disease, and health promotion. In OHCs for health promotion, such as communities interested in weight loss, informational support is more frequently expressed in initial posts of threads or in public channels, while emotional support is more popular in comments of threads or in private channels (Chuang & Yang, 2012; Mi Zhang & Yang, 2015). By contrast, social support is exchanged differently among users suffering from chronic health conditions. For example, informational support is emphasized more in a diabetes group (Greene et al., 2011; Y.

Zhang et al., 2013), while network support is more frequent in an Amyotrophic Lateral Sclerosis OHC, a disease characterized by stiff muscles, resulting in difficulty speaking, swallowing, and eventually breathing (Loane & D'Alessandro, 2013). Moreover, the geographical locations can also differentiate users in exchanging social support: urban users act more as suppliers of social support, while rural participants are recipients (Goh et al., 2016). Therefore, social support exchanged in the community depends on the nature of both users and OHCs.

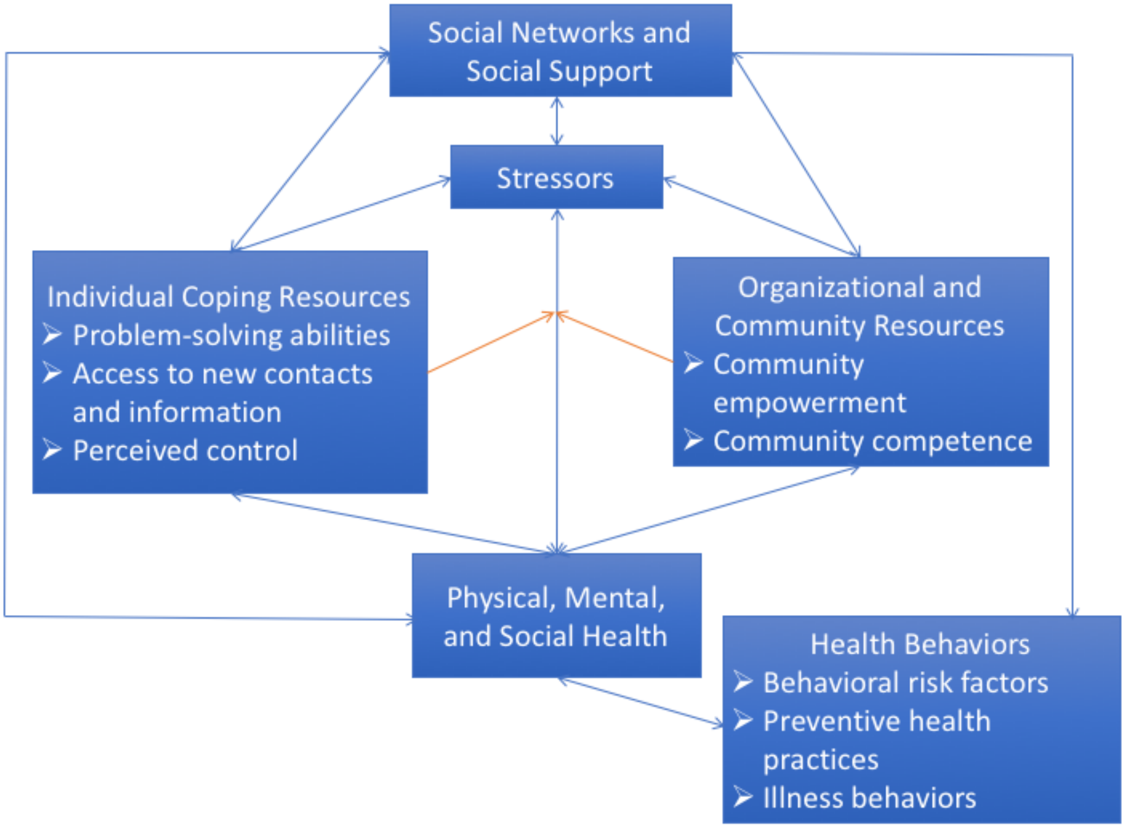

Social Networks and Social Support
Stressors
Individual Coping Resources
➢ Problem-solving abilities
➢ Access to new contacts and information
➢ Perceived control
Organizational and Community Resources
➢ Community empowerment
➢ Community competence
Physical, Mental, and Social Health
Health Behaviors
➢ Behavioral risk factors
➢ Preventive health practices
➢ Illness behaviors

Figure 1 The conceptual model for the relationship of social networks and social support to health

Overall, there is substantial research related to social support and health online. The aim of all these past studies is to understand users' requirements and preferences better. If a user were well satisfied by social support in the OHC, there would be a higher chance of this user continuously participating and contributing. Therefore, social support is also discussed a lot in observing users' participation in OHCs. For example, previous studies showed that OHC users' activities in receiving different types of social support are positively correlated with users' engagement in OHCs (Wang et al., 2015). I will cover more details about this topic in developing research goals of Chapter Five.

Topics of discussions

Driven by multiple motivations, users' posts in OHCs relate to various topics. In addition to traditional health information, other content, such as end-of-life content (JoAnn Coleman et al., 2005), gratitude, and fundraising (Bender et al., 2011) are also discussed in OHCs. Uncovering topics discussed in OHCs assists researchers in understanding users' needs and experiences. One application of such study is building recommender system to provide the user who needs help. Some research groups adopted rule-based analysis, summarized discussed topics by manual coding, while some others clustered topics automatically with the help of machine learning. For example, some researchers design topic modeling algorithms to reach higher degrees of accuracy in topic detection in OHCs (Huh et al., 2013; Mi Zhang & Yang, 2014).

Whichever method it takes, the findings of these studies show overlap among topics discussed. For example, Pfeil and Zaphiris (2007) manually reviewed 400 messages from SeniorNet and summarized seven topic categories, namely self-disclosure, light support, deep support, community building, medical facts, technical issues, and off topic chitchat. A. T. Chen (2012) implemented bisecting k-means to cluster posts from DailyStrength and summarized topics such as support, experiential knowledge and treatments. In addition, the topic of posts was highly correlated with a user's health trajectory as well. Patients who were suffering a chronic disease, such as diabetes, discussed more emotional-related content; however, the cancer survivors, who were threatened by limited survival time, were more eager to find treatment options to extend their lives (Lu, 2013; Lu et al., 2013; M. Wen & Rose, 2012).

Medical information retrieval

As user-generated content, posts in OHCs are diverse in linguistic styles. Compared with narratives offered by professionals, user-generated content contains more patient-centered words than medical jargon. Therefore, extracting valuable medical information from user-generated content and summarizing linguistic styles of users in OHCs are necessary.

Popular research goals related to medical information retrieval include adverse drug reaction detection, drug repositioning, drug-drug interaction detection, and deriving user-generated vocabularies (Jiang & Yang, 2013, 2015; Patki et al., 2014; Yang et al., 2012; M. Zhao & Yang, 2016). According to a research report (Sarker et al., 2015), similar methods have also been used in other social media to address alike problems (Ginn et al., 2014; O'Connor, Gaynes, Burda, Soh, & Whitlock, 2013). The outcome of these studies can provide important medical or other related knowledge and reduce the communication gap between consumers and medical professionals.

Meanwhile, due to a lack of visual and aural cues in OHCs (White & Dorman, 2001), the words or emoticons users select to express themselves become meaningful and informative. For example, words such as "aware," "know," and "realize" were discovered to be significantly positively associated with the OHC users' decreasing concerns on the disease itself and were helpful in reducing negative emotional feelings (Lieberman, 2007; Shaw, Hawkins, McTavish, Pingree, & Gustafson, 2006; Wentzer & Bygholm, 2013).

Sentiment analysis and emotional dynamics

Users in OHCs may show different emotional states. One individual may express a positive emotion in talking about an event in one post, and may also interpret negative sentiment toward a bad result of a medical test in another. Sentiment analysis and emotional dynamics of users in other social media have been widely studied, so many mature algorithms for building sentiment classifiers are implemented directly in OHCs' scenarios. However, the health-issue-centered OHCs require minor tweaks on implementing these methods, which are summarized from other social media platforms, such as methods designed for a movie ranking website specifically. Many factors have been discovered particularly impacting a user's sentiment in OHCs, such as active level in the community, age, and stage of disease of the user (Wu & Peng, 2015; S. Zhang, Bantum, Owen, & Elhadad, 2014). Based on these factors, the algorithms designed through other research are improved to judge sentiment for posts in OHCs (Ali, Schramm, Sokolova, & Inkpen, 2013; Qiu et al., 2011). In addition, users' emotions may also change over time within one thread. For example, the number of others' replies, the subsequent involvement of the thread originator, and the average sentiment of the other users all positively contribute to sentiment switching of the originator from negative to positive (Bui, Yen, & Honavar, 2015; Qiu et al., 2011; K. Zhao et al., 2014). Sentiment analysis in OHCs attracts attentions from many researchers, and the outcome of the analysis can be used as critical preparations in solving other research problems.

Overall, studies on posts analysis at a community level are critical to understand users' general requirements and emotions in OHCs. According to the outcome of the studies, operators of the community may have new insights in sustaining a healthy online ecosystem.

Individual-level Analysis

Individuals join in OHCs for similar experiences or concerns. Moving from the community-level to individual-level service, high-quality social support may enhance a user's ability and encourage the user fighting with health issues. Barak et al. (2008) summarized that others influence users in OHCs in a general way rather than presenting some specific changes. How can I quantify gain and loss of users from OHCs? Prior studies answer this question from the analysis of user role, demographics, social network, and user participation.

User role analysis

People participate in OHCs in different styles and may play multiple roles in such communities. A popular research goal is identifying users' roles and exploring influential factors changing those roles. For example, Faraj, Kudaravalli, and Wasko (2015) summarized that not only the sociability and active contribution behavior result in a user's role of leader, but also the structural positioning in the communication network matters. Previous research about identifying users' roles in OHCs was mainly based on aggregated data of users' behaviors, such as the number of log-ins, the number of posts, and the number of active days (Jones et al., 2011). However, social network centrality metrics are also important, such as in and out degrees (Cobb et al., 2010). While intuitive and easy to obtain, these metrics did not reflect what a user has sought from or contributed to an OHC. Thus, roles identified based on these metrics can only provide a coarse-grained view of a user's online activities. A more fine-grained view of users' roles would require analyzing the contents of a user's posts. Prior studies that examined such content for user role identifications either depended on manual content analysis of posts, or stayed at the lexicon level by counting the appearance of words or phrases (Füller, Hutter, Hautz, & Matzler, 2014; Sudau et al., 2014).

In addition to exploring how users play different roles in OHCs, some qualitative research also suggested that a user's role may shift over time (Pfeil & Zaphiris, 2007). For example, based on manual content analysis, Loane and D'Alessandro (2013) argued that many OHC users start as information seekers and some of them could switch to support providers after a time. Another study proposed that online community users may switch between the roles of reader, contributor, collaborator, and leader, but did not provide any empirical evidence (Jennifer Preece & Shneiderman, 2009).

Demographics analysis

Detecting demographic similarities and discrepancies of users in OHCs is another hot topic attracting researchers' attention. Monnier, Laken, and Carter (2002) suggested that in addition to developing Internet-based health service, fences of ethnicity, age, and education level should be addressed as well.

The demographic information contains many categories, such as gender, age, education attainment, and posting frequency. Prior studies found that age, race, education level or medical history are all factors associated with users' health-related Internet use (Castleton et al., 2011; Y. Zhang, 2010). Females dominate in publishing question-based posts, interaction and communication, whereas males are more active in responding to a request, information gathering and entertainment in OHCs (Eichhorn, 2008; Idriss et al., 2009; Kinnane & Milne, 2010). Users from the same OHC also have demographic features in common, such as the average woman who used an OHC for breast cancer was 46 years old, married, and held a professional occupations (Rodgers & Chen, 2005).

Most studies in this area collected users' demographic information through surveys or interviews. Because users answered a questionnaire or question by recalling their online behaviors, the results of statistical analysis have limitations. For example, sometimes respondents hide their thoughts or real behaviors to avoid judgments from others. In addition, expensive manpower and time requirements make it is impossible to be used in dealing with large-scale data.

Network pattern analysis

Users in OHCs express opinions under threads and discuss interested topics with others, forming an online social network. Such connections can be described as a co-participation network. The links can be either directed, such as A replies to B, or undirected, such as both A and B comment under one blog. The size of the network grows gradually with more users joining in the OHC, and meanwhile indicating more resources are available for the community.

A network of OHCs has been described as a “bottom-up” rather than a “top down” network, without the structure of “provider as authority” (Lester, Prady, Finegan, & Hoch, 2004). The degree distribution of OHC network usually follows a power law distribution. In other words, the majority of users in OHCs have a few ties with others, while the minority of users are well connected with all others. In addition, an OHC network is centralized, in which users prefer to connect with someone sharing similar characteristics, such as age and gender, and connections are usually weak ties rather than strong ties, which means users communicate on a daily basis but are not necessarily close friends (Centola & van de Rijt, 2014; Cobb et al., 2010; K. Wright & Bell, 2003).

Users are involved in multiple subgroups in one OHC and the network neighborhoods of a user can impact his or her behavior to some degree. Users exchange different types of support with surrounding neighbors, depending on the nature of the network neighbor. For example, posts containing sensitive content were shared merely in a patient–patient network, rather than a patient–physician network (Greene et al., 2011). Besides, informational support is more frequently delivered from senior OHC members to juniors and presented a lower density, while emotional support is exchanged between neighbors at the same membership status (H.-J. Chang, 2009; M. Zhang et al., 2013). In addition, the findings of network neighborhood studies can also be used for further analyses. For example, researchers have developed approaches to capture influential users in the OHC based on analyzing the neighborhood around individuals (J.-H. Tang & Yang, 2005; X. Tang & Yang, 2012; K. Zhao et al., 2014).

User participation analysis

User participation analysis is another popular research topic related to OHC studies. Ten years ago, no negative effects of OHC had been reported, but some researchers noticed that high dropout rates occurred (Eysenbach et al., 2004). Users' continued participation in an OHC is not only beneficial to OHC operators but can also be therapeutic to users themselves (Idriss et al., 2009). Mo and Coulson (2013) pointed out that participating in OHCs could help users gain more information, better understand the circumstance they are involved in, and become better able to make the decisions that may affect their lives.

OHCs anticipate more about users' posting behavior than reading or lurking behavior. However, users are usually driven by different motivations to contribute actively. For example, indirect social control is found more influential than direct rewarding (monetary prize), to spur users' posting behavior (Matzat & Rooks, 2014). The sense of belonging to the community might also be a reason encouraging users' active contributions (Y. Zhang et al., 2013).

Although people may use OHCs for a wide range of needs, seeking and obtaining social support is one of the key benefits of participation in OHCs (E. Kim et al., 2012; Rodgers & Chen, 2005). As mentioned earlier, different types of social support can impact the users' participation in the OHC. Receiving emotional support has been found to be positively related to users' long-term participation and activity level (Wang et al., 2012, 2015). Compared with active contributors, lurkers can also receive support from others, but are less satisfied with the social support they receive (Mo & Coulson, 2010;

Setoyama et al., 2011). Therefore, understanding users' participation is valuable for sustaining stable OHCs. More discussion about this topic will be provided in Chapter Four and Five.

Overall, this Chapter summarizes prior studies related to OHCs. To understand clearly what has been studied in this area, a bipartite network of topics studied and methods adopted is shown in Figure 2 . Apparently, machine learning, text mining, and manual coding methods are widely used for community-level analysis, while surveys and interviews accompanied with statistical analysis are implemented more in individual-level analysis.

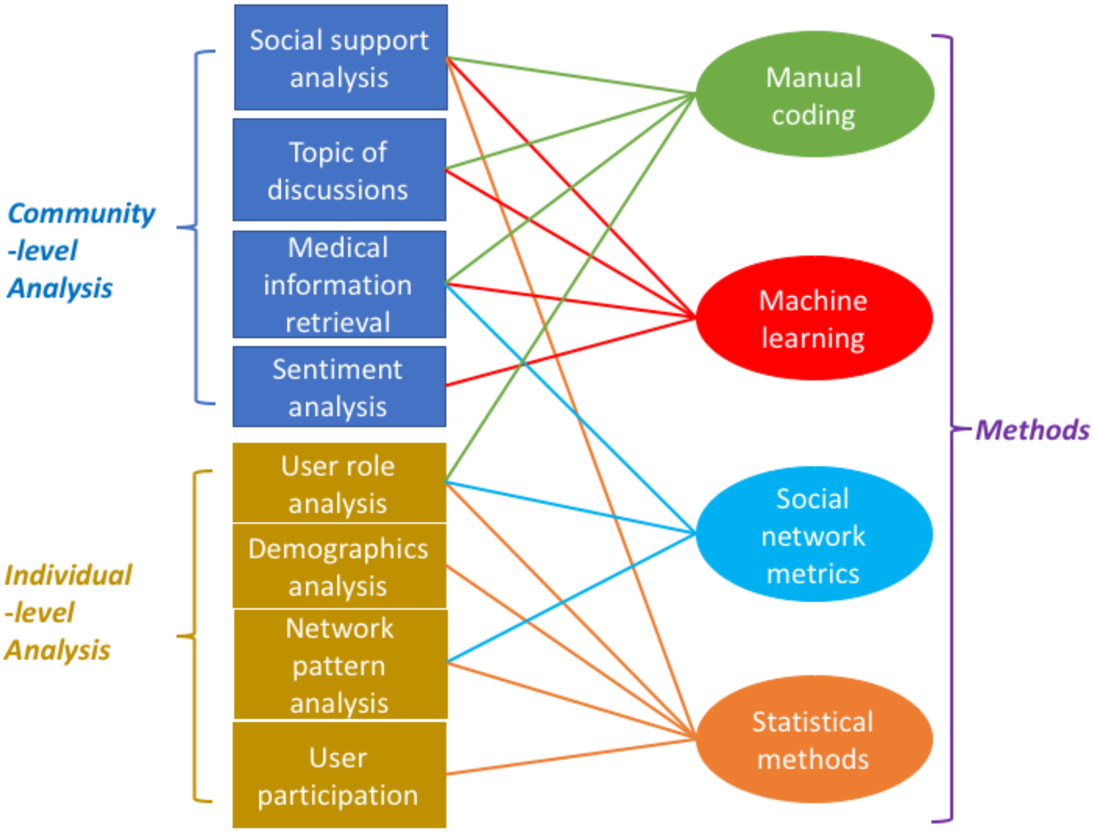


Figure 2 The bipartite network illustrating topics and methods in prior OHC studies

CHAPTER THREE
SOCIAL SUPPORT DETECTION

Prior studies related to OHCs have covered multiple research topics as mentioned above. However, compared with other well-studied social media, such as Twitter or Facebook, many pros and cons of engaging in OHCs remain unanswered. Because receiving different kinds of social support is the most usual goal of users who involve in OHCs, I propose research questions to address the problems from the perspective of social support.

The history of the linkage between social support and health can be traced back to 1897 (Durkheim, 1897). Cassel (1976) pointed out that social support is an essential "protective factor" that decrease one's vulnerability to the harmful effects coming from pressure on health. Several decades later, the emergence of OHCs provides new opportunities to study social support at unprecedented scales and granularities on health. Traditional studies about offline support communities relied heavily on data collected through ethnographical observations, interviews, questionnaires, or surveys (Campbell, Phaneuf, & Deane, 2004; Gorlick, Bantum, & Owen, 2014; Hambly, 2011; Lieberman, 2007; Setoyama et al., 2011). However, research using these data collection methods faces three challenges. First, the scale of the data is limited because observations and interviews are labor intensive and time consuming. Second, results may be biased due to the realities of sampling community members. For example, members who are active in or satisfied with their communities may be more likely to respond to questionnaires or surveys. Third, survey and interview methods typically have coarse temporal granularity and rely on members' recall of past events and associated feelings. This sometimes makes it very difficult accurately to track members' activities during an extended time in the community. By contrast, OHCs not only enable, but also record asynchronous and distributed social interactions among individuals, making the "big data" available for computational analysis (H. Chen, Chiang, & Storey, 2012). Such detailed data from users' online interactions (e.g., the amount, content, and time of interactions) contains valuable information on users' behaviors. Nevertheless, many previous studies about social support in OHCs did not take full advantage of the large-scale data and still examined a small sample of OHC users' social support activities. To study social support at such a large scale and fine granularity, I need to reveal the nature of social support embedded in users' contributions in an automated way. Hence the first research problem I want to address is detecting different types of social support activities from mining large-scale text data contributed by OHC users.

Research Goal of this Chapter: Detect the seeking and provision of different types of social support from unstructured text of large-scale distributed interactions among OHC users.

Dataset and the Taxonomy of Social Support

In this research, I used the data from a very popular peer-to-peer OHC for breast cancer (https://community.breastcancer.org). I designed a web crawler to collect data from its online forum. The dataset consists of all the public posts and user profiling information from October 2002 to August 2013. There are more than 2.8 million posts (including 107,549 threads) contributed by nearly 50,000 users. Although medical science advanced rapidly during this period of time, for example, more and more treatment options came out, the 11-year data is an adequate source to understand users' behavior in the context of OHCs.

Empirical studies suggested that informational support, emotional support, and companionship are common in many OHCs, but instrumental support is rare, as such support is often limited by geographical proximity (Coulson, Buchanan, & Aubeeluck, 2007; M. Zhang et al., 2013). In addition, the further exchange and arrangement of instrumental support may often occur via private or offline communication channels (e.g., setting a time for grocery shopping via cell phones, giving someone money, and sending food). Due to the instrumental support was expected to be uncommon in the Web-based context, to simplify the automated social support classification, I did not consider instrumental support in this thesis.

To understand users' behaviors, it needed to determine whether the post was seeking informational support (SIS), providing informational support (PIS), seeking emotional support (SES), providing emotional support (PES), or simply about companionship (COM). Note that there is no necessity to differentiate the seeking and provision of companionship because the nature of companionship is about participation and sharing. By getting involved in activities or discussions about companionship through posting, one is seeking and providing support at the same time. It is also possible that a post could belong to more than one of the five categories above. Table 3 lists example posts for each category and a post that belongs to two categories.

Social Support	**Examples**
Companionship (COM)	(1) *Kelly Have a wonderful time in Florida, enjoy the sun and fun. Heather* (2) *I'm loving her new CD. Didn't recognize any of the songs at first, but there are a few now that I find myself singing the rest of the day.* (3) *This game has the poster making a new 2 word phrase starting with the second word of the last post Example: Post : Hand out Next poster: Out cast Next poster: Cast Iron Next poster: Iron Age Now let's begin the game~ Age Old*
Seeking Informational Support (SIS)	*Where do you buy digestive enzymes and what are they called?*
Seeking Emotional Support (SES)	*I feel like everyone else's lives are going forward, they have plans, hopes, aspirations because they feel. I am one of those not yet out of the woods. I was also someone who could never get cancer. I was a good person, exercised, ate well. Good people don't get sick. I have taken the step of antidepressants, they mitigate the damage, but do not block the pain or sadness I feel.*
Providing Informational Support (PIS)	*I had surgery Aug05 for bc recurrance. B4 surgery I had 33 IMRT rads, prior to that had 4A/C &amp; 4 Taxol. I had bc in 2000 &amp; had 37 rads in same general area. Now, my surgery won't heal. Wound doc says there is adema or something on my sternum (shown on recent MRI). My wound has been draining since it broke open in Sept.*
Providing Emotional Support (PES)	*Hope you feel better soon, we are here! Prayers Hugs come from Massachusetts APPLE♥.*
Providing Informational Support (PIS) & Providing Emotional Support (PES)	*I am also the daughter of a 35 yrs BC survivor. Mom is just now going through some more Cancer - alas - they found it in her lung, but it is totally unlikely to be a follow-up of her old BC. I am 45, and was 43 at DX time, my mom was diagnosed at 38... and I am a BRCA2 carrier. Tina, one day at a time. Maybe you'll get good news - it is so hard to wait!!! It is also important to remember that - whatever it is, it is highly treatable, and that YOU WILL SURVIVE too!!! and life goes on after. It will take some time, but it goes on... see my picture? even the hair is back!!! Hugs to all. I am happy you all found your way here, it is a great site for exchanging information, learning and finding support.*

Table 3 Example posts for different types of social support

Annotations and Features

Because it is practically impossible to label all OHC posts manually (2.8 million posts), I used classification algorithms to decide what kind(s) of social support each post contains. To train the classification algorithm, the human annotations have been leveraged. I randomly selected 1,333 posts (54 initial posts and 1,279 comments) out of the dataset. After training on the definitions and examples of the aforementioned five categories of social supports (SIS, PIS, SES, PES, COM), five human annotators were asked to read each post and decide whether the post belonged to one or more categories of social support.

To control the quality of human annotations, I also added to the pool 10 posts that have been annotated by domain experts. For each post, I only accepted results from annotators whose performance on the 10 quality-control posts was among the top three. Results from the other two annotators were discarded. Then a majority vote among the top three annotators was used to determine whether a post is related to a category of social support. Table 4 shows the results of the annotation process.

Social Support Category	**Number**
Companionship (COM)	435
Seeking Informational Support (SIS)	96
Seeking Emotional Support (SES)	22
Providing Informational Support (PIS)	411
Providing Emotional Support (PES)	249

Table 4 The number of posts in each category of social support in the annotated dataset

Users in OHCs may have different writing styles or linguistic preferences to express themselves. To capture these characteristics, I examined each post and extracted various types of features for building the classifier: basic features, lexical features, sentiment features, and topic features. Table 5 summarizes these features. Many of the features were picked specifically for classification in this context of OHC. For example, I included "whether a post is an initial post" as a feature because many users sought support by starting a thread. Inside each post, the existences of URLs and emoticons are often related to informational and emotional supports, respectively. Similar to the approach used by Wang et al. (2012), I also checked the usage of phrases in the format of <you/he/she + MODAL verb > to express possibilities, such as "you should," and "she could." I considered "he" and "she" in addition to "you," because family members of cancer survivors created some of the posts. To identify the difference between "seeking" and "providing" support, I included words related to seeking behavior, such as "question," "wonder," and "anybody." The words concerning daily life topics and geographical locations were also included to discover COM posts. Meanwhile, I used OpinionFinder (Wilson, Wiebe, & Hoffmann, 2005) to find the overall sentiment, as well as subjectivity and objectivity of each post.

In addition to these handpicked or dictionary-based lexicons, I also wanted to capture whether the usage of other words and phrases can contribute to the classification. Using unigrams and bigrams was too fine-grained and leads to a feature set with very high dimension. Thus, I adopted an approach similar to a previous study (Wang et al., 2012) and applied the topic-modelling technique Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) to the content of all posts and generated 20 topics (top terms for LDA topics are shown in Table 6). For each post, LDA gave a topic probability distribution, indicating the probability of this post corresponding to each topic. Such a distribution was then included in the feature set as well.

Group	Features
Basic Features	Whether the post is an initial post in a thread Whether the post is a self reply Length of the post
Lexical Features	Whether the post contains URLs (Y or N) Whether the post contains emoticon(s) Number of numeric numbers Number of Pronouns (e.g., they, we, I) Whether the post contains the negation word(s) (e.g., not, never, no) Whether the post contains name(s) of city, state, country (U.S.A, Canada, etc.) Whether the post contains phrases related to possibility (you must, you might, she had better, etc.) Whether the post contains names of drugs related to breast cancer (From http://www.cancer.gov/cancertopics/druginfo/breastcancer) Whether the post contains breast cancer terminology (From http://www.breastcancer.org/dictionary) Whether the post contains verb related to advice (Need, require, recommend, etc.) Whether the post contains emotional words (Love, sorry, hope, worry, etc.) Whether the post contains words related to seeking behaviours (Anybody, question, wonder, etc.) Whether the post contains words related to daily life topics (Vacation, joke, run, walk, etc.)
Sentiment Features	Frequency of words with positive and negative sentiment Objectivity and subjectivity scores
Topic Features	Topic distributions derived from LDA

Table 5 Summary of features for the classifier

Topic	Top 10 terms
LDA-0	feel, pain, time, back, bad, felt, hurt, hard, normal, ca
LDA-1	week, wait, surgery, call, rad, back, start, dr, month, time
LDA-2	make, amp, life, people, thing, time, understand, deal, decision, give
LDA-3	breast, surgeon, biopsy, node, cancer, mastectomy, mri, lumpectomy, lump, dcis
LDA-4	cancer, woman, breast, risk, study, patient, treatment, research, cell, recurrence
LDA-5	year, family, friend, time, husband, live, life, sister, kid, love
LDA-6	eat, make, water, food, weight, drink, add, lot, diet, good
LDA-7	hair, back, head, grow, long, lose, start, wear, short, cut
LDA-8	year, side, tamoxifen, month, problem, effect, start, stop, hot, blood
LDA-9	god, pray, mom, prayer, love, friend, bless, peace, comfort, daughter
LDA-10	day, work, today, night, back, home, time, sleep, hour, walk
LDA-11	chemo, day, treatment, week, start, give, tx, taxol, Herceptin, port
LDA-12	love, great, dh, weekend, fun, enjoy, nice, hope, today, lol
LDA-13	good, ve, ll, thing, time, luck, lot, make, ca, feel
LDA-14	hope, good, hear, great, glad, happy, love, hugs, news, hug
LDA-15	post, read, find, thread, site, question, info, gt, board, information
LDA-16	room, watch, house, dog, put, laugh, guy, car, big, clean
LDA-17	stage, chemo, treatment, scan, year, cancer, bone, test, mets, onc
LDA-18	insurance, work, care, people, pay, call, medical, health, doctor, make
LDA-19	surgery, arm, le, ps, side, skin, implant, bra, reconstruction, drain

Table 6 Top terms for LDA topics

Evaluation of the Classifier and Predictions

Since I considered five categories of social support (SIS, PIS, SES, PES, COM) and a post might belong to more than one category, I trained a classifier for each category. For the classification of each category of social support, I applied various classification algorithms on annotated posts and picked the best-performing one using 10-fold cross-validation. Because SES posts accounts for only a small proportion of the annotated posts (22 out of 1,333), I oversampled positive posts when building the SES classifier. Among all the classifiers I tried, AdaBoost, with Naïve Bayesian as the weak learner, was chosen to classify COM, PES, PIS, and SIS, while logistic regression was the best choice for SES. Overall, the classifiers achieve decent performance with an accuracy rate above 0.8 in all five classification tasks (Table 7).

Social support	**Results**	**Naïve Bayes**	**Logistic Regression**	**SVM**	**Random Forest**	**Decision Tree**	**AdaBoost**
COM	Accuracy	0.696	0.787	0.783	0.771	0.767	**0.804**
	AUC	0.839	0.817	0.768	0.848	0.75	**0.852**
PES	Accuracy	0.713	0.830	0.840	0.830	0.81	**0.817**
	AUC	0.823	0.787	0.681	0.825	0.687	**0.817**
PIS	Accuracy	0.753	0.813	0.823	0.767	0.779	**0.801**
	AUC	0.824	0.83	0.783	0.837	0.717	**0.859**
SES	Accuracy	0.893	**0.901**	0.970	0.967	0.963	0.963
	AUC	0.749	**0.867**	0.656	0.851	0.671	0.668
SIS	Accuracy	0.851	0.880	0.943	0.931	0.937	**0.914**
	AUC	0.893	0.803	0.745	0.86	0.766	**0.869**

Table 7 Performance of classification algorithms for the five categories of social support

After applying the best-performing five classifiers on the remaining of the 2.8 million posts, each post received five labels, each of which indicated whether the post belongs to one of the five social support categories. The total numbers of posts in each category are listed in Table 8. Intuitively, there are more posts to provide support than to seek support. This is what one would most expect from a popular OHC with a large and active user base. About 37% of the posts provide informational support, making it the largest group among the five. In other words, providing informational support is the most popular activity in the OHC. COM posts constitutes the second largest group, which suggests that members of the OHC did form a strong sense of community and discussed many issues other than cancer. In addition, 197,956 posts are predicted as PIS and PES at the same time, representing the largest group with more than one category of social support.

Social support category	**Total number of posts**
Companionship (COM)	932,538
Seeking Informational Support (SIS)	284,027
Seeking Emotional Support (SES)	227,188
Providing Informational Support (PIS)	1,034,682
Providing Emotional Support (PES)	497,096

Table 8 Total numbers of posts in each category of social supports

Summary

In this Chapter, with machine learning methods, I was able to detect the seeking and provision of different types of social support automatically for 2.8 million posts. According to the distribution of social support, it is surprised to notice that companionship consists an indispensable part of this OHC. Even if the OHC is designed for providing information and support to the users with breast cancer concerns and the users may first join for some information, the off-topics are widely discussed in this community. Based on the results, I could keep working on the analysis of users' engagement in the OHC from the perspective of social support.

CHAPTER FOUR

USER ROLE DYNAMICS

As mentioned previously, participating in OHCs could help users gain more information, better understand their own circumstance, make decisions that may affect their lives, and acquire resources offline (Mo & Coulson, 2013). Despite all the benefits, many OHCs are still facing low level of activity from their users. For example, many users read lots of others' posts but never leave replies. Thus, it is important to understand the motivation of users' engagement and contributions in OHCs. Some online communities feature mechanisms to explicitly reward users' participation or contributions (e.g., virtual badges or stars), and these online rewards can sometimes have monetary implications too. For example, a programmer's badges on StackOverflow can potentially land her or his a well-paid job (Feffer, 2015). When such explicit mechanism is missing, which is the case for many OHCs, people's continued participation is often driven by altruism- people's behaviors to increase the welfare of others (Gintis, Bowles, Boyd, & Fehr, 2003).

In the context of OHCs, altruism can be interpreted as serving the interests of the community, such as providing social support to others without explicit rewards, whether it is informational support, emotional support or companionship. Users who contribute to such "community interests" have been associated with continued participation in online communities (Hsu & Lin, 2008; Prasarnphanich & Wagner, 2009; Wasko & Faraj, 2005), including OHCs (Oh, 2012). By contrast, some "self-interest" users focus on meeting their own needs from the OHC (K. B. Wright, 2015). For example, seeking social support is one of the key motivations for users to start using OHCs (E. Kim et al., 2012; Rodgers & Chen, 2005).

Due to users' different motivations, users may behave in different ways and play "community-interest" and "self-interest" roles. It is also worth noting that a user's altruism in a community could switch between “cooperating” and “self-centered” (Fehr & Fischbacher, 2003). In other words, the interest of “community” and “self” can shift over time. Thus, it would be interesting to explore if such changes do occur regarding users’ roles in an OHC, and if so, what factors drive such changes, because an OHC would need more “community-interest” users to sustain the community.

OHCs are essentially social networks among users, because they interact with each other online to seek, receive and provide social support. Social network researchers have observed the spread of innovations among individuals connected in a social network (Rogers, 1962). This phenomenon is usually referred to as “diffusion”. Projecting it onto the social networks among OHC users, would a user’s future role be influenced by roles of her or his network neighbors? More specifically, if a user is surrounded by many “community-interest” users in the online social network, would that increase the chance of the user becoming a “community-interest” user? Conversely, does having more social network neighbors who are “self-interest” users lead to a user more likely to become a “self-interest” user? I will try to address these questions by examining users’ roles based on their social support activities and analyze the role dynamics in inter-user social network. The framework is shown in Figure 3 .

Research Goal of this Chapter: Explore whether users play different roles with regards to social support activities. Detect users’ role dynamics in an OHC and capture influential factors for such dynamics.

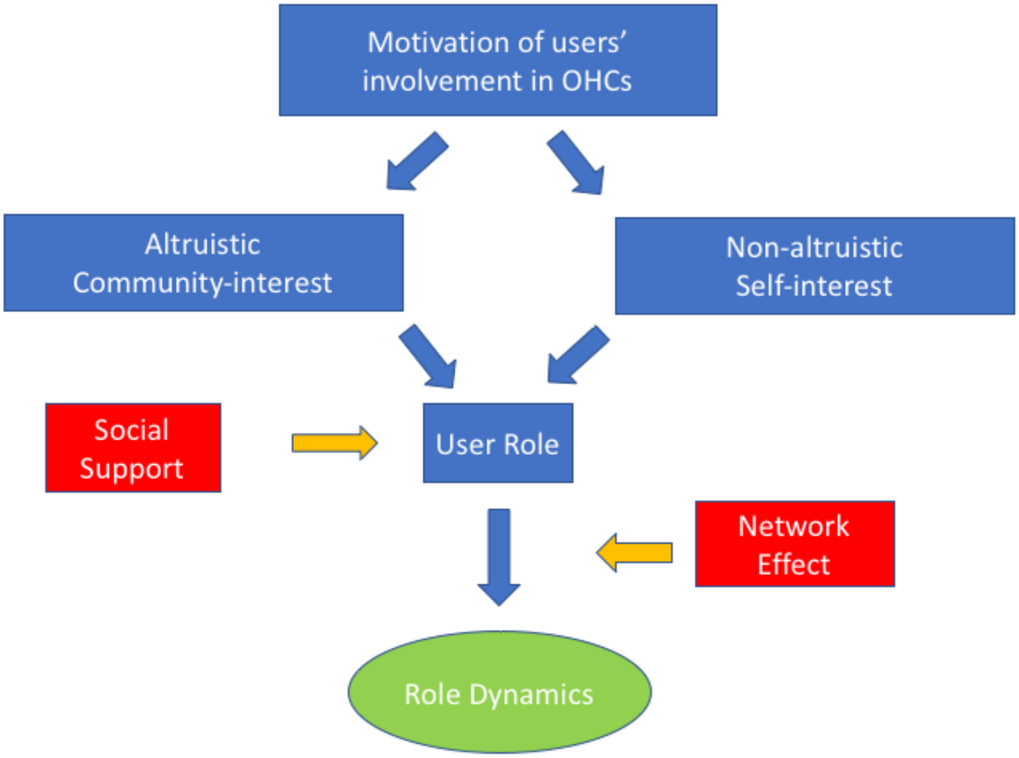

Motivation of users' involvement in OHCs
Altruistic Community-interest
Non-altruistic Self-interest
Social Support
User Role
Network Effect
Role Dynamics

Figure 3 Framework of user role dynamics

The Identification of User Roles

Prior research on users' roles in OHCs was mainly based on aggregated data of users' behaviors, such as the number of log-ins, the number of posts, and the number of active days (Jones et al., 2011), or social network centrality metrics, such as in and out degrees (Cobb et al., 2010; K. Zhao et al., 2016). However, these metrics do not reflect what a user has sought from or contributed to an OHC. Thus, roles identified based on these metrics can only provide a coarse-grained view of a user's online activities. A more fine-grained view of user roles would require analyzing contents of users' posts. Previous studies that examined such content for user role identifications either depended on manual content analysis of posts, which would not scale beyond a few thousand posts, or stayed at the lexicon level by counting the appearance of words or phrases (Füller et al., 2014; Sudau et al., 2014). Fortunately, in Chapter Three, I have already been able to overcome the problem and explore social support in each post. The outcome can be used in identifying user roles in a fine-grained view in this Chapter.

After estimating the nature of social support, I built a profile for each user by aggregating his or her posts by their social support categories. Each user's social support involvement was represented with a 1×5 a vector. Each element in the vector is the percentage of the user's posts in a social support category. For example, user Mary has published 10 posts, with 3 companionship posts, 4 posts providing emotional support, 2 posts providing informational support, 1 post seeking emotional support, and no posts seeking informational support. Then she would have a vector of <0.3, 0.4, 0.2, 0.1, 0>.

With social support distribution vectors of 47,581 users, I applied the classic K-means clustering algorithm to divide users into K groups, so that the users with similar social support distributions would belong to the same cluster. To find the best grouping of users, I tested various K values (from 2 to 20) and clustering results with Davies-Bouldin Index (DBI) (Davies & Bouldin, 1979). DBI is defined as Equation 1, where $D_{intra}(C_i)$ is the average distance from all the other users in cluster C_i to the centroid of C_i, and $D_{inter}(C_i, C_j)$is the distance between centroids of $\mathrm{C}_i$and C_j. Euclidean distance was used for this study. In general, DBI prefers smaller groups, for the value of intra-cluster distance is lower in the smaller group, and penalizes short inter-cluster distances. Therefore, the solution with the lowest DBI provides relative balance of small clusters and long distances between every pair of clusters.

Equation 1:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{j:i \neq j}\left\{\frac{D_{intra}(C_i) + D_{intra}(C_j)}{D_{inter}(C_i, C_j)}\right\}$$

I summarized the DBIs for different K values in Table 9. K=7 yielded the lowest DBI value and hence the best clustering results. Centroids for each of the 7 clusters are shown in Table 10.

K	DBI	K	DBI
2	1.485806117	12	0.932705779
3	1.183743056	13	0.914857805
4	1.147831469	14	1.148624229
5	1.002816698	15	0.94766141
6	0.962159462	16	0.915504995
7	0.89111499	17	0.895295641
8	0.977535018	18	0.907029696
9	0.960697173	19	0.935044276
10	0.940555275	20	1.001204328
11	0.904557568		

Table 9 The DBIs for the K-means clustering with various K values

Social Support	All users	Cluster 0 (IP)	Cluster 1 (CB)	Cluster 2 (AC)	Cluster 3 (IS)	Cluster 4 (EP)	Cluster 5 (IE)	Cluster 6 (ES)
COM	0.1126	0.0042	**0.6492**	0.1271	0.0154	0.0504	0.0408	0.0404
PES	0.1178	0.0074	0.0833	0.1511	0.0053	**0.6120**	0.0315	0.0351
PIS	0.4422	**0.9655**	0.1277	0.4762	0.0152	0.2394	0.4369	0.0325
SES	0.0743	0.0067	0.0349	0.1245	0.0107	0.0481	0.0494	**0.5868**
SIS	0.2531	0.0162	0.1049	0.1211	**0.9534**	0.0501	0.4414	0.3052
# of users	47,581	6,647	3,923	15,336	3,502	3,994	13,225	954
% of users		14%	8%	32%	7%	8%	28%	2%

Table 10 Centroids of user clusters – role of users

Intentionally or not, OHC users do have different patterns in social support activities and play different roles in the community. Some users' posts focused on one major category of social support. For example, Information Providers (IP) published an average of 96.55% of their social support posts to provide informational support. Similarly, Companionship Builders (CB) had 64.92% of their posts in companionship, and Emotional Support Providers (EP) were more interested in providing emotional support. The two smallest clusters are for Informational Support Seekers (IS), and Emotional Support Seekers (ES). Meanwhile, the largest cluster of the seven represents All-around Contributors (AC) with relatively balanced profiles in each social support category. There is also a group of Information Enthusiasts (IE), who focused mainly on informational support, both seeking and providing.

To investigate how users in different groups engaged in the OHC, the complementary cumulative distributions are provided in Figure 4 and Figure 5. Engagement levels were measured by two metrics: productivity (i.e., a user's total number of posts) and time span of activities (i.e., the number of days between a user's first and last post). Figure 4 compares the distributions of productivity for users in the 7 clusters. The curves suggest that CB in cluster 1, albeit a small group of users, and AC in cluster 2 are the most productive members. By contrast, those who mainly seek support (informational or emotional) in clusters 3 and 6 published fewer posts than others. Figure 5 points to similar conclusions: those in clusters 1 and 2 stayed with the community for the longest time, while support seekers in clusters 3 and 6 have relatively short time span of activities. Overall, those who are more actively involved in companionship tend to get engaged in the community, while those who only seek support are more likely to "churn". Also, EPs in cluster 4 are more engaged than IPs in cluster 0.

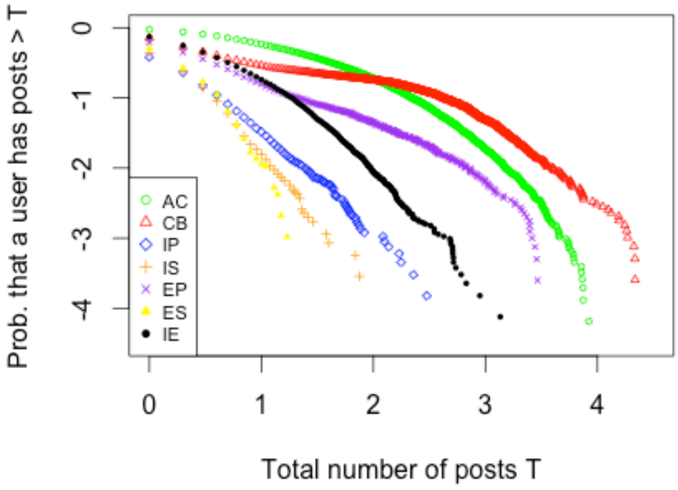


Figure 4 Complementary cumulative distributions of engagement metrics for the users in different clusters (productivity)

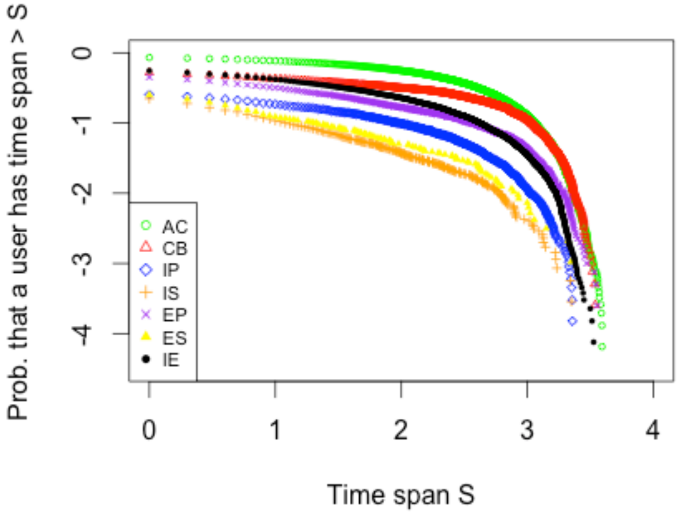
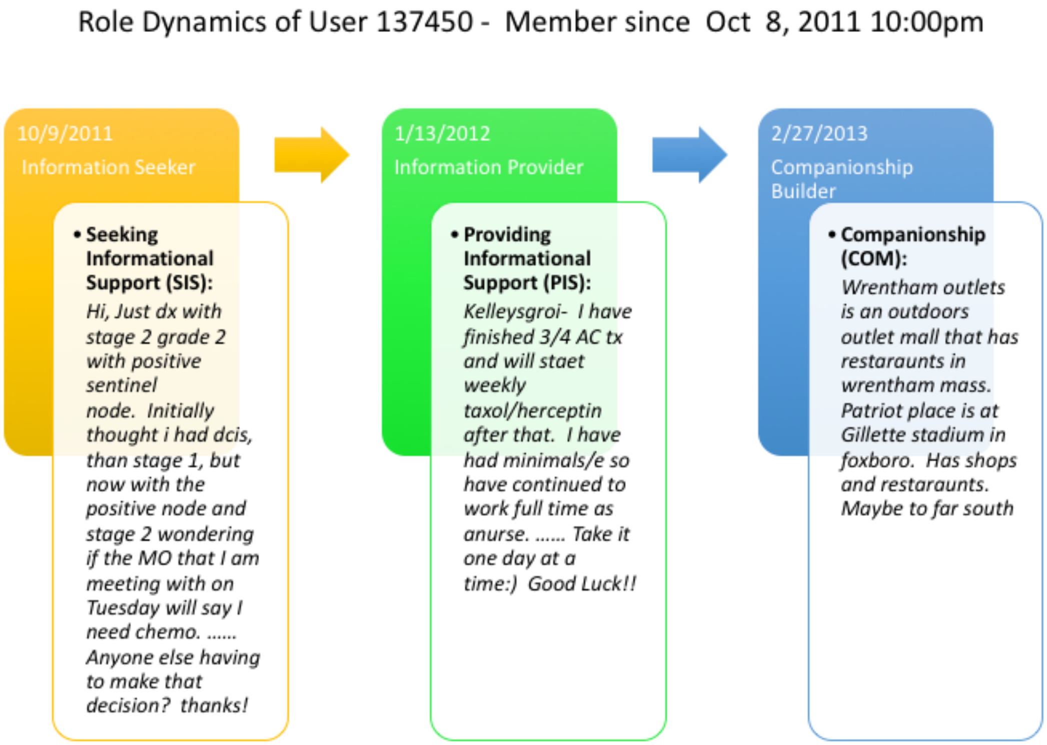


Figure 5 Complementary cumulative distributions of engagement metrics for the users in different clusters (time span of activities)

User Role Transitions

It’s important to note that users’ roles listed in Table 10 only represent a static snapshot of all the users based on users’ aggregated social support activities. In other words, each user was assigned to a role based on all his or her social support activities from his or her first post to the last, no matter how long the user was active in the OHC. As some qualitative literature suggested, a user’s role may shift over time in an OHC (Pfeil & Zaphiris, 2007). For example, based on manual content analysis, Loane and D’Alessandro (Loane & D’Alessandro, 2013) argued that many OHC users start as information seekers and some of them could switch to support providers. Therefore, one user may play multiple roles during his or her lifespan in an OHC, which means roles of a user can evolve over time. Figure 6 provides one example of such evolution from the dataset. Specifically, a user joined in the community in Oct 2011 as an information seeker, mainly sought informational support for self interests. After several months of participation, the user shifted to a community-interest role: she started to provide informational support to others as an information provider. Motivated by more community interests, this user finally shifted to a companionship builder, posting off-topic content related to daily life.

There is a lack of systematic and large-scale examination of how OHC users’ roles evolve over time. In addition, the factors that contribute to such role evolutions are not well understood. To capture the dynamics of user roles over time, I need to analyze user roles at a temporally more fine-grained level. Therefore, I examined the dynamics of user roles on monthly basis and calculated monthly social support profiles for each user. For example, if a user were active for only a month in this OHC, the user would have only one social support activity profile that represented his or her activities in the five social support categories during that month. Another user, who continuously contributed to this community for two years, would have 24 such profiles, each one for a month.



Role Dynamics of User 137450 - Member since Oct 8, 2011 10:00pm

10/9/2011
Information Seeker

- **Seeking Informational Support (SIS):** *Hi, Just dx with stage 2 grade 2 with positive sentinel node. Initially thought i had dcis, than stage 1, but now with the positive node and stage 2 wondering if the MO that I am meeting with on Tuesday will say I need chemo. ...... Anyone else having to make that decision? thanks!*

1/13/2012
Information Provider

- **Providing Informational Support (PIS):** *Kelleysgroi- I have finished 3/4 AC tx and will staet weekly taxol/herceptin after that. I have had minimals/e so have continued to work full time as anurse. ...... Take it one day at a time:) Good Luck!!*

2/27/2013
Companionship Builder

- **Companionship (COM):** *Wrentham outlets is an outdoors outlet mall that has restaraunts in wrentham mass. Patriot place is at Gillette stadium in foxboro. Has shops and restauraunts. Maybe to far south*

Figure 6 An example of user-role's dynamics in the OHC

To identify which role among the seven roles in Table 10 a user played during a specific month, I adopted a simple k-Nearest Neighbor (kNN) classification scheme. Given a user's monthly social support profile during a month in which the user was active in posting, I compared the monthly profile with the 47,581 aggregated social support activity profiles of all users (each profile is for one user, and has been assigned to a role in Table 10), and assigned his or her the role in that month based on a majority vote among the k nearest aggregated profiles. To check the robustness of the monthly role assignment, I varied the value of k from 1 to 10 and the role assignments were very consistent--the probability that an assignment changes with a different k value is less than 1%. Thus, I selected role assignments with k = 10 in subsequent analysis.

Among 384,423 monthly social support profiles, users' monthly roles distributions are shown in Table 11. Note that, in addition to the 7 roles in Table 10, some monthly social support profiles were labeled as "Lurkers" (LU), which means all elements of the user's social support profile were 0s. A user can become a lurker during a month mainly because the user did not post anything in the OHC during that month. It is also possible, but less likely, that the user published posts during that month, but these posts are not related to social support.

User Role	Number of monthly social support activity profiles in the role
Information Seeker (IS)	9,632
Information Provider (IP)	26,557
Emotional Support Seeker (ES)	3,027
Emotional Support Provider (EP)	19,123
Community Builder (CB)	36,626
All-around Contributors (AC)	63,282
Information Enthusiasts (IE)	31,355
Lurkers (LU)	194,821
Total	384,423

Table 11 The distribution of users' roles by their numbers of monthly social support activities

Based on users' roles in each month of their activities in the OHC, the temporal trajectory of a user's roles can be extracted, which shows that users' roles do evolve over time. Figure 7 draws a role transition network of the OHC to depict how users shift from one role to another. This weighted and directed network has 10 nodes. Seven of them correspond to the seven user roles I listed in Table 10. There is one node for the status of LU. I also included two more nodes, Registration (REG) and Churn (CHU), to represent the starting and ending points of one's posting activities in this OHC. Basically, all users started from Registration, but might ended at any other node in this graph as time goes by. I assumed that a user had left this OHC if the user had no posts during the last 12 weeks in the dataset. A directed link from node A to node B means that at least one user who played role A (or in status A) in one month switched to role (or status) B in the subsequent month. The weight of each link is computed as the probability of transiting from one node to another, with the sum of all is proportional to such weights. For example, users who acted as IS in this OHC have high probabilities to leave the community in the next month. By contrast, an IS was unlikely to become an IE later.

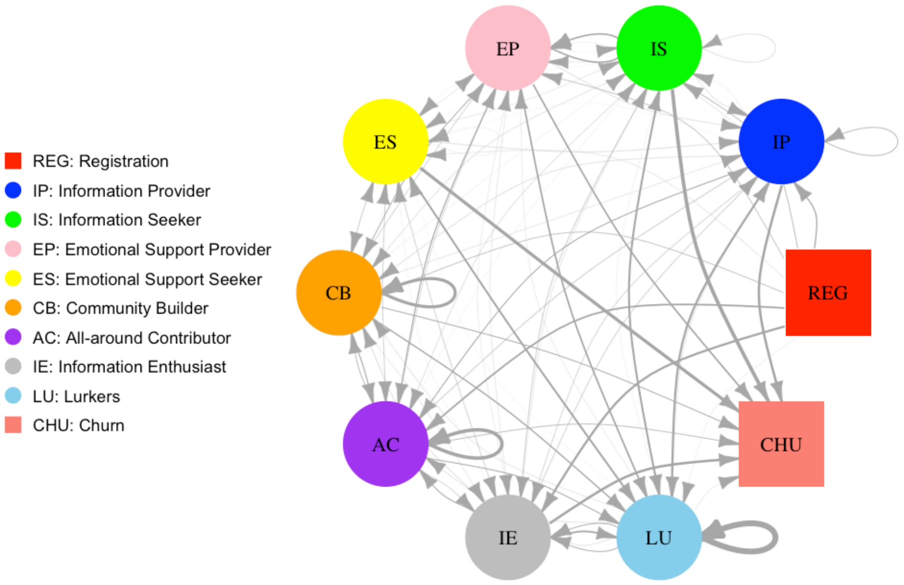


Figure 7 The transition of users’ roles in the OHC

Diffusion of User Role in the OHC

Diffusion is a common phenomenon in social networks-it describes the spread of certain “contagion” from one individual to another via social ties. Prior studies of network diffusion have found evidence for the spread of epidemics, innovations, opinions, information, and behaviors in social networks (Leskovec, Backstrom, & Kleinberg, 2009; Muchnik, Aral, & Taylor, 2013; Newman, 2002; Rogers, 1962; Valente, 1996; Watts & Dodds, 2007). The basic idea behind network diffusion is the effect from an ego's social network neighborhood on the ego. In other words, the more an ego's neighbors become "infected" or adopt a "contagion", the more likely the ego is to adopt the "contagion". For example, a person's obesity status is positively related to weight gain in his or her friends, siblings, spouse and neighbors (Christakis & Fowler, 2007). Political self-expression and voting behavior of an individual can be influenced by his or her friends or even friends of friends (Bond et al., 2012).

There could be various reasons for network diffusion. From the perspective of economics, network externality may contribute to the diffusion of some products or innovations (Shapiro & Varian, 1999), as the benefits of adopting such products or innovations increase as more adoptions occur in a potential adopter's social or business network (e.g., the adoption of social networking or messaging apps). From the perspective of social behaviors, diffusion can be attributed to homophily (a.k.a., birds of a feather) (Centola, 2011; McPherson, Smith-Lovin, & Cook, 2001), social influence (Aral & Walker, 2012; Muchnik et al., 2013), or exposure to the same external stimuli (Van den Bulte & Lilien, 2001).

After showing that user roles with regards to social support activities do evolve over time, I investigated if user role transitions are influenced by their neighbors (connected others) in the social network. The analysis was rooted in users' motivations to participate in an OHC. As mentioned earlier, OHC users' participation can be driven by community-interests and self-interests. Users' roles based on social support activities can map to these motivations. Providing social support is motivated by users' interest in helping the community, so support providers who are active in contributing support to the community (IP, EP, AC and CB) are considered as playing community-interest roles. By contrast, because seeking support is usually driven by one's own needs or interests, I assumed support seekers, namely IS and ES, and lurkers (LU) who are inactive in social support activities, as self-interest roles. It is worth noting that users who play the role "Information Enthusiasts" actively seek and provide informational support at the same time. Thus, IE is considered as a hybrid role that combines both community and self-interests.

From a social network perspective, I wanted to check if a user's transitions between community-interest roles and self-interest roles are correlated with the roles adopted by their current social network neighbors. Specifically, if a user's role is influenced by his or her neighbors, then the more interactions a user had with community-interest peers in his or her social network neighbors at month t , the more likely the user will play a community-interest role at month $t+1$, and vice versa for those who play self-interest roles. The logistic regression with fixed effects was used to examine the diffusion of user roles in social networks.

The model studied the transition among 3 types of user roles based on their motivations to participate in the OHC: community-interest roles, self-interest roles and hybrid roles. I created a social network among users of this OHC based on co-participation in threaded discussions. Each node in the network represents a user, and there is an undirected but weighted tie from user i to user j if they both contributed posts (initial posts or comments) to the same threaded discussion. The weight of the tie between i and j is the number of times they co-participated in threads.

For each user, at a particular month t, the dependent variable $Role_{i,t+1}$ is a nominal variable indicating whether user i plays one of the 3 roles at time $t+1$. Independent variables in the model try to capture the influence from a user's social network neighbors. User i's role at month t is represented by a probability distribution over the three roles $P_i(t) = (r_{i,1}, r_{i,2}, r_{i,3})$, where $r_{i,m}$ is the probability that user i's monthly social support profile for month t is classified by the kNN classifier $k = 10$ to be role m. For example, $P_i(t) = (0.9,0.1,0)$ means that among user i's 10 nearest neighbors at month t, 9 have community-interest roles, 1 has self-interest role, and none takes the hybrid role. Then I measured the influence from a user i's network neighbors on i at month t. $I_i(t)$as the weighted sum of such probabilities for all of i's social network neighbors: $I_i(t) = \sum_{j \in N_i} \omega_{i,j} * P_j(t)$, where $\mathrm{N_i}$ represents the set of social network neighbors for user i. By doing so, those who have stronger ties with a user would be able to exert more influence on the user.

In terms of control variables, I controlled a user's individual characteristics, including his or her current role at month t (i.e., probabilities of his or her playing each of 3 user roles), the user's level of online activities (measured by the number of posts from the user during month t), and the user's length of tenure in the OHC. In addition, the calendar year and month ($date$ in Table 12) is also included, which controls the overall trend or major events (i.e., external stimuli) in the OHC. For example, COM posts in the form of holiday greetings and plans may be more popular in November and December. Also, more informational support may appear after a new treatment becomes available. Table 12 summarizes variables of the model.

Variables	**Variable Name**	**Notation in Equation 2**	**Description**	**Data Type**
Dependent	**Future_Role**	$Role_{i,t+1}$	Whether user i plays one of the 3 roles at time $\mathrm{t}+1$	Nominal
Independent	**Inf_comm**	$I_i(t)$	The weighted influence from a user's network neighbors at month t to be a community-interest user at $\mathrm{t}+1$	Numerical
	Inf_self	$I_i(t)$	The weighted influence from a user's network neighbors at month t to be a self-interest user at $\mathrm{t}+1$	Numerical
	Inf_hybrid	$I_i(t)$	The weighted influence from a user's network neighbors at month t to be a hybrid user at $\mathrm{t}+1$	Numerical
Control	**current_comm_pr**	$currentRole_{i,t}$	A user's probability of having a community-interest role (AC+CB+IP+EP) at t	Numerical
	current_self_pr	$currentRole_{i,t}$	A user's probability of having a self-interest role (IS+ES) at t	Numerical
	current_hybrid_pr	$currentRole_{i,t}$	A user's probability of being a hybrid user (IE) at t	Numerical
	num_post	$controls_{i,t}$	The total number of posts a user published during month t	Numerical
	tenure	$controls_{i,t}$	The number of months elapsed since the user enrolled in the OHC	Numerical
	date	$controls_{i,t}$	Month t	Numerical

Table 12 Description of variables in the logistic regression based on 3 user roles

It should be noted that I included all user's monthly activities except the records of a user labeled as LU at month t. Because LU indicated this user had no social support activities at month t, which means this user would have no neighbor in the social network built based on co-participation in social support activities for that month. In this way, I was able to construct a panel data set with 142,053 observations for a total of 18,267 users, where the longest panel contains 107 time periods (months).

Equation 2 shows the model I used. Coefficients γ estimate the relationship between users' current and future roles; β estimates the influence from a user's neighbors to his or her future roles. Fixed effects of individual users' time-invariant attributes, e.g. gender, education level and personality, are controlled by α_i for a particular user i. I also included year and month dummy variables as controls.

Equation 2:

$$Role_{i,t+1} = \alpha_0 + \alpha_1 + \beta I_i(t) + \gamma currentRole_{i,t} + \delta controls_{i,t} + \epsilon_{i,t+1}$$

Results of the regression model are summarized in Table 13. Columns in the table represent destinations of user role transitions at time $t + 1$, i.e. $Role_{i,t+1}$. Coefficients in each row indicates the relationship between the independent or control variable and the future role of a user. Basically, the results verified that users' roles are indeed influenced by roles of their social network neighbors. Specifically, the influence one received from his or her neighborhood is associated with his or her subsequent roles in the OHC. For example, the positive and significant coefficient of $\beta = 0.12$ $(p < 0.001)$ between *Inf_comm* and *next_community_role* suggests that the more influence a user receives from his or her community-interest neighbors, the more likely the user becomes a community-interest user in the next month. In other words, if an ego interacted more with users who have community-interest roles, the likelihood of this user to adopt community-interest roles would increase. Similar results can be found for self-interest roles of ($\beta =$

$0.05,\ p < 0.001$) and hybrid roles ($\beta = 0.19,\ p < 0.001$). In sum, users' role dynamics are influenced by the roles of their social network neighbors: having a social network neighborhood full of altruistic users increases the chance that a user become altruistic, while a user surrounded by self-interest neighbors is more likely to become self-interested.

	next_community_role	**next_self_role**	**next_hybrid_role**
Inf_comm	0.12^{***}	-0.09^{***}	-0.19^{***}
Inf_self	0.00	0.05^{***}	-0.00
Inf_hybrid	0.06^{***}	-0.08^{***}	0.19^{***}
current_comm_pr	-0.02	-0.07^{**}	0.25^{***}
current_self_pr	-0.12^{**}	-0.22^{***}	0.31^{***}
num_post	0.30^{***}	-5.54^{***}	-0.46^{***}
tenure	0.01	-0.03	0.40^{***}
date	-0.03^{***}	0.17^{***}	-0.14^{*}
N	117559	115106	67118
BIC	93211.12	76157.66	37353.06

Table 13 Logistic regression results for the diffusion of 3 user roles
* p<0.05, ** p<0.01, *** p<0.001
current_hybrid_pr=1-current_comm_pr-current_self_pr
Individual user, year and month fixed effects are controlled

Among control variables, meta features and the current role of a user can be predictive of users' subsequent roles. Intuitively, active users, who contributed more posts to the OHC, were likely to have community-interest roles ($\delta = 0.30,\ p < 0.001$), instead of self-interest roles ($\delta = -5.54,\ p < 0.001$) and hybrid roles ($\delta = -0.46, p < 0.001$). However, the effect of users' current roles is a little bit surprising -- users who currently have community-interest roles do not tend to keep their community-interest roles ($\gamma = -0.02,\ p > 0.05$), neither do users who have self-interest roles ($\gamma = -0.22$, $p < 0.001$). To better understand this, I conducted further analysis.

Further Analysis

The previous model analyzed three types of user roles aggregated based on users' motivations to participate in the OHC. Community-interest roles actually combined 4 types of user roles, namely AC, CB, IP and EP. Self-interest roles include 3 types-ES, IS, and LU. To better understand user role dynamics, I studied the transition among 7 user roles I identified earlier in Table 10: AC, CB, IP, EP, IS, ES, and IE. To be consistent with the previous model, I also included LU as another destination role only, because those who are currently lurkers would have no connected others in the social network.

I still adopted logistic regression, and used the undirected but weighted co-participation network I created for the previous model. The difference from the previous model is to have 8 user roles, instead of 3. Specifically, for each user, at a particular month t, the dependent variable $\mathrm{RoleBinary}_{\mathrm{i,t+1}}$ is a binary variable indicating whether user i played one of the 8 roles at time $t + 1$. Independent variables in the model are still set to capture the influence from a user's social network neighbors. Different from the previous model, user i's role at month t is represented by a probability distribution over 7 roles, $p_i(t) = (r_{i,1},\ r_{i,2},\ r_{i,3},\ r_{i,4},\ r_{i,5},\ r_{i,6},\ r_{i,7})$, where $r_{i,m}$ is the probability that user i's monthly social support profile for month t is classified by the kNN classifier $k = 10$ to be role m. For example, $p_i(t) = (0.2, 0.2, 0.1, 0.1, 0.3, 0.1, 0)$ means that among user i's 10 nearest neighbors for month t, there are 2 AC, 2 CB, 1 IP, 1EP, 3 IS, 1 ES and no IE. Similar to the previous model, I measured the influence from user i's network neighbors on i at month t ($I_i(t)$) as the weighted sum of of such probabilities for all of i's social network neighbors: $I_i(t) = \sum_{j \in N_i} \omega_{i,j} * p_j(t)$, where N_i represents the set of social network neighbors for user i.

Control variables still include a user's current role $currentRoleVec_{i,t}$ at month t. Different from the previous model, this model used vectors with binary elements rather than probabilities to represent a user' s current role in this model, so that I could clearly capture the transition among roles. For example, a user with a current role vector (0, 0, 0, 0, 0, 0, 1) means this user is an IE at month t, and a user with a current role vector (1, 0, 0, 0, 0, 0, 0) means this user is an AC at month t. Other control variables are the same as the previous model, including a user' s level of online activities (measured by the number of posts from the user during month t), a user's length of tenure in the OHC, and date. All the variables are listed in Table 14.

Equation 3 summarizes the new logistic regression model. β estimates the influence from a user' s neighbors to his or her future roles. Coefficient vector γ captures the transition patterns from each of the roles to a future user role in consideration. Similar to the previous model, I also controlled time-invariant factors (α_0) and unobserved time-invariant individual effect (α_i).

Equation 3 :

$$RoleBinary_{i,t+1} = \alpha_0 + \alpha_1 + \beta I_i(t) + \gamma currentRoleVec_{i,t} + \delta controls_{i,t} + \epsilon_{i,t+1}$$

Table 15 shows results of the new model. It turns out, when analyze at a more fine-grained level, not all the role transitions are influenced by social network neighbors. On one hand, results in the top panel provide evidence for social network influence on the transitions to 4 roles: AC ($\beta = 0.51,\ p < 0.001$), CB ($\beta = 0.57,\ p < 0.001$), EP ($\beta = 0.30,\ p < 0.001$) and IE ($\beta = 0.11,\ p < 0.001$). For example, the more influence a user received from his or her social network neighbors who are All-around Contributors, the more likely the user will become an All-around Contributor. On the other hand, the other three roles-IP, ES, and IS-do not feature such patterns. For example, more interactions with support seekers (IS and ES) does not necessarily mean a user will seek more support in the future.

Coefficients for control variables also revealed some differences compared to the previous model based on aggregated user roles. For instance, although active users with many posts are still unlikely to have self-interest roles (IS, ES, and LU), their chances of taking specific community-interest roles vary. Active users tend to become AC $(\delta = 0.17,\ p < 0.001)$ and CB $(\delta = 0.31,\ p < 0.001)$, but not IP $(\delta = -1.81,\ p < 0.001)$ and EP $(\delta = -0.24,\ p < 0.001)$. In other words, although AC, CB, IP and EP all belong to the same group of community-interest roles because users in these roles are motivated by altruism, these roles' transition dynamics can be different from each other. In fact, transition probabilities among all the roles in Figure 7 confirmed that, even though AC and CB have high probabilities to stay at their current roles (47.9% and 42.2% respectively), IP and EP have probabilities of 28.8% and 22.6% respectively to switch to LU, one of the self-interest roles. Thus, I conjectured that the different dynamics of the four community-interest roles might have contributed to the instability of community-interest roles in general.

Summary

In this Chapter, I identified users' roles based on their social support behaviors, which is aggregated based on their motivations to participate. I found that user do play different roles with regards to social support, and their roles evolve over time. Through regression analysis of role evolution, I illustrated that users' roles are influenced by roles of their social network neighbors.

Like most online communities, OHCs would like to have more users who are motivated by the community's interests and actively contribute to the community. This is because having more community-interest users in an OHC means more support and benefits for those who are dealing with health problems. Nevertheless, it is difficult to control why a user starts his or her participation in an OHC. What I have found suggests that although OHC users' original motivation to get involved may vary, their behaviors and roles in social support activities can change over time and be influenced by their social network neighbors--more interactions with community-interest users can prompt a user to become a community-interest user. By contrast, users surrounded by self-interest users would have high probabilities of becoming a self-interest user in the community.

The findings make it possible for an OHC to intervene and facilitate a user’s evolution to community-interest users by surrounding his or her with other altruistic users. For example, the OHC can incorporate a recommender system or search engine that prioritize community-interest users and their contributions, so that those who are seeking support using the systems would have a higher chance to interact with them and take community-interest roles.

Variables	Variable Name	Notation in Equation 3	Description	Data Type
Dependent	**Future_Role**	$RoleBinary_{i,t+1}$	Whether user i plays one of the 8 roles at time $t+1$	Binary
Independent	**InfAC**	$I_i(t)$	The weighted influence from a user's network neighbors at month t to be a AC at $t+1$	Numerical
	InfCB	$I_i(t)$	The weighted influence from a user's network neighbors at month t to be a CB at $t+1$	Numerical
	InfEP	$I_i(t)$	The weighted influence from a user's network neighbors at month t to be a EP at $t+1$	Numerical
	InfES	$I_i(t)$	The weighted influence from a user's network neighbors at month t to be a ES at $t+1$	Numerical
	InfIE	$I_i(t)$	The weighted influence from a user's network neighbors at month t to be a IE at $t+1$	Numerical
	InfIP	$I_i(t)$	The weighted influence from a user's network neighbors at month t to be a IP at $t+1$	Numerical
	InfIS	$I_i(t)$	The weighted influence from a user's network neighbors at month t to be a IS at $t+1$	Numerical
Control	**currentAC**	$\mathrm{currentRoleVec}_{\mathrm{i,t}}$	Whether the user's role is *All-around Contributor*	Binary
	currentCB	$\mathrm{currentRoleVec}_{\mathrm{i,t}}$	Whether the user's role is *Companionship Builder*	Binary
	currentEP	$\mathrm{currentRoleVec}_{\mathrm{i,t}}$	Whether the user's role is *Emotional Support Provider*	Binary
	currentES	$\mathrm{currentRoleVec}_{\mathrm{i,t}}$	Whether the user's role is *Emotional Support Seeker*	Binary
	currentIE	$\mathrm{currentRoleVec}_{\mathrm{i,t}}$	Whether the user's role is *Information Enthusiast*	Binary
	currentIP	$\mathrm{currentRoleVec}_{\mathrm{i,t}}$	Whether the user's role is *Information Provider*	Binary
	currentIS	$\mathrm{currentRoleVec}_{\mathrm{i,t}}$	Whether the user's role is *Information Seeker*	Binary

Table 14 Description of variables for the logistic regression model based on 8 user roles

Variables	Variable Name	Notation in Equation 3	Description	Data Type
	num_post	$controls_{i,t}$	The total number of posts a user published during month t	Numerical
	tenure	$controls_{i,t}$	The number of months elapsed since the user enrolled in the OHC	Numerical
	date	$controls_{i,t}$	Month t	Numerical

Table 14 – continued

	AC	**CB**	**IP**	**EP**	**IS**	**ES**	**IE**	**LU**
InfAC	0.51***	-0.24***	0.12**	0.09*	0.27***	0.07	0.07	-0.19***
InfCB	-0.34***	0.57***	-0.02	-0.04	-0.16*	0.05	-0.27***	0.09***
InfIP	-0.02	-0.08***	0.04	-0.06	0.01	-0.01	0.08*	0.04*
InfEP	0.02	0.01	-0.12***	0.30***	-0.08	0.06	-0.02	-0.04*
InfIS	0.01	-0.06***	0.02	-0.02	-0.01	-0.05	0.03	0.04**
InfES	-0.02	0.02	-0.05**	0.00	-0.01	-0.11*	-0.03*	0.04**
InfIE	0.21***	-0.07**	0.13***	-0.12***	-0.14*	0.07	0.11***	-0.04*
currentAC		-0.40***	0.60***	0.06	0.84***	1.51***	0.29***	-0.31***
currentCB	-0.53***		0.34***	-0.05	0.86***	1.54***	0.11*	-0.11*
currentIP	-0.34***	-0.40***		0.07	0.68***	1.28***	0.30***	-0.06
currentEP	-0.32***	-0.25***	0.43***		0.68***	1.45***	0.23***	-0.16***
currentIS	-0.55***	-0.29***	0.40***	-0.15		1.40***	0.30***	
currentES	-0.44***	-0.17	0.54***	-0.20	0.99***		0.35***	-0.07
currentIE	-0.08**	-0.27***	0.48***	0.02	1.13***	1.75***		-0.19***
num_post	0.17***	0.31***	-1.81***	-0.24***	1.57***	-1.42***	-0.46***	-5.84***
tenure	0.36***	-0.26***	-0.08	-0.24***	-0.17	-0.26	0.37***	0.02
date	-0.07	0.05	-0.10	-0.01	-0.11	0.32	-0.15*	0.17***
N	114241	90516	74650	66812	22738	14824	67118	113956
BIC	98683.43	68182.37	40881.64	39330.06	9168.78	4619.51	37355.57	72743.14

Table 15 Logistic regression results for the diffusion of 7 user roles

* $p<0.05$, ** $p<0.01$, *** $p<0.001$

Individual user, year and month fixed effects are controlled.

CHAPTER FIVE
USERS' PARTICIPATION AND CHURN PREDICTION

According to the previous Chapter, user role dynamics in the OHC reveal that the motivation of users indeed change over time. Users' social support behaviors in an OHC can be impacted by neighbors in their social network. Due to the variety of their motivations in the community, users might engage with different active time span. For example, some of them involved for several months to provide social support, while some others only show up for several days to seek information and then slip away. This Chapter focuses on the analysis of the users' continuous participation and predicts users' churning behaviors in this OHC.

Value of Participation Analysis and Churn Prediction

Like other online communities, OHCs would like to encourage users' participation and prevent users' churn behaviors (i.e., leaving a community), because most online communities, whether they are for profit or not, want to be successful. An online community is successful when its members participate actively and develop lasting relationships (Kraut et al., 2012; Young, 2013). By contrast, poor participation and transient membership can lead to the termination or failure of an online community (Iriberri & Leroy, 2009).

What makes user retention in OHCs different from that in many other services, such as banking, gaming, and telecommunication, is that users' churn is not only harmful to the service operator, but may also bring negative impact to individual OHC members. This is because a user's participation in an OHC can be beneficial and therapeutic (Bouma et al., 2015; Campbell et al., 2004; Eysenbach et al., 2004; Hoey, Ieropoli, White, & Jefford, 2008; Idriss et al., 2009; S. Zhang et al., 2014). Receiving such support can empower (Burrows, Nettleton, Pleace, Loader, & Muncer, 2000) and help patients adjust to the stress of living with and fighting against their diseases (Dunkel-Schetter, 1984; Qiu et al., 2011). The support they received online can also improve their offline life and health management (Maloney-Krichmar & Preece, 2005; Yan & Tan, 2014). In addition to receiving support from others, staying in an OHC and providing support to others can be beneficial to providers as well (Dunkel-Schetter, 1984). There is actually a positive relationship between posting frequency and psychosocial well-being (Rodgers & Chen, 2005).

In other words, a user's continued participation in an OHC can help herself or himself as well as others. Admittedly, for some individuals who have received satisfactory support from an OHC or recovered from the health problem, leaving the OHC may not be a bad thing. However, even though user-generated information about a disease will still be available online to new OHC members, most of the psychosocial benefits for individual users cannot be achieved if the exodus of experienced users in the OHC keeps happening, leaving new members stranded (Rodgers & Chen, 2005). In fact, providing assistance for new members from experienced members and reminding members to participate continuously are also key factors for the success of online communities (Iriberri & Leroy, 2009).

As social support is a pillar of OHCs, a natural question to ask would be: when it comes to users' participation, are a user's online activities in various types of social support related to his or her participation in an OHC? If so, can I predict whether a user will churn from an OHC based on these social support activities? Despite the large amount of research on social support in OHCs, few studies have answered this question systematically by examining users' seeking, receiving, and provision of various types of social support from large-scale datasets. An explanatory model (Wang et al., 2012) suggested that receiving more emotional support is associated with users' longer stay in an OHC. However, the types of social support investigated were limited and only the receiving of support was considered, while I mentioned earlier that providing social support is also important and beneficial.

Theoretical studies regarding online community participation mainly focus on motivation, managerial principles, and social attachment. For example, besides altruism, researchers also identified other motivations for users' participation in online communities, such as anticipated reciprocity, increased recognition, trust, and sense of efficacy (Kollock, 1999; Leimeister, Ebner, & Krcmar, 2005). Principles to manage and run sustainable online communities include a clear vision, community leadership, offline interactions, moderations, and useful content (J. Chen, Xu, & Whinston, 2011; A. J. Kim, 2000; Williams & Cothrel, 2000).

Social attachment theories proposed two reasons people are engaged in their communities. First, they are attached to the community as a whole (i.e., social identity), or second, they are attached to individuals in the community (i.e., social bond) (Back, 1951). Participation in identity-based communities are driven by members' common social categories, tasks, or purposes. Thus content useful for encouraging participation is often related to such common identity, such as debating policies in political forums, reporting and fixing bugs in open-source software development, and discussions about information sources in Wikipedia. By contrast, in bond-based communities (e.g., online gaming), social and personal interactions as well as personal knowledge of others can help to build social bond among members and lead to participation. In other words, users participate in these communities because they have social ties with other members. As a result, exchange of personal life stories, and even other seemingly off-topic discussions can be useful.

As seeking and obtaining various types of social support is a key reason people participate in an OHC (E. Kim et al., 2012), which types of social support are more "useful" in keeping users engaged in the community? On one hand, OHCs have often been considered as identity-based communities, because all users in an OHC share a common identity as survivors or patients of a particular health condition or disease. Based on such a common identity, information about the condition or disease will be discussed and exchanged very often. As OHC users often suffer from emotional stress because of their common health problems, seeking and providing emotional support can also represent activities based on their common identity as survivors or patients. On the other hand, the exchange of emotional support and participation in companionship support, often in the form of seemingly off-topic discussions, can help OHC users get to know each other personally as they share things beyond health and diseases. Such interactions at the personal level can establish social bond among community members. Therefore, the first research goal of this Chapter tries to connect different types of social support with user participation. Figure 8 shows the conceptual framework of this goal.

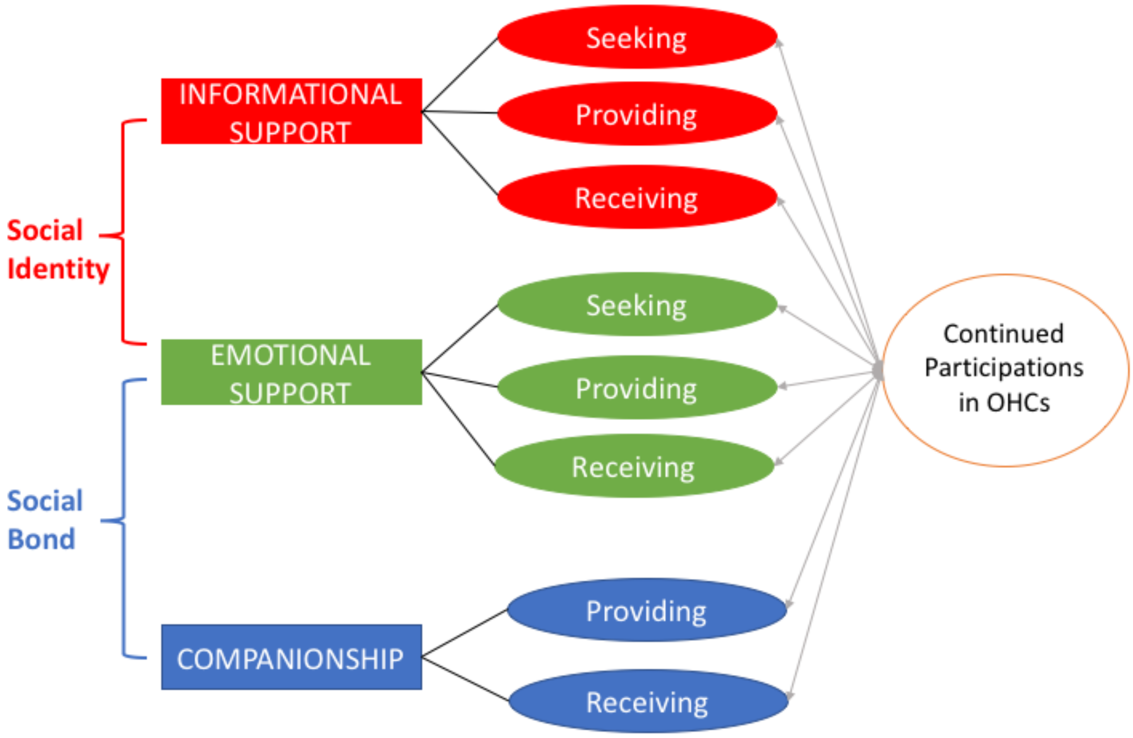


Figure 8 Conceptual framework of users' participation analysis

In addition to building explanatory models to understand factors related to users' participation, another key to sustain an online community is to predict user churn, so that the community can intervene when a user is about to leave the OHC and try to retain him or her. Implications for churn prediction are not limited to online communities, but also to other online and offline business. In the literature, predictive models for customer churn have been developed for telecommunication (Wei & Chiu, 2002), retail (Buckinx & Poel, 2005), Internet access service (Huang, Kechadi, & Buckley, 2009), online gaming (Kawale, Pal, & Srivastava, 2009), among other sectors. These models have leveraged different types of data about customers and the market, including those related to money, contracts, demographics, usage, products, complaints, competitions, and social networks (Backiel, Baesens, & Claeskens, 2014; Berson, Smith, & Thearling, 1999; X. Zhang, Zhu, Xu, & Wan, 2012). Various data mining and machine learning methods have been adopted for the prediction, such as decision trees (Bin, Peiji, & Juan, 2007; Xie, Li, Ngai, & Ying, 2009), logistic regression (Mozer, Wolniewicz, Grimes, Johnson, & Kaushansky, 2000), support vector machines (Coussement & Poel, 2008), and neural networks (Tsai & Lu, 2009).

When it comes to online communities, traditional churn prediction faces challenges as well as opportunities. On one hand, many of the features commonly used for churn prediction in for-profit business are not available or make no sense. For instance, users' demographic data (e.g., residential address, income and ethnicity) is usually unavailable or inaccurate in online communities. Also, because many online communities are based on voluntary participation and do not charge any fee, monetary and contractual issues become largely irrelevant. On the other hand, online communities provide more detailed data about users' behaviors for predictive analytics (Shmueli & Koppius, 2011). While previous churn prediction studies have leveraged structured data of users' activities, few have examined the unstructured content of users' interactions or contributions. By contrast, in many online communities, including OHCs, large amount of such content is publicly available from the Web. Previous research on online social networks and social media has suggested that content analysis can be helpful in areas such as personalized recommendation (Barbieri, Bonchi, & Manco, 2014), business intelligence (Chau & Xu, 2012), community discovery (Sachan, Contractor, Faruquie, & Subramaniam, 2012), and influential user identification (K. Zhao et al., 2014). Analyzing unstructured text contributed by online community users should provide new insights to churn prediction.

Moreover, many churn predictions for traditional business are limited to snapshot data- a model is learned from data for customers, who were active during a specific period (i.e., the training period, usually a couple of months to half a year), based on which customers churned in the subsequent testing period (often a few months). In other words, a model learns whether a customer will churn during the testing period based on his or her data in a predetermined training period. For an online community, data for a user's complete "life span" in the community can be available for analysis. Such complete data can provide valuable information, because those who churn after the first week may behave differently from those who churn after a month. In addition to building different models for different training periods, a unified model was also built that takes into consideration the length of time a user has been active and predict whether the user will churn based on all his or her historical data. Thus, the second research goal is about building predictive models using data of users' social support activities.

Research Goal of this Chapter: Explore whether users' activities in seeking, providing, and receiving different types of social support are related to their continued participation in an OHC. Leverage data about users' online social support activities to predict whether and when a user will churn from an OHC.

Analyzing Users' Continued Participation

After detecting the nature of social support in each post, I conducted survival analysis to study how different types of social support activities are related to users' participation. An individual may enter or exit a community not only based on his or her own expectations and behaviors, but also based on the community's reactions towards this individual (Levine & Moreland, 1994). Thus, in addition to users' own posting behaviors, I also examined whether the receiving or exposure to different types of social support would impact a user's participation.

The survival analysis was based on the Cox Proportional-Hazards Model (Cox, 1972), which assesses the importance of different independent variables on the "survival time" it takes for a specific event to occur. The hazard $h_i(t)$ represents the events occur to individual i at time t (defined in Equation 4),

Equation 4:

$$h_i(t) = h_0(t) * exp\{\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}\}$$

where the baseline hazard function $h_0(t)$ can be any function of time *t* as long as $h_0(t) > 0$. x_i and β_i represent independent variables and corresponding coefficients. Equation 4 can also be formulated as Equation 5, where the ratio of two individuals' hazard functions does not depend on time t.

Equation 5:

$$\frac{h_i(t)}{h_j(t)} = exp\{\beta_1(x_{i1} - x_{j1}) + \cdots + \beta_k(x_{ik} - x_{jk})\}$$

By using Maximum Likelihood Estimation, β can be estimated with regards to the hazard. $\beta_k = 0$ would indicate that independent variable x_k has no association with survival time; $\beta_k > 0$ means that independent variable x_k induces a higher hazard of event occurring, and vice versa. Correspondingly, $exp\{\beta_k\}$ is the hazard ratio of independent variable x_k.

Specifically, for the analysis, an "event" refers to a user's cessation of activities in the OHC (i.e., churn from the OHC). A user's survival time is measured by the difference between her last and first posts in the OHC. Similar to a previous study(Wang et al., 2012), I assumed that a user had churned from this OHC if the user had no post during the last 12 weeks in the dataset. For those who were active in the OHC during the last 12 weeks, their survival time is right-censored because they were still participating in this OHC.

Table 16 summarizes independent variables in the model. They reflect users' own posting behaviors in various social support categories, as well as the amount of social support they receive in threaded discussions in direct or indirect ways. A user receives support directly when the user initiates a thread to seek support and got support from others' replies to the thread. Meanwhile, social support can also be received indirectly when one replied to a thread started by another user, because the user may be exposed to support that other users provided to the original poster. In addition to these independent variables, I also included three control variables to reflect users' overall levels of activities.

To ensure the robustness of the results, I ran survival analysis in two experiments. The first experiment includes 24,604 users whose time spans of activities in the OHC exceeded one week. Values of control and independent variables are collected based on their behaviors in seeking, providing, and receiving social support in the first week of their participation. The second experiment has similar settings, but focuses more on long-term users and only included 19,165 users who are with the OHC for more than a month. To reduce the impact of multi-collinearity, I calculated the correlation coefficients for every pair of variables. I then removed TotalPost and NumThread from the model, as both are strongly correlated with the other control variable InitPost (with correlation coefficients greater than 0.8). Thus, the model for survival analysis includes 1 control variable and 10 independent variables.

Variables	Descriptions
TotalPost	The total number of posts a user has published
InitPost	The total number of threads a user initiated
NumThread	The number of threads a user contributed to (excluding those initiated by the user)
PES	The number of a user's posts that provided emotional support
PIS	The number of a user's posts that provided informational support
SES	The number of a user's posts that sought emotional support
SIS	The number of a user's posts that sought informational support
COM	The number of a user's posts that were related to companionship
$\boldsymbol{RIS_D}$	Direct informational support received--the number of informational support posts a user received after initiating a support-seeking thread.
$\boldsymbol{RES_D}$	Direct emotional support received--the number of emotional support posts a user received after initiating a support-seeking thread.
$\boldsymbol{RIS_I}$	Indirect informational support received--the number of informational support posts a user was exposed to in threads that she/he did not initiate but contributed to.
$\boldsymbol{RES_I}$	Indirect emotional support received--the number of emotional support posts a user was exposed to in threads that she/he did not initiate but contributed to.
RCOM	Companionship received--the number of companionship posts a user was exposed to in threads that she/he did not initiate but contributed to.

Note: for $\boldsymbol{RIS_I}$, $\boldsymbol{RES_I}$, and *RCOM*, I assumed that a user read others' replies that were posted within 7 days before the user's replies in the same thread.

Table 16 Control variables (in gray shade) and independent variables in the survival analysis

Table 17 shows the results of the model. Variables with hazard ratios lower than 1 contribute positively to the "survival" (i.e., continued participation) of users, whereas those with hazard ratio higher than 1 are considered "hazardous" to keep users in this OHC. Three independent variables have hazard ratios that have high statistical significance ($p<0.001$) and are consistent in the two experiments: the hazard ratios of COM in both experiments are lower than 1, meaning that users who post more companionship have longer time spans of activities in the OHC. More specifically, a hazard ratio of 0.919 for COM in Experiment 2 means that a user's "survival" rate after one month is 8.1% higher (100% - 91.9%) if his or her number of companionship posts is one standard deviation higher than the average. Similarly, those who posted more to provide informational support (PIS) tended to stay with the OHC for longer. By contrast, those who sought more informational support (SIS) often left the OHC earlier. Other variables are either significant in only one experiment (e.g., SES and RISD) or not significant in both experiments (e.g., PES). To further check the robustness of the model, I also ran the same analysis for two different time periods: Period 1 (Jan 2006 to Jan 2009), and Period 2 (Feb 2009 to Aug 2013). Period 1 was when the OHC saw significant growth in the number of active users, while the OHC's number of active users stayed stable during Period 2. The results are basically consistent with main findings in Table 17.

Variables	Hazard Ratio (Experiment 1)	Hazard Ratio (Experiment 2)
InitPost (Control)	1.098***	0.995
PES	1.012	1.000
PIS	0.966***	0.948***
SES	0.991	0.972*
SIS	1.034***	1.050***
COM	0.956***	0.919***
$\boldsymbol{RIS_D}$	0.974	1.047*
$\boldsymbol{RES_D}$	1.008	0.997
$\boldsymbol{RIS_I}$	0.992	1.053*
$\boldsymbol{RES_I}$	0.987	0.964
RCOM	0.983	0.983

Table 17 Results from two survival analysis experiments

According to the analysis, the OHC features both social identity-based and social bond-based participation. First, informational support is the most popular social support being sought and provided. This is common for communities based on common social identities, because the large amount of information about a disease and the common identity as patients of the disease are probably why many users come to the OHC in the first place. While providing more informational support is positively correlated with longer participation, seeking informational support is negatively associated with participation and receiving informational support is not a consistently significant factor. In other words, those who focus on seeking information may not stay in the long run, even after they receive informational support.

Second, companionship has the strongest correlation with users' participation. Recall that companionship includes discussions of offline events, sharing daily life stories, birthday wishes, and playing online games. This is a very interesting finding—even though this is an OHC about cancer, discussions of non-cancer-related issues is the key to keeping users engaged in the community. This highlights the importance of social bond-based participation in the OHC, as off-topic discussions in the form of sharing personal stories about life or having fun together can strengthen the social bond among users more than informational support, which often lacks the personal touch. The role of companionship has significant implications for the management of an OHC. Although some OHCs may discourage off-topic discussions in order to achieve a "cleaner" environment with only relevant content, these discussions turn out to be a good way to bond users and keep them engaged, and OHC managers may want to encourage, or even initiate, more of these activities. Companionship posts may also be good candidates to be included in email reminders or post recommendations, as an intervention to retain users and encourage participation.

Third, although I expected emotional support to be positively related to user participation as suggested by Wang et al. (2012), the results are mixed based on whether emotional support was being sought, provided, or received. The hazard ratio of SES is below 1 in both experiments, and is statistically significant in Experiment 2. This contradicts the effect of SIS and suggests that seeking emotional support can be a sign of longer participation, especially for those who have been with the OHC for a while.

However, providing and receiving emotional support are not significant factors. I suspect that a fair amount of emotional support in the OHC can be generic and a mere formality (e.g., "I will pray for you", "Love you and Hug"). Such emotional support can still be valuable for those who seek support, but activities in providing and receiving such support are not related to users' continued participation.

Predicting User Churn

Knowing that different types of social support activities are related to users' participation in OHCs, the next step is to predict user churn from OHCs by utilizing users' social support activities. As mentioned previously, the traditional method for churn prediction is to train a model based on data from a specific period. However, to predict user churn at different times, multiple predictive models need to be trained. Thus, one unified model is proposed here to predict user churn given the length of time the user has been active. In this section, I will describe, evaluate and discuss the predictive models.

Basic features for predictive models are derived from the 13 independent variables I used for survival analysis (see Table 16). Because these features aggregate users' activities during the training period, I also measured how users' values on the 13 features vary over time using four types of temporal features. Specifically, for each user, I divided the users' activities measured by each of the 13 basic features into weeks, and used four additional metrics to capture how the value of each feature changes over the weeks:

(1) The overall slope of a feature—a positive slope suggests a user's weekly activities was on the rise during the training period, and vice versa;

(2) The Shannon entropy (Shannon, 1948) of users' weekly activities, with lower entropy indicating more stable activities for the corresponding feature during the training period. For instance, if a users' total number of posts across 4 weeks are 1, 2, 1, and 3 respectively, the probability of publishing 1 post in a week is ½. The probabilities of

publishing 2 posts in a week is ¼, and so is the probability of publishing 3 posts in a week. Based on Equation 6, the entropy of total number of posts published would be $-\left(\frac{1}{2} * \log_2 \frac{1}{2} + \frac{1}{4} * \log_2 \frac{1}{4} + \frac{1}{4} * \log_2 \frac{1}{4}\right) = 1.5$. However, this metric only considers the appearance of different numbers, instead of numeric values of these numbers. For instance, another user with 1, 5, 1, and 6 posts in 4 weeks will have the same entropy as the previous user with 1, 2, 1, and 3 posts.

Equation 6:

$$Entropy = -\sum p * \log p$$

(3) The new metric of stability is used to address the problem of Shannon entropy. Its calculation is similar to Shannon entropy, as defined in Equation 7, but p_i' represents the proportion of activities from week i compared to the total activities from all weeks. The higher the stability metric is, the more stable a user's activities over time.

Equation 7:

$$Stability = -\sum p_i' * \log p_i'$$

To handle cases when all the values are 0 during a time period, I also adopted Laplace Smoothing (a.k.a., Add-one Smoothing). The same example for Shannon Entropy is used to illustrate how stability is calculated. Note that the total activities are 1+2+1+3=7. $\mathrm{p'}_1 = (1+1)/(7+4), \mathrm{p'}_2 = (2+1)/(7+4), \mathrm{p'}_3 = (1+1)/(7+4), \mathrm{p'}_4 = (3+1)/(7+4)$, and the stability for this user across 4 weeks would be $-\left(\frac{2}{11} * \log_2 \frac{2}{11} + \frac{3}{11} * \log_2 \frac{3}{11} + \frac{2}{11} * \log_2 \frac{2}{11} + \frac{4}{11} * \log_2 \frac{4}{11}\right) = 1.936$;

(4) The temporal variation (TV) of features, which extends entropy and stability by considering the fluctuation in a temporal sequence of data (K. Zhao & Kumar, 2013). For instance, if two users' values of a feature across 4 weeks are 1,3,1,3 and 1,1,3,3 respectively, they will share the same Shannon entropy and stability while the second user has less fluctuation on this feature. User i's TV on feature f, is defined in Equation

8, where $f_{i,t}$ measures user i's activity (e.g., total number of posts) during time interval t; S_i and E_i are the starting and ending time of the training period. Basically, TV measures the average variation between successive time intervals (e.g., weeks) during a given time period (e.g., a month), normalized by the average value across the given time period. The higher the value of TV, the more fluctuated a temporal sequence is.

Equation 8:

$$TV_{f,i} = \frac{\frac{1}{E_i - S_i} \sum_{t=S_i}^{E_i - 1} \left| f_{i,t} - f_{i,t+1} \right|}{\frac{1}{E_i - S_i + 1} \sum_{t=S_i}^{E_i} f_{i,t}}$$

In addition to cumulative values for each basic feature during the training period, I also conjectured that a user's intention to churn may be better captured during the last week of his or her online activities. Thus I also included values for basic features during the last week of the training period, if the training period is longer than one week. Each basic feature for the last week also has four corresponding features to reflect its temporal patterns (i.e., slope, Shannon entropy, stability, and TV), although the unit of time is day instead of week. I also added into the feature set the time difference between a user's registration time and the time of his or her first post, because it may reveal what brought the user to the OHC for the first time. A user who is eager to find some information may have a low gap between the registration time and the time of first posting.

A user is said to churn in his or her k-th week if his or her last online activity occurred during his or her k-th week in the OHC. Similar to the hazard model, users whose last online activities occurred during the last 12 weeks in the dataset are not considered as churned. I first built separate classification models for different time periods (referred to as time-dependent models). To predict whether a user will churn in the k-th week of his or her online activities, I focused on all users that are still active before the k-th week and extracted data based on their k weeks of activities. For example, the dataset for predicting user churn during the 3rd week contains users who were still active in the OHC before their 3rd week of online activities. Data of their behaviors during their first two weeks is collected for training. Users who churn in their 3rd week and never came back were labeled as “positive” instances in the dataset.

I built four models based on four datasets to predict user churn during the 1st week, 3rd week, 5th week, and 13th week of users’ online activities. For each dataset, I randomly chose half of the users as the training dataset and the rest of them as a hold-out testing dataset. I measured the performance of classifiers using various metrics, including precision, recall, F1 score and Area under the ROC (AUC).

After comparing the performance of different classification algorithms (Naïve Bayes, logistic regression, and SVM with poly kernel) with 10-fold cross validation on the training set, I picked logistic regression as the best performer. For training sets, I applied various instances of 10-fold CV and summarized the AUCs of the four time-dependent models in Table 18 (“T.Dep.” columns). Table 19 lists their performance on hold-out testing sets (“T.Dep.” columns). In addition to classification, community managers may also want to know the users who are mostly likely to churn. Thus, among users in each testing dataset, I also ranked the probability of a user being “positive” (i.e., churn), and show recall@K(rec.) and precision@K (pre.) of the four time-dependent models in Table 20.

	Churn@1st wk		Churn@3rd wk		Churn@5th wk		Churn@13th wk	
	T.Dep.	**Unif.**	**T.Dep.**	**Unif.**	**T.Dep.**	**Unif.**	**T.Dep.**	**Unif.**
Mean	0.981	0.972	0.923	0.903	0.912	0.907	0.856	0.922
Std.D	0.003	0.004	0.023	0.032	0.034	0.033	0.090	0.063
P-val	p<0.001		p<0.05		p>0.05		p<0.01	

Table 18 Comparing AUCs of time-dependent (T.Dep.) and unified (Unif.) models on training sets using various instances of 10-fold CV
The “P-value” row indicates whether the two models’ AUCs are significantly different

	Churn@1st wk		Churn@3rd wk		Churn@5th wk		Churn@13th wk	
	T.Dep.	**Unif.**	**T.Dep.**	**Unif.**	**T.Dep.**	**Unif.**	**T.Dep.**	**Unif.**
Precision	0.948	0.950	0.869	0.872	0.794	0.880	0.578	0.838
Recall	0.938	0.937	0.592	0.534	0.547	0.511	0.545	0.504
F1	0.943	0.943	0.704	0.662	0.648	0.647	0.561	0.629
AUC	0.981	0.972	0.921	0.901	0.913	0.909	0.854	0.929

Table 19 Comparing the performance of time-dependent and unified models on hold-out testing sets (Precision and recall are for the positive class)

	Churn@1st wk		Churn@3rd wk		Churn@5th wk		Churn@13th wk	
	Pre.	**Rec.**	**Pre.**	**Rec.**	**Pre.**	**Rec.**	**Pre.**	**Rec.**
K=50	0.940	0.004	0.940	0.086	0.940	0.131	0.800	0.325
K=100	0.940	0.008	0.920	0.168	0.900	0.250	0.650	0.528
K=200	0.970	0.017	0.905	0.331	0.850	0.472	0.400	0.650
Testing set	23,581 ins. 11,360 pos.		10,549 ins. 547 pos.		9,496 ins. 360 pos.		7,541 ins. 123 pos.	

Table 20 Precision@K (pre.) and Recall@K(rec.) of the four time-dependent models on hold-out testing sets (for the positive class)

While each of the four time-dependent models achieves decent performance, building a different predictive model for each time period may not an efficient solution. If the OHC wants to know who will churn in their 2nd week, another new model is needed. Inspired by (Street, Mangasarian, & Wolberg, 1995), I attempted to consolidate all these models into one unified model by leveraging a user’ s data across his or her complete “online life span”. Specifically, in addition to all the features used by time-dependent models, I added one feature-time stamp t. An instance in the new dataset will reflect a user’s historical activities until t. As the unit of t is the same for all users (a week in the experiment), one user can correspond to multiple instances in the dataset. For example, a user who churn in his or her 3rd week of activities has three instances in the dataset- one instance for his or her activities and features until the end of his or her 1st, 2nd, and 3rd week respectively. The first two instances are labeled as “negative” as the user is still active during these two weeks, while the third instance is labeled as a “positive” instance because the user churn at his or her 3rd week. In other words, the unified model tries to capture the complete life span of a user in the OHC.

To train the unified model, 24,000 users are randomly selected from 47,581 users in the OHC to be included in the training dataset, while others are placed in the hold-out testing dataset. It is worth noting that the unified model with time stamps as a feature greatly increased the amount of training data, as a loyal user who has been active for a long time will have many instances in the dataset. While 24,000 users in the training dataset would mean 24,000 instances for a time-dependent model, the unified model uses a training dataset with 132,341 instances. I built the training dataset and trained the model on a high-performance computing cluster. I also confirmed that instances for the same user must belong to the same fold in cross validation. Again, logistic regression has the best performance on the training dataset (using 10-fold cross validation). Because the dataset for the unified model is organized differently, I cannot directly compare the classification performance of the unified model with time-dependent models. Instead, I divided instances of the testing set into many subsets based on their values of time stamp t, so that each subset includes unique users who are still active until the same week t. Then I could apply the learned model to individual users in each subset and compare the unified model's performance with time-dependent models with the same t. In Table 18 Table 19, the unified model's classification performance on training and testing sets is compared with that of time dependent models. Table 21 lists its recall@K and precision@K for four different t values.

Compared with time-dependent models, the unified model offers comparable performance: time-dependent models are slightly better for predictions of short-term churns, while the unified model is better at predicting churns of long-term users. Specifically, on training sets with 10-fold CV, the unified model trails time-dependent models in predictions for the 1st and 3rd weeks, but outperforms time-dependent models for in the prediction for the 13th week. The similar results are observed on hold-out testing sets. Time-dependent models for the 1st, 3rd, and 5th week lead in AUC by very small margins: 0.009, 0.02, and 0.004 respectively, while the unified model's AUC is higher for the 13th week by 0.075.

	Churn@1st wk		Churn@3rd wk		Churn@5th wk		Churn@13th wk	
	Pre.	**Rec.**	**Pre.**	**Rec.**	**Pre.**	**Rec.**	**Pre.**	**Rec.**
K=50	0.940	0.004	0.900	0.082	0.900	0.125	0.880	0.358
K=100	0.940	0.008	0.920	0.168	0.930	0.258	0.740	0.602
K=200	0.945	0.017	0.910	0.333	0.890	0.494	0.465	0.756
Testing set	23,581 ins. 11,360 pos.		10,549 ins. 547 pos.		9,496 ins. 360 pos.		7,541 ins. 123 pos.	

Table 21 Precision@K (pre.) and Recall@K(rec.) of the unified model on hold-out testing sets (for the positive class)

When it comes to identifying the users who are most likely to churn, the unified model also offers satisfactory performance. Time dependent models perform only slightly better for the 1st week, but the unified model catches up quickly: it offers better results when K=200 for the 3rd week, and when K=100 and 200 for the 5th week. It then dominates time dependent models for all the three K values in the 13th weeks.

In summary, by incorporating the complete time spans of users' online activities, the unified model can predict churn across different time periods with performance that is close to or even better than time-dependent models for each period. The single unified model can provide churn prediction for different time periods and is very handy for OHC managers. For any user, the model can track data of his or her activities from the user's first day in the OHC to present. Then based on the data and the length of time the user has been in the OHC, the model provides OHC managers with a prediction on whether the user is about to churn. Besides predicting user churn at the individual level, the unified model also makes it easier to plot the hazard curve for users' participation across the community. One can simply apply the trained unified model to a group of users who are still active at a specific time and get the probability of churn for the group. Figure 9 plots three hazard curves: one based on real data, one based on predictions from time dependent models and one based on predictions from the unified model. The horizontal axis represents weeks, and the vertical axis refers to the probability of users' churn at specific weeks. It is clear that the hazard curve predicted by the unified model is very close to the curve based on the real data. I only showed five data points for time-dependent models because plotting such a hazard curve using time-dependent models requires 13 different models.

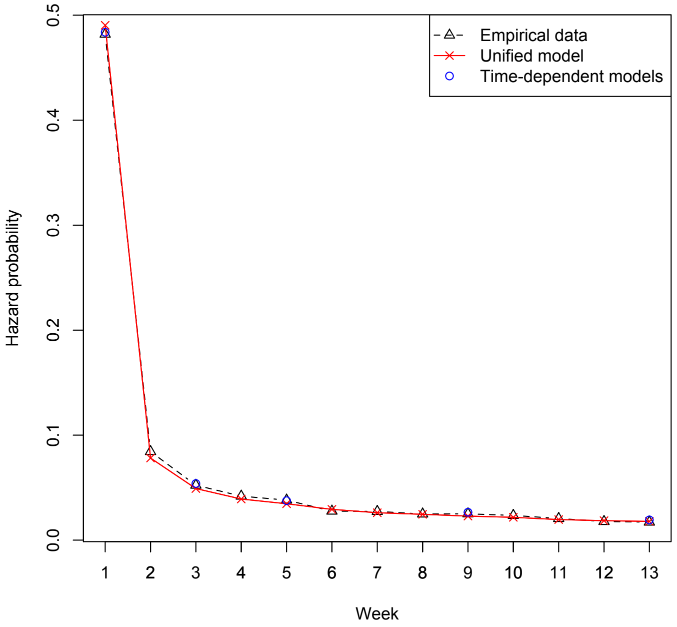


Figure 9 Hazard curves for user participation.

To further understand which features are more important for the unified model's predictive power, I ranked the 145 features used in the model using information gain (Guyon & Elisseeff, 2003). The top 20 features are listed in Table 22. Among the top 20 features for the unified model, 18 of them are features that reflect the temporal dynamics in users' activities, especially the stability during the last week of the training period. Intuitively, the correlation analysis between these stability metrics and churn behaviors suggests that the more stable a user's activities are, the less likely a user will churn from the OHC. Also, 11 features among the top 20 are made possible only after the classification of different types of social support. Overall, this shows that users' temporal changes both in activity volume (e.g., publishing posts and participating in threads), as well as their activities in seeking, providing, and receiving different types of social support, can greatly enhance churn predictions in OHCs.

In addition, all the models' performance deteriorates when the prediction is for churn after a longer time of online activities. All models achieve the best AUC for predictions at the 1st week and the worst AUC for churn prediction at the 13th week. I conjectured that two reasons may have contributed to the decrease: first, the rate of churn is much lower when a user has been active for a longer time. Thus the dataset becomes much more unbalanced: only 1.7% of users who were active before the 13th week churned during the 13th week, compared to 48.2% for churn in the 1st week. Such a problem may be ameliorated by carefully handling the unbalanced dataset. Second, the churn of a long-term user may be inherently more difficult to predict. Some may pass away because of the breast cancer, and some may have become cancer free and left.

Rank	Feature
1	Stability of the total number of threads a user initiated during the last week of the training period
2	Stability of the number of threads a user participated in during the last week of the training period
3	Stability of the number SIS a user posted during the last week of the training period
4	Stability of the number SES a user posted during the last week of the training period
5	Stability of the total number of posts from a user during the last week of the training period
6	Stability of the number of PIS a user posted during the last week of the training period
7	Stability of the number of PES posts a user received directly during the last week of the training period
8	Stability of the number of PES a user posted during the last week of the training period
9	Stability of the number of COM a user posted during the last week of the training period
10	Stability of the number of PIS posts a user received directly during the last week of the training period
11	Stability of the number of COM a user was exposed to during the last week of the training period
12	Stability of the number of PES posts a user received indirectly during the last week of the training period
13	Stability of the number of PIS posts a user received indirectly during the last week of the training period
14	Total number of posts from a user during the last week of the training period
15	The number of threads a user participated in during the last week of the training period
16	Stability of the number of threads a user participated in across weeks
17	Stability of the total number of posts from a user across weeks
18	Entropy of the total number of posts from a user during the last week of the training period
19	Stability of the number of PIS posts a user received indirectly across weeks
20	Stability of the total number of threads a user initiated a user across weeks

Table 22 Top 20 features by information gain for the unified model

Summary

In this Chapter, I analyzed the impact of seeking, providing and receiving different types of social support on users' continued participation. Getting involved in COM posts as an unexpected factor, is positively associated with users' long term engagement. I also developed predictive models to forecast users' churning behaviors.

The outcome of this Chapter should be valuable in improving website design. For example, traditionally, an OHC will send reminder emails to a user who has been inactive for a while, hoping to raise the user’s interests in coming back. With the help of the churn prediction model, an OHC can find at an early stage whether a user is about to leave. Then, it can intervene proactively and try to retain the user via email reminders. More importantly, instead of including a generic reminder or some random recent posts from the community, such emails can be designed based on the results of the survival analysis. For example, because companionship is a key predictor of users’ continued participation, including some of these companionship posts (e.g., birthday wishes, holiday plans, and online scrabble games) in reminder emails may be more effective to keep users engaged than having random posts or just informational posts.

CHAPTER SIX

SOCIAL CAPITAL AND USERS' FUTURE CONTRIBUTIONS

The success of OHCs depends largely on sustained participation and voluntary contributions from users (Burke, Marlow, & Lento, 2009). Therefore, besides predicting and preventing user churn from an OHC, it is also important to encourage users to contribute more to the community. Motivating user contributions is considered the greatest challenge to such virtual communities (Chiu, Hsu, & Wang, 2006). Although knowledge sharing in online communities has long been studied (Wasko & Faraj, 2005), few have investigated OHCs. Some factors such as trust and shared language are verified that impact OHC users' future contributions in OHCs (J. Zhao, Ha, & Widdows, 2016), but few conclusions are summarized based on content analysis.

According to the findings in Chapter Four, users' motivations are influenced by social network neighbors through social interactions, and the content users create in the OHC may be driven by different motivations. Those with community interest in mind tend to provide social support to others without explicit rewards. By contrast, "self-interest" users focus only on meeting their own needs when using the OHC. A user's social network may change the user's motivation, for example, from community-interest users to self-interest users. From the perspective of knowledge contribution, this means social network can influence users' contributions in different categories, including community-interest posts and self-interest posts.

In the social network of an OHC, through which social support is exchanged, users may gain social capital, a special resource within a social network. Social capital is a resource within a social network. In general, social capital facilitates people's access to sources of knowledge. Several definitions of social capital have been provided in the past, whether from the perspective of social structure (Bourdieu & Wacquant, 1992) or from the viewpoint of its function (James Coleman, 1990). Differing perspectives aside, all the definitions agree that social capital is the sum of resources that can create for particular individuals or groups a modest advantage in reaching their final goals, so that "better connected people enjoy higher returns" (Burt, 2001). In other words, the more social capital and relationships one has, the more knowledge can be created or shared in one's social network. In this Chapter I would like to explore social capital factors related to OHC users' community-interest and self-interest contributions.

Social Capital and Knowledge Contribution

Social capital can serve as a measure of the quality of a group, which includes the rule of law, social integration and trust (Borgatti, Jones, & Everett, 1998). However, in this thesis, social capital is considered as the value of an individual's relationships, as connecting with more people can help the individual to access needed resources (James Coleman, 1990; Lin, 2001; Putnam, 2001). Prior studies verified that accessing to intellectual capital can impact the combination and exchange of knowledge, and further influence the dynamics of individuals' knowledge contributions (Bouty, 2000). Therefore, understanding how social capital works is beneficial for the study of information flow and users' behaviors in an OHC. According to Coleman (1988), certain kinds of social structure are especially valuable in facilitating social capital in some specific forms.

From a social structure perspective, social capital can be divided into three categories: structural social capital, relational social capital and cognitive social capital (Nahapiet & Ghoshal, 1998). Structural social capital refers to the connections among individuals in social interactions, specifically, who can be connected and how they can be connected. By contrast, the relational social capital examines the assets created by historical relationships, such as respect, trust and friendship. As for cognitive social capital, it describes shared vision, such as shared norms, values, attitudes, and beliefs. Earlier studies examined correlations between these three types of social capital and knowledge contribution in different online communities and presented interesting findings. For example, Wasko and Faraj (2005) concluded that structural and cognitive social capital play vital roles in knowledge contribution, while relational social capital holds relatively weak predictive power.

By contrast, from a network perspective, social capital can be divided into two major categories: bridging social capital and bonding social capital. Bridging social capital is related to recourses within the network to novel information. Network members who can help information spread more quickly and efficiently are the most important ones in terms of bridging social capital. These members span multiple clusters and close "structural holes" between unconnected groups (Burt, 1995; N. B. Ellison & Vitak, 2015). By contrast, members with high bonding social capital are close friends or family members, derived from an individual's inner cluster of connections. Prior studies have also pointed out that strong and weak ties in the online social network are associated with positive bonding and bridging social capitals, respectively (N. Ellison, Lampe, Steinfield, & Vitak, 2011). Specifically, weak ties are more valuable for informational benefits, such as holding diverse views and accessing new information, while strong ties are generally considered as a way of spreading emotional support, especially between family members or friends (Burt, 1995; Putnam, 2001).

Social capital has long been studied to explain knowledge sharing in online communities. Some empirical studies show that the benefits of social capital are positively correlated to users' knowledge sharing activities in an online community, while the risks of social capital have a negative impact (H. H. Chang & Chuang, 2011; Chiu-Ping Hsu, 2015; Sheng & Hartono, 2015). Moreover, because social capital is associated with knowledge contribution, it is also a predictor of leadership in online communities (Faraj et al., 2015). It is unclear however, whether this will apply to OHCs.

OHCs' features make them different from other online communities. First, OHC users have no monetary incentives to contribute. Specifically, some online communities feature mechanisms to explicitly reward users' participation or contributions (e.g., virtual badges or stars), and these online rewards can sometimes have monetary implications too. For example, a programmer's badges on StackOverflow can potentially land him or her a well-paid job (Feffer, 2015). The community-interest behavior of users in such online scenarios can be the outcome of self-interested reasons. By contrast, when such incentives are missing in the OHCs, users contribute with considering what they really need or be interested in. OHCs are used to share community- and self-interest posts at the same time, indicating users' intention of benefiting others and benefiting themselves, respectively. Second, different from other non-monetary online communities, such as Wikipedia, contributions in OHCs are also about social support, instead of limiting in knowledge. Knowledge contributions may have certain requirements on a user's inherent knowledge level or experience, while the emotional support and companionship can be provided by almost everyone in OHCs. Therefore, whether valuable social capitals are positive indicators of users' contributions in OHCs is an interesting question to explore.

Inspired by frameworks in prior studies, which illustrated how social capital is related to users' knowledge contribution (Chiu-Ping Hsu, 2015; Faraj et al., 2015; Wasko & Faraj, 2005), I propose a framework to investigate social capital factors impacting users' future contributions in the OHC as shown in Figure 10. Rather than classifying social capital based on the network perspective or the structural perspective mentioned above, I included bonding social capital and bridging social capital as two measurements of structural social capital. Thus, both the social network and social structure perspectives can be covered.

Research Goal of this Chapter: Explore what types of users' social capital are related to users' future contributions to an OHC.

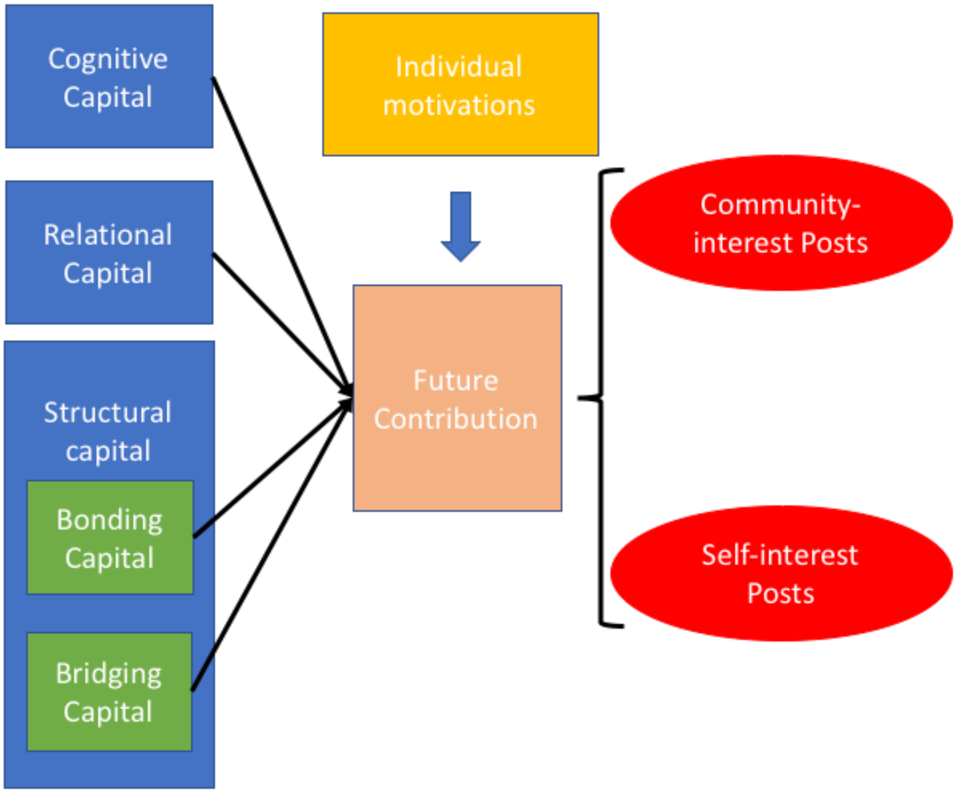


Figure 10 Framework of social capital study in the OHCs

Explanatory Model

In measuring users' social capital and its impact on their future contributions, I adopted the monthly data entries of users as same as Chapter Four. However, I have only included in the dataset users whose participation had a life span more than 1 month. Among 384,423 data entries in Table 11, 336,842 remained within the dataset. For each user, at a particular month t, the dependent variables $CommPosts_{i,t+1}$ and $SelfPosts_{i,t+1}$ are two numerical variables indicating the number of the user i's community-interest posts and self-interest posts at time t. The independent variables $Socialcapital_{i,t}$ capture the user i's social capital in the three categories of structural, relational and cognitive capital. Specifically, I included both betweenness and closeness to measure the user’s bridging capital and bonding capital in the structural social capital category.

To maintain consistency with the last two Chapters, the co-participation network has been used here for calculating variables in the model. Meanwhile, I selected mutual responses and shared visions to represent relational and cognitive social capital, based on the definitions of those terms. Mutual responses measures instances where A replies to B’s thread and B also replies to A’s thread. Shared visions counts similar opinions held by two users in one thread. The users who share visions both provide their opinions to a subjective initial thread. For example, in Figure 11, where B provides an informational support to A’s thread, and D also posts informational support to validate B’s opinion (sharing the same polarity), then B and D are classified as users who are sharing a similar vision. Although E also provides informational support, the polarity of E’s post is opposite to B’s, which means B and C did not share the same vision. In addition, because I am measuring the opinions of the users, only PIS are considered here.

Rather than simply counting the number of mutual responses or vision-shared interactions a user had, I captured different dimensions of the variable. I implemented RFM model (Fader, Hardie, & Lee, 2005) to measure relational and cognitive social capital. RFM analysis has been used to determine the most valuable customer from a pool of candidates by examining their recency (i.e. how recently a customer has purchased), frequency (i.e. how many items the customer has purchased), and monetary value (i.e. how much the customer has spent). For example, to a retailor, customers who purchased more recently, more often and spent more are more likely to buy again than the ones who did not come for a while, bought infrequently and spent less. The former group of users features higher value to the retailor than the latter group.

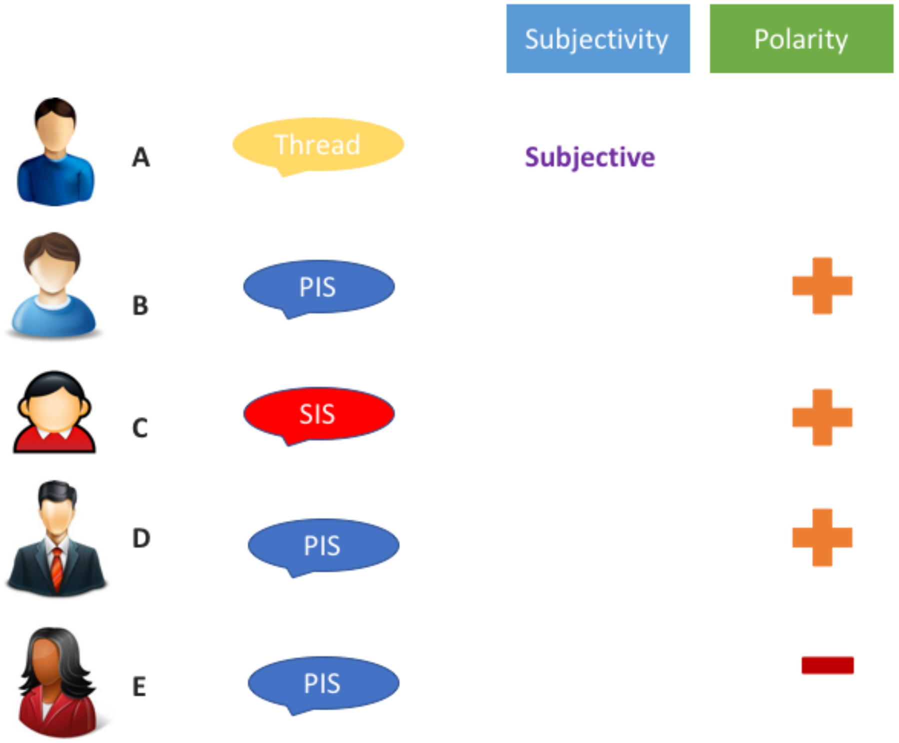


Figure 11 Visualized example of shared visions calculation

Specifically for the OHC, to discover which users have higher social capital, without changing the settings of recency and frequency, I tweaked the definition of the last dimension, monetary value, to fit the OHC scenario. Since OHCs are non-monetary online communities, it is difficult to measure how much benefit a mutual response or shared vision can create for a user. Because users' interactions and connections with strangers are formed in the social network, this can help users gain access to more resources as well as social support. As a result, the number of connected users created by the social capital can be used as a measure of the value of that capital. Therefore, taking the example of mutual response for a user during month t, I computed the most recent time of a mutual response (recency), the number of mutual interactions the user had (frequency), and the number of unique persons with whom the user interacted with (monetary value). Notice that there might be a lot of mutual interactions between two users, so “the number of mutual interactions the user had” is always a value not less than “the number of unique others the user interacted with”. A similar approach used to measure cognitive social capital.

Control variables included a user’s current contributions at month t (including both community-interest posts and self-interest posts), and their active months until month t. The active months captures the actual active behavior of a user. For example, if Mary was involved in the community from January to May, but only contributed some social support in February and March, then the active months of her at May would be 2. The variables are summarized in Table 23.

Variables		**Description**	**Name**	**Type**
Dependent		number of PIS+PES+COM @ month t+1	D1	Numerical
		number of SIS+SES @ month t+1	D2	Numerical
Control		number of PIS+PES+COM @ month t	C1	Numerical
		number of SIS+SES @ month t	C2	Numerical
		Active Months	C3	Numerical
Independent	**Structural Social Capital**	Betweenness of a user @month t	SSCB	Numerical
		Closeness of a user @month t	SSCC	Numerical
	Relational Social Capital	Recency of mutual responses @month t	RSCR	Numerical
		Frequency of mutual responses @month t	RSCF	Numerical
		Number of users having mutual responses @month t	RSCM	Numerical
	Cognitive Social Capital	Recency of shared visions @month t	CSCR	Numerical
		Frequency of shared visions @month t	CSCF	Numerical
		Number of users with shared visions @month t	CSCM	Numerical

Table 23 Variables of the explanatory model

Equation 9 below shows the model I used for this study. Coefficient γ estimates the relationship between users’ current and future contributions, and β estimates the influence of a user’s social capital on his or her future contributions. Fixed effects of individual users' time-invariant attributes, e.g. gender, education level and personality, are controlled by α_i for user i.

Equation 9:

$$\left.\begin{matrix} CommPosts_{i,t+1} \\ SelfPosts_{i,t+1} \end{matrix}\right\} = \alpha_0 + \alpha_i + \beta Socialcapital_{i,t} + \gamma \left\{\begin{matrix} CommPosts_{i,t} \\ SelfPosts_{i,t} \end{matrix}\right\} + \delta Controls_{i,t} + \epsilon_{i,t+1}$$

Variables	**C2**	**SSCB**	**SSCC**	**RSCR**	**RSCF**	**RSCM**	**CSCR**	**CSCF**	**CSCM**	**C3**
C1	0.67	0.45	0.34	0.24	0.27	0.25	0.34	0.36	0.37	0.12
C2		0.56	0.45	0.34	0.36	0.35	0.47	0.50	0.52	0.09
SSCB			0.36	0.27	0.28	0.27	0.38	0.37	0.40	0.09
SSCC				0.22	0.20	0.21	0.48	0.45	0.48	0.23
RSCR					0.89	0.90	0.24	0.21	0.25	0.09
RSCF						0.95	0.24	0.21	0.25	0.09
RSCM							0.24	0.21	0.26	0.10
CSCR								0.91	0.94	0.11
CSCF									0.97	0.09
CSCM										0.11

Table 24 Pearson correlation of independent variables

To avoid decreasing the influence of highly scaled data entries, I transformed all the variables following power-law distributions to the log scale. To check the multi-collinearity, a pair-wise person correlation test was conducted among all independent and control variables. The outcome (Table 24) shows that, besides internal variables recency, frequency and monetary dimensions of mutual responses and shared visions being highly correlated to each other, correlations between any two variables are less than 0.67. I adopted linear models for the panel data, and ran separated models for each social capital group to compare.

Preliminary Results

Figure 12 and Figure 13 shows the results of all the regression models. The x-axis represents the estimated coefficients, and the y-axis indicates different variables, including control variables (Cont.), structural social capital (S_SC), relational social capital (R_SC), and cognitive social capital (C_SC). Only the significant results with p-value <0.1 are shown in the figures. The coefficients are presented as dots and the confidence intervals are shown as whiskers (Kastellec & Leoni, 2007). The outcomes of all the models are consistent for either type of dependent variables. However, setting two categories of user's contribution (community-interest posts and self-interest posts) as the dependent variables returns altered coefficients in terms of social capitals.

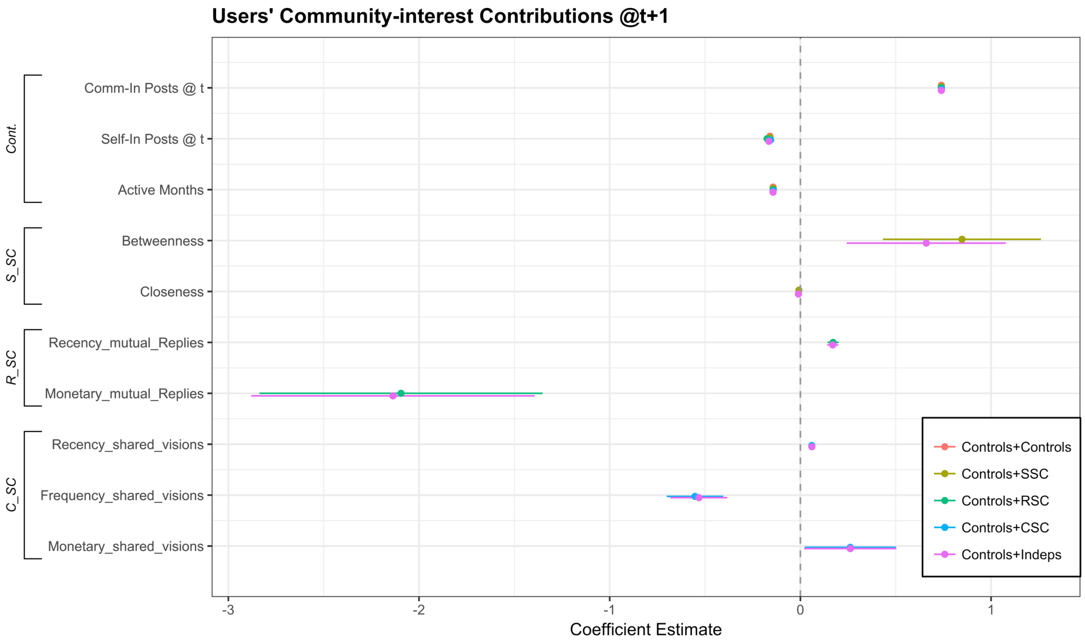


Figure 12 Regression results for community-interest contributions

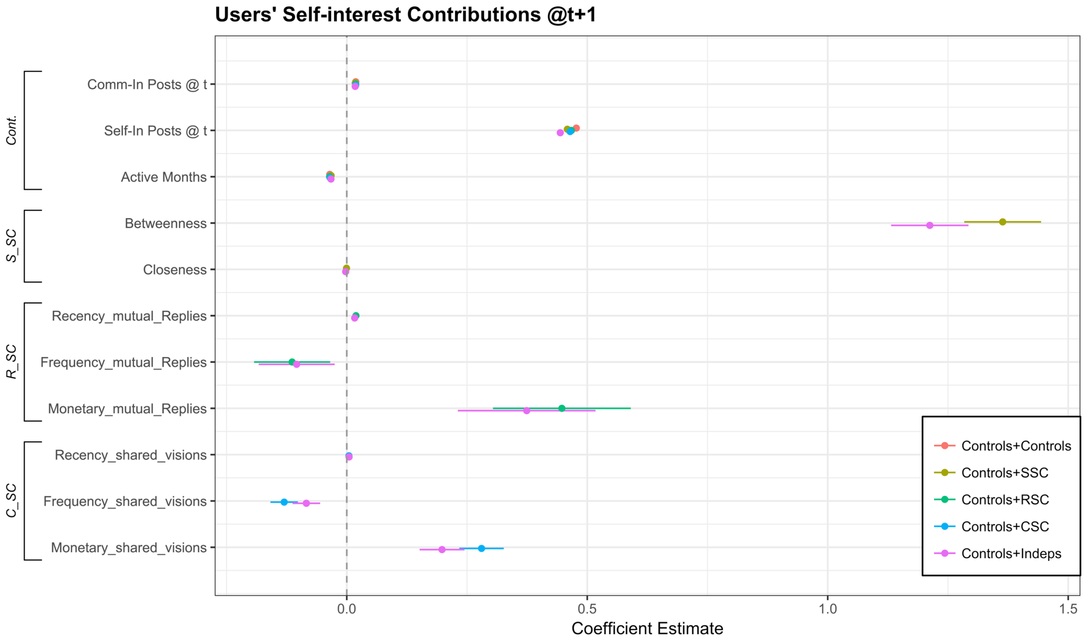


Figure 13 Regression results for self-interest contributions

To better understand the differences between models for two dependent variables, I summarized the outcome of the two models with combining control and all independent variables in Table 25.

In general, structural and cognitive social capital have similar impacts on a user's future community-interest and self-interest contributions. Relational social capital on the other hand, influenced the two categories in different ways. As for control variables, the increasing number of current self-interest posts will result in fewer community-interest posts published the next month.

Dependent **Independent + Controls**	**Community-interest Posts @ t+1**	**Self-interest Posts @ t+1**
Community-interest Posts @ t	0.739***	0.017***
Self-interest Posts @ t	-0.165***	0.444***
Active Months	-0.143***	-0.033***
Betweenness	0.660***	1.212***
Closeness	-0.010***	-0.002***
Recency of Mutual Responses	0.170***	0.016***
Frequency of Mutual Responses	0.047	-0.104***
Number of users having Mutual Responses	-2.137***	0.374***
Recency of Shared Visions	0.060***	0.005***
Frequency of Shared Visions	-0.532***	-0.084***
Number of users having Shared Visions	0.262***	0.198***

Table 25 Results of the explanatory model

Discussion

According to the preliminary outcomes of the model, a user's current interest in the community will extend over time and impact the user's contributions in the following month. If the user is currently driven by community interest, the user will contribute in both seeking and providing categories, i.e. self-interest posts and community-interest posts. On the other hand, if the user focuses solely on his or her own needs, i.e. only seeks support for herself, the community-interest posts of the user will decrease in the following month (-0.165).

In terms of the structural social capital of a user, a positive coefficient of betweenness indicates that the more resources a user can access, the more contributions the user will make in the following month, whether in terms of community-interest or self-interest posts. Unexpectedly, closeness returns the opposite result. The values -0.01 and -0.002 show the negative effect of the closeness to a user's future contribution. In other words, co-participating behavior itself would not trigger more contributions from a user.

As for relational social capital, the positive value of recency of mutual response for both contributing categories suggests that the more recent a user's last action was, the higher the probability of continued contributions from the user in the next month. However, the frequency and number of interacting users of mutual responses produced different results for the two contributing categories. It seems that the number of mutual responses is not a positive indicator of a user's self-interest contributions, whereas the pool of unique users is. In other words, compared to getting a bulk of information from just one person, communicating with a larger number of users provides more valuable stimulus for a user to seek additional support.

By contrast, the more unique users with whom one has mutual contact, the lower the probability of the user contributing to the community will be (-2.137). However, if more others share similar visions with the user, it might trigger more community-interest posts (0.262). It is very unlikely that a set of occasional interactions between people will be the reason for a user contributing to the community, however, meeting users who share similar visions might trigger his or her desire to support others.

In summary, the social capital of a user may have an impact on the level of his or her future contributions in one or more ways. Multiple dimensions of one type of social capital have different impacts on a user's contributions in the following month. The preliminary results suggest that having access to more people, especially those who share similar visions, is positively related to a user's contribution in both community-interest and self-interest posts. In addition, users often intend to extend their original motivation for involvement, but their active behaviors attenuate over time.

Summary

This Chapter shows some preliminary results of the analysis of the relationship between a user's social capital and his or her future contributions in an OHC. The results reveal that different forms of social capital can impact a user's contribution to the community in various ways. Even among one category of the social capital, different aspects of the same variable may lead to altered results. I have to admit that the results are puzzling. The possible reasons include the multi-collinearity of the variables and the measurements I selected for the calculation of social capitals. After eliminating these puzzles, I will run the model one more time. The next stage of my research will involve building a predictive model for forecasting users' future contributions. My long-term research goal is to formulate interventions that will keep high-value users active in the OHC.

CHAPTER SEVEN

CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

Discussions and Conclusions

After analyzing large-scale data from a real-world OHC with various data analytics techniques, including text mining, survival analysis, fixed effects regression, and predictive modeling, I can tell a story of why, when, and how users actively participate in OHCs from the perspective of online social support.

First, users engaged in this breast cancer OHC is driven by different motivations, and as a result, they behave differently in seeking and providing various types of social support. Users motivated by altruism act as community-interest roles and provide social support to benefit others and community construction, while the self-interest users focus more on social support seeking. Such motivations and behaviors of users can evolve over time and be impacted by connected others in the social network. The findings are consistent with behavior or innovation diffusion of users in other social networks proposed by previous studies (Newman, 2002; Rogers, 1962), and also provide website managers suggestions in how to trigger more users' active community-centered behaviors.

Second, users' continued engagement in the OHC is related to both social identity and social bond (Back, 1951; J. S. Coleman, 1988). Specifically, both providing informational support and being involved in companionship posts are positive indicators of users' long-term engagement. Although the OHC is designed for sharing breast cancer related topics, the off-topics act an important role to produce cohesions within the community. The findings have implications for website operators in making interventions to retain OHC users.

Third, accessing to various social resources can affect users' contributions in different ways. Generally, different current social capital has distinct impacts on users' community-interest or self-interest contributions in the following month. The findings are partly consistent with past studies showing positive correlation between social capital and users' knowledge sharing in online communities (H. H. Chang & Chuang, 2011; Chiu-Ping Hsu, 2015; Sheng & Hartono, 2015). I also found that even if multiple dimensions of one type of social capital a user can acquire may also influence the users' contribution in one or more ways. The outcome is valuable in recommender system design of the OHC, such as recommending users who share similar visions to a target user may facilitate more community-interest content contributed the target.

The contributions of the thesis can be summarized from three perspectives:

(1) From a theoretical perspective, my study is motivated by and supports previous conceptual models or social science theories related to human behaviors. For example, by showing the connection between various types of social support activities and users' engagement in OHCs, I verified that OHCs combine both identity-based and bond-based participation, and social resources can affect a user's level of contributions to OHCs. In addition, the study finds for the first time that users' motivation for involvement is "contagious", and spreads through social network ties.

(2) From a methodological perspective, my research represents the first to differentiate the seeking and providing of various social support types from large-scale OHC data. Distinguishing the seeking and provisions of social support provides an opportunity to better understand the OHC users' engagement from multiple aspects. Meanwhile, these studies based on large-scale data of OHC users' actual behaviors complements studies based on traditional surveys and interviews.

(3) From a managerial perspective, the outcome of the study can provide OHC managers with suggestions on how to motivate user contributions and provide decision support that will assist with retaining users through interventions (e.g., information recommendation and email reminders). A sustainable and successful OHC will eventually be of benefit to those with the health concerns.

Limitations

I also would like to mention some limitations of my study. First, the sampled population may produce bias for the outcome of the explanatory model. Specifically, I conducted a case study on a breast cancer OHC, which might lead to the findings be tempered by the fact that only reflect what happens largely with older and female users in an OHC. The generalizability of the findings need to be verified by using the same approach on another dataset. As for the user role dynamics, I only examined two-step sequences of users' role transitions, i.e., switching from one role to another, but did not address longer sequences of user role evolutions from registration to churn (e.g., registration- IS-IP-AC-churn). Analyzing such sequences of consecutive transitions would help to reveal the full trajectories of users' behaviors in an OHC. In terms of users' continued participation analysis, I assumed that a user received indirect support when she or he read a thread initiated by another user and other users' replies to the thread. This approach of capturing indirect support received can be inaccurate: on one hand, I might underestimate the amount of support because I limited the calculation to threads a user replied to. In fact, a user can get indirect support by reading a thread without posting a reply. On the other hand, my approach can also overestimate such indirect support, because when posting to a long thread, a user may not have time to read all the previous replies, even though they were published within 7 days before the user's reply. This limitation can be addressed by analyzing users' click streams, but such data is not available for the public or this thesis. As for the correlation between social capital and users' future contributions. The findings might be influenced by the way I cut off users' profiles. Considering users' behavior in the OHC is an accumulative process, it seems more reasonable to measure aggregated resources a user can access in analyzing users' future contributions rather than only based on monthly data.

Future Work

There are also several other interesting directions for future research. First, detecting users' health status from their posts will be an interesting endeavor, as it not only can help understanding why a user becomes involved in an OHC, but also could potentially improve the recommendation and retrieval of online information. In addition, social support in OHCs can help users adjust to their life and survive, but is this kind of contribution only limited in their online performance? In other words, does receiving online support impact users' offline life? How does it influence users' offline behavior? To answer this question, it is hard to capture offline status only based on post and profile analysis, additional work such as survey and interview need to be done. Due to difficulties in implementing among large number of users, few studies showed the influence of online support for offline events. Therefore, associating online data analysis and offline behavior would be another interesting topic in the future as well.

In closing, an important reason for studying OHCs is to create valuable outcomes that will enable OHC operators to better design and manage these communities. This will mean the OHCs can better serve the users who suffer from their health issues. Although this study has answered some questions, there is much more work to be done. I intend to continue my research in this area to keep building bridges between information science and human beings.

REFERENCES

Ali, T., Schramm, D., Sokolova, M., & Inkpen, D. (2013). Can I Hear You? Sentiment Analysis on Medical Forums. In *IJCNLP* (pp. 667–673). Retrieved from http://www.anthology.aclweb.org/I/I13/I13-1077.pdf

Aral, S., & Walker, D. (2012). Identifying Influential and Susceptible Members of Social Networks. *Science*, *337*(6092), 337–341.

Back, K. W. (1951). Influence through social communication. *The Journal of Abnormal and Social Psychology*, *46*(1), 9–23. https://doi.org/10.1037/h0058629

Backiel, A., Baesens, B., & Claeskens, G. (2014). Mining Telecommunication Networks to Enhance Customer Lifetime Predictions. In L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, & J. M. Zurada (Eds.), *Artificial Intelligence and Soft Computing: 13th International Conference, ICAISC 2014, Zakopane, Poland, June 1-5, 2014, Proceedings, Part II* (pp. 15–26). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-07176-3_2

Bambina, A. (2007). *Online Social Support: The Interplay of Social Networks and Computer-Mediated Communication*. Cambria Press.

Barak, A., Boniel-Nissim, M., & Suler, J. (2008). Fostering empowerment in online support groups. *Computers in Human Behavior*, *24*(5), 1867–1883. https://doi.org/10.1016/j.chb.2008.02.004

Barbieri, N., Bonchi, F., & Manco, G. (2014). Who to Follow and Why: Link Prediction with Explanations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1266–1275). New York, NY, USA: ACM. https://doi.org/10.1145/2623330.2623733

Barrera, M., & Ainlay, S. L. (1983). The structure of social support: a conceptual and empirical analysis. *Journal of Community Psychology*, *11*(2), 133–143.

Bender, J. L., Jimenez-Marroquin, M.-C., & Jadad, A. R. (2011). Seeking Support on Facebook: A Content Analysis of Breast Cancer Groups. *Journal of Medical Internet Research*, *13*(1). https://doi.org/10.2196/jmir.1560

Berson, A., Smith, S., & Thearling, K. (1999). *Building Data Mining Applications for CRM* (1st ed.). McGraw-Hill Professional.

Bin, L., Peiji, S., & Juan, L. (2007). Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service. In *2007 International Conference on Service Systems and Service Management* (pp. 1–5). https://doi.org/10.1109/ICSSSM.2007.4280145

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421

Borgatti, S. P., Jones, C., & Everett, M. G. (1998). Network measures of social capital. *Connections*, *21*(2), 27–36.

Borko, H. (1968). Information science: What is it? *American Documentation*, *19*(1), 3–5. https://doi.org/10.1002/asi.5090190103

Bouma, G., Admiraal, J. M., de Vries, E. G. E., Schröder, C. P., Walenkamp, A. M. E., & Reyners, A. K. L. (2015). Internet-based support programs to alleviate psychosocial and physical symptoms in cancer patients: A literature analysis. *Critical Reviews in Oncology/Hematology*, *95*(1), 26–37. https://doi.org/10.1016/j.critrevonc.2015.01.011

Bourdieu, P., & Wacquant, L. J. D. (1992). *An Invitation to Reflexive Sociology* (1st edition). Chicago: University Of Chicago Press.

Bouty, I. (2000). Interpersonal and Interaction Influences on Informal Resource Exchanges between R&D Researchers across Organizational Boundaries. *The Academy of Management Journal*, *43*(1), 50–65. https://doi.org/10.2307/1556385

Buckinx, W., & Poel, D. V. den. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual {FMCG} retail setting. *European Journal of Operational Research*, *164*(1), 252–268. https://doi.org/http://dx.doi.org/10.1016/j.ejor.2003.12.010

Bui, N., Yen, J., & Honavar, V. (2015). Temporal Causality of Social Support in an Online Community for Cancer Survivors. In N. Agarwal, K. Xu, & N. Osgood (Eds.), *Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 13–23). Springer International Publishing. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-16268-3_2

Burke, M., Marlow, C., & Lento, T. (2009). Feed me: motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 945–954). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1518847

Burrows, R., Nettleton, S., Pleace, N., Loader, B., & Muncer, S. (2000). VIRTUAL COMMUNITY CARE? SOCIAL POLICY AND THE EMERGENCE OF COMPUTER MEDIATED SOCIAL SUPPORT. *Information, Communication & Society*, *3*(1), 95–121. https://doi.org/10.1080/136911800359446

Burt, R. S. (1995). *Structural Holes: The Social Structure of Competition* (1st Paperback Edition edition). Cambridge, Mass.: Harvard University Press.

Burt, R. S. (2001). Structural Holes versus Network Closure as Social Capital. *Social Capital: Theory and Research*, 31–55.

Campbell, H. S., Phaneuf, M. R., & Deane, K. (2004). Cancer peer support programs—do they work? *Patient Education and Counseling*, *55*(1), 3–15. https://doi.org/10.1016/j.pec.2003.10.001

Caplan, S. E., & Turner, J. S. (2007). Bringing theory to research on computer-mediated comforting communication. *Computers in Human Behavior*, *23*(2), 985–998. https://doi.org/10.1016/j.chb.2005.08.003

Cassel, J. (1976). The Contribution of the Social Environment to Host Resistance the Fourth Wade Hampton Frost Lecture. *American Journal of Epidemiology*, *104*(2), 107–123.

Castleton, K., Fong, T., Wang-Gillam, A., Waqar, M. A., Jeffe, D. B., Kehlenbrink, L., … Govindan, R. (2011). A survey of Internet utilization among patients with cancer. *Supportive Care in Cancer*, *19*(8), 1183–1190. https://doi.org/10.1007/s00520-010-0935-5

Centola, D. (2011). An experimental study of homophily in the adoption of health behavior. *Science*, *334*(6060), 1269–1272.

Centola, D., & van de Rijt, A. (2014). Choosing your network: Social preferences in an online health community. *Social Science & Medicine*.

Chang, H. H., & Chuang, S.-S. (2011). Social capital and individual motivations on knowledge sharing: Participant involvement as a moderator. *Information & Management*, *48*(1), 9–18. https://doi.org/10.1016/j.im.2010.11.001

Chang, H.-J. (2009). Online supportive interactions: Using a network approach to examine communication patterns within a psychosis social support group in Taiwan. *Journal of the American Society for Information Science & Technology*, *60*(7), 1504–1517. https://doi.org/10.1002/asi.21070

Chau, M., & Xu, J. (2012). Business Intelligence in Blogs: Understanding Consumer Interactions and Communities. *MIS Q.*, *36*(4), 1189–1216.

Chen, A. T. (2012). Exploring online support spaces: Using cluster analysis to examine breast cancer, diabetes and fibromyalgia support groups. *Patient Education and Counseling*, *87*(2), 250–257. https://doi.org/10.1016/j.pec.2011.08.017

Chen, H., Chiang, R., & Storey, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *Management Information Systems Quarterly*, *36*(4), 1165–1188.

Chen, J., Xu, H., & Whinston, A. B. (2011). Moderated Online Communities and Quality of User-Generated Content. *Journal of Management Information Systems*, *28*(2), 237–268. https://doi.org/10.2753/MIS0742-1222280209

Chiu, C.-M., Hsu, M.-H., & Wang, E. T. G. (2006). Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decision Support Systems*, *42*(3), 1872–1888. https://doi.org/10.1016/j.dss.2006.04.001

Chiu-Ping Hsu. (2015). Effects of social capital on online knowledge sharing: positive and negative perspectives. *Online Information Review*, *39*(4), 466–484. https://doi.org/10.1108/OIR-12-2014-0314

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, *357*(4), 370–379.

Chuang, K. Y., & Yang, C. C. (2012). Interaction Patterns of Nurturant Support Exchanged in Online Health Social Networking. *Journal of Medical Internet Research*, *14*(3), e54. https://doi.org/10.2196/jmir.1824

Cobb, N. K., Graham, A. L., & Abrams, D. B. (2010). Social Network Structure of a Large Online Community for Smoking Cessation. *American Journal of Public Health*, *100*(7), 1282–1289. https://doi.org/10.2105/AJPH.2009.165449

Coleman, J. (1990). Foundations of Social Theory. Retrieved from http://www.citeulike.org/group/1702/article/1338351

Coleman, J., Olsen, S. J., Sauter, P. K., Baker, D., Hodgin, M. B., Stanfield, C., … Nolan, M. T. (2005). The effect of a Frequently Asked Questions module on a pancreatic cancer Web site patient/family chat room. *Cancer Nursing*, *28*(6), 460–468.

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology*, *94*, S95–S120.

Coulson, N. S. (2005). Receiving social support online: an analysis of a computer-mediated support group for individuals living with irritable bowel syndrome. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, *8*(6), 580–584. https://doi.org/10.1089/cpb.2005.8.580

Coulson, N. S., Buchanan, H., & Aubeeluck, A. (2007). Social support in cyberspace: a content analysis of communication within a Huntington's disease online support group. *Patient Education and Counseling*, *68*(2), 173–178. https://doi.org/10.1016/j.pec.2007.06.002

Coulson, N. S., & Shaw, R. L. (2013). Nurturing health-related online support groups: Exploring the experiences of patient moderators. *Computers in Human Behavior*, *29*(4), 1695–1701. https://doi.org/10.1016/j.chb.2013.02.003

Coussement, K., & Poel, D. V. den. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, *34*(1), 313–327. https://doi.org/http://dx.doi.org/10.1016/j.eswa.2006.09.038

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227.

Dunkel-Schetter, C. (1984). Social support and cancer: Findings based on patient interviews and their implications. *Journal of Social Issues*, *40*(4), 77–98.

Durkheim, E. (1897). Suicide: A study in sociology.

Eichhorn, K. C. (2008). Soliciting and Providing Social Support Over the Internet: An Investigation of Online Eating Disorder Support Groups. *Kontaktanbahnung Und Soziale Unterstützung Mit Hilfe Des Internets: Eine Untersuchung von Online-Selbsthilfegruppen Zum Thema Essstörungen.*, *14*(1), 67–78. https://doi.org/10.1111/j.1083-6101.2008.01431.x

Ellison, N. B., & Vitak, J. (2015). Social Network Site Affordances and Their Relationship to Social Capital Processes. In S. S. Sundar (Ed.), *The Handbook of the Psychology of Communication Technology* (pp. 203–227). John Wiley & Sons, Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118426456.ch9/summary

Ellison, N., Lampe, C., Steinfield, C., & Vitak, J. (2011). With a little help from my friends: How social network sites affect social capital processes. *A Networked Self: Identity, Community, and Culture on Social Network Sites*, 124–145.

Eysenbach, G., Powell, J., Englesakis, M., Rizo, C., & Stern, A. (2004). Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *BMJ*, *328*(7449), 1166. https://doi.org/10.1136/bmj.328.7449.1166

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using Iso-Value Curves for Customer Base Analysis. *Journal of Marketing Research*, *42*(4), 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

Faraj, S., Kudaravalli, S., & Wasko, M. (2015). Leading Collaboration in Online Communities. *Management Information Systems Quarterly*, *39*(2), 393–412.

Feffer, M. (2015, July 23). Using Stack Overflow in Your Job Search. Retrieved September 12, 2016, from http://insights.dice.com/2015/07/23/using-stack-overflow-in-your-job-search/

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791. https://doi.org/10.1038/nature02043

Ferguson, T. (1996). *Health Online*. Reading, Mass: Addison Wesley.

Fox, S. (2014, January 15). The social life of health information. Retrieved June 27, 2016, from http://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-of-health-information/

Füller, J., Hutter, K., Hautz, J., & Matzler, K. (2014). User Roles and Contributions in Innovation-Contest Communities. *Journal of Management Information Systems*, *31*(1), 273–308. https://doi.org/10.2753/MIS0742-1222310111

Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., O'Connor, K., Sarker, A., … Gonzalez, G. (2014). Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.672.9123&rep=rep1&type=pdf

Ginossar, T. (2008). Online participation: a content analysis of differences in utilization of two online cancer communities by men and women, patients and family members. *Health Communication*, *23*(1), 1–12. https://doi.org/10.1080/10410230701697100

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, *24*(3), 153–172. https://doi.org/10.1016/S1090-5138(02)00157-5

Glanz, K., Rimer, B. K., & Viswanath, K. (Eds.). (2008). *Health Behavior and Health Education: Theory, Research, and Practice* (4 edition). San Francisco, CA: Jossey-Bass.

Goh, J. M., Gao, G. G., & Agarwal, R. (2016). The creation of social value: Can an online health community reduce rural–urban health disparities? *Management Information Systems Quarterly*, *40*(1), 247–263.

Gorlick, A., Bantum, E. O., & Owen, J. E. (2014). Internet-based interventions for cancer-related distress: exploring the experiences of those whose needs are not met. *Psycho-Oncology*, *23*(4), 452–458. https://doi.org/10.1002/pon.3443

Greene, J. A., Choudhry, N. K., Kilabuk, E., & Shrank, W. H. (2011). Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. *Journal of General Internal Medicine*, *26*(3), 287–292. https://doi.org/10.1007/s11606-010-1526-3

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hambly, K. (2011). Activity Profile of Members of an Online Health Community After Articular Cartilage Repair of the Knee. *Sports Health*, *3*(3), 275–282. https://doi.org/10.1177/1941738111402151

Ho, Y.-X., O'Connor, B. H., & Mulvaney, S. A. (2014). Features of Online Health Communities for Adolescents With Type 1 Diabetes. *Western Journal of Nursing Research*, *36*(9), 1183–1198. https://doi.org/10.1177/0193945913520414

Hoey, L. M., Ieropoli, S. C., White, V. M., & Jefford, M. (2008). Systematic review of peer-support programs for people with cancer. *Patient Education and Counseling*, *70*(3), 315–337. https://doi.org/10.1016/j.pec.2007.11.016

House, J. S. (1981). *Work stress and social support*. Addison-Wesley Longman, Incorporated.

Hsu, C.-L., & Lin, J. C.-C. (2008). Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation. *Information & Management*, *45*(1), 65–74. https://doi.org/10.1016/j.im.2007.11.001

Huang, B. Q., Kechadi, M.-T., & Buckley, B. (2009). Customer Churn Prediction for Broadband Internet Services. In T. B. Pedersen, M. K. Mohania, & A. M. Tjoa (Eds.), *Data Warehousing and Knowledge Discovery: 11th International Conference, DaWaK 2009 Linz, Austria, August 31–September 2, 2009 Proceedings* (pp. 229–243). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-03730-6_19

Huh, J., Yetisgen-Yildiz, M., & Pratt, W. (2013). Text classification for assisting moderators in online health communities. *Journal of Biomedical Informatics*, *46*(6), 998–1005. https://doi.org/10.1016/j.jbi.2013.08.011

Idriss, S., Kvedar, J., & Watson, A. (2009). The role of online support communities: Benefits of expanded social networks to patients with psoriasis. *Archives of Dermatology*, *145*(1), 46–51. https://doi.org/10.1001/archdermatol.2008.529

Iriberri, A., & Leroy, G. (2009). A Life-cycle Perspective on Online Community Success. *ACM Comput. Surv.*, *41*(2), 11:1–11:29. https://doi.org/10.1145/1459352.1459356

Jayanti, R. K., & Singh, J. (2010). Pragmatic learning theory: An inquiry-action framework for distributed consumer learning in online communities. *Journal of Consumer Research*, *36*(6), 1058–1081.

Jiang, L., & Yang, C. C. (2013). Using Co-occurrence Analysis to Expand Consumer Health Vocabularies from Social Media Data. In *2013 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 74–81). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6680463

Jiang, L., & Yang, C. C. (2015). Expanding Consumer Health Vocabularies by Learning Consumer Health Expressions from Online Health Social Media. In N. Agarwal, K. Xu, & N. Osgood (Eds.), *Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 314–320). Springer International Publishing. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-16268-3_36

Jones, R., Sharkey, S., Smithson, J., Ford, T., Emmens, T., Hewis, E., … Owens, C. (2011). Using metrics to describe the participative stances of members within discussion forums. *Journal of Medical Internet Research*, *13*(1), e3. https://doi.org/10.2196/jmir.1591

Kastellec, J. P., & Leoni, E. L. (2007). Using Graphs Instead of Tables in Political Science. *Perspectives on Politics*, *5*(4), 755–771. https://doi.org/10.1017/S1537592707072209

Kawale, J., Pal, A., & Srivastava, J. (2009). Churn Prediction in MMORPGs: A Social Influence Based Approach. In *2009 International Conference on Computational Science and Engineering* (Vol. 4, pp. 423–428). https://doi.org/10.1109/CSE.2009.80

Keating, D. M. (2013). Spirituality and support: a descriptive analysis of online social support for depression. *Journal of Religion and Health*, *52*(3), 1014–1028. https://doi.org/10.1007/s10943-012-9577-x

Kim, A. J. (2000). *Community Building on the Web: Secret Strategies for Successful Online Communities* (1st ed.). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Kim, E., Han, J. Y., Moon, T. J., Shaw, B., Shah, D. V., McTavish, F. M., & Gustafson, D. H. (2012). The process and effect of supportive message expression and reception in online breast cancer support groups. *Psycho-Oncology*, *21*(5), 531–540. https://doi.org/10.1002/pon.1942

Kinnane, N. A., & Milne, D. J. (2010). The role of the Internet in supporting and informing carers of people with cancer: a literature review. *Supportive Care in Cancer: Official Journal of the Multinational Association of Supportive Care in Cancer*, *18*(9), 1123–1136. https://doi.org/10.1007/s00520-010-0863-4

Kirk, S., & Milnes, L. (2016). An exploration of how young people and parents use online support in the context of living with cystic fibrosis. *Health Expectations: An International Journal of Public Participation in Health Care and Health Policy*, *19*(2), 309–321.

Kling, R. (2007). What Is Social Informatics and Why Does It Matter? *The Information Society*, *23*(4), 205–220. https://doi.org/10.1080/01972240701441556

Kollock, P. (1999). The Economies of Online Cooperation: Gifts and Public Goods in Cyberspace. In P. Kollock & M. Smith (Eds.), *Communities in Cyberspace* (pp. 220–239). 11 New Fetter Lane, London EC4P 4EE: Routledge. Retrieved from http://www.sscnet.ucla.edu/soc/faculty/kollock/papers/economies.htm

Krause, N. (1986). Social support, stress, and well-being among older adults. *Journal of Gerontology*, *41*(4), 512–519.

Kraut, R. E., Resnick, P., Kiesler, S., Burke, M., Chen, Y., Kittur, N., … Riedl, J. (2012). *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press. Retrieved from https://books.google.com/books?id=lIvBMYVxWJYC

Leimeister, J. M., Ebner, W., & Krcmar, H. (2005). Design, Implementation, and Evaluation of Trust-Supporting Components in Virtual Communities for Patients. *Journal of Management Information Systems*, *21*(4), 101–131. https://doi.org/10.1080/07421222.2005.11045825

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 497–506). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1557077

Lester, J., Prady, S., Finegan, Y., & Hoch, D. (2004). Learning from e-patients at Massachusetts General Hospital. *BMJ*, *328*(7449), 1188–1190. https://doi.org/10.1136/bmj.328.7449.1188

Levine, J. M., & Moreland, R. L. (1994). Group Socialization: Theory and Research. *European Review of Social Psychology*, *5*(1), 305–336. https://doi.org/10.1080/14792779543000093

Lieberman, M. (2007). The role of insightful disclosure in outcomes for women in peer-directed breast cancer groups: a replication study. *Psycho-Oncology*, *16*(10). Retrieved from http://escholarship.org/uc/item/8rw0v48t

Lin, N. (2001). *Social Capital: A Theory of Social Structure and Action*. Cambridge University Press.

Loane, S. S., & D'Alessandro, S. (2013). Communication That Changes Lives: Social Support Within an Online Health Community for ALS. *Communication Quarterly*, *61*(2), 236–251. https://doi.org/10.1080/01463373.2012.752397

Lu, Y. (2013). Automatic topic identification of health-related messages in online health community using text classification. *SpringerPlus*, *2*(1), 309. https://doi.org/10.1186/2193-1801-2-309

Lu, Y., Zhang, P., & Deng, S. (2013). Exploring Health-Related Topics in Online Health Community Using Cluster Analysis. In *2013 46th Hawaii International Conference on System Sciences (HICSS)* (pp. 802–811). https://doi.org/10.1109/HICSS.2013.216

MacLean, D., Gupta, S., Lembke, A., Manning, C., & Heer, J. (2015). Forum77: An analysis of an online health forum dedicated to addiction recovery. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1511–1526). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2675146

Maloney-Krichmar, D., & Preece, J. (2005). A multilevel analysis of sociability, usability, and community dynamics in an online health community. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *12*(2), 201–232.

Matzat, U., & Rooks, G. (2014). Styles of moderation in online health and support communities: An experimental comparison of their acceptance and effectiveness. *Computers in Human Behavior*, *36*, 65–75. https://doi.org/10.1016/j.chb.2014.03.043

McClellan, W. M., Stanwyck, D. J., & Anson, C. A. (1993). Social support and subsequent mortality among patients with end-stage renal disease. *Journal of the American Society of Nephrology*, *4*(4), 1028–1034.

McLeroy, K. R., Bibeau, D., Steckler, A., & Glanz, K. (1988). An ecological perspective on health promotion programs. *Health Education Quarterly*, *15*(4), 351–377.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 415–444.

Mo, P. K. H., & Coulson, N. S. (2010). Empowering Processes in Online Support Groups Among People Living with HIV/AIDS: A Comparative Analysis of “Lurkers” and “Posters.” *Comput. Hum. Behav.*, *26*(5), 1183–1193. https://doi.org/10.1016/j.chb.2010.03.028

Mo, P. K. H., & Coulson, N. S. (2013). Online support group use and psychological health for individuals living with HIV/AIDS. *Patient Education and Counseling*, *93*(3), 426–432. https://doi.org/10.1016/j.pec.2013.04.004

Monnier, J., Laken, M., & Carter, C. L. (2002). Patient and Caregiver Interest in Internet-Based Cancer Services. *Cancer Practice*, *10*(6), 305–310. https://doi.org/10.1046/j.1523-5394.2002.106005.x

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, *11*(3), 690–696. https://doi.org/10.1109/72.846740

Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science*, *341*(6146), 647–651.

Nahapiet, J., & Ghoshal, S. (1998). Social Capital, Intellectual Capital, and the Organizational Advantage. *Academy of Management Review*, *23*(2), 242–266. https://doi.org/10.5465/AMR.1998.533225

Newman, M. E. (2002). Spread of epidemic disease on networks. *Physical Review E*, *66*(1), 16128.

O'Connor, E., Gaynes, B. N., Burda, B. U., Soh, C., & Whitlock, E. P. (2013). Screening for and treatment of suicide risk relevant to primary care: a systematic review for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, *158*(10), 741–754. https://doi.org/10.7326/0003-4819-158-10-201305210-00642

Oh, S. (2012). The characteristics and motivations of health answerers for sharing information, knowledge, and experiences in online environments. *Journal of the American Society for Information Science and Technology*, *63*(3), 543–557. https://doi.org/10.1002/asi.21676

Patki, A., Sarker, A., Pimpalkhute, P., Nikfarjam, A., Ginn, R., O'Connor, K., … Gonzalez, G. (2014). Mining adverse drug reaction signals from social media: going beyond extraction. Retrieved from http://phenoday2014.bio-lark.org/pdf/2.pdf

Pearson, J. L., Amato, M. S., Wang, X., Zhao, K., Cha, S., Cohn, A. M., … Graham, A. L. (2017). How US Smokers Refer to E-cigarettes: An Examination of User-Generated Posts From a Web-Based Smoking Cessation Intervention, 2008–2015. *Nicotine & Tobacco Research*, *19*(2), 253–257. https://doi.org/10.1093/ntr/ntw206

Petrovčič, A., & Petrič, G. (2014). Differences in intrapersonal and interactional empowerment between lurkers and posters in health-related online support communities. *Computers in Human Behavior*, *34*, 39–48. https://doi.org/10.1016/j.chb.2014.01.008

Pfeil, U. (2009). Online Support Communities. In *Social Computing and Virtual Communities* (Vols. 1–0, pp. 121–150). Chapman and Hall/CRC. Retrieved from http://www.crcnetbase.com/doi/abs/10.1201/9781420090437-c6

Pfeil, U., & Zaphiris, P. (2007). Patterns of Empathy in Online Communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 919–928). New York, NY, USA: ACM. https://doi.org/10.1145/1240624.1240763

Prasarnphanich, P., & Wagner, C. (2009). The Role of Wiki Technology and Altruism in Collaborative Knowledge Creation. *The Journal of Computer Information Systems*, *49*(4), 33–41.

Preece, J. (2000). *Online Communities: Designing Usability and Supporting Sociability* (1 edition). New York: Wiley.

Preece, J., & Shneiderman, B. (2009). The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation. *AIS Transactions on Human-Computer Interaction*, *1*(1), 13–32.

Putnam, R. D. (2001). *Bowling Alone: The Collapse and Revival of American Community* (1st edition). New York: Touchstone Books by Simon & Schuster.

Qiu, B., Zhao, K., Mitra, P., Wu, D., Caragea, C., Yen, J., … Portier, K. (2011). Get Online Support, Feel Better – Sentiment Analysis and Dynamics in an Online Cancer Survivor Community. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)* (pp. 274–281). https://doi.org/10.1109/PASSAT/SocialCom.2011.127

Rodgers, S., & Chen, Q. (2005). Internet community group participation: Psychosocial benefits for women with breast cancer. *Journal of Computer-Mediated Communication*, *10*(4), 0–0.

Rogers, E. M. (1962). *Diffusion of innovations.* New York: Free Press of Glencoe.

Rupert, D. J., Moultrie, R. R., Read, J. G., Amoozegar, J. B., Bornkessel, A. S., O'Donoghue, A. C., & Sullivan, H. W. (2014). Perceived healthcare provider reactions to patient and caregiver use of online health communities. *Patient Education and Counseling*, *96*(3), 320–326. https://doi.org/10.1016/j.pec.2014.05.015

Sachan, M., Contractor, D., Faruquie, T. A., & Subramaniam, L. V. (2012). Using Content and Interactions for Discovering Communities in Social Networks. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 331–340). New York, NY, USA: ACM. https://doi.org/10.1145/2187836.2187882

Sarker, A., Nikfarjam, A., O'Connor, K., Ginn, R., Gonzalez, G., Upadhaya, T., … Smith, K. (2015). Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2015.02.004

Setoyama, Y., Yamazaki, Y., & Namayama, K. (2011). Benefits of Peer Support in Online Japanese Breast Cancer Communities: Differences Between Lurkers and Posters. *Journal of Medical Internet Research*, *13*(4). https://doi.org/10.2196/jmir.1696

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shapiro, C., & Varian, H. R. (1999). *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business Press.

Shaw, B. R., Hawkins, R., McTavish, F., Pingree, S., & Gustafson, D. H. (2006). Effects of insightful disclosure within computer mediated support groups on women with breast cancer. *Health Communication*, *19*(2), 133–142. https://doi.org/10.1207/s15327027hc1902_5

Sheng, M., & Hartono, R. (2015). An exploratory study of knowledge creation and sharing in online community: a social capital perspective. *Total Quality Management & Business Excellence*, *26*(1–2), 93–107. https://doi.org/10.1080/14783363.2013.776769

Shmueli, G., & Koppius, O. R. (2011). Predictive Analytics in Information Systems Research. *MIS Q.*, *35*(3), 553–572.

Shumaker, S. A., & Brownell, A. (1984). Toward a theory of social support: Closing conceptual gaps. *Journal of Social Issues*, *40*(4), 11–36.

Sproull, L., & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science*, *32*(11), 1492–1512.

Sudau, F., Friede, T., Grabowski, J., Koschack, J., Makedonski, P., & Himmel, W. (2014). Sources of Information and Behavioral Patterns in Online Health Forums: Observational Study. *Journal of Medical Internet Research*, *16*(1). https://doi.org/10.2196/jmir.2875

Tang, J.-H., & Yang, H.-L. (2005). User role and perception of requirements in a web-based community of practice. *Online Information Review*, *29*(5), 499–512.

Tang, X., & Yang, C. C. (2012). Ranking user influence in healthcare social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(4), 73.

Tang, X., Zhang, M., & Yang, C. C. (2012). User Interest and Topic Detection for Personalized Recommendation. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01* (pp. 442–446). IEEE Computer Society. Retrieved from http://dl.acm.org/citation.cfm?id=2457627

Tsai, C.-F., & Lu, Y.-H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, *36*(10), 12547–12553. https://doi.org/http://dx.doi.org/10.1016/j.eswa.2009.05.032

Turnock, B. J. (2015). *Public Health: What It Is and How It Works* (6 edition). Burlington, Massachusetts: Jones & Bartlett Learning.

Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. *Social Networks*, *18*(1), 69–89.

van der Eijk, M., Faber, M. J., Aarts, J. W., Kremer, J. A., Munneke, M., & Bloem, B. R. (2013). Using Online Health Communities to Deliver Patient-Centered Care to People With Chronic Conditions. *Journal of Medical Internet Research*, *15*(6), e115. https://doi.org/10.2196/jmir.2476

Van den Bulte, C., & Lilien, G. L. (2001). Medical Innovation Revisited: Social Contagion versus Marketing Effort. *American Journal of Sociology*, *106*(5), 1409–1435. https://doi.org/10.1086/320819

Walther, J. B. (1996). Computer-Mediated Communication Impersonal, Interpersonal, and Hyperpersonal Interaction. *Communication Research*, *23*(1), 3–43. https://doi.org/10.1177/009365096023001001

Walther, J. B., Van Der Heide, B., Ramirez, A., Burgoon, J. K., & Peña, J. (2015). Interpersonal and Hyperpersonal Dimensions of Computer-Mediated Communication. In S. S. Sundar (Ed.), *The Handbook of the Psychology of Communication Technology* (pp. 1–22). John Wiley & Sons, Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118426456.ch1/summary

Wang, Y.-C., Kraut, R. E., & Levine, J. M. (2015). Eliciting and receiving online support: using computer-aided content analysis to examine the dynamics of online social support. *Journal of Medical Internet Research*, *17*(4), e99. https://doi.org/10.2196/jmir.3558

Wang, Y.-C., Kraut, R., & Levine, J. M. (2012). To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 833–842). New York, NY, USA: ACM. https://doi.org/10.1145/2145204.2145329

Wasko, M. M., & Faraj, S. (2005). Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice. *MIS Quarterly*, *29*(1), 35–57.

Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, *34*(4), 441–458.

Wei, C.-P., & Chiu, I.-T. (2002). Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications*, *23*(2), 103–112. https://doi.org/http://dx.doi.org/10.1016/S0957-4174(02)00030-1

Weitzman, E. R., Cole, E., Kaci, L., & Mandl, K. D. (2011). Social but safe? Quality and safety of diabetes-related online social networks. *Journal of the American Medical Informatics Association*, *18*(3), 292–297. https://doi.org/10.1136/jamia.2010.009712

Wen, K.-Y., McTavish, F., Kreps, G., Wise, M., & Gustafson, D. (2011). From Diagnosis to Death: A Case Study of Coping With Breast Cancer as Seen Through Online Discussion Group Messages. *Journal of Computer-Mediated Communication*, *16*(2), 331–361. https://doi.org/10.1111/j.1083-6101.2011.01542.x

Wen, M., & Rose, C. P. (2012). Understanding Participant Behavior Trajectories in Online Health Support Groups Using Automatic Extraction Methods. In *Proceedings of the 17th ACM International Conference on Supporting Group Work* (pp. 179–188). New York, NY, USA: ACM. https://doi.org/10.1145/2389176.2389205

Wentzer, H. S., & Bygholm, A. (2013). Narratives of empowerment and compliance: Studies of communication in online patient support groups. *International Journal of Medical Informatics*, *82*(12), e386–e394. https://doi.org/10.1016/j.ijmedinf.2013.01.008

White, M., & Dorman, S. M. (2001). Receiving social support online: implications for health education. *Health Education Research*, *16*(6), 693–707. https://doi.org/10.1093/her/16.6.693

Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., … Heywood, J. (2010). Sharing Health Data for Better Outcomes on PatientsLikeMe. *Journal of Medical Internet Research*, *12*(2). https://doi.org/10.2196/jmir.1549

Williams, R. L., & Cothrel, J. (2000). Four smart ways to run online communities. *MIT Sloan Management Review*, *41*(4), 81.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 347–354). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1220575.1220619

Wright, K. B. (2002). Social support within an on-line cancer community: an assessment of emotional support, perceptions of advantages and disadvantages, and motives for using the community from a communication perspective. *Journal of Applied Communication Research*, *30*(3), 195–209. https://doi.org/10.1080/00909880216586

Wright, K. B. (2015). Computer-Mediated Support for Health Outcomes. In S. S. Sundar (Ed.), *The Handbook of the Psychology of Communication Technology* (pp. 488–506). John Wiley & Sons, Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118426456.ch22/summary

Wright, K., & Bell. (2003). Health-related Support Groups on the Internet: Linking Empirical Findings to Social Support and Computer-mediated Communication Theory. *Journal of Health Psychology*, *8*(1), 39–54. https://doi.org/10.1177/1359105303008001429

Wu, B., & Peng, Y. (2015). Sentiment Analysis in the Online Health Community. Atlantis Press. https://doi.org/10.2991/ic3me-15.2015.254

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, *36*(3, Part 1), 5445–5449. https://doi.org/http://dx.doi.org/10.1016/j.eswa.2008.06.121

Yan, L., & Tan, Y. (2014). Feeling Blue? Go Online: An Empirical Study of Social Support Among Patients. *Information Systems Research*, *25*(4), 690–709. https://doi.org/10.1287/isre.2014.0538

Yang, C. C., Yang, H., Jiang, L., & Zhang, M. (2012). Social media mining for drug safety signal detection. In *Proceedings of the 2012 international workshop on Smart health and wellbeing* (pp. 33–40). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2389714

Young, C. (2013). Community Management That Works: How to Build and Sustain a Thriving Online Health Community. *Journal of Medical Internet Research*, *15*(6). https://doi.org/10.2196/jmir.2501

Zhang, M., & Yang, C. C. (2014). Classification of Online Health Discussions with Text and Health Feature Sets. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*. Retrieved from http://www.aaai.org/ocs/index.php/WS/AAAIW14/paper/view/8739

Zhang, M., & Yang, C. C. (2015). Using content and network analysis to understand the social support exchange patterns and user behaviors of an online smoking cessation intervention program. *Journal of the Association for Information Science and Technology*, *66*(3), 564–575. https://doi.org/10.1002/asi.23189

Zhang, M., Yang, C. C., & Gong, X. (2013). Social Support and Exchange Patterns in an Online Smoking Cessation Intervention Program. In *2013 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 219–228). https://doi.org/10.1109/ICHI.2013.37

Zhang, S., Bantum, E., Owen, J., & Elhadad, N. (2014). Does Sustained Participation in an Online Health Community Affect Sentiment? *AMIA Annual Symposium Proceedings*, *2014*, 1970–1979.

Zhang, X., Zhu, J., Xu, S., & Wan, Y. (2012). Predicting Customer Churn Through Interpersonal Influence. *Know.-Based Syst.*, *28*, 97–104. https://doi.org/10.1016/j.knosys.2011.12.005

Zhang, Y. (2010). Contextualizing Consumer Health Information Searching: An Analysis of Questions in a Social Q&#38;A Community. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 210–219). New York, NY, USA: ACM. https://doi.org/10.1145/1882992.1883023

Zhang, Y., He, D., & Sang, Y. (2013). Facebook as a Platform for Health Information and Communication: A Case Study of a Diabetes Group. *Journal of Medical Systems*, *37*(3), 9942. https://doi.org/10.1007/s10916-013-9942-7

Zhao, J., Ha, S., & Widdows, R. (2016). The Influence of Social Capital on Knowledge Creation in Online Health Communities. *Inf. Technol. and Management*, *17*(4), 311–321. https://doi.org/10.1007/s10799-014-0211-3

Zhao, K., Greer, G. E., Yen, J., Mitra, P., & Portier, K. (2015). Leader identification in an online health community for cancer survivors: a social network-based classification approach. *Information Systems and E-Business Management*, *13*(4), 629–645.

Zhao, K., & Kumar, A. (2013). Who blogs what: understanding the publishing behavior of bloggers. *World Wide Web*, *16*(5–6), 621–644. https://doi.org/10.1007/s11280-012-0167-3

Zhao, K., Wang, X., Cha, S., Cohn, A. M., Papandonatos, G. D., Amato, M. S., … Graham, A. L. (2016). A Multirelational Social Network Analysis of an Online Health Community for Smoking Cessation. *Journal of Medical Internet Research*, *18*(8), e233. https://doi.org/10.2196/jmir.5985

Zhao, K., Yen, J., Greer, G., Qiu, B., Mitra, P., & Portier, K. (2014). Finding influential users of online health communities: a new metric based on sentiment influence. *Journal of the American Medical Informatics Association: JAMIA*, *21*(e2), e212-218. https://doi.org/10.1136/amiajnl-2013-002282

Zhao, M., & Yang, C. C. (2016). Mining Online Heterogeneous Healthcare Networks for Drug Repositioning. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 106–112). https://doi.org/10.1109/ICHI.2016.18