Social Support And User Engagement In Online Health Communities

Xi Wang, Kang Zhao, and Nick Street
{xi-wang-1, kang-zhao, nick-street}@uiowa.edu
The University of Iowa

Abstract. Online health communities (OHCs) have become a major source of social support for people with health problems. Members of OHCs interact online with those who face similar problems and are involved in different types of social supports, such as informational support, emotional support and companionship. Using a case study of an OHC among breast cancer survivors, we first use machine learning techniques to reveal the types of social support embedded in each post from an OHC. Then we generate each user's contribution profile by aggregating the user's involvement in various types of social support and reveal that users play different roles in the OHC. By comparing online activities for users with different roles and conducting survival analysis on users' time span of online activities, we illustrate that users' levels of engagement in an OHC are related to various types of social support in different ways.

Keywords: Social Support, Online Health Communities, User Engagement, Survival Analysis, Text Mining.

1 Introduction

Nowadays more and more people use the Internet to satisfy their health-related needs. According to a study by the Pew Research Centre, 80% of adult Internet users in the U.S. use the Internet for health-related purposes. Among them, 34% read health-related experiences or comments from others [1]. Compared with traditional health-related websites that only allow users to retrieve information, online health communities (OHCs) increased members' ability to interact with peers facing similar health problems and as a result better meet their immediate needs for social support. It is estimated that 5% of all Internet users participated in an OHC [2].

While people use OHCs for a wide range of needs, obtaining psychosocial support is one of the key benefits of engagement in OHCs [3,4]. Research has found that such support can help patients adjust to the stress of living with and fighting against their diseases [5,6,7] and is a consistent indicator of survival [8]. An OHC also serves as an outlet for users' emotional needs and improve their offline life [9]. Thus active engagement in an OHC has been found to be therapeutic to users [10] and it is important to keep users engaged in the community.

Literature on social support suggests that OHCs mainly feature three types of social support: informational support, emotional support, and companionship (a.k.a., network support) [11,12]. *Informational support* is the transmission of information, suggestion or guidance to the community users [13]. 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. Such support can help one reduce the levels of stress or anxiety. Companionship or network support consists of chatting, humour, teasing, as well as discussions of offline activities and daily life that are not necessarily related to one's health problems. Examples include sharing jokes, birthday wishes, holiday plans, or online scrabble games. Companionship helps to strengthen group members' social network and sense of communities.

Previous studies of OHCs have examined social support among OHC members. For instance, the qualitative study by Zhang et al. [14] found that users exchange *informational support* more frequently than other types of social support in an OHC for smoking cession. Nambisan [15] indicated *informational support* and social support existing in OHCs, but the information seeking effectiveness affects patient's perceived empathy. By contrast, Ahmed et al. [16] suggested that peer-to-peer information support is the key aspect for a Facebook group related to concussion.

Then is a user's involvement in and exposure to different types of social support related to her/his engagement in an OHC? Few studies have answered this question systematically by examining users' seeking, receiving, and provision of various types of social support. A previous study showed that users who received more emotional

support tended to stay longer, while receiving more informational support does not keep a user engaged [17]. However, the study did not consider companionship or users' roles and behaviors in seeking and providing support.

In this research, we addressed three research questions regarding social support and user engagement in OHCs: (1) Can we use machine learning techniques to detect the seeking and provision of three major types of social support embedded in interactions among users; (2) Are there any patterns of users' involvement in different types of social support activities? Or in other words, do users play different roles when it comes to seeking and providing social support? And (3) Are the seeking, providing, and exposure to different types of social support related to users' engagement in an OHC? The outcome of this research has implications for building and sustaining an active OHC through better thread/post recommendations and community management.

2 Detecting Social Supports From Texts

2.1 Dataset and the Taxonomy of Social Support

In this research, we used Breastcancer.org as a case study. It is a very popular peer-to-peer OHC among breast cancer survivors and their caregivers. With more than 140,000 registered users, the website provides various ways for its members to communicate, including discussion forums, private messaging, friend subscriptions, listserv, etc. We designed a web crawler to collect data from its online forum, which has 73 discussion boards. Our dataset consists of all the public posts and user profile information from October 2002 to August 2013. There are more than 2.8 million posts, including 107,549 initial posts. These posts were contributed by 49,552 users.

Table 1 Example posts for types of social support

Social Support Category	Examples
Companionship (COM)	*Kelly Have a wonderful time in Florida, enjoy the sun and fun. Heather* *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.* *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 & 4 Taxol. I had bc in 2000 & 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.*

As we mentioned earlier, informational support, emotional support, and companionship are the three major types of social supports in OHCs. Thus for each post, we need to determine whether it was seeking informational support (SIS), providing informational support (PIS), seeking emotional support (SES), providing emotional

support (PES), or simply about companionship (COM). Note that we did not differentiate seeking and provision of companionship, because the nature of companionship is about participation and sharing. By getting involved in activities or discussions about companionship, 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 categories above. Table 1 lists example posts for each category and a post that belongs to two categories.

2.2 Annotations and Features

As it is almost impossible to label all 2.8 million posts manually, we used classification algorithms to decide what kind(s) of social support each post contains. To train the classification algorithm, we leveraged human annotated data. We randomly selected 1,333 (54 initial posts and 1,279 comments) out of our dataset. After basic training on 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 is related to one or more categories of social supports.

To control the quality of human annotations, we also added to the pool 10 posts that have been annotated by domain experts. For each post, we only accepted results from annotators whose performance on the 10 quality-control posts is among top 3. Results from the other two annotators were discarded. Then a majority vote was used to determine whether a post is related to a category of social support. Table 2 shows the results of the annotation process.

Table 2 The number of posts in each category of social support in the annotated dataset.

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

Users in OHCs may have different writing styles or linguistic preferences to express themselves. To capture these characteristics, we examined each post and extracted various types of features for the classifier: basic features, lexical features, sentiment features, and topic features. Table 3 summarizes these features. Many of the features were picked specifically for classification in this context. For example, we included whether a post is an initial post as a feature because many users seek 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 [17], we also checked the usage of phrases in the format of <you/he/she + MODAL verb > to express possibilities, such as “you should”, “she could”. We considered “he” and “she” in addition to “you”, because some posts were created by family members of cancer survivors. To identify the difference between “seeking” and “providing” support, we included words related to seeking behaviour, such as “question”, “wonder” and “anybody”. We also hoped that words related to daily life topics and geographical locations can effectively detect companionship. Meanwhile, we used OpinionFinder [18] to find the overall sentiment, as well as subjectivity and objectivity of each post. Besides these handpicked or dictionary-based lexicons, we also wanted to capture whether the usage of other words and phrases can contribute to the classification. Using unigrams and bigrams is too fine-grained and leads to a feature set with very high dimension. Thus we adopted an approach similar to [17] and applied topic-modelling technique Latent Dirichlet Allocation (LDA, with k=20) [19] to the content of all posts and generated 20 topics. For each post, LDA gave a topic probability distribution, indicating the probability of this post corresponding to each topic. Such a distribution for each post was then included in the feature set.

2.3 Evaluation of the classifier

Because there are five categories of social supports and a post may be related to more than one category, we built a classifier for each category. For the classification of each category of social support, we applied various classification algorithms on annotated posts and picked the best performing algorithm (using 10-fold cross-validation). Because posts seeking emotional support accounted for only a small proportion among annotated posts (22 out of 1,333), we oversampled posts seeking emotional support when building the SES classifier. Table 4 compares the performance of different algorithms for the five categories of social support. AdaBoost was

chosen to classify COM, PES[1], PIS and SIS, while logistic regression was the best choice for SES. Overall, our classifiers achieved decent performance with accuracy rate over 0.8 in all five classification tasks.

Table 3 Summary of features for the classifier.

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 4 Performance of classification algorithms for five categories of social supports.

Social support	Results	Naïve Bayes	Logistic Regression	SVM (Poly Kernel)	Random Forest	Decision Tree	AdaBoost
COM	Accuracy	0.696	0.787	0.783	0.771	0.767	**0.804**
	ROC Area	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**
	ROC Area	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**
	ROC Area	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
	ROC Area	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**
	ROC Area	0.893	0.803	0.745	0.86	0.766	**0.869**

After applying the best-performing five classifiers on the remaining of the 2.8 million posts, each post received 5 labels, each of which indicates whether the post belong to one of the five social support categories. The total numbers of posts in each category are listed in Table 5.

Intuitively, there are more posts to provide support than to seek support. This is what most would expect from a popular OHC with a large and active user base. About 37% of the posts provided informational support, making it the largest group among the five. In other words, providing informational support is the most popular activity in the OHC. Companionship posts constitute 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

[1] Although the results of accuracy and ROC area of random forest are slightly better than AdaBoost for the PES classifier, the random forest classifier has much worse recall and f-measure. Thus we decided to choose AdaBoost.

were predicted to provide informational and emotional support at the same time, representing the largest group with more than one category of social support.

Table 5 Total numbers of posts in each category of social supports.

Social support category	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

3 User Profiling And Engagement

After estimating the nature of social support in each post, we can then build a profile for each user by aggregating her/his posts by their social support categories. We represented each user's social support involvement 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 will have a vector of <0.3, 0.4, 0.2, 0.1, 0>.

With social support distribution vectors of 47,581 users, we 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, we tested various K values (from 2 to 20) and clustering results with Davies-Bouldin Index (DBI) [20]. 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 $C_i$and C_j. Euclidean distance was used for this study. Generally speaking, 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.

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

We summarized the DBIs for different K values in Table 6. K=7 yielded the lowest DBI value and hence the best clustering results. Centroids for each of the 7 clusters are shown in Table 7.

From Table 7, we can see that, intentionally or not, OHC users do have different patterns in social support involvement and thus play different roles in the community. Some users' posts focused on one major category of social support. For example, users in cluster 0 published an average of 96.55% of their social support posts to provide informational support. They obviously act as *information providers* in the community. Similarly, cluster 1 is for *community builder*s with 64.92% of supports in companionship, and cluster 4 consists of *emotional support providers.* The two smallest clusters are for seekers: cluster 3 for *information seekers* and cluster 6 for *emotional support seekers*. Meanwhile, users in cluster 2, the largest cluster of the seven, are *all-around contributors* with relatively balanced profiles in each social support category. Cluster 5 represents a group of *information enthusiasts*, who focus mainly on informational support, both seeking and providing.

Next we investigated how users in different groups engaged in the OHC. 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). Fig 1(a) compares the distributions of productivity for users in the 7 clusters. The curves suggest that community builders in cluster 1, albeit a small group of users, and all-around contributors 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. Fig 1(b) 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, emotional support providers in cluster 4 are more engaged that information providers in cluster 0.

Table 6 The DBIs for the K-means clustering with various K values

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		

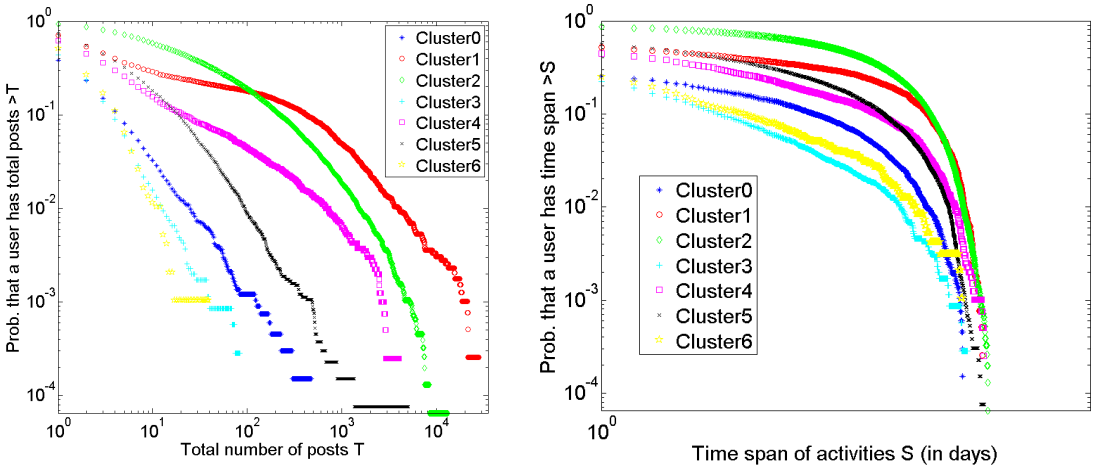
Table 7 Centroids of user clusters

Social Support	All users	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
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	0.612	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	47581	6647	3923	15336	3502	3994	13225	954
% of users		14%	8%	32%	7%	8%	28%	2%

To better clarify the differences among the clusters, we use two-sample Kolmogorov-Smirnov test (K-S test) to compute the statistical gaps between every two clusters in both productivity and time span of activities. Two-sample K-S test, which is used to compare whether two one-dimensional probability distributions are different, is defined as

$$D_{n,n'} = \sup\left|F_{1,n}(x) - F_{2,n'}(x)\right| \quad (2)$$

where $F_{1,n}(x)$ and $F_{2,n'}(x)$ are the empirical distributions of a metric for two groups. The closer the result is to 0, the more likely the two samples are drawn from the same distribution. In Table 8, the upper triangular matrix (shaded area) shows the K-S statistics for comparing productivities between every pair of user clusters, and the lower triangular matrix shows the K-S statistics for time spans. For both metrics, the difference between clusters 2 and 3 is the greatest, which is consistent with what we observed in the Fig 1. In addition, p-values for all K-S tests are less than to 0.001, suggesting statistically significant differences among all clusters' distributions of both engagement metrics.



(a) Numbers of posts (b) Time span of activities

Fig. 1. Complementary cumulative distributions of engagement metrics for the users in different clusters

Table 8 K-S statistics on engagement metrics for each pair of user clusters. The shaded area is for the comparison on number of the posts. The unshaded area is for time span of activities. All values are significant at p=0.001.

	Cluster 0	C 1	C 2	C 3	C 4	C 5	C 6
Cluster 0	-	0.329	0.660	0.056	0.235	0.358	0.062
C 1	0.278	-	0.348	0.321	0.148	0.178	0.301
C 2	0.602	0.348	-	0.673	0.465	0.409	0.653
C 3	0.080	0.320	0.665	-	0.219	0.317	0.071
C 4	0.213	0.148	0.457	0.230	-	0.124	0.191
C 5	0.330	0.112	0.363	0.334	0.117	-	0.286
C 6	0.062	0.305	0.650	0.041	0.200	0.306	-

4 Survival Analysis

Our analysis on users' roles has revealed that the level of users' engagement in an OHC is related to their posting behaviors on various types of social supports. The goal of conducting survival analysis in this section is to more systematically study social support factors related to users' engagement. In addition to users' posting behaviors, we also wanted to examine whether the exposure to different types of social support would impact a user's engagement. An individual may enter or exit a community not only based on his/her own expectation and behavior, but also based on the community's expectations and behaviors regarding this individual [21].

Our survival analysis was based on the Cox Proportional-Hazards Model [22,23], which assessed 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 as Equation-3),

$$h_i(t) = h_0(t) * exp\{\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}\} \quad (3)$$

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 represent independent variables and corresponding coefficients. Equation-3 can also be formulated as Equation-4, where the ratio of two individuals' hazard functions does not depend on time t.

$$\frac{h_i(t)}{h_j(t)} = exp\{\beta_1(x_{i1} - x_{j1}) + \cdots + \beta_k(x_{ik} - x_{jk})\} \quad (4)$$

With Maximum Likelihood Estimates (MLE), β can be estimated with regard 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 our analysis, an "event" refers to a user's cease of activities in the OHC (i.e., "leaving the OHC"). A user's survival time was measured by the difference between her/his last and first posts in the OHC. Here we assumed that a user had left this OHC if she/he had no post during the last 12 weeks in our dataset. For those who were still with the OHC during the last 12 weeks, their survival time was right-censored because they were still participating in this OHC.

Table 9 summarizes independent variables in our model. Values of these variables were based on users' activities within the first month of their online activities. Data was collected for 19,135 users whose time spans of activities exceeded one month. To reduce the impact of multi-collinearity, we calculated the correlation coefficients for every pair of independent variables. We then removed *TotalPost* and *NumThread* from the model, as they are strongly correlated with several other independent variables (with correlation coefficients greater than 0.8). Thus our full model for survival analysis included 11 independent variables.

Table 10 shows the results of the full model. Independent variables with hazard ratio less than 1 contribute positively to the "survival" (i.e., engagement) of users, whereas those with hazard ratio higher than 1 are considered "hazardous" to keep users in this OHC. For example, the hazard ratio of 0.907 for *COM* means that a user's "survival" rate after one month is 9.3% higher (100%-90.7%) if her num-

ber of companionship posts is one standard deviation higher than the average. Similarly, those who posted more to seek emotional support (*SES*) tended to stay with the OHC for longer. By contrast, those who sought and received more informational support (*SIS* and RIS_D) often left the OHC earlier. Besides the four, other independent variables are not significant predictors of users' time span of activities.

Table 9 Independent variables in survival analysis

Indep. Variables	Descriptions
TotalPost	The total number of posts a user has published
InitPost	The total number of threads a user initiates
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
RIS_D	Direct informational support received--the number of informational support posts a user received after initiating a support-seeking thread.
RES_D	Direct emotional support received--the number of emotional support posts a user received after initiating a support-seeking thread.
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.
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 RIS_I, RES_I, and *RCOM*, we assumed that a user read others' replies that were posted within 7 days before the user's replies.	

Table 10 Full model of survival analysis

Independent Variables	Hazard Ratio	Std. Err.
InitPost	0.990	0.0171
PES	1.015	0.0137
PIS	0.977	0.0162
SES	0.958***	0.0117
SIS	1.055***	0.0134
COM	0.907***	0.0131
RIS_D	1.048*	0.0192
RES_D	0.993	0.0137
RIS_I	1.040	0.0221
RES_I	0.970	0.0236
RCOM	0.968	0.0212

*:p<0.05, ***: p<0.001

To evaluate the robustness of the full model, we conducted the same analysis using backward sequential elimination [24]. Specifically, for the full Cox model with 11 independent variables, we removed the least significant variable in each iteration of the Cox model, until all independent variables left in the model were statistically significant. The four independent variables that were statistically significant in the full model were still significant and with similar hazard ratios after the last iteration of backward sequential elimination (shown in Table 11).

Table 11 Coefficients of 4 independent variables after backward sequential elimination

Independent Variables	Hazard Ratio	Std. Err.
SES	0.955***	0.0109
SIS	1.049***	0.0112
COM	0.906***	0.0122
RIS_D	1.032***	0.0103

***: p<0.001

5 Discussions

According to our survival analysis, those who started the first month of their online activities in the OHC with a lot of information seeking posts may not get engaged in the long run, even though they may also receive plenty of informational support as a result. This is in accordance with what a previous study found about informational support [17] and our analysis on users' roles in Section 3 (users in Cluster 3). In other words, informational support seekers have a higher chance of "churn" after they get the information they want from the community. This suggests that although community members have spent a lot of effort in providing informational support, as evidenced by the large number of *PIS* posts in Table 5, informational support does not seem to be the key to keep users engaged.

Conversely, those who were involved in companionship activities are more likely to stay. The positive effect of companionship on user engagement is even stronger than seeking emotional support, which was suggested to be a strong indicator of engagement by [17]. The exposure to companionship *RCOM* also has a hazard ratio below 1, although it is not statistically significant. The hazard ratios of both companionship-related independent variables indicate the importance of companionship. Similarly, in our user role analysis, community builders in Cluster 1 are very active. This is a very interesting finding—even though this is an OHC about cancer, it was the discussions of non-cancer-related issues (e.g., everyday family life and online games) that kept users engaged in the community. Recall that companionship includes discussions of offline events, sharing daily life stories that are more personal, and playing online games. We conjecture that companionship can strengthen the ties among users more than informational support that often lacks the personal touch, or emotional support, which can sometimes be generic and a mere formality (e.g., "I will pray for you", "Love you and Hug").

Another observation from the survival analysis is that none of the three independent variables for indirect support (*RIS_I, RES_I, and RCOM*) was statistically significant. While it might be true that indirectly received support is not related to users' engagement, this may also be caused by our inaccurate measure of indirect support a user received. In our model, we assumed that a user received indirect support when she read a thread initiated by another user and other users' replies to the thread. This can be problematic: on one hand, we may underestimate the amount of support because we limited our calculation to threads a user replied to. In fact, a user can get indirect support from a thread without posting a reply. On the other hand, our 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. Additional data of users' click streams will be needed to address this problem.

6 Conclusions and Future Work

This research analyzed users' behavioral patterns related to different types of social support and how such patterns are related to their engagement in an OHC. Using an OHC for breast cancer as a case study, we built classification models to detect the nature of social support in each post. After aggregating each user's posts, we grouped users based on their social support behavioral patterns and discovered seven different user roles in the OHC. Through comparisons between different roles and more systematic survival analysis, we found that those with high level of engagement in the OHC are actively involved in companionship. In other words, sharing stories from personal daily life and activities that are not directly related to cancer are the key for keeping this community together. This is followed by seeking emotional support, which can also keep user engaged. On the other hand, simply seeking and receiving informational support makes a user vulnerable to churn.

The outcome of our study can shed light on the design and management of an OHC. For example, to keep an OHC active and sustainable, community managers may want to initiate and promote companionships activities, such as holiday plan discussions, gardening tips, online scrabble games, offline gatherings, etc. Also, a thread/post recommender that leverages users' roles in the community can help proficient providers of certain support quickly find those who that are seeking such support. For future research, we would like to build a predictive model of user engagement based on our findings. Exploring whether a user's role changes over time would also be an interesting direction.

References

1. Fox, S. 2011. "The Social Life of Health Information, 2011," Pew Research Center's Internet & American Life Project

2. Chou, W. S., Hunt, Y. M., Beckjord, E. B., Moser, R. P. & Hesse, B. W. Social Media Use in the United States: Implications for Health Communication. J Med Internet Res 11, (2009)

3. Kim, E. et al. The process and effect of supportive message expression and reception in online breast cancer support groups. Psycho-Oncology 21, 531–540 (2012)

4. Rodgers, S. & Chen, Q. Internet Community Group Participation: Psychosocial Benefits for Women with Breast Cancer. Journal of Computer-Mediated Communication 10, 00–00 (2005)

5. Dunkel-Schetter, C. Social Support and Cancer: Findings Based on Patient Interviews and Their Implications. Journal of Social Issues 40, 77–98 (1984)

6. Qiu, B., Zhao, K., Mitra, P., Wu, D., Caragea, C., Yen, J., Greer, G. E., and Portier, K. : Get Online Support, Feel Better – Sentiment Analysis and Dynamics in an Online Cancer Survivor Community. in Privacy, security, risk and trust (passat), in Proceedings of the Third IEEE Third International Conference on Social Computing (SocialCom'11), 274–281 (2011)

7. Zhao, K., Yen, J., Greer, G., Qiu, B., Mitra, P., and Portier, K. "Finding influential users of online health communities: a new metric based on sentiment influence," Journal of the American Medical Informatics Association: JAMIA. (2014) (online first).

8. McClellan, W. M., Stanwyck, D. J. & Anson, C. A. Social support and subsequent mortality among patients with end-stage renal disease. JASN 4, 1028–1034 (1993)

9. Maloney-Krichmar, D. & Preece, J. A Multilevel Analysis of Sociability, Usability, and Community Dynamics in an Online Health Community. ACM Trans. Comput.-Hum. Interact. 12, 201–232 (2005)

10. Idriss SZ, Kvedar JC & Watson AJ. The role of online support communities: Benefits of expanded social networks to patients with psoriasis. Arch Dermatol 145, 46–51 (2009)

11. Bambina, A.: Online social support: the interplay of social networks and computer-mediated communication, Youngstown, N.Y.: Cambria Press. (2007)

12. Keating, D. M. Spirituality and Support: A Descriptive Analysis of Online Social Support for Depression. J Relig Health 52, 1014–1028 (2013)

13. Krause, N. Social Support, Stress, and Well-Being Among Older Adults. J Gerontol 41, 512–519 (1986)

14. Zhang, M., Yang, C. C. & Gong, X. Social Support and Exchange Patterns in an Online Smoking Cessation Intervention Program. in 2013 IEEE International Conference on Healthcare Informatics (ICHI) 219–228 (2013)

15. Nambisan, P. Information seeking and social support in online health communities: impact on patients' perceived empathy. J Am Med Inform Assoc 18, 298–304 (2011)

16. Ahmed, O. H., Sullivan, S. J., Schneiders, A. G. & Mccrory, P. iSupport: do social networking sites have a role to play in concussion awareness? Disabil Rehabil 32, 1877–1883 (2010)

17. Wang, Y.-C., Kraut, R. & Levine, J. M. 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 833–842 (2012)

18. Wilson, T., Wiebe, J. & Hoffmann, P. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. in Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing 347–354 (2005)

19. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)

20. Davies, David L.; Bouldin, Donald W.: "A Cluster Separation Measure". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224–227. (1979)

21. Levine, J. M. & Moreland, R. L. Group Socialization: Theory and Research. European Review of Social Psychology 5, 305–336 (1994).

22. Cox, D. R.: Regression models and life tables. Journal of the Royal Statistical Society, Series B, 34(2), 187–220. (1972)

23. Fox, J.: Cox proportional-hazards regression for survival data. Retrieved from http://socserv.mcmaster.ca/jfox/books/companion/appendix/Appendix-Cox-Regression.pdf (2002)

24. Cotter, S. F., Kreutz-Delgado, K. & Rao, B. D. Backward sequential elimination for sparse vector subset selection. Signal Processing 81, 1849–1864 (2001)