Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities

Diyi Yang
Language Technologies Institute
Carnegie Mellon University
diyiy@andrew.cmu.edu

Robert E. Kraut
Human-Computer Interaction Institute
Carnegie Mellon University
robert.kraut@cmu.edu

Tenbroeck Smith
Behavioral Research
American Cancer Society
tenbroeck.smith@cancer.org

Elijah Mayfield
Language Technologies Institute
Carnegie Mellon University
elijah@cmu.edu

Dan Jurafsky
Department of Linguistics
Stanford University
jurafsky@stanford.edu

ABSTRACT

Participants in online communities often enact different roles when participating in their communities. For example, some in cancer support communities specialize in providing disease-related information or socializing new members. This work clusters the behavioral patterns of users of a cancer support community into specific functional roles. Based on a series of quantitative and qualitative evaluations, this research identified eleven roles that members occupy, such as *welcomer* and *story sharer*. We investigated role dynamics, including how roles change over members' lifecycles, and how roles predict long-term participation in the community. We found that members frequently change roles over their history, from ones that seek resources to ones offering help, while the distribution of roles is stable over the community's history. Adopting certain roles early on predicts members' continued participation in the community. Our methodology will be useful for facilitating better use of members' skills and interests in support of community-building efforts.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; **Collaborative content creation**; *Computer supported cooperative work*; • **Computing methodologies** → *Cluster analysis*; Discourse, dialogue and pragmatics.

KEYWORDS

Social Roles; Social Support; Online Health Communities; Natural Language Processing; Machine Learning

ACM Reference Format:
Diyi Yang, Robert E. Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019. Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3290605.3300574

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300574

1 INTRODUCTION

A wide body of literature studying online health communities has developed and tested hypotheses on how these communities differ from the internet at large, how users support each other, and how communities thrive over time. For example, [52] studied how social support exchange in an online cancer support group affects the length of people's participation, and [19] examined support exchange around behavior changes in online weight loss communities. Using descriptive statistical models, this research modeled characteristics of user behavior to show that early actions result in differential long-term membership trends. For instance, users self-disclose more personal information in online health communities than in parallel technical support communities, like Stack Overflow [5, 34]. Not all users display these behaviors, though: for instance, many users join when facing crucial healthcare events, like the start of chemotherapy, and are seeking information for decision-making rather than hoping to join a community [56]. Early actions and interactions can be predictive of commitment. Newcomers looking for informational support are significantly less likely to transition into long-term community membership, and those who receive support are more likely to continue than those who

do not [52, 58]. Yet 10% of support-seeking messages get no replies, and many of the replies do not provide the support sought, as when long-time members provide emotional support when the new user was seeking information [53].

Interaction in health support communities is in part the products of the *roles* that members occupy [49]. For example, some members might specialize in seeking support, providing disease-related information or socializing new members. In contrast to roles in conventional organizations, where roles are often assigned and come with defined responsibilities, roles in most online communities are emergent. For example, a user can assume an "expert" role in the community without seeking permission from others. Researchers have clustered lower-level behavior to identify roles in some online communities like Wikipedia [54, 57]. However, few studies have applied similar approaches to online health communities [30].

The goal of the current paper is to study members' participation and coordination in online health communities, and develop a taxonomy of the emergent roles that are observed in these communities, linking individual behaviors with community-level outcomes. Identifying emergent roles can be beneficial for sustaining communities. Understanding the roles that are important for a community and the roles particular people are likely to occupy can help to optimize user experiences. For example, information experts can be matched to information seekers, giving the expert fulfilling work to do while helping the seeker get timely responses; welcomers can be matched to newcomers to ensure they receive timely support that will help them become integrated into the community.

To this end, we propose a framework for defining social roles in online communities together with a general modeling methodology. We use data from an online cancer support community to identify behavioral features associated with different facets of social roles. We then build an unsupervised Gaussian mixture model from the data to discover 11 roles that members occupy. We validate these roles through a series of quantitative robustness checks of the modeling procedure, followed by confirmatory interviews with domain experts in the community. To demonstrate the utility of the role model, we examine how roles predict the stability of activities on the site and participation by users as they enter the community and evolve from being newcomers to old-timers. (1) We find that occupying socially positive roles, such as private communicator and story sharer, is associated with members staying in the community longer, while members occupying roles such as informational support seeker are associated with lower long-term participation in the community. (2) While the distribution of roles in the community is relatively stable over time, members change their roles frequently across their participation. As members stay longer in the community, they are more likely to occupy the roles of emotional support provider and welcomer and less likely to occupy roles such as story sharer and informational support seeker. A closer look at members' role transitions suggests that they frequently change their roles from seeking resources to roles that offer help to others. (3) Both the tendency of certain roles' occupants to drop out of the community and the trajectory of roles in users' lifecycle in the community follow consistent patterns. These findings suggest the value of the role framework as the basis for intervention in online health communities, opening a new opportunity for socio-technical systems to support users and communities in their healthcare needs.

2 ROLES IN ONLINE COMMUNITIES

Self-organized online communities are a novel area for theoretical exploration of emergent roles. In contrast to most empirical studies of roles, which have looked at "formal" roles like leaders or moderators [17, 39], our work examines members' emergent roles in online health communities, which are not structurally defined or constrained, but rather emerge from common patterns of members' behaviors. Theory on coordination in groups and organizations has emphasized role differentiation along with the division of labor associated with roles as major mechanisms through which members coordinate complex activities [10, 32, 33].

In the *Structural* perspective [22], the traditional model for describing offline organizations, roles are generally formally assigned, often in terms of a formal job title and prescribed activities needed to fulfill the role well. These roles are mainly based on formal and informal social expectations and norms along with positive and negative sanctions to enforce the norms. In online environments, the structural model sometimes applies, including moderator roles in many online discussion sites or administrator roles in Wikipedia. In these cases, members have formal assignment to those roles and clear expectations of responsibilities [1].

However, in the vast majority of online communities, roles are emergent, self-selected and are often not formally recognized [2, 57]. As a result, although these emergent roles constitute consistent patterns of behavior, neither the role occupant nor other community members may have a clear understanding of who is occupying which role or how role occupants should behave. This pattern more closely matches the *interactionalist* view of roles, which has built on several decades of sociological theory research [12, 25, 50]. While explicit roles have been studied in depth in online communities [18], the consequences of the more fluid, sociologically-informed definition of emergent roles has received relatively little attention in studies of behavior in online communities. The little research that does exist has largely focused

on production roles in collaborative projects like Wikipedia [3, 54, 57, 60].

To begin to fill this gap, here we define **social role** as a set of interaction patterns regulated by explicit or implicit expectations and adopted by people in a social context to achieve specific social goals. Our definition hangs on four core facets of roles:

- **Goal**: Roles are associated with specific social *goals*. Goals may serve the individual interests of the role occupant, role partners or the groups in which the roles are embedded [39]. For example, specific roles may be adopted to facilitate collective effort toward the completion of a task, such as a devil's advocate role in a course project team[47]. Roles can also be oriented toward the long-term functioning of the group as a whole, such as "Vandal Fighter" in Wikipedia [54]. Finally, people may take on some roles to satisfy their individual needs or desires, such as newcomers acting as information seekers to understand what the group has to offer or senior members experiencing pleasure in mentorship.
- **Interaction**: Roles are based on role holders' characteristic *interactions*, which can happen when role holders engage with other persons or objects, within or outside the context where the role is enacted. These interactions make up the core content of online communities. In discussion-oriented communities, these are the threads-starting messages and comments through which discussion takes place. But these interactions also take place when role holders interact with the user interface of the community's website, or when they speak with their spouse or friends outside. Such interactions are observed by role holders, repeated over time [50], and whether or not each interaction is expected or approved by a role holder, each interaction shapes the roles they may enact in the future.
- **Expectation**: Roles also involve *expectations* about typical interaction patterns of persons [25, 29]. Adherence to or departure from these understandings can result in positive or negative sanctions from others [15, 37]. Expectations are bidirectional: both the role holders and the others with whom they interact often have expectations about how the role holders should behave and what they should believe. In conventional organizations offline where roles are assigned, they are generally associated with strong expectations; managers in corporations speak differently when speaking to their employees than they do when speaking to bosses, for instance [16]. In many online communities, though, roles are emergent. In these cases, there may exist informal or implicit "negotiated understandings" about how role occupants should conduct themselves or they may come with no expectations at all. Because these understandings may be implicit or known only to long-time members, they can create barriers to community participation; for instance, on Stack Overflow, fear of hostile feedback for improperly meeting expectations of information seekers can prevent new users from asking questions or joining the community in the first place [24].
- **Context**: Roles can be very broadly applicable or limited to specific *contexts*. These contexts set boundaries for role holders, i.e. delimiting the perimeter or setting the scope of roles. For example, *information provider* is a common role in many groups, including social Q&A websites, health discussion forums, and problem-solving groups; In contrast the *committer* role [51] is limited to open-source development communities. Within a community, roles may be based on the privacy of the context, with people taking on a set of roles in public while enacting others in private discussions.

Note that roles are performed by *people* [12]. Sometimes people's non-behavioral attributes such as their demographics like gender or race may be related to the roles they occupy. Except in specialized cases, these characteristics may not be an intrinsic part of roles, but they are often entwined with expectations. For example, although Wikipedia bills itself as is the encyclopedia that anyone can edit, men are much heavier contributors than women [27].

The current research investigates members' emergent, behavioral roles when participating in online health communities independently of the demographics of the people who occupy them. For example, any member can assume the role of emotional support provider, no matter their gender, age or cancer type. Our goal is to design a model that can ultimately be deployed in online interventions, in environments where both technical constraints and user privacy dictate that demographics should not be a factor in the technical system. Thus, we do not model personal attributes of members in our research. Future studies in constrained, privacy-aware contexts may extend this work to directly cross the behavioral roles identified with some of members' personal attributes (e.g., informational support provider × cancer type).

3 RESEARCH SITE

Our research was conducted on the American Cancer Society's Cancer Survivor Network[1] (CSN), which is the largest online support community for people suffering from cancer and their caregivers. The CSN discussions boards are public places where registered members can participate by starting new threads or commenting on other members' existing threads. Registered members of CSN can also communicate

[1] https://csn.cancer.org/

directly with each other using a function called "CSN Email". Conversations between two people are recorded in a format like email or private chat messages and are only visible to individuals addressed in the message headers. We were provided access to all public posts and comments, private chats as well as the profile information for users registered between Dec 2003 and Mar 2018. During this period, there were a total of 66,246 registered users who exchanged 139,807 private messages, 1,080,260 comments and 141,122 threads. This work was approved by Carnegie Mellon University's Institutional Review Board (IRB).

4 METHOD FOR ROLE IDENTIFICATION

Our method of identifying emergent social roles in online communities is *a repeated cycle of role postulation, definition, automated processing and evaluation.* When participating in the community, a user takes on one or more implicit roles for their activities. In their future interactions, they may take on the same roles or shift roles. To model this, we define a Gaussian mixture model [36], a statistical model that clusters heterogeneous user-session representations into a set of coherent, discovered user roles. Unlike traditional unsupervised learning such as k-means clustering, in which an object can only be a member of a single cluster, a mixture model allows users to occupy multiple roles during a session (e.g., a welcomer and information provider).

The model assumes that user activities can be described by a set of observable behaviors X, and there exist k components per role $\{c_{i=1}^{k}\}$. Each component c_i has an associated vector μ_i of average values for each feature in X. A user's activity is generated from a mixture of these components and a covariance matrix Σ_i, representing the likelihood of each role co-occurring with each other role. Formally, Gaussian Mixture models are a linear combination of Gaussians, with a probability density function as follows:

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot N(x|\mu_k, \Sigma_k), \text{ where } \sum_{k} \pi_k = 1$$

Here, $\{\pi_{i=1}^{K}\}$ are called mixing coefficients, and each user will be assigned a coefficient π_i for each role c_i. The coefficient represents the proportion of a user that was associated with a particular role; each user unit is modeled as a mixture of roles, which enables us to capture participants' versatility and dynamics in the online community. When building this model, we need to learn mixing parameters $\{\pi_1, \pi_2, \ldots, \pi_K\}$, means $\{\mu_1, \mu_2, \ldots, \mu_K\}$ and covariances $\{\Sigma_1, \Sigma_2, \ldots, \Sigma_K\}$ from data $\{x_i\}_{i=1}^{N}$. Here, each x_i is a heterogeneous vector of features extracted from each user, while N represents the total number of user units in our corpus. Given a large corpus of data, we can estimate the covariance matrices by positing that each component has its own general covariance matrix.

This model has three key parameters that need to be set by researchers: the behavior features X, the length of user representation l, and the number of implicit roles K. In the following, we describe the procedures used to set each parameter and the steps taken to design robust models.

Operationalizing Behavioral Features

To extract the emergent roles that members take on when participating on CSN, we identified a set of behavioral features that operationalize the four components in our definition of role definition described above: *goal*, *interaction*, *expectation* and *context*.

Recently, deep learning based techniques have been proposed to learn user embeddings based on their interactions in an end-to-end manner [26, 28, 44]. Although that approach requires less domain knowledge and manual feature construction, it suffers from lack of interpretability especially about the nature of discovered roles and the people who occupy them. In terms of techniques for identifying social roles online, most research employed clustering analysis or principal component analysis to cluster each user into one or more clusters [54, 57]. To make the derived roles interpretable, we followed this common practice to construct explainable patterns to capture members' role-relevant behaviors.

Goal (9 features). Many people with chronic illnesses, including cancer patients and survivors, participate in online health support groups. Ridings and Gefen found that 76% of people who joined online health groups were looking for two types of social support [45] - informational support and emotional support. Informational support contains information, advice, or knowledge, and emotional support refers to the provision of empathy, sympathy or encouragement. Building on prior studies on social support [13, 52], we operationalized a set of goal-oriented actions that members exchange in the context of support groups. This resulted in 4 features of linguistic behaviors: *seeking informational support*, *providing informational support*, *seeking emotional support*, and *providing emotional support*.

We observed from our data that people tend to employ very specific language strategies when providing emotional support to others. Some choose to show empathy, saying that they understand what the recipient is going through and identify with their emotional reactions and feelings. Some express encouragement and hope that others' situations will improve. Others show appreciation for others' accomplishments to increase others' senses of worth, value and competence. To capture these nuanced intentions, we differentiated three finer-grained sub-categories of providing emotional support: *providing empathy*, *providing encouragement*, and *providing appreciation*. In addition to exchanging social support, members also share their experiences and stories to help others understand who they are and to provide social

Goal-oriented conversational acts	ICC	Correlation
seeking informational support	0.91	0.73
providing informational support	0.92	0.79
seeking emotional support	0.83	0.64
providing emotional support	0.92	0.75
providing empathy	0.74	0.72
providing encouragement	0.68	0.64
providing appreciation	0.73	0.67
self-disclosing positively	0.90	0.72
self-disclosing negatively	0.90	0.71

Table 1: The intra-class correlation and correlations between human decisions and predictions for 9 conversational acts

comparison information [21]. Thus, we also considered the language people use to self-disclose via two additional features: *self-disclosing positively* and *self-disclosing negatively*.

Automatic text analysis techniques can accurately measure the amount members' messages contain each of these nine features. Four trained nursing students rated a sample of 1,000 messages threads and their first responses for degree they represented these nine goal-oriented conversational acts. Using previously developed procedures [13, 52], we built machine learning models to predict the students' assessments of the nine conversational acts in messages. These machine learning models map a set of linguistic features, as described in [52, 59], to a set of continuous output values, indicating how much informational support, emotional support, positive self-disclosure, and negative self-disclosure a thread-starting message conveys as well as how much informational support, emotional support, empathy, encouragement, appreciation, positive self-disclosure, and negative self-disclosure responses provided. Human annotation agreement on a training dataset was high (mean ICC=.84), and the machine learning models achieved reasonable correlation with the average of the human judgments (mean Pearson r=.71; see Table 1). We then applied these models to estimate the nine conversational acts in all messages in our corpus.

Separate from these automatic annotations, we also extracted 2 features measuring raw activity count for users - the number of threads initialized, and the number of comments.

Interaction (53 features). The actions members take toward achieving their goals are essential for understanding the roles they occupy. In this part we use two methodologies to extract interaction features: *linguistic* and *network-based*.

We developed linguistic indicators of members' topical interests by comparing each person's word usage with semantic categories provided by the psycho-linguistic lexicon LIWC [42]. The presences of affective expressions such as anxiety, sadness, or anger related words, were used as indicators of members' emotional orientation. To figure out whether members talked about their personal relationships, we counted their usage of words related to family and friends via corresponding dictionaries in LIWC. Similarly, members' religious orientations and emphasis on themselves vs others (interpersonal pronouns) were calculated via related dictionaries. In total, 16 features were extracted via using corresponding LIWC categories. Topic modeling [14] was conducted to derive topics that members discuss with others on CSN, resulting in 25 topics including prayer, surgery, radiation, clinical trials, and chemotherapy side effects. One feature is included for each topic. We also incorporated domain knowledge from Freebase to capture 4 features counting members' use of words related to disease, medicine, ingredients, and symptoms in their messages when providing information to others. To identify potentially knowledgeable CSN members, we extracted two features: the number of external links and the number of words in messages.

We then looked at interaction patterns that emerge from users' social networks in the online community. Previous studies demonstrated methods for revealing network structure and people's relationships with other users [23, 54, 55]. For this purpose, we constructed a user-reply network and extracted features through network analysis, where the vertices represent members who have participated in at least one messages, and edges represent replies. For example, an edge from user u to user v means that u replied to v's messages. From this graph, we extracted six network-based features: (1) To capture the centrality of members' role in the social structure, we calculated their (1) in-degree and (2) out-degrees. To capture tenure effects we measured (3) members' ratio of talking to newcomers and (4) being talked to by old-timers. (5) To measure whether users talk mainly to several specific users or broader audiences, we calculated the entropy of the user-user interaction distribution. Here, a higher entropy means users talking to broader audiences. Finally, to measure a user's breadth of interests, we measured the number of sub-forums a person has posted in, where each sub-forum represents one cancer type.

Expectation (2 features). Emergent roles may be associated with informal implicit "negotiated understandings" among individuals about what persons should do if they seem to occupy such roles. Members on CSN might indicate such positive or negative evaluations of others via their language choices such as complaining to administrators or telling others what to do. To this end, we extracted two features: (1) the number of messages members exchanged with moderators and (2) their usage of modal words such as "*should*", "*could*", and "*must*". Here, modality in members' messages may convey their suggestions, request or advice to others.

Context (17 features). The context of communication matters. For the purposes of this study, we focused on public vs private conversations as the context. Members may talk to others in private chats to protect their personal information or interact with them on the public discussion board. To capture members' potential concerns of privacy, we differentiated all 9 *Goal* features and their 6 network-based *Interaction* features into separate values for communication in private chats and in the public forum. For example, *seek informational support* will have two features: *seek informational support in private chats*[2] and *seek informational support in the forum*. Similarly, *being talked to by oldtimers* becomes *being talked to by oldtimers in private chats* and *being talked to by oldtimers in the forum*. Note that this domain differentiation is a common practice in text representation for statistical modeling [35] as well as in social computing research [8, 9]. Finally, we calculated 1 feature that measures the ratio of members' private communication to all their private and public activities to capture their preferences for different contexts.

Determining the Granularity of User Activity

Determining the unit of analysis for appropriately representing members' activity is key decision in modeling social roles. Treating users as an aggregation of all their historical actions on CSN prevents one from examining the evolution of roles or transitions between them. On the other hand, employing very small time intervals, such as a single user action, might miss important larger constructs like a cluster of actions needed to achieve a goal. In this analysis we use aggregated data from each **user session**, which is defined as a time interval in which the time gap between any two adjacent actions is less than a threshold (24 hours). Within sessions, users' behaviors were regarded as consistent. We operationalized the 83 features described above to capture members' behaviors within each session.

To test the robustness of the role models, we explored the degree to which they varied across different temporal units–all activity within each calendar day, week, or month. We found that frequently-occurring roles were consistent across different settings. The roles that emerged using a calendar day as the unit of analysis were very similar model to those emerging from session-level modeling, likely due to the similar time-scale. As the temporal unit increased from a day to a week to a month, the derived roles became harder to interpret. This suggests that unlike assigned roles in offline organizations (e.g., professor in a university), emergent roles in this community are more variable over time. This variability led us to examine transitions between roles, described in more detail below.

Role theory also states that role are based on multiple interactions [50], suggesting that detection of roles based on only one observed action is impossible. To address this, we conducted a sensitivity analysis removing sessions that had fewer than t actions ($t \in \{1, 2, 3\}$). We did not observe any significant changes in the derived roles. For all analyses below, we follow the 24-hour inactivity threshold to define sessions and include all sessions, without removing ones with few actions. In total, this resulted in 517,272 user-sessions from 66,246 users.

Determining the Number of Roles

Quantitative Setting of Upper and Lower Bounds. The number of roles K in this model is a free parameter and is the element most susceptible to over-tuning [46]. We used the Bayesian Information Criterion (BIC) to select the number of components in the Gaussian mixture model (GMM). We trained Gaussian mixture models on the user-session corpus and experimented with K ranging from 2 to 20 to determine the optimal number of components/roles. We found that models with $K \in [10, 15]$ seemed to be a good fit.

Qualitative Validation of Final Setting. Validating these behavioral role components inferred from unsupervised methods is challenging. Existing work on similar tasks such as LDA topic modeling has tried to validate the derived components by asking people to provide summary labels for each component [14, 41] or by measuring the purity of the clusters or components [20, 38]. However, interpreting topics or components by researchers themselves might introduce biases, and defining the purity of components that consist of member behaviors rather than simpler features, like bag-of-words representations of topics, is hard to operationalize.

To overcome these problems, we followed a qualitative protocol to finalize the number for user roles and their names. We ran the Gaussian mixture model with our behavior features and user-session length for different values of K. We then discussed the extracted components with 6 domain experts (5 moderators from CSN and a senior researcher familiar with the site). We used their input to help interpret the latent components. We showed the domain experts the top ranked features associated with each role as well as three users who were most representative of each role (i.e., the three users from each role component whose behaviors were closest to the centroid representation of that component). The details about our semi-structured interview with domain experts is here[3]. Based on their input, we set K=11.

[2]For privacy concerns, annotators are not allowed to view and annotate private messages. In these cases, we applied the trained regression models from public forum posts to predict 9 conversational acts in private messages. Accuracy may be lower in these contexts, as this prediction requires transferring the model to a slightly different domain.

[3]http://www.cs.cmu.edu/~diyiy/docs/csn_role_interview_instruction.pdf

Role Name	%	Typical Behaviors
Emotional support provider	33.3	Provide emo support in the forum, provide appreciation in the forum, provide encouragement in the forum, # subforums a user participated, provide empathy in the forum, provide info support in the forum
Welcomer	15.9	out-degree in forum, # replies in the forum, the ratio of talking to newcomers in the forum provide encouragement and provide empathy in the forum, the entropy of user-user interaction in the forum
Informational support provider	13.3	Provide info support in the forum, provide empathy in the forum, provide encouragement in the forum, mention symptom related words, mention drug related words, mention anxiety related words
Story sharer	10.2	# threads in the forum, self-disclose positively in the forum, seek emo support in the forum, self-disclose negatively in the forum, seek info support in the forum, use interpersonal pronouns
Informational support seeker	8.9	# threads in the forum, seek info support in the forum, self-disclose negatively in the forum, seek emo support in the forum, mention disease related words, mention symptom related words
Private support provider	5.3	Provide emo support in private chats, provide appreciation and provide empathy in private chats, provide info support and provide encouragement in private chats, self-disclose positively in private chats
Private communicator	5.3	Preference for using private chats, provide encouragement and provide info support in private chats, provide emo support in private chats, provide empathy in private chats, seek info support in private chats
All-round expert	2.5	# messages in private chats, provide appreciation in private chats, provide emo support in private chats, provide encouragement in private chats, # replies in the forum, self-disclose positively in the forum
Newcomer member	2.4	# threads in the forum, seek info support in the forum, self-disclose positively in the forum, self-disclose negatively in the forum, seek emo support in the forum, mention diagnostic test related words
Knowledge promoter	2.2	# urls/links per message, mention ingredient related words, provide info support in the forum, mention drug related words, mention anxiety related words, mention death related words
Private networker	0.8	The entropy of user-user interaction in private chats, out-degree in private chats, in-degree in private chats, # messages in private chats, the ratio of being talked to by oldtimers, # private conversation initialized

Table 2: Derived roles and their representative behaviors ranked by their frequency (%) in descending order.

5 DISCOVERED FUNCTIONING ROLES

After final parameter tuning and validation from discussions with domain experts, we have evidence that the model is effective in identifying latent roles that members occupy. Once these parameters were set, we worked with the 6 domain experts to co-develop short names and interpretable descriptions of each component in the model, describing the roles that emerged. These roles, their frequency in the corpus, and highest-probability features are described in Table 2.

(1) **Emotional Support Provider**: people who respond to others with empathy, encouragement and emotional support. These active forum members participate in a number of sub-forums, in contrast to most users on CSN who only participate in one sub-forum most relevant to their cancer type.
(2) **Welcomer**: people who respond to newcomers after they first post on CSN. These higher-tenured members interact with newcomers frequently and provide supportive empathy and encouragement.
(3) **Informational Support Provider**: people who offer information and advice to others in the discussion board. This group of members discusses cancer-specific issues by mentioning symptoms and ingredient-related words, and provides information to others.
(4) **Story Sharer**: people who disclose personal information and emotions in order to receive support. They share their own experiences and stories in an introspective and verbose manner, which might help similar users and/or inform potential support providers about their situations.
(5) **Informational Support Seeker**: people who ask questions and seek information from others in public forums. Members with this role initialize more threads, and seek around 1.7 standard deviations more informational and emotional support than average. They also talk more frequently about metastasis and other aspects of their disease.
(6) **Private Support Provider**: people who use private chats to provide social support to others. People in this role provide emotional support, encouragement, appreciation and information to others in private chats, as well as self-disclose in a positive manner to encourage others.
(7) **Private Communicator**: people who are protective of their personal details and only choose to participate in private chats. They seek and provide different types of support such as informational support, empathy and encouragement, and have strong tendency to

communicate privately (3.7 standard deviations more frequently than the average level).

(8) **All-round Expert**: people who engage in a large set of support exchange behaviors in both public discussion board and private chats. This group of members active engages and performs various kinds of actions such as providing appreciation in private chats, replying to others and self-disclosing positively in the forums.

(9) **Newcomer Member**: people who ask questions and seek support shortly after joining CSN. Most members in this group stay at CSN for less than one month. They use the discussion board to ask for both informational and emotional support, and emphasize the uncertainty associated with cancer diagnosis results (0.8 standard deviation more than average).

(10) **Knowledge Promoter**: users who post links and information from outside CSN. Those users present themselves as knowledgeable about what they are talking about and recommend external research pointers to members in need of help. Compared to regular members, knowledge promoters share two standard deviations more links in their replies to others.

(11) **Private Networker**: people who seem to be network hubs in private chats. Although they participate in the discussion forum and exchange social support in private chats from time to time, they talk to a larger set of members in private chats and exchange more messages compared to other members.

After discussion with domain experts, we obtained agreement on the name and characteristics of 10 of the 11 derived roles. However, we failed to achieve consensus for *all-round expert*[4]. Despite this, domain experts agreed that the set of behavioral roles we identified were comprehensive:

> *"It seems very comprehensive and there are so many different examples, so I feel like it is covered very well with your different roles and labels."*

Domain experts did point out roles that our model did not capture. For instance, they identified "Guardian" or "Defender" role - people who fight with spammers or violate norms on CSN, trying to regulate others' behaviors. One of the domain experts described the defender role this way:

> *"The one that I think did not emerge is the policeman, these people complain to moderators when some people are doing things wrong or tell other people that they are violating norms. They shouldn't be diagnosing the way that they are diagnosing or other sorts of problems.".*

[4]We urge readers to interpret our follow up analyses about *all-round expert* with caution.

Role	HR	Std.Err
Emotional support provider	0.984	0.027
Welcomer	0.883***	0.028
Informational support provider	1.060	0.034
Story sharer	0.872***	0.034
Informational support seeker	1.324***	0.023
Private support provider	0.842***	0.033
Private communicator	1.031	0.022
All-round expert	0.869***	0.028
Newcomer member	1.054***	0.025
Knowledge promoter	1.091***	0.028
Private networker	0.916*	0.035

Table 3: Survival Analysis predicting how long members continue to participate in the community. p<0.001: *; p<0.01**; p<0.05*. Number of users = 66,246. Number of user-session records = 522,429**

> *"there are not a lot of them, but they stick in your memories since they are telling others what to do."*

The defender role likely does exist on CSN, but our model did not capture it, either because the behaviors that characterize the defender role occur infrequently or the features we used to characterize user-sessions did not reflect these behaviors.

6 INFLUENCE OF ROLES ON COMMITMENT

Members' patterns of activities and roles can influence their contribution and commitment to the community. Although previous research has investigated members' commitment to both offline and online organizations [6, 31, 58], no computational research has examined how members' assumption of emergent roles relates to commitment in online health communities. This section examines how emergent roles help predict continued participation of members on CSN. Doing so will allow us to better understand members' engagement, as well as demonstrate the utility of our derived roles.

We use survival analysis to investigate how members' occupation of social roles correlates with the length of their participation on CSN. Survival analysis is a type of regression analysis for estimating influences on the time to an event of interest, especially for censored data. In our context, the event is defined as members dropping out of CSN. We used Stata survival command with a Weibull distribution of survival times in order to perform this analysis [48], with the unit of analysis being the user-session. Control variables included the member's gender, whether the member had cancer, and his/her tenure (i.e., how many months they have stayed at CSN). Since the continuous explanatory variables were standardized, the Hazard Ratio (HR) is the predicted change in the probability of dropout from CSN for a standard

deviation increase in the predictor. A hazard ratio greater than one means the role is associated with a higher than average likelihood of dropping out, while a hazard ration less than one means a lower than average likelihood of dropping out. Because of the correlations between different roles, and correlations among roles and tenure, we built separate survival models for each role, resulting in 11 models.

Results of the survival analyses are shown in Table 3. The analyses show that members occupying certain roles - *knowledge promoter, informational support seeker* and *newcomer member* - are less likely to continue in CSN (i.e., lower survival rates). Specifically, members who were one standard deviation more likely to occupy *informational support seeker* roles were 32.4% more likely to leave the community after that session. Similarly, members who were one standard deviation more likely to be *newcomer-seekers* were 5.4% more likely to drop out from the community, while members who share external knowledge with others on CSN (*knowledge promoters*) were 9.1% less likely to continue their participation. These results suggest that roles related specifically to information-sharing are associated with higher rates of drop-out, possibly because researching disease or treatment relevant information is a distinct, time-consuming use of online resources, separate from community-building goals. These members may see CSN as a more transactional resource, either giving or receiving information, and represent a less committed user.

In contrast, occupying roles such as private networker, private support provider, newcomer welcomer, and story sharer are associated with members staying at CSN longer. This may be because being support-providers to others encourages members to interact with other members time after time, developing stronger relationships. People who respond to newly registered members with support were 12% more likely to stay on CSN; members who were willing to self-disclose their experiences to seek support or benefit others had a 13% higher survival rate.

7 STABILITY AND DYNAMICS OF ROLES

As members go through their life cycles, they might choose to drop out or stay on CSN. The roles of those who stay might change over time. For example, as previously described by the Reader-Leader framework [43], people may change from being peripheral to core members of the community. In this section, we examine whether members' emergent roles vary over their tenure at CSN, and we test the stability of users' emergent roles at both individual- and community- levels.

Community Level Stability

We first investigated the mixture of roles in the forum overall over a thirteen years period (see Figure 1). The frequency of the majority of the behavioral roles on CSN did not change

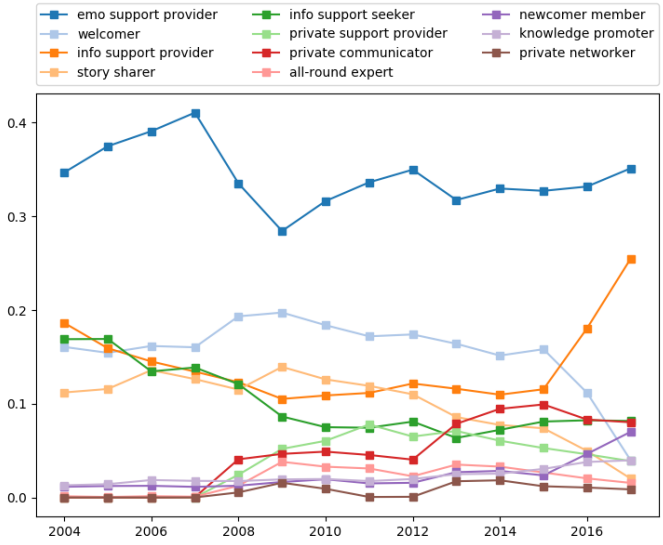

Figure 1: The percentage of different role occupations

substantially over time. This demonstrates that although new members join and old members leave, organization-level compositions in terms of emergent role behaviors remain stable. A closer look at the year-by-year role composition revealed that informational support provider increased to 25.5% in 2017 from 11%-13% in earlier years (2004 2015). We also observed a weak increase for newcomer seekers, likely due to large increase in active forum users after 2015. In contrast, the percentage of welcomers in the community decreased to 4% in recent years, perhaps suggesting that old-timers, who dominate the welcomer role, are becoming less welcoming to newcomers or less polite over time.

Individual Level Dynamics

Changes in Role Occupation Over the User Lifecyle. When members first join CSN, they may have high uncertainty about the type of people who are members and the group's norms [7]. Over time those who stay may accumulate experience in terms of both domain knowledge related to their diseases and the group and its norms. This knowledge may increase people's ability to give back to the community. To investigate whether higher tenured members occupy a different set of roles than newcomers, we compared role associated with members' tenure in CSN, as described in Figure 2. Specifically, we looked at members' role occupation in their first month - (0, 1], from their second month to six months - (1, 6], from six months to a year - (6, 12], and after one year - (12, +]. Among 66,246 members, 93% of users participated in CSN in their first month after registering. Figure 2 shows that emotional support providers, welcomers, informational support providers, story sharers and informational support

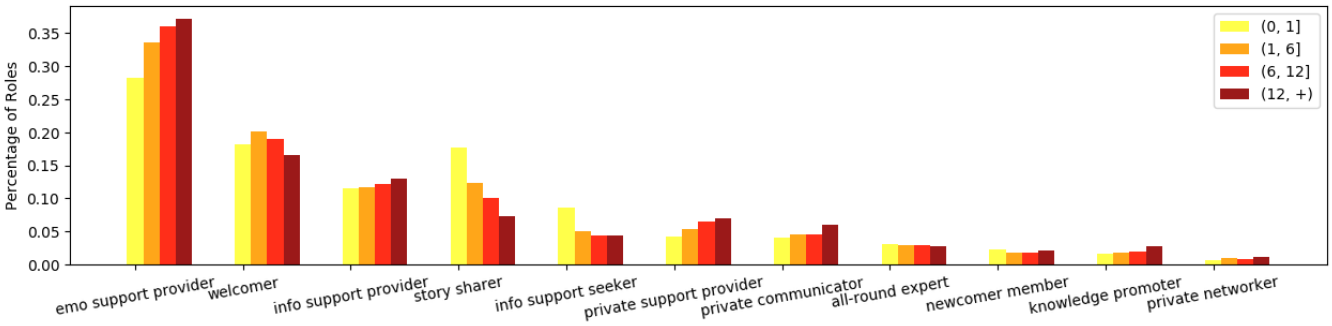

Figure 2: The percentage of role occupation for users by their CSN tenure among user who participated in CSN for at least a year. (0, 1] refers to members' first month in CSN, (1, 6] refers to their second to sixth month, (6, 12] refers to their six months to one year and (12, +) refers to after one year.

seekers were the most common roles. During members' first month on CSN, roughly 20% of them occupied the role of information support seeker, and 15% choose to share their experiences and stories to start their conversations. As tenure increases, members were more likely to occupy the role of emotional support provider, private support provider and private networker. In contrast, members are less likely to occupy the story sharer and information support seeker roles the longer they stayed on CSN, while they were more likely to be newcomer welcomers after their first month. Although Figure 2 includes only users who have been at CSN for a year, similarity results obtain for users with who have been at CSN for less than 12 months or less than 6 months.

Role transition pattern	Prob
private communicator → private communicator	0.413
info support provider → emo support provider	0.362
emo support provider → emo support provider	0.336
welcomer → emo support provider	0.335
newcomer member → emo support provider	0.330
info support seeker → emo support provider	0.326
private networker → private communicator	0.315
story sharer → emo support provider	0.312
story sharer → welcomer	0.207

Table 4: The top 9 most frequent role transition patterns.

Role Transition Processes. These results suggest that members assume different roles in different stages of participation. To further investigate role evolution, we examined the process of members' moving from one role to another across sessions. Specifically, we model users' role transitions as a Markov process, i.e., if a user assumed a particular role during session i, what is the probability that he or she would take on any specific one of the eleven roles in session $i + 1$? We calculated the presence of each role transition pattern by looking at members' roles in any adjacent sessions. Here, a user is said to occupy a role in a session if that role had the largest weight across the 11 roles. We also model a user's likelihood of dropping out (i.e., discontinuing participation in CSN) after occupying a role. This produces 132 total possible transitions (11 x 12, where the one added transition probability leads to dropout).

We described the most common transitions overall in Table 4. Since 70% members dropped out of CSN after 30 days, we calculated this transition pattern only for members who stay on CSN longer than that. We found that private communicators are the most stable role, at 41.3% carryover from session to session; users who take on this role are more likely to maintain it in their next session compared to any other role. Not only do users who provide emotional support in one session tend to continue in that role in the next session, but it is the most common role for users to transition into from other roles - 33.5% of welcomers, 36.2% of informational support providers, 32.6% of information support seekers and 31.2% of story sharers. The conditional probability of transiting from informational support seekers to emotional support providers is 0.326, confirming the typical transitions from outside observers into core members of the community [43]. This also reflects the rule of reciprocity that members who seek resources eventually give back to their communities. This showed that members transit from roles that seek for resources to roles that offer help to others. The *emotional support provider* role derives its stability partially from being a role associated with longer-term users, rather than newcomers. We show this by next deriving transition matrices *conditioned on session*. Figure 3 shows the results for two particular session transitions: from session 1 to session 2 (left side), indicating the first step of users from newcomers to group membership; and from session 10 to session 11 (right

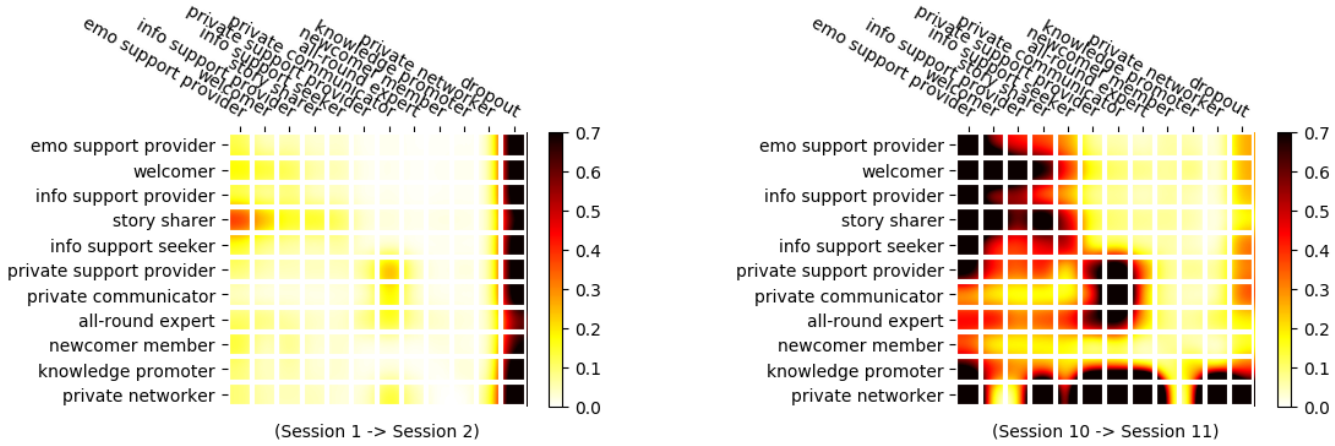

Figure 3: Conditional probability of role transitions from one session (row) to another (column) after the first (left) and tenth (right) session.

side), as an example of the more stable matrix that emerges as users become long-term members.

We found three distinct groups of newcomers. The first group does not follow any of the public roles that engage in broader discussion forum, but instead use the site primarily as a vehicle for private conversations, such as from *private communicator* to *private communicator* (25.4%). The second group is primarily *information seekers*, who then transition into providers (of both informational and emotional support) and welcomers in their follow-up sessions. The third common group, *story sharers*, are notable for their very low dropout - 64.2% of story sharers return for a second session on CSN, compared to 35.5% of first-time users that assume all other roles combined.

As tenure increases in the 10th session transition matrix, members are likely to transition out of the role of *information support seeker* and *story sharer*, and more likely to transition into the role of *emotional support providers* and *welcomers*. These roles are common and "sticky" - users have high probabilities of maintaining that role from session to session. *Private support providers* and *private networkers* were present at high rates among longer-term users, and maintain their roles over time. While support providers transition into their roles over time, *private networkers* were more likely to have taken on this role early in their tenure.

Note that for role transition analyses, we used a heuristic rule and treated each user in a session as occupying a single role - the role with the highest weight - to model the process of role transition. Since users can occupy hybrid roles, it is possible that co-occurring roles might affect our role transition results. For example, users transit from one set of roles to another set of roles in their next sessions or dropout if they did not have a next session. Future work could address this multiple role transition by modeling the mapping from 2^K roles to 2^K roles and dropout, resulting in a $2^K \times (2^K + 1)$ matrix compared to a $K \times (K + 1)$ matrix in Figure 3. For example, how do people transit from informational support seeker, newcomer seeker to emotional support provider, welcomer. However, such a complete approach might run into challenges with data sparsity, so the right course of action will likely be to investigate the tradeoffs in representation.

8 DISCUSSION

This research investigated the functional roles that members occupy in an online cancer support community, and how such role occupation influences their engagement within their communities. We first introduced a generic framework to define emergent roles in online communities with four components - goal, interaction, expectation and context. We operationalized a set of behavioral features to represent each component and then employed unsupervised models to extract the functioning roles that members occupy, which discovered 11 interpretable roles in online cancer support groups.

Among the few studies that investigated emergent roles in online communities, most have paid attention to platforms such as Wikipedia [2, 4, 57]. Previous research in online health communities suggested that there are distinct subsets of users with different "roles" [58], but had no formal methods of modeling what those subsets were. We extend this line of work into another type of community - to the best of our knowledge, the first work to use data-driven methods to identify behavioral roles in online health communities. Some of the prototypical behaviors associated with the roles we derived correspond to roles in conceptual frameworks; for

instance, our "informational support seeker" and "informational support provider" correspond to "information seeker" and "information giver" [11]. The role of "emotional support provider" seems to reflect the role of "encourager" [39, 40], which involves showing understanding and acceptance of others' ideas and suggestions.

In addition to helping define these roles, this generative model to describe subsets of users can both identify a user's assumption of a role in real time, and model how an individual member is likely to transition across roles over time. Most earlier research on role identification used limited metrics in evaluating roles, and statistical models more well-suited to analysis of static datasets, rather than real-time prediction in a machine learning architecture. These models also required metrics of success such as model fit or manual labeling, suffering from potential biases and lack of domain knowledge. To overcome such issues, in addition to quantitative validation of model fit, we followed through with in-depth interviews with 6 domain experts who have a deep understanding of CSN. The results of these interviews support the validity and quality of our derived roles. We believe that most existing empirical methods for identifying roles in other domains [2, 57] can be abstracted into this generic methodology, which can be applied to any other types of community, both online and offline.

Our studies on how roles influence members' survival revealed that socially positive roles such as support providers and newcomer welcomers were associated with staying longer at CSN. It may be that to take on these socially positive roles, members have to stay in the group for a while to be familiar with the group norms and other members; occupying such roles may also indicate that members already have relationships with and attachment to others or the group as a whole. The role transition analyses illustrate that members on CSN enact emergent roles and frequently transit to other roles, confirming prior work that such roles are transient [2].

Implication

Our research sheds light on how to build more successful online communities from both practical and theoretical perspectives. Theoretically, our work contributes to the understandings of emergent roles by introducing a general, four-component role framework. The iterative role identification process described here is reproducible broadly within the HCI community, as are our mixed-methods (quantitative/qualitative) criteria for evaluating the quality of derived roles. Practically, our role modeling methods can be employed to develop tools that detect members' needs, track their activities, and offer them help and task of interests. Such identified roles can better help patients know themselves and others. Future work should focus on incorporating this information into profile pages and other interface affordances. The derived roles can be incorporated as additional features for connecting users to other users, content and tasks based on their roles along with other information about them (e.g., their disease, expertise or, emotional support needs). In addition to the potentials in boosting the recommendation performance, members' functioning behavioral roles can also be used as explanations to users about why such recommendations are made. For example, instead of "*You might be interested in ...*," the recommendations can be explained like "*This is an information expert who can help you with breast cancer.*" Online communities could also introduce some of these derived roles as badges to encourage users to assume these roles and reward those who do.

Limitations

This research has significant limitations. While it is an initial step towards understanding emergent roles in online support groups, we do not have self-reported evaluations from CSN members about their perceived role occupations. Although we validate our derived roles with a set of domain experts, future work surveying members who tend to occupy such roles will allow us to compare model predictions with user-perceived role occupation. Second, while we make correlative descriptions of members' role occupation and their engagement on CSN, our work is not causal. Thus occupying socially positive roles may motivate users to stay longer, but alternatively, new users who were more likely to maintain membership may be more likely to perform such roles, reversing the causal link. While this research looks at one online cancer support group, we cannot necessarily generalize findings to other online health communities without further work. Finally, the opportunity to use role predictions to alter user experiences and make recommendations has important ethical considerations. We have developed a model with the potential to predict users' future behaviors in online communities, and adjust their user experience based on those predictions. However, such models have the potential to become a self-fulfilling prophecy, shepherding users into a particular activity path without giving them the full breadth of opportunity to explore other roles. As this research evolves into interventions, a crucial element for analysis will be interviews with members, observation of changes in their behaviors compared to baseline conditions, and an interdisciplinary analysis on the changed outcomes for users - particularly vulnerable, healthcare-seeking users - in these and similar communities.

ACKNOWLEDGEMENT

This work was supported by NIMH grant R21 MH106880-01 and a grant from Google to Robert Kraut. Diyi Yang was supported by Facebook Fellowship. The authors would like to thank Zheng Yao and the reviewers for their feedback.

REFERENCES

[1] George A. Akerlof and Rachel E. Kranton. 2000. Economics and Identity. *Quarterly Journal of Economics* 115, 3 (August 2000), 715–753.

[2] Ofer Arazy, Johannes Daxenberger, Hila Lifshitz-Assaf, Oded Nov, and Iryna Gurevych. 2016. Turbulent stability of emergent roles: The dualistic nature of self-organizing knowledge coproduction. *Information Systems Research* 27, 4 (2016), 792–812.

[3] O. Arazy, H. Lifshitz, O. Nov, J. Daxenberg, M. Balestra, and C. Cheshite. 2017. On the How and Why of Emergent Role Behaviors in Wikipedia.. In *Proceedings of the ACM SIGCHI Conference on Computer Supported Cooperative Work*. ACM.

[4] Ofer Arazy, Felipe Ortega, Oded Nov, Lisa Yeo, and Adam Balila. 2015. Functional Roles and Career Paths in Wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. 1092–1105.

[5] Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1373–1378.

[6] Thomas S Bateman and Stephen Strasser. 1984. A longitudinal analysis of the antecedents of organizational commitment. *Academy of management journal* 27, 1 (1984), 95–112.

[7] Talya N Bauer, Todd Bodner, Berrin Erdogan, Donald M Truxillo, and Jennifer S Tucker. 2007. Newcomer adjustment during organizational socialization: a meta-analytic review of antecedents, outcomes, and methods. *Journal of Applied Psychology* 92, 3 (2007), 707.

[8] Natalya N Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. 2015. Social sharing of emotions on Facebook: Channel differences, satisfaction, and replies. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM, 154–164.

[9] Natalya N Bazarova, Jessie G Taft, Yoon Hyung Choi, and Dan Cosley. 2013. Managing impressions and relationships on Facebook: Self-presentational and relational concerns revealed through the analysis of language style. *Journal of Language and Social Psychology* 32, 2 (2013), 121–141.

[10] Beth A Bechky. 2006. Gaffers, gofers, and grips: Role-based coordination in temporary organizations. *Organization Science* 17, 1 (2006), 3–21.

[11] Kenneth D Benne and Paul Sheats. 1948. Functional roles of group members. *Journal of social issues* 4, 2 (1948), 41–49.

[12] Bruce Jesse Biddle. 1979. *Role theory: Expectations, identities, and behaviors*. Academic Press New York.

[13] Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014. Identifying emotional and informational support in online health communities. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 827–836.

[14] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[15] Herbert Blumer. 1986. *Symbolic interactionism: Perspective and method*. Univ of California Press.

[16] Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 773–782.

[17] C Shawn Burke, Kevin C Stagl, Cameron Klein, Gerald F Goodwin, Eduardo Salas, and Stanley M Halpin. 2006. What type of leadership behaviors are functional in teams? A meta-analysis. *The Leadership Quarterly* 17, 3 (2006), 288–307.

[18] Moira Burke and Robert E. Kraut. 2008. *Mopping up: Modeling Wikipedia promotion processes*. ACM Press, New York.

[19] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms Matter: Contrasting Social Support Around Behavior Change in Online Weight Loss Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 666.

[20] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*. 288–296.

[21] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity.. In *ICWSM*.

[22] Helen Rose Fuchs Ebaugh. 1988. *Becoming an ex: The process of role exit*. University of Chicago Press.

[23] Danyel Fisher, Marc Smith, and Howard T Welser. 2006. You are who you talk to: Detecting roles in usenet newsgroups. In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, Vol. 3. IEEE, 59b–59b.

[24] Denae Ford, Justin Smith, Philip J Guo, and Chris Parnin. 2016. Paradise unplugged: Identifying barriers for female participation on stack overflow. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. ACM, 846–857.

[25] Erving Goffman. 1959. Presentation of self in everyday life. (1959).

[26] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*. 1024–1034.

[27] Eszter Hargittai and Aaron Shaw. 2015. Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia. *Information, Communication & Society* 18, 4 (2015), 424–442.

[28] Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Sugato Basu, Leman Akoglu, Danai Koutra, Christos Faloutsos, and Lei Li. 2012. Rolx: structural role extraction & mining in large graphs. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1231–1239.

[29] Isa Jahnke. 2008. Knowledge Sharing through Interactive Social Technologies: Development of Social Structures in Internet-based Systems over time. In *Building the knowledge society on the Internet: Sharing and exchanging knowledge in networked environments*. IGI Global, 195–218.

[30] Ray Jones, Siobhan Sharkey, Janet Smithson, Tamsin Ford, Tobit Emmens, Elaine Hewis, Bryony Sheaves, and Christabel Owens. 2011. Using metrics to describe the participative stances of members within discussion forums. *Journal of medical Internet research* 13, 1 (2011).

[31] Amy Jo Kim. 2000. *Community building on the web: Secret strategies for successful online communities*. Addison-Wesley Longman Publishing Co., Inc.

[32] Aniket Kittur and Robert E. Kraut. 2010. Beyond Wikipedia: Coordination and Conflict in Online Production Groups. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. ACM, New York, NY, USA, 215–224.

[33] Steve WJ Kozlowski and Katherine J Klein. 2000. *A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes*. Jossey-Bass., San Francisco:, 3–90.

[34] Elijah Mayfield, Miaomiao Wen, Mitch Golant, and Carolyn Penstein Rosé. 2012. Discovering habits of effective online support group chatrooms. In *Proceedings of the 17th ACM international conference on Supporting group work*. ACM, 263–272.

[35] Andrew McCallumzy, Kamal Nigamy, Jason Renniey, and Kristie Seymorey. 1999. Building domain-specific search engines with machine learning techniques. In *Proceedings of the AAAI Spring Symposium on Intelligent Agents in Cyberspace*. Citeseer, 28–39.

[36] Geoffrey J McLachlan and Kaye E Basford. 1988. *Mixture models: Inference and applications to clustering*. Vol. 84. Marcel Dekker.

[37] George Herbert Mead. 1934. *Mind, self and society*. Vol. 111. Chicago University of Chicago Press.

[38] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 889–892.

[39] Troy V Mumford, Michael A Campion, and Frederick P Morgeson. 2006. Situational judgment in work teams: A team role typology. *Situational judgment tests: Theory, measurement, and application* (2006), 319–343.

[40] Troy V Mumford, Chad H Van Iddekinge, Frederick P Morgeson, and Michael A Campion. 2008. The Team Role Test: Development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology* 93, 2 (2008), 250.

[41] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics* 3 (2015), 299–313.

[42] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. *UT Faculty/Researcher Works* (2015).

[43] J Preece and B Shneiderman. 2009. The Reader-to-Leader Framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction* 1, 1 (2009), 13–32.

[44] Leonardo FR Ribeiro, Pedro HP Saverese, and Daniel R Figueiredo. 2017. struc2vec: Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 385–394.

[45] Catherine M Ridings and David Gefen. 2004. Virtual community attraction: Why people hang out online. *Journal of Computer-Mediated Communication* 10, 1 (2004), 00–00.

[46] Peter Schlattmann. 2003. Estimating the number of components in a finite mixture model: the special case of homogeneity. *Computational statistics & data analysis* 41, 3-4 (2003), 441–451.

[47] David M Schweiger, William R Sandberg, and James W Ragan. 1986. Group approaches for improving strategic decision making: A comparative analysis of dialectical inquiry, devil's advocacy, and consensus. *Academy of management Journal* 29, 1 (1986), 51–71.

[48] LP StataCorp et al. 2007. Stata data analysis and statistical Software. *Special Edition Release* 10 (2007).

[49] Greg L Stewart, Ingrid S Fulmer, and Murray R Barrick. 2005. An exploration of member roles as a multilevel linking mechanism for individual traits and team outcomes. *Personnel Psychology* 58, 2 (2005), 343–365.

[50] Ralph H Turner. 1990. Role change. *Annual review of Sociology* 16, 1 (1990), 87–110.

[51] Patrick Wagstrom, Corey Jergensen, and Anita Sarma. 2012. Roles in a Networked Software Development Ecosystem: A Case Study in GitHub. *Department of Computer Science & Engineering, University of Nebraska-Lincoln, Technical Report* (2012).

[52] Yi-Chia Wang, Robert Kraut, and John M. Levine. 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 (CSCW '12)*. ACM, New York, NY, USA, 833–842.

[53] Yi-Chia Wang, Robert E Kraut, and John M Levine. 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 (2015).

[54] Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. 2011. Finding Social Roles in Wikipedia. In *Proceedings of the 2011 iConference (iConference '11)*. ACM, New York, NY, USA, 122–129.

[55] Howard T Welser, Eric Gleave, Danyel Fisher, and Marc Smith. 2007. Visualizing the signatures of social roles in online discussion groups. *Journal of social structure* 8, 2 (2007), 1–32.

[56] Miaomiao Wen and Carolyn Penstein Rosé. 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*. ACM, 179–188.

[57] Diyi Yang, Aaron Halfaker, Robert E Kraut, and Eduard H Hovy. 2016. Who Did What: Editor Role Identification in Wikipedia.. In *ICWSM*. 446–455.

[58] Diyi Yang, Robert Kraut, and John M Levine. 2017. Commitment of newcomers and old-timers to online health support communities. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. ACM, 6363–6375.

[59] Diyi Yang, Zheng Yao, and Robert E Kraut. 2017. Self-Disclosure and Channel Difference in Online Health Support Groups.. In *ICWSM*. 704–707.

[60] Haiyi Zhu, Robert Kraut, and Aniket Kittur. 2012. Effectiveness of shared leadership in online communities. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 407–416.