Thesis Proposal

Computational Social Roles: Identify, Recommend and Configure Emergent Social Roles in Online Communities

Diyi Yang

January 13, 2018

Language Technologies Institute
School of Computer Science
Carnegie Mellon University

Thesis Committee:

Robert E. Kraut, Co-Chair	(Carnegie Mellon University)
Eduard Hovy, Co-Chair	(Carnegie Mellon University)
Brandy L. Aven	(Carnegie Mellon University)
Dan Jurafsky	(Stanford University)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Copyright © Diyi Yang

Abstract

Millions of people participate in online communities, exchange expertise and ideas, and collaborate to produce complex artifacts, such as Linux and Wikipedia. They engage in a variety of roles, which strongly influence the amount and types of work they do, and how they coordinate their activities. Better understanding members' roles benefits members by clarifying how they should behave to participate effectively and also benefits the community overall by encouraging members to contribute in ways that best use their skills and interests. Social sciences have provided rich theoretic taxonomies of social roles within groups, while natural language processing techniques enable us to automate the identification of social roles in online communities. However, most social science work has focused on generic roles without accommodating the activities associated with tasks in specific contexts or automating the process of role identification. While there has been work to date about automatic role inference, identification of social roles has not had a corresponding strong emphasis in the language technologies community. A variety of methods were developed to extract specific "roles", patterns, or components in different contexts, lacking of generalized definitions about what are roles and systematic methods about how to extract roles. Moreover, how roles change over time and how the awareness of roles influence role holders' performance and the group production, have not been adequately researched in both fields.

In this thesis, I advocate for both theories of social science and models of text analysis to better define roles, develop ways to extract roles, optimally recommend roles to users, and configure roles within the community. Concretely, I focus on four perspectives. The first work constructs profiles of users from what they do and with whom they interact in online cancer support groups, from which we then extract social roles in an unsupervised manner. The second work predicts when and how members transit from one role to another and examines how role contrast helps explain the occurrences of different transitions. Third, I model how the presence of different types of roles and their interaction with task level, group tenure, and group type, predicts group performances. The last perspective investigates whether making role expectation explicit increases group performance.

I will produce both theoretical and computational results: this thesis will develop new algorithms to mine behavioral data to extract social roles and recommend roles, tasks, and groups to target an optimal group configuration; it can also advance the development of social science theories on how roles and role collaboration affect participation in online communities.

Contents

1 Introduction

We have many concepts but few confirmed theories;
many points of view, but few theorems; many "approaches"; but few conclusions.
Perhaps a shift in emphasis would be all to the good.
– Robert King Merton

Online production communities like Wikipedia aggregate the efforts of hundreds of millions of volunteers to product complex artifacts such as the largest encyclopedia in human history and the software that runs the internet. Despite their proliferation into diverse aspects of life, such communities are not always successful in soliciting contributions and producing anticipated outcomes. Two major challenges are: how to sustain members' engagement and how to coordinate users' activities to contribute to public goods and community needs. For instance, lack of appropriate contributions has left over 92% of the roughly 4.5 million articles in the English Wikipedia at "stub" or "start" quality levels, and 60% new editors do not come back (Halfaker et al., 2013). Not only in Wikipedia, in health support groups, around 10% thread-starting messages get no replies and many of the replies are not relevant to thread-starting posts, for example providing emotional support when people were seeking information (Wang et al., 2015).

In order for such complex socio-technical organizations to succeed, online communities have to depend equally on the technical infrastructure on which they rest, the policies that govern participants to behave in ways consistent with community goals, and *the behavior, roles, and coordination of their members*. The goal of this thesis is to study members' participation and coordination in online production communities, focusing on the social roles they enact, which link individual contributions with community-level coordination and outcomes (Stewart et al., 2005). To this end, I propose to integrate computational methods with insights from social science to study social roles and the optimal organization of online communities.

Social sciences have provided rich taxonomies of social roles within production groups. They range from 27 roles that Benne's and Sheats identified as fulfilling a group's needs to accomplish its production tasks, to maintain itself as a functioning group and to meet the needs of individual members (Benne and Sheats, 1948) to more recent taxonomy of ten group roles covering a similar set of functions (Mumford et al., 2006). Natural language processing research provides us with a variety of techniques to automate the identification of social roles in online communities. For example, Bamman et al. (2013, 2014) leveraged probabilistic graphical models to learn personas in movie plot summaries and English novels. Previous work also discovers roles in social networks based on the network structure, and typically focus on roles such as centers of stars, members of cliques, peripheral nodes, etc. For example, *RoleX* introduced a unsupervised approach to extract features for each node, group features and interpret clusters (Henderson et al., 2012). And *struc2vec* uses heuristic to construct a multi-layered graph based on topological metrics and simulates random walks on the graph to capture structural information (Ribeiro et al., 2017). Other examples include such models as mixed membership stochastic block models (Airoldi et al., 2008), unsupervised matrix factorization methods (Hu and Liu, 2012), or semi-supervised role inference models (Zhao et al., 2013). Another line of work formulated predefined roles as classification problems. For example, Welser et al. (2011a) identified four roles in Wikipedia: substantive

experts, technical editors, vandalism fighters and social networkers. Fazeen et al. (2011) classified Twitter users into leaders, lurkers, associates, and spammers. Other common roles identified in online media include experts (Zhang et al., 2007), opinion leaders (Bodendorf and Kaiser, 2009), and influential bloggers (Agarwal et al., 2008).

However, most social science work has focused on roles that are designed to be generic without accommodating the activities associated with tasks in specific contexts or automating the process of role identification. Although utilizing network guarantees generalizability when discovering structural roles, the crux is how to construct a network that can reflect user interactions in a meaningful and representative manner. While there has been work mining semantic actions to date for automatic role inference, identification of social roles has not had a corresponding strong emphasis in the language technologies community. Moreover, a variety of methods were developed to extract specific "roles", patterns, or components in different contexts, largely ignoring the relevant social theories on roles and lacking generalized methods about how to extract roles, let alone examining how roles change over time and how the awareness of roles influences role holders' performance, the expectations of others, and the production as a whole.

I argue that a systematic identification of social roles from a combined view from both social science and NLP is needed, and has to take into account four major challenges: (1) In contrast to roles in conventional organizations, roles in online communities are often self-selected and emergent, without explicit expectations associated with them, and limited literature to date has provided consistent definition and methodology. (2) Members' participation in online production communities are recorded in what they do, to whom, and why. Although numerous studies have discussed how to identify roles based on users' behavioral regularities, most research classified users based on their repeating patterns of activities or social network signatures, failing to capture what type of work were actually performed and for what purposes users conducted such interactions. (3) Moreover, members move upward or downward, vertically or horizontally within the community, making their roles change as a function of the tasks they perform, their tenure and audience in these communities. Understanding the mobility and stability of roles requires accurately delineating the dynamics (paths, directions, and strengths) of role transitions. (4) While some specific tasks associated with roles such as copy-editing in Wikipedia or providing support in support groups are community dependent, there are certain kinds of roles that are essential to the success of any community. The effectiveness of such roles on community success might also be different under different situation contexts. How to extract such cross-community or community-specific roles and configure roles within a specific context are still in its infancy.

Thesis Statement: In this thesis, I advocate for both theories of social roles and models of text analysis to better identify roles, develop ways to optimally recommend roles to users, and configure roles within the community. I argue that developing computational models together with theories of social roles can capture the complexity of users' interaction and answer socially important scientific questions that give us a deeper understanding of roles and communities.

Concretely, I will explore four different vintage points: The first work infers profile representation of individuals - such as *age, gender, political affiliation* in Twitter and *age, gender, disease type, cancer stage* in CSN, from what they say and what others say about them, in an end-

to-end architecture. Through this we can construct profiles and define categories of users. The second work predicts when and how members transit from one role to another through the lens of fine-grained taxonomies of actions and intentions, and examines how roles change as a function of members' tenure in Wikipedia. Third, I propose to model how the presence of different types of roles and their interaction with various context factors including task level, group tenure and group type, predicts the group performance. And last, I propose to investigate whether making role expectation explicit increases group performance, and use these knowledge as the basis to build interventions for real world benefits, such as building recommender systems that match users to roles and tasks to improve online production communities. Note that instead of analyzing the structural properties in members' social networks, this thesis focuses more on understanding roles' behavioral cues and social actions in the content of user interaction.

2 Role Definition

The ruler rules, the minister ministers, the father fathers, and the son sons.
– Confucius

As in conventional organizations, members in online communities engage in a variety of emergent and informal *roles* that define the set of activities they perform. Before we conduct any particular studies, we must be very clear about the meaning of "role". Here, I define **role** as a set of interaction patterns regulated by explicit or implicit expectations and adopted by one or more persons in a social context to achieve one or more specific social goals. This definition hangs on five terms of roles: context, person, interaction, goal and expectation.

Roles are performed by *Persons*, i.e. behaviors of human beings or organizations in some extreme cases (Biddle, 1979). Any number of persons may be studied for the roles they exhibit. Persons are characterized by their attributes or non-behavioral attributes, such as their demographics, physical features or background experiences. Person's attributes such as sex, race and age might be related to the roles that he/she occupy. For example, African-American may be more likely to do well than Asian in sports. Person's characteristics or attributes do not have intrinsic effects about roles, but sometimes we use them to predict behavior and posit such associations between roles and person attributes.

Roles are associated with specific social *Goals*. Some common social goals are: for the task, for the context as a whole, and for oneself. Roles could be related to tasks which the context is planning to take or had already taken, thus the goal here is to facilitate collective effort in the solution of the task, such as a leader in a course project team. Roles can be also oriented towards the functioning of the context (e.g., group, community) as a whole, such as Vandal Fighter in Wikipedia. Goals can be irrelevant either to the task or the functioning of the context as a whole. That is, roles can exist for the satisfaction of the person's individual needs, such as for the pleasure of mentorship or reciprocity.

Roles are persons' *Interactions* that can be observed and systematic (Turner, 1990), possibly repeated and goal-oriented. Specifically included in this definition are what persons characteristically do specifically toward the goals, to/with/for some agents/role observers.

Interactions can happen between persons (role holders) and other persons, between role holders and objects within the systems or context, or even between role holders and agents outside the context. These interactions may or may not be expected, valued, or approved either by role holders or other persons. In terms of the granularity of interaction, a role may be occupied by a single person, in one specific context, or even exhibit one single type of behavior. But a role must be based on at least multiple characteristic observed goal-oriented interactions, otherwise it is impossible for the follow-up detection. Of course, these characteristics might consist of both core/central interactional features and peripheral interactional features.

Roles are normally limited in *Contexts* or social systems, like in a group, organization or community in relation to other members. These contexts set boundaries for persons, i.e. delimiting the perimeter or setting the scope of roles. These contexts might also have their shared values or norms via which persons interact with others.

Roles also involve informal *Expectations* associated with typical interaction patterns of persons (Goffman, 1959; Jahnke, 2008). In a given online context, roles are very likely to be emergent, and there might exist some informal implicit "negotiated understandings" among individuals about what persons should do if they seem to occupy such roles, which results in positive or negative sanctions from others towards their behaviors (Blumer, 1986; Mead, 1934). Expectations are bidirectional: from role holders to others, and from others towards role holders. Expectations are beliefs or cognition held by individuals, and are part of people's internal planning processes, such as their estimation or belief of the social and personal outcomes of their behaviors. If people do not keep to such informal expectation, they might somehow become excluded by this environment.

2.1 Online Emergent Social Roles

Roles are major mechanisms through which project members, including volunteers in large online communities, coordinate complex activities (Kozlowski and Klein, 2000). Theory on coordination in groups and organizations emphasized role differentiation, division of labor and formal and informal management (Kittur and Kraut, 2010). In traditional offline organizations, most roles are **assigned roles** that are appointed formally to a social position and whose activities performed while fulfilling their roles are mainly based on social expectations, norms, and status positions. However, in online environment, only in a few cases, roles are formally defined and both the group and individual role incumbents are likely to have clear expectations of what the incumbents should do (Akerlof and Kranton, 2000), such as the moderator role in many online discussion sites or administrator roles in Wikipedia. In many other cases, roles are **emergent** and **self-selected**, with no explicit expectations associated with them. As a result, although these emergent roles constitute consistent patterns of behavior, neither the role occupant nor other community members may have clear expectations of who is occupying which role and how role occupants should behave to do their jobs well. While much attention has been paid to assigned roles such as moderators and leaders in both offline and online settings (Burke et al., 2006a; Mumford et al., 2006; Arazy et al., 2017; Zhu et al., 2012), less is known about the nature of emergent roles, its dynamics, and the consequences of role expectation and coordination on the overall communities' success. In the following, I will investigate each dimension of the five components in the definition of roles, in order to unpack the black box of **emergent** and **behavioral** roles in online communities.

3 Role Extraction

It's not who I am underneath, but what I do that defines me.
– Batman Begins

Millions of volunteers participate in online production communities, exchange their expertise and ideas, and collaborate to produce complex artifacts. They engage in a variety of roles in the process of helping the public at large. For example, in Wikipedia, editors take up different responsibilities, when editing articles, based on their interest and expertise. Some, for example, might add substantive new content to articles while others may focus on copy-editing. In online health support groups, cancer patients or survivors often give each other information and advice about the disease and treatments, emotional support, social comparison or even companionship. This raises our first ontological question: How can we build computational models to identify members' behavioral roles? What are some typical **interaction** patterns that help us recognize these roles?

3.1 Completed Work

In this part, I mainly investigate the *interaction* dimension of emergent social roles. Completed work has begun to answer this question by designing finer-grained taxonomies to capture members' actions and intentions, building on which we infer the roles that people occupy.

3.1.1 ICWSM 2016

This work (Yang et al., 2016) develops new methods to identify roles that editors exhibit when contributing to Wikipedia and then tests whether work done by editors occupying different roles affects article quality. The problem of identifying editors' roles in Wikipedia has attracted significant attention. Numerous studies have discussed how to identify roles based on users' behavioral regularities and social network signatures (Welser et al., 2007). Most research classifies editors based either on their edits in different namespaces (Welser et al., 2011b) or via the user attributes such as access privileges (Arazy et al., 2015), personalized barnstars (Kriplean et al., 2008), etc. Classification based on users' attributes is relatively accurate, but this information is not available for many active editors and is insufficient in explaining the nature of an editor's work. While classification based on edit histories can be constructed for most active editors, approaches focus on simple edit counts and access privileges fail to provide a finer grained description of the work actually performed in an edit. For example, it cannot tell the difference between an editor who copy-edits or rephrases a paragraph and an editor who inserts markup, template to an article.

To overcome such challenges, building on Daxenberger and Gurevych (2012), we proposed a fine grained taxonomy of edit types to differentiate editors who occupy different editing roles in Wikipedia. In our taxonomy, edits are distinguished contextually in terms of the object being edited (e.g. information, template, reference, etc.) and functionally, in terms of the edit operation (e.g. insert, delete, modify, etc.). Specifically, we developed 24 edit categories to understand how different users perform the editing task collaboratively. Figure 1 provides an overview of our edit taxonomy. The two top-level layers summarize whether these edit categories are meaning-preserving or meaning-changing. We then developed methods to automatically identify

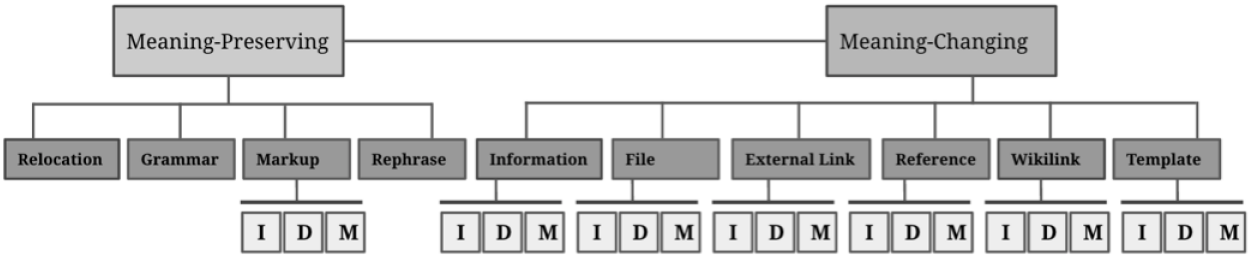

Figure 1: **The Taxonomy of Edit Categories**. Note: Insertion is abbreviated as I, Deletion as D and Modification as M

these edit categories revealed in users' edits. Based on the difference of content change between parent and current article revisions, the comments editors used to describe their edits, and characteristics of the edit, we built a RAKEL ensemble model for this multi-label classification, which randomly chooses a small subset with k categories from the overall set of categories as described in Tsoumakas and Vlahavas (2007). This gave classifications that agreed with the human judgments, indicated by the AUC score of 0.865.

Building on this automated measurement of editors' low level edits, we then used a graphical model analogous to LDA topic modeling analysis to identify the latent roles editors occupy, much as documents comprise topics. Just as documents are mixtures of topics, editors are mixtures of roles. The roles that editors occupy generate the edits they perform, just as the topics that comprise a document determine the works in it. In contrast to studies that employed either clustering analysis or principle component analysis to extract user roles (Liu and Ram, 2011), our role modeling treats an editor as comprising multiple roles at the same time. We ended up with eight roles that occur frequently in Wikipedia, including *Social Networker, Fact Checker, Substantive Expert, Copy Editor, Wiki Gnome, Vandal Fighter, Fact Updater*, and *Wikipedian*.

3.1.2 EMNLP 2017

While the above taxonomy categorizes edits into low level actions such as file deletion, image modification, simply understanding the syntactic revision operations does not provide the information we seek: *why do editors do what they do? How do their actions say about them?* For example, edit type taxonomies cannot tell the difference between simplifying a paragraph and maliciously damaging that paragraph, since both involve deleting a sentence. This nuance can largely affect our role identification, e.g., the former could be a *Simplifier* while the latter is a *Vandal*.

To address this, I modeled another dimension of emergent social roles - **goals** behind roles' interactions. Thus, we focused explicitly on revision intention by introducing a fine-grained taxonomy of the reasons why an author in Wikipedia made an edit (Yang et al., 2017a). Our semantic taxonomy of edit intentions builds on prior literature on collaborative writing (Faigley and Witte, 1981; Fitzgerald, 1987), research on document revision analyses (Bronner and Monz, 2012), studies on edit categories (Daxenberger and Gurevych, 2012), work on purpose/intention classification (Zhang and Litman, 2016), and feedback from Wikipedia editors. This categorization leads to 13 distinct semantic intentions, and Table 1 provides detailed descriptions. We define a top

Label	Description	α	Dist
Clarification	Specify or explain an existing fact or meaning by example or discussion without adding new information	0.394	4.1%
Copy Editing	Rephrase; improve grammar, spelling, tone, or punctuation	0.800	14.8%
Counter Vandalism	Revert or otherwise; remove vandalism	0.879	1.5%
Disambiguation	Relink from a disambiguation page to a specific page	0.401	1.8%
Elaboration	Extend/add substantive new content; insert a fact or new meaningful assertion	0.733	12.0%
Fact Update	Update numbers, dates, scores, episodes, status, etc. based on newly available information	0.744	5.2%
Point of View	Rewrite using encyclopedic, neutral tone; remove bias; apply due weight	0.629	2.2%
Process	Start/continue a wiki process workflow such as tagging an article with cleanup, merge or deletion notices	0.786	5.8%
Refactoring	Restructure the article; move and rewrite content, without changing the meaning of it	0.737	2.9%
Simplification	Reduce the complexity or breadth of discussion; may remove information	0.528	4.6%
Vandalism	Deliberately attempt to damage the article	0.894	2.0%
Verification	Add/modify references/citations; remove unverified text	0.797	9.8%
Wikification	Format text to meet style guidelines, e.g. add links or remove them where necessary	0.664	33.6%
Other	None of the above.	0.952	1.2%
Corpus Size		4,977	7,177

Table 1: A taxonomy of edit intentions in Wikipedia revisions, Cronbach's α agreement and the distributions of edit intention. The percentage in each row represents what percentage of revisions are labeled with this edit intention. The percentages do not sum up to 100% because one revision could belong to multiple categories.

level layer for the revision intention taxonomy: intentions that are common in general revisions: **General Revision Intentions**, and intentions that are specific in Wikipedia: **Wikipedia Specific Intentions**. Specifically, general revision intentions include: *Clarification*, *Copy Editing*, *Elaboration*, *Fact Update*, *Point of View*, *Refactoring*, *Simplification* and *verification*, and can be applicable to other contexts like academic writing. *Counter Vandalism*, *Disambiguation*, *Process*, *Vandalism*, and *Wikification* are edit intentions related to Wikipeida. Similar to the automation of identifying edit types, we introduced a labeled corpus of revisions, and developed machine learning models to recognize these edit intentions, with a micro-averaged F1 score of 0.621.

As an initial effort to uncover patterns of behaviors associated with editors, we used this model to investigate the editing work from two types of informal *roles* in Wikipedia: **survivors** and **non-survivors**. Among 100,000 randomly sampled Wikipedia users, 4,407 were survivors (i.e., made an

edit two months after registering) and 16,689 were non-survivors. We applied our edit intention model to revisions in users' first sessions, and compared the percentages of different types of edit intentions between survivors and non-survivors. We found that survivors tend to do more *Copy-editing* (Δ_+=2.3%) and more *Wikification* (Δ_+=6.5%), while non-survivors seem to perform more *Simplification* and *Vandalism*. A closer looking at the rejection of different editing work among survivors and non-survivors revealed that non-survivors compared to survivors get reverted more when performing *Wikification, Verification* and *Refactoring*, suggesting that sophisticated types of work might not be suitable for beginners.

3.2 Proposed Work: *Profile-Role Machine*

Building successful role extractors in online communities requires techniques to infer members' identity attributes and identify their roles from a representation of users' profiles and interactions, based on which we can match members with proper tasks that best fit their attributes and roles. The completed work so far has focused on developing finer-grained taxonomies of actions (*Interaction*) and intentions (*Goal*), through which we define roles. In addition to using specific actions associated with finishing tasks, I argue that obtaining a good knowledge of members such as knowing their demographic (*Person*) attributes can largely facilitate the process of role identification in social contexts. To this end, I propose to build a large-scale Profile-Role Machine to automatically construct users' profiles (profile-machine) and extract the behavioral roles they occupy (role-machine).

3.2.1 Profile Machine

Members signal their roles and membership via their language choice and style. In online communities, messages that members write can reveal substantial information about their identity, such as their age, gender, personality, location, native language, or socioeconomic status (Jurgens et al., 2017; Marwick et al., 2011). Their interactions or social relations with others also provide chances for an accurate profiling of members' identity. For instance, in Twitter, building on users' tweets, retweets, comments, likes, followings and followers, we can infer a person's age, gender, religion, etc. Numerous studies have been proposed in this context (Sakaki et al., 2014; Huang et al., 2014), however, no generalized framework exists. Most of them focused on a specific dimension such as gender, nationality or spouse (Li et al., 2014), or approach the inference of multiple identity dimensions individually (Jurgens et al., 2017), which fails to take into account potential correlations between different attributes.

The **profile-machine** uses textual messages and social network data as input representation, and predicts users' identity attributes such as age, gender, and cancer type. This enables us to take into account 1) the association between members' language and identity and 2) the correlation among different dimensions of members' profiles. Specifically, we will employ a neural network architecture to obtain input representation of users, which first uses a convolutional network to process a user's posts and then merges it with his/her social activity measures to create a vector representation for each user. We plan to apply a structured prediction energy networks (SPENs) (Belanger and McCallum, 2016) to capture the dependencies between labels (attributes of users), which will be achieved by defining an energy function of candidate labels. The predictions will be produced by using back propagation to iteratively optimize the energy with respect to the labels.

We plan to test our methods using data generated in two societally important contexts: Twitter and American Cancer Society's Cancer Survivor Networks (CSN) [1]. On Twitter, following Jurgens et al. (2017), we can construct "groundtruth" profile attributes for each user, including *age*, *gender*, *religion*, *extroversion*, and *diet*. On CSN, user profile attributes include *age*, *gender*, *cancer type*, and *cancer stage* based on patients' self-reported information. There were a total of 48,317 registered users who exchanged 1,073,020 messages belonging to 131,237 threads. We intent to predict profile attributes of users in a held-out set, and evaluate the accuracy of model predictions in both a user-dimension level and user level.

3.2.2 Role Machine

The **role-machine** analyzes users' profile attributes, text messages and social network structure in a repeated cycle of role postulation, definition, annotation, automated processing and evaluation, in order to establish a set of roles that are well-defined and enable both human and automated identification and analysis. Specifically, role-machine first identifies characteristics of each user, making a vector representation that might include profile attributes, actions, language patterns, and possible reactions from others. Then it finds assignment of behaviors to roles to identify which people function in which role with what strengths, and evaluates the derived roles.

I propose to build a role machine to identify the implicit roles that members play in an online cancer support group - CSN, with the goal of matching patients with proper information providers and caregivers. We will first build machine learning classifiers to automatically identify and measure at scale some of the important types of actions that participants perform, such as seeking and providing informational and emotional support and helping with technical problems. Building on members' profile attributes (via profile-machine), supportive actions exchanged and social network communication such as interacting with newcomers, we can then extend unsupervised graphic models to cluster those sets of heterogeneous features into coherent social roles that individuals adopt.

Our preliminary results revealed a set of roles that members occupy in this online cancer support group, such as *Short-term Seeker, Welcomer, Support Provider, All Rounder*, etc. To evaluate derived roles, we have already asked CSN moderators to judge the extracted roles and received positive feedback. In the next step, we plan to send out surveys to ask CSN users to what extend they identify themselves as the roles we generated. Then we can quantitatively compare our role predictions towards them with their self-reported scales.

3.3 Generalization

The above role identification reveals a generic process that we use to identify emergent social roles in online communities, which is a repeated cycle of **role postulation, definition, annotation, automated processing and evaluation**. This generic approach can help establish a set of roles that are well-defined and enable both human and automated identification and analysis. Each stage of role identification requires not only appropriate statistical approaches to extract roles in a valid and reliable way, but also social theoretical groundings to guide us what to measure and

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

what to expect. For instance, Benne and Sheats (1948) classified functional roles that members might occupy when participating in offline groups into three categories: group task roles such as *information seeker*, *coordinator* and *evaluator-critic*, group building and maintenance roles such as *harmoizer, gate-keeper* and *commentator*, and individual roles such as *aggressor, blocker* and *self-confessor*. Mumford et al. (2006) also identified similar role topology in work teams. These can be utilized as seeds to guide what to expect for role postulation. Similarly, for role definition and annotation, extensive social science work has already produced a set of role-based actions that are generalized across different contexts. For example, Burke et al's meta-analysis attempting to evaluate the types of leadership behaviors that are functional in teams (Burke et al., 2006b). It collapsed the hundreds of behaviors identified in 65 taxonomies of leadership behavior into a much smaller two categories of task-focused and person-focused behaviors. Task-focused behaviors deal with task accomplishment, e.g., initiating structure, acquiring task information through boundary spanning, distributing praise and sanctions, and making decisions. Person-focused behaviors are designed to facilitate team interaction or development, e.g., proposing a direction of subordinates, maintaining social relationships with team members, and resolving conflicts. I argue that these theoretical groundings will be of great value as insights for unsupervised techniques on what to measure. Moreover, automated role identification in different online contexts can be used in turn to validate existing roles and discover new ones, contributing to the social science field.

4 Role Evolution

No man ever steps in the same river twice, for it's not the same river and he's not the same man.
– Heraclitus

Changes in identity and roles are universal experiences in offline world like turning to the adult role from a child or to the PhD role from a freshman. Members of online communities also go through a life cycle of participation, such as making transitions from outside observers into core members of the community, as specified in the Reader-Leader framework (Preece and Shneiderman, 2009). Formally, we can view role transition as a process of moving from one role to another through a series of events or episodes (Holt, 2008). As members go through their life cycles, they might move upward, downward, vertically or horizontally. As a result, their roles are changing as a function of the tasks they perform, their tenure and their audiences. How does one move through a community, via which paths in which directions? What constructs or components from the communities (*Context*) affect their role transition? In this section, I focus on answering these questions to better understand role transition processes in online communities.

4.1 Completed Work

4.1.1 CHI 2017

We began with investigating how members' commitment to their social contexts change as members stay longer in an online cancer community (Yang et al., 2017b). To survive and prosper, online communities must recruit, socialize and retain successive generations of new members. If a community does not replace members who leave, it will eventually die. New members may be especially valuable because they are often a source of new ideas, diversity, and innovation

for the community (Kraut et al., 2010). In contrast, higher tenured members may serve other important roles (Preece and Shneiderman, 2009). However, evidence from a wide variety of online communities indicates that a substantial number of new members drop out before they could plausibly receive many benefits from the community or contribute benefits to others (Arguello et al., 2006; Halfaker et al., 2013; Wang et al., 2015). Moreover, higher tenured members can drop out as well, and their loss may have an especially large impact on the community.

Thus, in order to understand the health of online communities, it is necessary to understand the factors that influence both new and more established members' commitment. Although others have investigated commitment to offline and online organizations, little research in either setting has examined how the factors influencing commitment change with participants' tenure in the community. We use Levine and Moreland's model of group socialization to theoretically frame our research (Lee and Allen, 1982; Levine and Moreland, 1994). According to this model, members engage in an evaluation process to determine how well the group can satisfy their needs. In so doing, they consider how rewarding the group has been in the past, how rewarding it currently is, and how rewarding it is likely to be in the future. The outcome of this evaluation process determines members' commitment to the group, which in turn affects the likelihood that they will remain in it and expend effort to achieve collective goals.

We focused on the relationship between members' commitment to health support groups and their exchange of informational and emotional support. We first built machine learning models to automatically identify the extent to which messages exchanged in the discussion forums contained emotional or informational support. We operationalized commitment using three correlated metrics: members' self-reported attachment, their willingness to help others in the community by providing them social support, and the length of time they continued to participate in the community. We then used hierarchical regression models to predict respondents' self-reported commitment and to predict the number of comments they provided to others in the next month using the amount and content of comments participants received in a month. Survival analysis was also used to assess how tenure and the amount & type of support members receive jointly predict the length of their subsequent participation. Through these analyses, we found that: (1) The extent to which members sought informational and emotional support on the site *declined* with tenure. Although members initialized more threads in their first month on the site than they did subsequently, initiating threads declined with tenure after the first month. This suggests that seeking resources on the site was less important for users as they gained experience, either because people with short-term support needs dropped out of the community, motives for participating in the community changed with experience, or higher tenured users didn't need to probe the community to establish how well it fits their needs. (2) Receiving comments from others was generally associated with increases in members' commitment to the community. However, receiving informational or emotional support in the comments was associated with declines in commitment. (3) Receiving comments was generally associated with larger increases in commitment for higher tenured members than for lower tenured ones. However, receiving informational or emotional support was associated with more commitment for lower tenured members than higher tenured members.

4.2 Proposed Work: *Role Evolution*

The above work differentiated members based on their tenure in online communities and concluded that members of different tenure react differently to their received communication and support. However, it has not answered *how* members' behavioral roles evolve over time. As we mentioned above, online production communities are characterized by movements from all levels of members among periphery, immediate or even core positions of responsibility and influence at the community. Although many studies have explored the various activities and different ways in which users participate (Dahlander and O'Mahony, 2011; Butler et al., 2002; Burke and Kraut, 2008), few work has provided a systematic transition analysis that involves a broader set of roles and their interaction patterns in online production communities.

I investigate the process of how members' prior role enactment contributes to their transition to other roles. As a starting point, I propose to study the transition processes of how editor candidates get promoted into administrators in Wikipedia, i.e., *Requests for adminship*[2] (RfA) by designing a series of hypotheses. Instead of utilizing our derived latent roles as units of analyses, we turn to assigned roles in Wikipedia since it provides strong supervision to our role transition modeling, and avoids potential error cascade introduced by automated role inference.

4.2.1 Method

In Wikipedia, anyone can request adminship ("RfA") from the community, regardless of their experience. Users can either submit their own requests for adminship (self-nomination) or may be nominated by other users. The Wikipedia community decides who will become administrators – who are users with access to additional technical features that aid in maintenance. Any registered editor can comment on a request, and each editor will assess each candidate in their own way. An RfA remains open for seven days, after which a bureaucrat will review the RfA and determine consensus on it. To date, the English Wikipedia has 1,239 administrators.

This proposed work predicts who will be promoted to administrator status in Wikipedia based on their prior role enactment. In Wikipedia, administrators are considered as trusted custodians of the successful encyclopedia, and an experienced or biased administrator might cause significant damage to the encyclopedia and demotivate other editors. Burke and Kraut (2008) operationalized a set of explicit criteria that many RfA evaluators look for nominees based on the Guide to RfA[3], including: strong edit history, varied experience, user interaction, trustworthiness, helping with chores, high quality of articles, observing consensus, edit summaries, etc. They found that extensive and diverse experience, as well as coordination on talk pages and edit summaries are good predictors of promotion, while editors who help with chores and observe consensus are not more likely to be promoted, not line with criteria on the Guide to RfA. Although this work operationalized factors based on the community's stated criteria and correlated them with promotion success, it used easy-to-measure behavioral data as proxies for abstract criteria, and did not look into the quality of contribution, debate process and other factors identified in the organizational behavior literature (Ng et al., 2005).

[2] https://en.wikipedia.org/wiki/Wikipedia:Requests_for_adminship

[3] https://en.wikipedia.org/wiki/Wikipedia:Guide_to_requests_for_adminship

In addition to explicit criteria that Wikipedia editors look for in candidates for promotion, we are also interested in whether voters also use implicit cues beyond what was defined in the explicit criteria. While an understanding regarding the nature of role promotion is beginning to form, less is known about how these emergent roles that community members occupy contribute to their promotion process. Thus, I propose to investigate the transition processes of how editor candidates get promoted into administrators. I argue that our derived roles in Wikipedia such as *Substantive Expert*, *Copy Editor*, *Wiki Gnome* can better capture the wider contribution of candidates. As specified in the Guide to RfA, strong edit history and varied experience are two positive indicators of promotion success. Compared to simplistic edit count and the number of involved name-spaces, roles based on users' low-level edit types may provide more power in reflecting "varied experience". Roles such as *Fact Checker, Social Networker* might reveal other latent cues like network relations that evaluators look for. Trust and recognition from other Wikipedian towards a candidate's expertise or roles may suggest to evaluators that he/she will use administrators rights carefully and properly to avoid irreversible damage. Another reflection of candidates' qualification is that candidates are already performing most work that administrator do.

To sum up, candidates' prior role enactment, recognition from others towards their skills and social network relations should all positively correlate with their successful transition to administrators. Specifically, we expect that (1) The more roles candidates have performed and get recognized by others, the higher their chances of getting promoted to administrators. (2) The more administrators candidates have interacted with, the higher their chances of getting promoted to administrators. (3) The smaller the behavioral contrast between candidates and administrators, the higher their chances of getting promoted to administrators.

Building on our work on identifying editor roles in Wikipedia, we first extract the emergent roles that RfA candidates occupy. We then measure the number of roles candidates perform, their received recognition such as barnstars[4] or others' trust shown in the discussions of the RfA evaluators, the amount of interactions candidates had with administrators, and their behavioral similarities and differences. We can then utilize such factors in a statistical framework to predict promotion decisions.

5 Role Configuration

The self is not something ready-made, but something in continuous formation through choice of action.
– John Dewey

As discussed previously, active volunteers in online communities often take on consistent patterns of behavior that define their social roles. For example, some forum members of cancer support groups routinely provide informal support while other provide emotional support. In Wikipedia some editors focus on substantive writing, while others specialize in enforcing style and citation standards (Yang et al., 2016). These roles help the volunteers themselves decide what to do and help others interact with them to anticipate how they will behave. How can we formalize roles and appropriately recommend them to members to improve online production communities? Does there exist an optimal role configuration? Can we extract such configuration

[4] `https://en.wikipedia.org/wiki/Wikipedia:Barnstars`

patterns associated with successful collaborations to provide guidance for unsuccessful teams? To answer these question, in this section, I conduct empirical analyses, build new theories of optimal community design based on that and evaluate them by conducting randomized lab and field experiments.

5.1 Completed Work

5.1.1 ACL 2015

Type	Behavior Definition	Example Messages
Team Building	Invite or accept users to join the group	*Lauren, We would love to have you.* *Jill and I are both ESL specialists in Boston.*
Task Management	Initiate a task or assign tasks to a team member	*Housekeeping Task 3 is optional but below are the questions I summarize and submit for our team.*
Collaboration	Collaborate with teammates, provide help or feedback	*I figured out how to use the Google Docs.* *Let's use it to share our lesson plans.*

Table 2: Three Different Types of Team Member Behaviors

To begin with, we first showed a possible way to extract a set of roles associated with successful role collaboration outcomes. Specifically, we examined the interaction between team members as they work together to achieve instructional goals in their project work in a team based Massive Online Open Courses (MOOCs) platform called NovoEd[5] (Yang et al., 2015). Our modeling goal is to identify behavior profiles that describe the emergent roles that team members take up in order to work towards a successful group grade for their team project. Our role identification process is iterative and involves two stages. The first stage adjusts the weight (role) vectors to predict the teamwork quality, given a fixed role assignment that assumes students are well matched to roles; the second stage iterates the possible assignments and finds a matching to maximize our objective measure. The two stages run iteratively until both role assignment and teamwork quality prediction converge. One essential component in our teamwork role identification models is student behavior representation. We first identified three main team members behaviors based on messages sent between team members as shown in Table 2. Such collaboration behaviors together with members' communication language choices and linguistic styles are characterized to represent the behaviors of team members. The experimental results on two MOOCs show that our proposed role identification models can not only perform accurate predictions of teamwork quality, but also provide interpretable student role assignment results, like that our identified leading role has substantial overlap with team leaders.

5.1.2 RecSys 2014

As we discussed, online communities are not always successful in soliciting contributions and producing anticipated outcomes. For example, in Massive Open Online Courses (MOOCs), although thousands of students can register for courses to learn at their convenience with no monetary cost, they suffer from high rates of attrition due to the rare interaction within courses. Problems students struggle with in the discussion forums, such as difficultly in finding

[5] `https://novoed.com/`

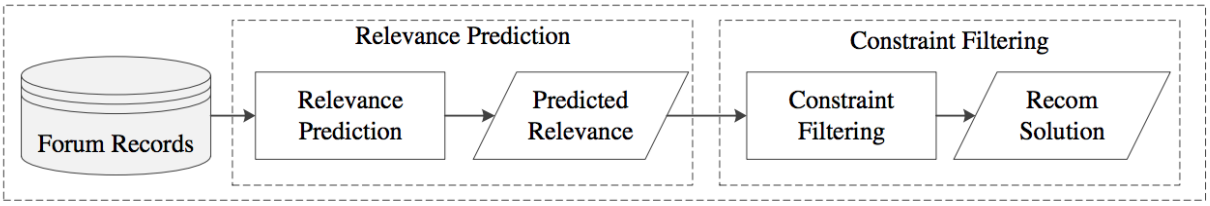

Figure 2: **The Framework of Constrained Question Recommendation**

interesting discussion opportunities or attracting helpers to address posted problems, provide new opportunities for recommender system to suggest appropriate answerers to help student seekers. In contrast to one-way product recommendation, question recommendation in the MOOC context requires the recommendation to be reciprocal and attractive to both the asker and the helper side, which existing techniques are inadequate in modeling. For this purpose, I formulated a novel constrained question recommendation problem to take into account two constraints: (1) Load Balancing - students should not be over-burdened with too many requests; and (2) Expertise Matching - students should not be requested to address problems they are not capable of addressing (Yang et al., 2014). Then I designed a two-phase framework to solve it as shown in Figure 2: the first phase estimates students' preferences over questions via a context-aware matrix factorization model, and the second phase optimizes community benefit under multiple constraints via a max cost flow model. Experimental results conducted on three MOOC datasets demonstrate that our method significantly outperforms baseline methods in optimizing overall forum welfare, and in predicting which seekers student helpers might be interested in.

5.2 Proposed Work: *Role Interaction*

The collaborative nature of the online production communities lead to naturally forming groups within those communities, such as individuals working on the same article in Wkipedia or the core participants in a subreddit in Reddit. However, within these naturally occurring teams, members are not often aware of roles they are playing towards achieving their common goals. This lack of awareness can also contribute to sub-optimal coordination on how most effectively distribute the efforts of each members of the group. Existing research shows that adhoc groups organized to improve a Wikipedia article are more effective when they are composed of a mixture of editor roles, including one or two editors who take the lead by defining the structure for the article combined with less committed editors who fill in detail (Kittur and Kraut, 2008). Thus, I hypothesize that *these naturally occurring teams will be more successful if it has clear role assignment, an appropriate role configuration, and a role coordination mechanism.*

5.2.1 Method

This part attempts to determine *what roles* are needed for optimal community production and at *when*. Drawing upon Biddle (1979), Stewart (2005), Mumford (2006) and Burke (1967), we first categorize roles into task oriented roles and social-emotional oriented roles. **Task roles** are identified in relation to functions of facilitation and coordination of group problem-solving activities, such as *information seeker*, *evaluator critic*. **Social-emotional roles** serve the building of

group-centered behavior and fulfill emotional community needs, such as *encourager*, *facilitator*.

The configuration of roles within a team are a major mechanism through which individual behavior is linked to group-level team coordination (Kozlowski and Klein, 2000). Benne and Sheats (1948) suggested that groups in different stages of an act of problem selection and solution might have different role requirements. For example, based on our prior work (Yang et al., 2016), Wikipedia articles in "stub" status might require less of *Cleanup Workers* than more mature articles. The stage of the group - the level of group maturity might also affect the composition of roles, since groups in different stages might have different goals. As suggested by Benne and Sheats (1948), a young group might require less of the role of the "standard setter" than a more mature group. Similarly, an "information giver" might be perceived as more important compared to an "encourager" in a course project team, while the latter might be more useful for support groups. Therefore, an appropriate role configuration should take into account these factors.

I propose that the effectiveness of roles on group success depends on the level of task, the level of group tenure, and the type of group. Specifically, we might expect that (1) task roles might be required more in the beginning of task stage. (2) Social-emotional roles are needed more in the beginning of group development. (3) Task roles are needed more in task-oriented groups, and social-emotional roles are required more in relation-oriented groups.

This work will be conducted in three different types of Facebook groups, including support groups, study groups and relation groups. Building on our role identification method, we plan to extract members' roles based on their interaction patterns in their groups. Such emergent roles will be further divided into task oriented roles and social-emotional roles. Then I will design statistical analyses to test how the presence of different types of roles predicts the changes of group collaboration outcomes such as group activity or time engagement, together with the interaction between roles types and task stage, group stage and group type.

5.3 Proposed Work: *Role Expectation*

Based on research in social psychology and organizational behavior on role formalization, when roles are formed, they become clearer. Moreover, the behaviors associated with roles shift from being simply descriptive – i.e., statistical regularities – to becoming prescriptive – obligations people have towards themselves and each other (Turner, 1990). When prescriptive norms are highlighted, people are more likely to act consistently with them (Cialdini et al., 1991). In addition, formalizing roles turns the implicit role patterns into an explicit identity (*Expectation*), which increases people's likelihood of performing the central behaviors associated with these roles.

Thus, I propose to establish the casual relationship between formalizing social roles and contributions in online contribution communities. Specifically, I hypothesize that labeling members with titles that reflects their informal roles will lead them to act more consistently with this behavioral patterns than members who are not labeled. It will have a greater effect on members' behavior if they and other community members can observe the role occupation of each other. The effect of role labels on members' behavior will be greater if the expectations about role-specific behavior are clearer.

5.3.1 Method

I use the below experiment to explicitly test the impact of assigning distinct roles to participants in a production task, by making role expectations explicit (i.e., role clarity) and making the identity of the role occupants visible to other members of the production group (i.e., role visibility). This experiment will build on previous experimental research we have conducted in which teams of Amazon Mechanical Turk workers improved a Wikipedia article (Tauscik et al., Under review). In our proposed research, we will conduct a $2 \times 2 \times 2$ experiment crossing *role assignment* (participants are explicitly assigned a role or not) by *role clarity* (participants receive guidelines for the duties associated with each role or not) by *role visibility* (all members of the team know who is occupying each role or only the role occupant is knows). We will conduct this research by assigning Turkers to improve a Wikipedia article. We will introduce problems in the article (e.g., typos, poor structure, non-sourced claims and inclusion of personal opinions) and give the participants brief training about editing Wikipedia articles and resources to improve the article (e.g., new content to include in the article; relevant and relevant sources). We will measure the quality of their work in terms of the extent their final article meets guidelines covered in training, as well as evaluations by trained coders of content completeness and writing quality. We predict that quality will be improved most when participants are assigned to roles, when role responsibilities are clear, and when all participants are informed about who should do what work.

6 Conclusion

There is no real ending. It's just the place where you stop the story.
– Frank Herbert

This thesis presents four work to identify, recommend and configure social roles in online communities from a combined view of social science, natural language processing and human-computer interaction. Each piece of work interacts with others in complex ways and pushes forward the research of social roles to different extent. The first work introduces an end-to-end architecture to learn profile and role representation of members in online communities. Building on those, the second one predicts how and why members' roles evolve as they move through their life cycles by focusing on a specific transition instance – Wikipedian's promotion into administrators. The third work proposes to model how the presence of different types of roles and their interaction with various context factors including task level, group maturity and group type predicts the group performance. The last one tests our theories on whether making role expectation explicit increases group performance via a set of lab experiments. By understanding roles theoretically via designing social science theories, computationally via proposing machine learning algorithms, and practically via conducting causal natural experiments, this thesis will demonstrate the possibilities of studying complex and systematic human social behavior.

6.1 Timeline

The overall goal is to complete this thesis by the Spring of 2019. The below presents a tentative timeline for my next two years.

Jan 2018 - May 2018	Chapter 3: Role Extraction
May 2018	EMNLP 2018 Deadline
May 2018 - Sep 2018	Chapter 4: Role Evolution
Sep 2018	CHI 2019 Deadline
Sep 2018 - Nov 2018	Chapter 5: Role Interaction
Dec 2018 - Feb 2019	Job Search and Applications
Nov 2018 - Feb 2019	Chapter 5: Role Expectation
Feb 2019	Writing
Feb 2019	Defense

References

Nitin Agarwal, Huan Liu, Lei Tang, and Philip S Yu. Identifying the influential bloggers in a community. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 207–218. ACM, 2008.

Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. volume 9, pages 1981–2014. JMLR.org, June 2008.

George A. Akerlof and Rachel E. Kranton. Economics and identity. *Quarterly Journal of Economics*, 115(3):715–753, August 2000.

O. Arazy, H. Lifshitz, O. Nov, J. Daxenberg, M. Balestra, and C. Cheshite. On the how and why of emergent role behaviors in wikipedia. In *Proceedings of the ACM SIGCHI Conference on Computer Supported Cooperative Work*. ACM, 2017.

Ofer Arazy, Felipe Ortega, Oded Nov, Lisa Yeo, and Adam Balila. Functional roles and career paths in wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1092–1105, New York, NY, USA, 2015. ACM.

Jaime Arguello, Brian S Butler, Elisabeth Joyce, Robert Kraut, Kimberly S Ling, Carolyn Rosé, and Xiaoqing Wang. Talk to me: foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 959–968. ACM, 2006.

David Bamman, Brendan O'Connor, and Noah A. Smith. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

David Bamman, Ted Underwood, and Noah A Smith. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 370–379, 2014.

David Belanger and Andrew McCallum. Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992, 2016.

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

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

Herbert Blumer. *Symbolic interactionism: Perspective and method*. Univ of California Press, 1986.

Freimut Bodendorf and Carolin Kaiser. Detecting opinion leaders and trends in online social networks. In *Proceedings of the 2nd ACM workshop on Social web search and mining*, pages 65–68. ACM, 2009.

Amit Bronner and Christof Monz. User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'12, pages 356–366, 2012.

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

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

Moira Burke and Robert Kraut. Mopping up: modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 27–36. ACM, 2008.

Peter J Burke. The development of task and social-emotional role differentiation. *Sociometry*, pages 379–392, 1967.

Brian Butler, Lee Sproull, Sara Kiesler, and Robert Kraut. Community effort in online groups: Who does the work and why. *Leadership at a distance: Research in technologically supported work*, pages 171–194, 2002.

R.B. Cialdini, C.A. Kallgren, and R.R. Reno. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in experimental social psychology*, 24:201–234, 1991.

Linus Dahlander and Siobhan O'Mahony. Progressing to the center: Coordinating project work. *Organization science*, 22(4):961–979, 2011.

Johannes Daxenberger and Iryna Gurevych. A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *Proceedings of COLING 2012*, pages 711–726, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

Lester Faigley and Stephen Witte. Analyzing revision. *College composition and communication*, pages 400–414, 1981.

Mohamed Fazeen, Ram Dantu, and Parthasarathy Guturu. Identification of leaders, lurkers, associates and spammers in a social network: context-dependent and context-independent approaches. *Social Network Analysis and Mining*, 1(3):241–254, 2011.

Jill Fitzgerald. Research on revision in writing. *Review of educational research*, 57(4):481–506, 1987.

Erving Goffman. The presentation of self. *Life as theater: A dramaturgical sourcebook*, 1959.

Aaron Halfaker, R. Stuart Geiger, Jonathan Morgan, and John Riedl. The rise and decline of an open collaboration system: How wikipedia's reaction to sudden popularity is causing its decline. *American Behavioral Scientist*, 57(5):664–688, May 2013.

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

Ian GS Holt. Role transition in primary care settings. *Quality in primary care*, 16(2), 2008.

Xia Hu and Huan Liu. Social status and role analysis of palin's email network. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 531–532, New York, NY, USA, 2012. ACM.

Wenyi Huang, Ingmar Weber, and Sarah Vieweg. Inferring nationalities of twitter users and studying inter-national linking. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 237–242. ACM, 2014.

Isa Jahnke. 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*, pages 195–218. IGI Global, 2008.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. Writer profiling without the writer's text. In *International Conference on Social Informatics*, pages 537–558. Springer, 2017.

Aniket Kittur and Robert E. Kraut. *Harnessing the wisdom of crowds in Wikipedia:: Quality through coordination*, pages 37–46. ACM Press, New York, 2008.

Aniket Kittur and Robert E. Kraut. Beyond wikipedia: Coordination and conflict in online production groups. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 215–224, New York, NY, USA, 2010. ACM.

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

Robert Kraut, Moira Burke, John Riedl, and P Resnick. Dealing with newcomers. *Evidencebased Social Design Mining the Social Sciences to Build Online Communities*, 1:42, 2010.

Travis Kriplean, Ivan Beschastnikh, and David W. McDonald. Articulations of wikiwork: Uncovering valued work in wikipedia through barnstars. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pages 47–56, New York, NY, USA, 2008. ACM.

Denis MS Lee and Thomas J Allen. Integrating new technical staff: Implications for acquiring new technology. *Management Science*, 28(12):1405–1420, 1982.

John M Levine and Richard L Moreland. Group socialization: Theory and research. *European Review of Social Psychology*, 5(1):305–336, 1994.

Jiwei Li, Alan Ritter, and Eduard H Hovy. Weakly supervised user profile extraction from twitter. In *ACL (1)*, pages 165–174, 2014.

Jun Liu and Sudha Ram. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Trans. Manage. Inf. Syst.*, 2(2):11:1–11:23, July 2011.

Alice E Marwick et al. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133, 2011.

George Herbert Mead. *Mind, self and society*, volume 111. Chicago University of Chicago Press., 1934.

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

Thomas WH Ng, Lillian T Eby, Kelly L Sorensen, and Daniel C Feldman. Predictors of objective and subjective career success: A meta-analysis. *Personnel psychology*, 58(2):367–408, 2005.

J Preece and B Shneiderman. The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction*, 1(1):13–32, 2009.

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

Shigeyuki Sakaki, Yasuhide Miura, Xiaojun Ma, Keigo Hattori, and Tomoko Ohkuma. Twitter user gender inference using combined analysis of text and image processing. *V&L Net*, 2014:54, 2014.

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

Yla Tauscik, Rosta Farzan, Robert E. Kraut, and John Levine. Consequences of socializing newcomers collectively in online communities. *Transaction on Social Computing*, Under review.

Grigorios Tsoumakas and Ioannis Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *Machine learning: ECML 2007*, pages 406–417. Springer, 2007.

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

Yi-Chia Wang, Robert E Kraut, and John M Levine. 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.

Howard T. Welser, Eric Gleave, Danyel Fisher, and Marc Smith. Visualizing the Signatures of Social Roles in Online Discussion Groups. *The Journal of Social Structure*, 8(2), 2007.

Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, iConference '11, pages 122–129, New York, NY, USA, 2011a. ACM.

Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, iConference '11, pages 122–129, New York, NY, USA, 2011b. ACM.

Diyi Yang, David Adamson, and Carolyn Penstein Rosé. Question recommendation with constraints for massive open online courses. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 49–56, New York, NY, USA, 2014. ACM.

Diyi Yang, Miaomiao Wen, and Carolyn Rose. Weakly supervised role identification in teamwork interactions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1671–1680, Beijing, China, July 2015. Association for Computational Linguistics.

Diyi Yang, Aaron Halfaker, Robert E Kraut, and Eduard H Hovy. Who did what: Editor role identification in wikipedia. In *ICWSM*, pages 446–455, 2016.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1990–2000. Association for Computational Linguistics, 2017a.

Diyi Yang, Robert Kraut, and John M. Levine. Commitment of newcomers and old-timers to online health support communities. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 6363–6375, New York, NY, USA, 2017b. ACM.

Fan Zhang and Diane Litman. Using context to predict the purpose of argumentative writing revisions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1424–1430, San Diego, California, June 2016. Association for Computational Linguistics.

Jing Zhang, Jie Tang, and Juanzi Li. Expert finding in a social network. In *Advances in Databases: Concepts, Systems and Applications*, pages 1066–1069. Springer, 2007.

Yuchen Zhao, Guan Wang, Philip S. Yu, Shaobo Liu, and Simon Zhang. Inferring social roles and statuses in social networks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 695–703, New York, NY, USA, 2013. ACM.

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