Advancement Proposal:
Social Role Temporal Dynamics and Interactions in Online Communities: How are Leaders and Members Different?

by

Ryan Compton

A proposal submitted in partial satisfaction of the

requirements for the advancement to

Ph.D. Candidacy

in

Computer Science

in the

Graduate Division

of the

University of California, Santa Cruz

Committee:
Professor Marilyn Walker
Professor Steve Whittaker
Professor Lise Getoor
Professor Sri Kurniawan

Spring 2016

Chapter 1

Introduction

1.1 Why communities are important

In 2012, a global survey found that 96 percent of internet users use the internet at least once a day. Of those internet users, 90 percent use it to connect with other people, and a majority of users (60 percent) interact with others online daily [24]. Online communities emerge from such online connections through things like common interests or circumstances. These communities typically form among individuals who are unknown to each other offline. Over the past 15 years, online communities have been among the most popular applications on the internet [35]. Usenet, a worldwide distributed internet forum, had over 160,000 newsgroups active in 2006, and Yahoo claims to host over a million online groups [35], Wikipedia reportedly hosts over 28 million users with around 100,000 participating daily[1], and GitHub, a collaboration platform founded in 2008, has accumulated more than 3.4 million developers and 5.9 million repositories as of 2014 with about 10,000 new users and 20,000 new repositories every day[2].

There are many definitions of an online community [35, 42, 66], and though there are subtle differences, these definitions have a shared core. In this paper I will adopt the common definition of online communities adopted by Preece [66] and Kraut and Resnick [35], who state that online communities are "*any virtual social space where people come together to get and give information or support, to learn or to find company"*.

People within online communities have a wide range of benefits that are similar to the benefits of offline communities. Participants within a community have opportunities for information sharing and learning, for companionship, social support, and for entertainment [35, 45, 83]. These benefits even stretch beyond the community to benefit others by providing goods like open source software, product reviews, and encyclopedia pages [35]. These products benefit from the diversity and range that a community can reach online. As Kraut and Resnick [35] argue, the promise of online communities is that they break the barriers of time, space, and scale that limit offline interactions.

Online communities can vary vastly in size, ranging from just a few people to millions of users. These communities have changed the way people interact by eliminating boundaries that constrain offline interactions. Online communities differ from social networks, as social networks tend to be bound together by pre-established interpersonal connections. In contrast, online communities typically connect networks of strangers around a common interest, topic, or goal. Online interactions give people access to the knowledge of others whom they would not typically encounter offline. Some examples are health communities, like breastcancer.org, providing support for those dealing with similar circumstances [35, 83], curation groups on pinterest.com filtering vast ever-growing sets of domain-specific information available online [56], GitHub providing a resource sharing platform where people

join teams to contribute to building software [20], and intranet enterprise communities allowing communication between workgroups [44]. With the rise of social software provided on the web, online communities also provide many different tools for collaboration. Historically, communities used simple communication tools such as email or Q/A forums [32] but newer *social tools* such as wikis, file repositories, and blogs have created diverse methods for communication and information sharing to occur [43].

There is extensive research studying various aspects of online communities, including what makes communities successful [3, 28, 35, 67] and how communities change over time [5, 14, 28, 35, 37]. Communities have also been found to vary in type, from large groups with common interests or practices to smaller task-based groups which have a shared goal for a particular project or function[52]. Researchers studying communities have defined community types in terms of their social attributes [36], supporting technology [36], relation to physical communities [36], functional characteristics[88], members' needs [38], and sponsoring organizations [16]. Some typical types found across multiple contexts are those of Communities of Practice and Teams. Communities of Practice are considered groups of people who have a shared interest or practice, where within the community they share information and network [52, 88]. Teams are communities that are working on a common goal, project, or function [52]. They typically are found to be working toward a well-defined "deliverable", and are common in multiple contexts from enterprise based communities [52] and out in the open internet within Open Source Communities like those on GitHub [39].

Other studies have examined communities at the user level, specifically the behaviors and formation of social roles [11, 42, 48, 62, 92, 94] and user dynamics over time [15, 18, 47, 50, 57, 58, 67, 70].

1.2 Social Roles in Online Communities

Online communities are inherently about collaboration and, as such, communities involve social interactions that have also been studied in offline settings [86]. One critical characteristic of communities is the concept of *social role* [23]. Online communities were one of the first Web 2.0 applications, where content is generated by users themselves leading to a more democratic style of interaction and governance. This is in contrast to a media or news site where content is largely crafted by the site owners. And unlike social media applications like Facebook, participants do not usually know one another well. This leads to questions about how community leaders incentivize participants to help others when they have no strong interpersonal ties to the community. The same issue provokes questions concerning the trustworthiness of community contributions: how are participants who do not know each other able to judge the reliability of others' posts? Current community models [28, 67] assume that participants are initially drawn to the community by an interest or specific question and that the community provides multiple different roles in an ecosystem that supports different levels of participation: from simply reading/lurking, occasional posts, active contribution, through to active stewardship of the community. These community models argue that social roles are critical in answering these questions about incentives to contribute, information quality, and social norms through the conventional behaviors that each role assumes.

For instance, a key factor in encouraging people to work together online is trust and empathy. People often trust each other because they see similarities between themselves and other people, so they in turn encourage others to participate [67]. This is where communities find roles beginning to develop and some researchers label such trust-enhancing behaviors as signifying a specific role that users adopt within that community, this role is defined as one that promotes trust and empathy, with the role labeled as collaborator (users who contribute to discussion, cooperation, and working together to create or share information) [67] or motivator (users that keep conversations going)[53]. More diverse behaviors occur in long term stewards, who reveal behaviors ranging from offering reliable answers to community questions, to more unambiguous leadership roles such as organizing community policies.

A major binary distinction in research on social roles is between member and leaders. Members provide the majority of content and interactions within a community, while leaders perform the much needed meta-community management such as establishment of community norms and explicit policies [10, 67]. Leaders also facilitate coordination among members [59]. Much research has been conducted on the benefits of leadership for a community [59], defining key leader behaviors [11, 19, 47, 67, 94], as well as distinguishing leader and member behaviors [11, 45]. While there are clear distinctions between members and leaders in both theory [28, 67] and observations [11, 45, 59], other work has suggested that the picture is more complex, observing that behaviors can be shared between roles [67, 94], with members sometimes enacting leadership behaviors. Certain behaviors that are typically associated with leaders, such as offering directive, positive and negative feedback, and person-focused leadership styles, have been observed in both leaders and members indicating a presence of non-formal leadership [94]. Directive leadership is an act of requesting or pushing forth goals and actions toward members of a community and Zhu et al. [94] finds that members exhibit this style of leadership more often than formal leaders. The authors found more person-focused styles, which encompass such behaviors as welcoming new members or simply posting positive emoticons, to be more formal leadership behaviors.

While the existence of apparently overlapping behaviors between members and leaders make it seem like a difficult problem to distinguish categories of social roles, there are still benefits to having roles being defined and labeled as best argued by Gleave et al.[23]. The authors make the argument that discovering social roles benefits social science research through: (1) encapsulating the differences in behavioral and influential factors of different types of users, (2) showing the alternative actions and imposed social structure between users, and (3) allowing for an understanding on an individual's choice of interaction with others given certain conditions. Social roles have been a window for social scientists to study the underlying structure of social interactions. For example, Nolker and Zhou [53] define a user to be a motivator if they perform an action of motivating other users to participate within a community. This label identifies a user who exhibits a behavior of interest (keeping a conversation going), and it shows that this behavior is not shared across all users thus making it a distinct role observation. With this in mind, I am proposing to further *study the behaviors, interactions, influences, and changes that individuals of certain social roles go through in an online community*. As it has already been stated, defining and understanding social roles is

key for community functions. It is therefore critical that we examine how roles differ, how they combine and change over time, and how they contribute to community success.

1.3 Research Questions

There are many outstanding questions in online community research at the user level. To study functions and change in social roles within communities, there are certain fundamental questions that must first be addressed. To simplify, I will begin by exploring two very common and well-studied roles: leaders and members.

One question is:

1. ***Can we reliably distinguish member and leader roles?***

This question builds on prior literature that studies users and their social roles. Of course there are many different aspects of community behavior, so the question can be broken down into multiple sub-questions, including:

- *Do members and leaders differ in overall posting behaviors? Do leaders post more overall?*
- *Is the content of posts different across members and leaders? Are members more likely to post questions and leaders to reply to them?*
- *Do members and leaders vary in the community tools they use? Are leaders more likely to create wikis and members more likely to use forums?*
- *Do members and leaders reference third party content differently?*
- *Do members and leaders vary in the social structure that they build for themselves? Are leaders more broadly networked across the entire community?*

These questions outline a plan for how the main question can be addressed. Throughout this work I will primarily be using the terms roles or social roles in reference to the two main roles of members and leaders. While there are many other types of roles defined, in order to make progress in understanding more complex processes of network influences and time dynamics, it is necessary to have a strong precise grounding of particular roles. For this proposal I will first define and test role metrics and models. Examining simple behaviors such as posting and where posting is made mainly reproduces previous work [19], but this ground work will be expanded upon in the later questions. Keeping in mind the types of contributions that a user can make, from different social tools [43] to referencing content through hyperlinks [56], it will also be important to examine the content of posts[15, 70]. It is expected that members will show less overall posting activity [67] and vary in the type of content being contributed, for example it is expected that members post more questions and more leadership roles answer questions more often [28]. Further exploration in posting behaviors is related to the type of social tool that a user can utilize. As some previous research has examined some social tool differences [43], it is yet to be examined how social roles use tools. However, it can be expected due to observed and theorized role behaviors that roles will utilize the tool that suits the purpose of the role. For example, leaders may be utilizing tools that are focused on community logistics, such as wikis, while members will be using more questions and answering tools such as forums.

Post are not limited to only the text present. Other types of content can exist in the forum as hyperlinks, which give a reference to content existing either inside or outside of the community. This is another avenue to

explore as linking can be a role specific behavior. For instance, building a community directory can only be useful with links involved; therefore links could be highly present within leader contributions. Furthermore there are tool effects that need to be taken into account i.e. do roles post links within different tools?

Users adopt varying roles [11, 42, 48, 62, 92, 94], and not all role behaviors are exclusive [94]. I will therefore explore the extent to which behaviors overlap between roles. It is expected that leaders have more person-focused leader styles (such as thanking others for work) as well as have broader social connections within the community [53, 67, 94], but members will display more directive leadership than that of formal leaders [94].

I will also examine how roles relate to one another. While this is a more complex question and will be dealt with in later sections, the groundwork of understanding individual networking behaviors within an online community will be addressed. Specifically in the expectation that members have fewer overall connections within the social network [67], but rather may be connected with sub-groups formed within the community[28].

Although I will begin by exploring these two very common and well-studied roles, once these roles have been examined, I will secondarily explore finer grained role definitions such as readers/lurkers, contributors, collaborators, motivators, experts, and social networkers. These more nuanced roles can provide a richer picture of different user behavior within a community. These extensions of role types can help extend work that has theorized that social roles sometimes share responsibilities leading to considerable overlap between behaviors [53, 67, 94]. However, this is only a secondary exploration of other role definitions. It isn't certain if there are clear metrics or methods to identify these types of roles, so it is not a main research goal since future questions are heavily reliant on precise role definitions and observations.

2. What is the ecology of social roles?

Research on social roles has generally focused on individuals and their roles rather than determining the ecology of different roles within the community. In other words prior models fail to explore how different roles co-exist and relate within a community. This second research question addresses whether there are typical configurations of roles at a community level. For example we might ask whether communities need a minimum number of leaders or whether there is a minimum ratio of leaders to members, as well as examining if there are common interaction patterns associated with roles. Relevant sub-questions here are:

- *Are a certain number of leaders or members necessary for a community?*
- *Are members and leaders interacting within or between their respective roles?*
- *Do community types influence the ecology of social roles?*
- *Are leadership styles more interleaved within different types of communities?*

Prior work has observed some aspects of community ecology, mostly noting that community members are more prevalent than leaders [52], but observing few cases where leaders are not present to help invigorate community behaviors [59]. However, such configuration of member to leader ratios may depend on community type and goals [28]. For instance, it can be expected in more team communities that directive leadership is more prominent and

necessary in order to accomplish the goal of producing a product, while communities of practice are present with higher social or person-based leadership.

Some expectations from prior work can be that leaders will be infrequent overall, but a minimum level of leader activity is necessary as communities cannot function without leader behaviors [59]. This minimum level of leadership activity is expected to serve as a catalyst for the community to be more active, thus increasing the level of member posting activity. Furthermore, communities types such as Communities of Practice are expected to have large member populations compared to the number of leaders and the leaders will have less influence on the members, but smaller Team-based communities will have higher member to leader interactions [52].

An expansion of this question deals with understanding roles and their relation to online community social norms. Online community norms are similar to social norms observed in an offline setting. Social norms are considered to be a consensus of perceptions and actions within a group setting. Some community norms that have been observed are those of community jargon, conventions for content attribution [15], interaction styles (paralinguistic cues), and message content [65]. Researchers studying online community norms have found that increasing an individual's social identification with a group does increase their conformity on prevailing attitudes and behaviors with that online group [64, 71]. Addressing the question of social roles and community norms is complex and the main work of understanding who is enacting norms can be addressed here. Further work into role dynamics and community norms will be address later in question 4.

3. Are role behaviors successful?

This is a key question that is addressed in prior work on role behaviors[45, 66, 70, 83]. Many possible measures of community success have been proposed[28, 35, 66] and while this work can and needs to be heavily studied further, this proposed work will stick with well documented measures of success. Success is a critical question, as it helps determine the effects of the various role, ecology, and behavioral differences that we have previously described. Previous work will be examined to elicit known behavioral measures of success obtained from scraping online communities. This will be supplemented through community surveys. Questions related to success will be:

- *Are social role specific behaviors influential to success?*
- *How do community ecologies relate to success?*
- *What interactions between roles are needed for a successful community?*

When answering these questions, it will be necessary to examine success in the behaviors which have been documented as successful. For this work, success will be presented as the indicated successful behaviors found within prior work. As it comes to hypothesizing which role behaviors may be more successful, it is intuitive to think leaders will be more influential[11, 28, 59, 67, 94]. However not all leadership styles are expected to be beneficial, for example styles such as positive reinforcement and interpersonal social style may predict an increase in user's number of contributions, while negative feedback styles are negatively predictive [94]. In relation to populations,

too many leaders active may actually be a hindrance on certain community types, as work by [45, 52] implies that too many leaders led to in-action.

Final expansions on this work is related to research that has identified key community members through observation, in particular key members are defined as those who drive community norms (community jargon or tool usage) or promote user activity and conversations [45, 53]. Finally

4. How do roles change over time?

Research has explored both areas in how communities and user change over time. There are clearly defined models of user change [62, 67] but community change models takes into account the dynamics of both the roles and the community examining how individuals change according to community overall behaviors defined as community norms [15, 70]. The outline plan of this question is rather complex as temporal analysis is adding time which is inherently influential to all previously mentioned phenomena. Since many theoretical models of community and user change give specific stages to where the progress of change is being made, many of these models are hard to operationalize. As proposed by Iriberri and Leroy [28], community lifecycles go through the stages of Inception, Creation, Growth, Maturity, and Death. Many of the stages that are contiguous are hard to differentiate from each other, such as Creation and Growth, which are both defined by the increase in community membership and the creation and use of a common vocabulary. However, the stages that have some distance in time, like those of Creation and Maturity, have more distinct behaviors allowing for easier categorization between the two. For instance, the Creation stage expects more direct recruitment and cultural formation behaviors (writing of community rules and regulations), while the Maturity stages sees more trust and lasting relationships as well as subgroups existing. This is also true in user lifecycle models[67], where new members typically contribute very lightly with either voting, rating, or simple replies to already existing forums, while leaders enact stronger motivational contribution and enforce community rules. When examining temporal effects, this work will remain simple, at first examining purely early and late behaviors, and only moving to more precise trend analysis when differences are found between beginning and late behaviors. To simplify this it will be addressed as such:

- *Do users typically change as hypothesized in models, particularly in the early and late stages of said models* [67]*?*
- *As members become more experienced, do they find themselves stagnate within a role or do they continue to gain responsibilities and become more leader-like* [47]*?*
- *As the community ages and leadership and member roles are interchanged between users, is there ever a lacking of a role, in particular a lack of leadership, for a community or are roles constantly filled when there is a need* [28, 67]*?*
- *Is there a relationship to collective user development over time that shifts the community norms or do community norms influence user development and influence whom* [15, 28, 42]*?*

Previous theoretical work states that members should increase their responsibilities over time, taking over tasks that are initially performed by leaders [28, 67]. This shift in responsibilities is considered to be a product of the

community becoming more mature [28]. Iriberri and Leroy [28] theorize that the community goes through various levels of development, from inception and creation, to maturity and eventually death. They argue that it is the creation and maturity stage in which most user change occurs as that is when the highest amount of user activity is present. However contrary to some of these models, it has been observed by some studies that members can actually deviate from community norms as their stewardship increases within a community [15, 70].

It is expected to see that members will increase in responsibilities over time, as theorized by existing models [28, 67]. This shift is expected to be seen when examining simple differences in early and late behaviors. It is further expected that leaders will change less over time, allowing members to develop leadership skills and behave more so as significant influencers in the community [28, 94]. In relation to community ecology, member to leader ratios are also expected to change over time corresponding to the early and late levels of the community lifecycle as theorized by Iriberri and Leroy [28]. Adding to the complexity as stated before, while members do become more active, some their behaviors can diverge from community norms [15, 70], this may indicate users becoming experts of their own and less reliant on the community to provide content. However, this needs to be considered in caution to those who lead community norms. There is some work on understanding who leads community norms [28, 53] as well as who is being influenced by norms [64, 65, 71], allowing for work to be done in examining if there are shifts in norm leaders and followers.

Finally, we need to examine these hypotheses in different community contexts. I will begin by exploring these questions in the context of online enterprise communities. I chose this context first because it is an understudied context, as most research has explored communities of practice on the open internet. Second I have a dataset which has three significant properties: (1) it has leader and member roles already labeled from participants self-declaring their roles, (2) the dataset has hashed identifier information, so that while members are anonymized it is still possible to track the same user's behavior across multiple communities and (3) I have access to subjective data from participants about the perceived success of their community.

While the above questions are important to examine within that single enterprise community context, it is important to see whether results generalize outside that domain of study. I will therefore examine a second community context: Open Source Online Communities where the goals are similar to enterprise team communities, in that communities are commonly formed around teams who have a specific goal in mind or are focused on a deliverable, however, they are likely to include fewer communities of practice. I also anticipate being able to develop survey instruments to allow me to assess subjective perceptions of success within these communities.

This work will benefit our understanding of online community behaviors and can utilize this knowledge to further develop better community tools in order to accomplish not only community goals, but individuals' goals. For instance, if a particular behavior is present in the community that is known to be of a negative effect to the community. Then a leader can be informed in where this is occurring and they can be informed how to deal with such interactions. Another example is in that of a newcomer to a community, suppose they are looking for a questioned answered but don't know how to attract the right people for this question. Community tools can take advantage of role labels and inform those who can help where their help is needed.

1.4 Challenges

Some issues with using the enterprise community data set on hand is that it only can look into a specific timeframe, roughly 2009 till 2014. This is rather limiting in its sampling, as no further observations can be made if need be. This gives a need to find an additional similar sample, which is addressed by the plan discussed above in collecting data from open source communities.

Many of the research questions proposed above are set up in a sequential order, meaning they are dependent on the results of the prior question and all are dependent on the ground work of differentiating social roles between leaders and members. This is why I am proposing to focus mainly on examining just members and leaders and only expanding to other social roles if high precision models are made to detect those given roles. If roles cannot be reliably distinguished [13, 67, 87, 94], then it will be difficult to identify the role ecology, role behaviors, and role dynamics within a community. This won't be an issue within the enterprise communities' data set as that role is already defined in the data. However, this may be a challenge when work is being conducted outside of enterprise communities.

Other common challenges with online community research relate to availability and noise of data. As noted by previous work [67] typical users contribute little or nothing at all, leading to a power distribution of very few highly contributing users. While this creates an opportunity for identifying significant individuals[53], nevertheless general user trends are going to be difficult to discover with low amounts of data available. Having said this, even with this skew present in the majority of community contexts, noticeable trends have been found in user behaviors [5, 11, 15, 39, 70] as in leadership styles being classified at the individual contribution level [94].

Examining individual effects at a community level is difficult as there is a potential of confounding variables in a theoretical model. For example, it is hard to say if specifically a leader is influencing members to post more often, when it could potentially be the members increase in relationships with the community. Many confounding variables can obscure studies as they introduce an unobserved, possibly unexplainable, or even hidden truth to understanding the effect of interest. Various empirical methods have been proposed to take into account such a possibility [78] and will need to be adopted into this work.

Chapter 2

Background and Related Work

2.1 Distinguishing member and leader roles

Social roles typically are defined in relational terms; i.e. a role only exists in relation to others who are likewise enacting social roles [23]. As we have seen, online communities provide a mechanism allowing different modes of participation through different roles. One major distinction in communities which is heavily what my work will be focused on is between leaders and members. Members can contribute in many different ways that have different effects on the community, the default of which being a single post. Members can also be divided into subcategories, such as those discussed above in user lifecycle models: readers, contributors, chatters, collaborators, and motivators [28, 53, 67]. As described in lifecycle models [67], readers, or lurkers, are 'entry level' users who only consume information, which they find through browsing or searching, only when users begin evaluating or creating content through rating, tagging, reviewing, posting, or uploading does the user become a contributor [67]. Contributors are simply distinguished from other roles by users contributing content to the community and the content is not focused on a social aspect such as networking, encouraging behaviors from other users, or creating social policies and norms. Chatters is sub area of contributor but they are identified through conversation patterns that are not community supportive [53] such as the behaviors mentioned above that influence other users, but these conversations can be seen as social networking behaviors which build up relationships. Collaborators and motivators are much more prevalent roles when it comes to social relationships [67]. Collaborators do develop relationships just as chatters, however collaborators do more by working together with others and even set goals, while motivators encourage conversation [53]. Figure 1 shows a separation between a Motivator and a Chatter as proposed by Nolker and Zhou [53].

Attributes and Measures of Motivators

Attributes	Measures
The average distance to all other members, puts this individual in the middle	High closeness
High posting count spread evenly over lots of threads	Low thread IDF and a low one- way conversation IDF
Has a mix of responses, both direct and indirect two way.	Moderate discussion ratio

Attributes and Measures of Chatters

Attributes	Measures
Talk a lot but only to a few people.	High TF*IDF in two way conversations
Majority of their two way conversations are direct	High discussion ratio

Figure 1. Attributes and Measures of Motivators and Chatters

Within figure 1, we can see that Motivators and Chatters are differentiated through various metrics such as behavioral and structural network measures like closeness, which is a measure of the average distance to all other members within a social network [84], indicating that Motivators interact with many users as opposed to Chatters which are frequently interacting with only a few select individuals [53].

Figure 2 shows the expected evolution that a user would undertake through such roles, chatter and motivator are not included as they are defined separately, but are subcategories of contributor and collaborator respectively. The arrows show that users are not confined to roles and can move back and forth between them. An example by Preece and Shneiderman (2009) brings up how some users may be a collaborator for a small instance by working with another user to create content (such as building a wiki) and then move back to being simply a contributor. These distinctive behaviors can assist in creating a user profile, in which can help in understanding the community ecology of roles through examining how many roles exist and are taken up by users.

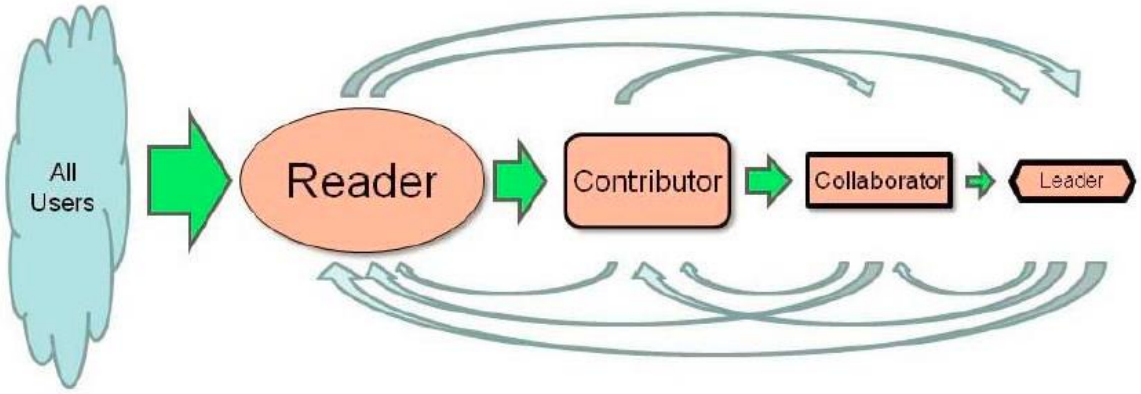

Figure 2. User Role Lifecycle Model, Preece and Shneiderman [67].

Not all roles are predefined by theoretical models. Previous work by Welser et al. [87] examined Wikipedia communities finding which were common behaviors across users and then defined these behaviors to a high level categorization, i.e. they labeled each behavior as a social role. Roles they found were not too different from theorized roles. Four main roles were found: Substantive experts, Technical editors, Counter vandalism, and Social networkers. Substantive experts are similar to experience contributors, in that they contribute by providing substantive content to the community. Technical editors are those that find small errors such as spelling, grammar, hyperlink format, or out of date facts. These editors are contributing in small but necessary aspects in the community, which these behaviors are described in the model proposed by Preece and Shneiderman [67] as being acts of contribution that users can conduct to start their role as a contributor. Counter vandalism is a more specific role to that of Wikipedia as they find vandalized articles, correct them, and sanction vandals, but these are not different from leadership behaviors like enforcing community norms and policies. Lastly social networkers are very

much like collaborators and motivators in that they are building ties with other users through channels other than typical content creation.

Welser et al. further examined how each role varies in the social network they surround themselves within the community. Figure 3 shows the results found by Welser et al. indicating that each role has a specific configuration in how they fit into the network. The most noticeable differences are that of social networks and substantive experts. Social networks keep to their group of friends and talk with one another while experts show large communities that develop relationships with fellow experts and outsiders of that interconnected subgroup.

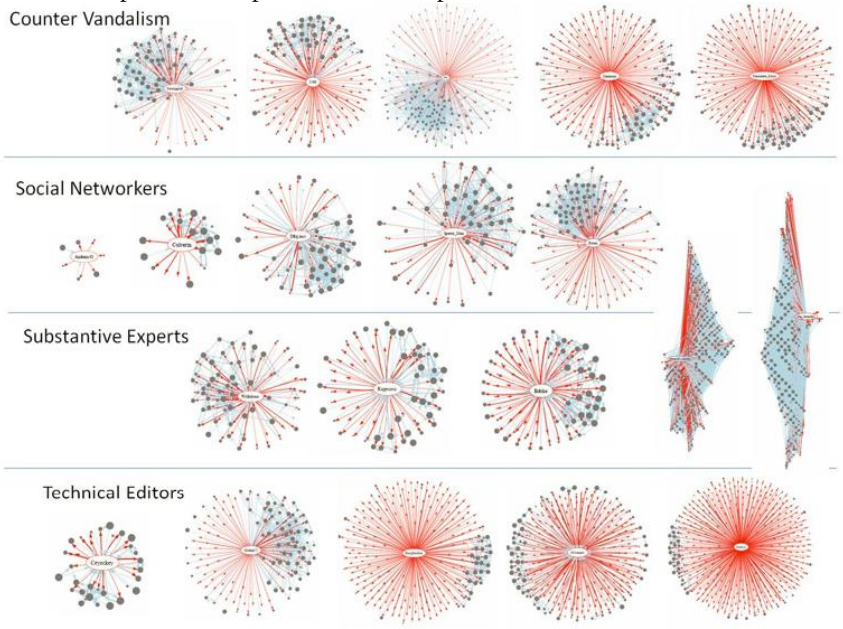

Figure 3. Egocentric network graphs found by Welser et al.[87]

These role definitions show that although there are specific behaviors to community domains, overall a universally defined model of social roles may be obtainable.

2.1.1 Leaders and their typical behaviors.

Leaders are more formally defined as the key role that promotes participation, mentoring, and setting and enforcing norms and policies [67, 94]. Previous work has identified a range of leadership behaviors: Transactional, Aversive, Directive, and Person-based leadership [94]. Table 1 shows these leadership styles as categorized by a machine learning model built by Zhu et al. [94]. Transactional leadership is when the interaction between the leader and member is considered a transaction or exchange, where the leader is providing a praise or reward for the member, "Great job, thanks for the work!" and in some cases even withhold from punishment. Aversive leadership is in contrast to transactional; instead the leader uses intimidation and reprimands to decrease undesired behaviors from targets ("If you continue in this manner you will be blocked"). Directive leadership involves the issuing of instructions and commands for members specifying their responsibilities ("Please finish this task as soon as possible"). Directive leaders can also be involved with the assignment of goals for members. Person-based leaders are defined by placing emphasis on the member being a person and forming a personal relationship with that

member. Person-based leadership works through encouragement, inspiration, intellectual stimulation and empowering, where the leaders focus on developing self-management skills of the member and team work.

Machine learning (ML) categories & Leadership type	Sample messages
ML category: Positive feedback **Leadership type:** Transactional leadership (Task-focused)	*"I award this barnstar[1] to XXX for your help and assistance in getting the WikiProject user warnings to the review phase, and to let you know your work has been appreciated."*
	"Thanks for all your work on the Survivor articles"
ML category: Negative feedback **Leadership type:** Aversive leadership (Task-focused)	*"If you continue in this manner you will be blocked from editing without further warning."*
	"...there is a concern that the rationale you have provided for using this image under "fair use" may be invalid. ... If it is determined that the image does not qualify under fair use, it will be deleted within a couple of days according to our criteria for speedy deletion."
ML category: Directive message **Leadership type:** Directive leadership (Task-focused)	*"Please read the instructions at... Using one of the templates at..., but remember that you must complete the template..."*
	"... one of these days do you think you could take some pictures at Mission Mill? I'd like to spruce up the article but it really needs some photos..."
ML category: Social message **Leadership type:** Person-focused leadership	*"Hi XX. Welcome to WikiProject XXX! I saw your name posted on the members list and wanted to welcome you... Anyway we are glad to have you. If I can help at all let me know :) ..."*
	"[[Image:Smiley.svg]] has smiled at you Smiles promote WikiLove and hopefully this one has made your day better... Happy editing"

Table 1.Sample messages from Zhu et al. [94] studying Wikipedia editors and their corresponding machine learning categories and leadership types

These defining characteristic of a leadership style give grounding for a leadership profile opposite to a member profile. Such a foundation grants the ability to define various actions of leaders and members allowing for stronger models of roles and their measurable influences on communities.

2.1.2 More complex models of leaders.

Pluempavarn et al. [62] showed that formal roles of members and leaders vary across community contexts, while Zhu et al. [94] and D'Innocenzo et al. [17] suggested a shared leadership framework in which there are less formally descriptive roles but responsibilities and behaviors to explain leadership in online communities. The shared leadership framework [17, 94] better explains the existence of members being collaborators and motivators who are sharing the responsibilities of promoting participation and mentoring typically associated with leaders. It is at this level that social roles are assumed to have the highest responsibilities that are affecting communities and create a complex social ecology[23]. These subcategories remain to be fully understood in their effect on communities.

Other work has examined different ways in which leaders presence influence a community. Panteli [59] examined four different forms of presence a leader can enact. The four types of presence are interactive (the extent

of leader interactions with their followers in terms of frequency and responding in an engaging manner), instructive (leaders taking on a more formal role such as a moderator), stimulating (leaders exerting an inspiring influence on members), and silent (leader is made available to members, but are not expected to interact with members on a frequent basis) as shown in Table 2.

Forms of leaders' online presence	Categories of online leaders	Characteristics – key features of leader behaviour	Examples
Interactive	Emergent leaders	Frequent role enactment through posts, responses and comments to other users; they arise to the leadership role due to their expertise and enthusiasm in the subject matter	'BL is a leader who is friendly, caring and active enough to regularly reply or interact with her followers' (BC interviewee 6)
Instructive	Appointed leaders	A form of emergent leaders; they are people who are recruited or elected to the post; frequent role enactment as expected by their assigned role and this is exercised through warnings, rules, enforcement and facilitation	'Hello Everyone … a **reminder** to all participants that we **should follow the rules** when posting and interacting with each other. Please **debate the issues** and not the poster, **respond kindly** to each other posting on topic and **with proof** if necessary … ' (moderator, IC1).
Stimulating	Community founder	Leader introduces topics for discussion; leader makes minimum intervention in discussions	We are again over 2 million unique visitors in March … I **don't intend to publish the stats every month** here, but I want you all to know that we have stabilised at a higher level. (SL's post to the community, March 2011)
Silent	Sustaining leaders	A leader is mainly silent; minimum input to the community; solidarity among members	'Recently, **BL does not contribute** to BC as much as she used to … her **disappearance** has **not affected** the way I follow the site' (interviewee/BC member 10)

Table 2. Summary of Findings by Panteli [59]

While the first three categories are not different from previously defined leadership behaviors, the fourth category, silent presence, is a unique influence. This presence was identified by Panteli [59] through the decrease in posting behaviors by leaders, however the limited number of posts that are made attracted high attention from members. This decrease in leader posting behaviors was found to be related to higher member interactions, indicating that the leader is allowing for more member interactions by interfering less. However, it is necessary for the leader to show that they are still involved, allowing members to believe that leaders are still actively reading the content being produced.

All these complex role behaviors need a framing to how they interact with the community as a whole. If this work sticks to the similar research approach of only identifying role behaviors, then it won't be able to answer fully how role interactions and populations influence community success.

2.2 Ecology of social roles

Addressing the larger concern of role interaction will need the perspective of examining the community as an ecosystem of social roles. Social role ecologies involve the balance and interaction of roles within a population [23] and previous work has examined multiple communities as an ecosystem being a collection of communities sharing a topic within a technology platform or organization [95]. An ecological perspective gives a missing influence factor that lifecycles only assume and state to its simplest form (typically defining only dyad relationships), but there are much more complex relationships that can exist within a community.

One form of ecological perspective can be based on organization ecological research [95] which is more community centric. Organization ecology research creates two ecosystem mechanisms: competition and complementary. Competition is where organizations compete with others in the same ecosystem for common resources as well as the intensity of competition through how similar the resource requirements are. Complementarity describes the benefits organizations get from the existence of competitors, i.e. more competitors of a business within a given location cause more customers to gather in one area. An ecological perspective can be obtained and is already being addressed with structural network methods [5, 33, 37, 50, 60, 70].

Some work has been accomplished on understanding population dynamics within Open Source Online Communities. Loyola and Ko [39] adapted biological models called Lotka-Volterra models, which are used for describing host-parasite interactions, to use in understanding how the population of contributors within a GitHub community evolve. They find this adaptation was able to explicitly providing a method of understanding population dynamics within a community over time. Other studies have been conducted on understanding the linguistic ecology of online gaming communities [77] and exploring the information structures of hyperlinks and how community ecologies are organized [21].

2.2.1 Role Interactions

As social roles are defined in relational terms, many roles cannot exist without the presence of other roles. Using an example from Gleave et al. [23], in a support group “question people provide the base material that stimulates answer people to generate replies”. Within online communities these interactions can be measured by the influence one role has over another. Using the example above, a definition of a question person is discovered by their influence within a community, i.e. do they answer questions? Finding influential users is a common research topic [13, 53, 59, 92, 94] and such methods to find influential users are to examine conversations between users [3, 59, 94], social network methods identifying central nodes within a network [13, 37, 53, 92], and examining difference between the user behaviors and community norms [15, 70].

Work in role interactions take two different approaches, either they examine the user-community interaction [3, 12, 15, 60, 70] or focus on smaller dyadic interactions between two users [14, 92]. Both approaches are essential to fully understanding interactions taking place within an online community. Community effects can take place by an aggregation or accumulation of common feedback that one can receive from the contributions they make, as found by Cheng et al.[12] where community negative feedback for a user promotes the user to produce more content but that content is actually of lower quality. Structural methods, models using a social network perspective and measures of individuals relationships, influences, and position with the network, have been used to study influences on users' joining, relationship forming, and communication behaviors [5, 33, 37, 60]. Further work shows how either structural features within a social network predict how likely a user is to conform to community norms [92] or linguistic deviations from community linguistic norms can predict lower levels of user participation [15] or even predict the user to leave a community [70]. Individual effects are just as important as shown by Zhao et al.[93], finding conversation sentiment can identify influential users which conduct successful behaviors like community building and information retrieval. The work discussed previously by Zhu et al.[94] further shows the influence of individuals by examining which leadership styles promotes contributions within Wikipedia communities.

Role interactions are a key part of social role research, as social roles are defined by their interactions with others. This proposed work will use previous methods to examine social role influence and ecosystems to further expand on the understanding of role dynamics.

2.3 Online Community Success

2.3.1 Success: Definitions and approaches.

Much of the prior work on online communities relates the phenomena of interest to either improved or declining community quality. Typically this is referred to as online community success. There are varying degrees to which online community success has been defined, from counting simple activities such as posts, voting, and editing [3, 12, 28, 66] to satisfying users intentions for joining a community or user retention [3, 12, 28, 66]. While there is agreement that some simple quantified metrics such as volume of messages posted indicate a more success online community, they have been criticized to be ignoring the content, quality, and community response that may be taking place [28, 45] for example, some messages may be considered spam or negative feedback. More precise measures targeting user perceptions such as sociability measure of member satisfaction, reciprocity, and trustworthiness can be seen to account for what is lacking [28, 45, 66].

2.3.2 Perspectives on Success.

Many of the success factors play a key role that interacts with each other; they only take a different perspective. Factors such as community growth, user retention, topic, interaction with other communities, and quality of information all take a community level perspective, while many other factors look at the individual level, such as response time for a user, quality responses, and community feedback (that through rating systems of a user's post).

All of these measures can lead to a common theme of how well does a community satisfy its user base, but this is not exclusive. At the community level, growth is in terms of how many users have an interest in the content present [28] as well as how much novel content is provided rather than content that is recurring in multiple communities [95]. Content quality is intuitively related to user's interest and for a community that has bad content, users won't see the value that the community brings and thus the user either never joins or leaves [12]. This is theoretically representative by a measure of user satisfaction. A study by Matthews et al.[46], conducted a survey of users and asked "how well this community is meeting your needs" where they would answer on a Likert scale. This feedback encapsulates many theories around community success; however it is not an all-encompassing measure of success as Preece [66] points out that they are usability factors that need to be taken into account as well as other sociability factors such as reciprocity and trustworthiness.

2.3.3 Who is successful? Variation in Goals of Social Roles

Member satisfaction does match one interesting point brought up by Preece [66] and Matthews et al.[45], when asking about community success, who is being successful? This brings up the factor that success for one individual does not necessarily indicate a successful community. A study by Matthews et al.[45] found that goals between members and leaders vary. For example leaders may be focused on community development; therefore a highly active community can be successful, while a member may be looking to answer a specific question. These goals can be in contention with each other, as in the example stated above a member's question may go unanswered in dense communities which can ignore non-important members.

There is also an additional level of complexity to this question, in that there are a variety of community types. Communities are going to have different goals and needs which they are going to achieve those in different ways [28, 66]. Q/A communities are looking for quality answers, but communities of practice may be looking to build social connections around a given topic. Small behavioral and community level metrics can examine aspects of success with these factors in mind, however once again they aren't a universal measure across all communities. Directed measures of user reciprocity, satisfaction, and trustworthiness appear to be the best option, as a survey measure of these is less susceptible to the variation of roles and community types. Working in unison simple measures can be indicators of community success where community and user goals are well defined, and when those aren't well defined more precise measures of sociable factors are needed.

There is another perspective to take on online community success and that is included within the lifecycle model by Iriberri and Leroy [28]. They theorize that success factors vary over time depending on where the community is within the lifecycle, for example a community early in development will have different goals and needs that a community already focusing on existing members. Some metrics proposed above work well the life cycle degrees of success. Many activity measures (number of posts, content quality, active contributors, response time) declining over time can indicate that a community is approaching death, while an increase in such metrics can indicate strong maturation of a community. However even with the factor of time included, Iriberri and Leroy still are examining what the community and user's goals are, which leads to the same conclusion that success can be seen as how well a community is satisfying it's and the user base's goals. While this is an important factor to take

into account for future work, this current proposed work will need to first address the questions at hand before moving into possible lifecycle stage success metrics.

This proposed work will continue in using primarily a surveyed measure of success, however in conjunction many of the lessor success factors will be used as they can be indicative to context specific success, i.e. Q/A communities can be measured by how well they are responding to questions in terms of rate of responses and speed. This work will focus on success for the whole life of a community and not look heavily into varying types of communities as well as a very detailed set of roles.

2.5 Roles changing over time

2.5.1 Community dynamics over time.

Iriberri and Leroy[28] formulate a community level lifecycle through the following stages: Inception, Creation, Growth, Maturity, and Death. Each stage is defined by the community of members and operators behaviors and needs for information, support, recreation, or relationships, as shown in figure 4.

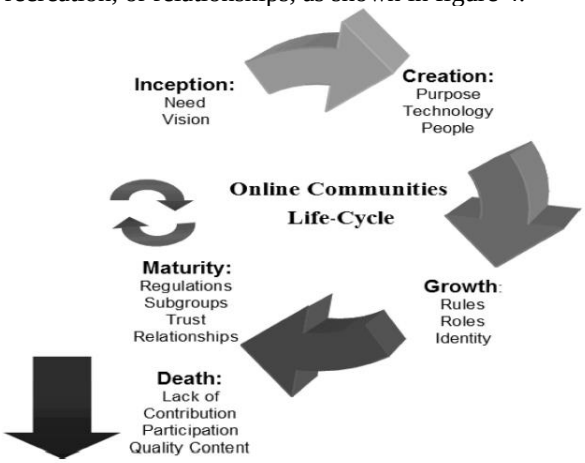

Figure 4. Community Lifecycle model as proposed by Iriberri and Leroy [28].

Within inception, an idea for the online community arises from these needs and depending on the type of need, users can form the intended goal of a community. With this goal, communities also create a baseline of rules to help maintain the focus. The creation stage begins once the online technical components are in place, tools such as listservs, discussion forums, wikis, chats, and community blogs. The beginning group of users can now interact and further increase membership through recruitment. The growth stage is defined as the development of the culture and identity of the community. As more users join, they will start to emulate a role and identity within the community. Users will support others by leading discussions, while simply some other users may only be looking for support and information. The maturity stage comes when a formal organization becomes a need from the high amount of activity. The community then creates more regulations, contribution incentives, sub-communities, and wide range of discussion topics. While some older users leave with their goals being satisfied, new users join bringing new ideas for discussion and fit into a community role. Iriberri and Leroy state that in maturity communities thrive for long periods as communities are strengthened from trust and lasting relationships between users. Some communities then

can iterate through the lifecycle again to maintain user interest, however they can approach the death stage when the community loses momentum and member interest leading to poor contributions and transient membership. Iriberri and Leroy define each stage through general users' needs or behaviors, even though they keep the perspective at a community level. User centric models provide an individualistic look into how a user forums an identity within a community.

2.5.2 *User Lifecycle:*

Preece and Shneiderman [67] generated a framework for how users progress through various roles within an online community. Figure 1 demonstrates this progression of user activities and roles. The decreasing size of each role indicates the total number of users that progress forward, i.e. there are fewer leaders than any other role. Each arrow shows that this is not a linear progression through roles, but many different paths. There is a time dimension with these roles but the first steps of a user transitioning into a different role are still uncertain. Preece and Shneiderman describe user transition occurring through repeat visits that can mature into a growing sense of confidence and increased activity. These authors further state that there are two paths of maturation for a user: they become more active within one stage or they move on to begin another stage. Users start looking to satisfy their curiosity or needs by joining a community and begin reading, searching, and returning if they feel they are satisfied. Users become contributors when they conduct an individual act that adds to a larger communal effort. Contributors can start small with simple corrections or ratings, but they can also tag, review, post, or upload to the community. When users find a mutual understanding or shared beliefs, they can then become collaborators. Collaboration in this context involves two or more contributors discussing or working together to create or share content. Collaborators work together, for example in Wikipedia collaborators will share information and correct each other with the goal of producing a wiki of consumable information for other users. If users are further motivated to improve the community that is where they become leaders. Leaders establish community norms and policies. Furthermore, leaders can promote participation and mentor other users. Leaders typically contribute the largest number of comments and are the most active which can be seen from the observation in certain communities 90 percent of the comments come from less than 10 percent of the participants [54].

2.5.3 Deducing Lifecycle Behaviors

However with these community and user centric lifecycle models there is little observed data supporting the evolution of such roles or the effects of role dynamics on community success. These two areas of literature have vital assumptions about the respective areas of study but future studies must be conducted to connect these two important aspects of online community research.

As implied by figure 1, user's behaviors are changing and this user dynamics is a central aspect of communities [28, 67], however some empirical studies differ from these proposed models. Various studies on membership roles have been conducted using a combination of behavioral or social structural network analysis methods. Researchers tend to start by defining a social role, such as Panciera et al., [58] defining a Wikipedian to be a significant contributor within Wikipedia, then they discover the users within a community that fit this role (Panciera et al.[58] finding users with at least 250 edits on Wikipedia), and finally examine how those users

behavior change over time (Panciera et al.[58] making observations of those users edits over time). Miritello et al.[50] examined user dynamics over time focusing on the number of social connections they can maintain and the level of activity with other users, finding that user's' social circle and interactions decreased with time. Danescu et al.[15] observed the linguistic changes of users through modeling the differences in term distributions in comparison to overall community norms, finding users to first adapt to community norms but then reframed from conforming after a certain time. Such findings are contradictory to theorized lifecycle models which believe interactions tend to increase over time. However these behaviors are not fully unpredicted as such behaviors may fit into the larger community lifecycle models predicting sub-community formation. Further research has examined how users expertise in the content of a community increases as their time with the community gets longer [37, 47] supporting the aspects of user lifecycle models which state that users mature within a community gaining the skills to assist other users. Other work has focused on the user-community interaction which takes the perspective of measure how users' behaviors differ from community norms. Rowe [70] used a combination of linguistic and structural features to show that users mimic the community linguistic behavior early in community lifecycles, but then diverge in language use toward the later lifecycle, verifying much of this previous work. It remains to be seen if such user dynamics are the same for all types of roles. Prior work is filled with key points that highlight what aspects of user and community lifecycle models are accurate but they more so shed light on the dynamics and complexity of users and communities that is missing from said models. This shows a need for modification to such models.

A major point that is lacking in these user centric models is the perspective of what is the user's relationship/role with the community. There has been some work focusing on the community level of roles [15, 53, 62, 70, 94] and the structural formation of online communities at the user level [14, 25], but there is a lack of consensus in role dynamics and only a few of these studies examine specific role impact, i.e. roles relationship and effects on other users and community success [53, 62, 94]. There is a need for macro orientation of online community roles which encompasses both user level models and the community ecology [23]. The obvious step is to update and verify these user centric theoretical models. This creates a common ground theory based on observations and can give a stronger direction to understanding role dynamics and their influence in a community. Applications of this include designing tools to enhance community benefits through maintaining and promoting online community success. A stronger knowledge of user behaviors will be influential in helping community designers/engineers as different roles require different levels of support and help community leaders understand participation allowing for meta-level management of communities[23, 67].

Chapter 3

Completed Work

My initial work has already included some results of modeling formal roles, results presented in table 3. Within this data, member and leader roles are labeled as reported from scraping enterprise online communities.

Confusion Matrix	Predicted Member	Predicted Leader	All
True Member	3075	151	3226
True Leader	0	2853	2853
All	3075	3004	6079

	Precision	Recall	F-Score	Support
Member	1.00	0.95	0.98	3226
Leader	0.95	1.00	0.97	2853
Avg / Total	0.98	0.98	0.98	6079

Table 3. Preliminary Results on Classification of Members and Leader

These are the results of running a binary classification model on predicting members and leaders. A Gradient Boosted Regression Tree[4] was fit on a 50-50 train test split of 6434 members and 477 leaders. Oversampling was used within the leader class in order to account for the class imbalance. This produces a balanced accuracy[8] of 95.1%. The feature set used for this model were all focused on the content of posts only using a bag-of-words feature set from the Linguistic Inquiry and Word Count tool[74] and the frequency of tagged parts of speech[79]. Significant features and the means for members and leaders are listed within table 4.

	Feature Importance	Medians	
		Member	Leader
Possessive endings	0.00507	0.00	0.09
Singular or mass Nouns	0.00544	13.54	13.41
Exclamation points	0.00557	0.00	0.24
Singular Proper Nouns	0.00569	14.06	21.83
>= 6 letter words	0.00628	25.56	30.94
Work	0.00701	3.82	6.09
LIWC Dictionary Words	0.00739	68.69	63.20
Plural Proper Nouns	0.00742	0.00	0.22
Symbols	0.0093	0.00	0.22
Word Count	0.01148	37.50	90.80

Table 4. Feature importance for classification model and feature medians of each role

While this work clearly shows that formal roles are distinguishable, there is future work in examining beyond these simple linguistic features, such as the frequency of contributions made by roles and the types of relationships being crafted. Table 5 shows the differences in social tool use by roles, shown by the average number of posts made within a specific tool. While an individual leader does show more posting overall on average than an individual member, this does have an interesting difference when examining accumulated posting over the entire population of members or leaders. This will be highlighted in the section on community lifecycle observations.

	Members	Leaders
Blog Posts	0.303	3.891
Blog Comments	0.601	1.399
Forum Posts	0.410	1.922
Forum Comments	1.570	4.593
Bookmarks	0.237	3.671
Files	0.230	4.081
Wikis	0.231	8.047
Total	3.699	27.813

Table 5. Differences in posting by roles across social tools

As expected from previous work on shared leadership [17, 94], 151 members were predicted as leaders by this model (shown in Table 3). While this is the only error from the model, this isn't theoretically an error. This group of members could be an existing population of non-formal leaders and therefore needs to be further examined to see how they are possibly exhibiting similar leadership traits that aren't found within the remaining 3075 members. Further examination is going to rely on similar quantitative methods looking for which linguistic differences are occurring, but qualitative methods such as coding for leadership expressions and phrases are going to be needed.

Additional work has been conducted in three different areas: community lifecycle differences between members and leaders, modeling of emotional content within posts, and adaptation of leadership style models from Zhu et al. [94].

3.1 Community Lifecycle

This work is a quantitative study from multiple workplace communities to explore how communities change over time. This study had two main goals, to examine who is creating and curating content and how is that changing over time. This study addresses this questions by looking at who is creating content, whether content is organized by the use of hyperlinks, who is organizing this content, and which social tools are they using to do so.

3.1.1 Methods

To explore curation and creation, I examined posts and links for multiple active online enterprise communities documenting when they were they occurring throughout the community lifecycle. I also explore role differences to examining whether members and leaders behave differently, as well as whether roles shift over time, and how different social media tools are invoked to support content creation and curation.

2,010 communities met the criteria of being an active community totaling in 428,476 posts and 1,246,570 links. All leader- or member-posted text was crawled in these communities (excluding text inside attached Files) to identify posts (content added) and links (URLs) from June 2011 to July 2013. For each post, the following was captured:

- Community ID (Where it was posted)
- Author ID (Unique identifier)
- Date (Time stamp when post was made)
- Tool (Which tool the post was in)
- Author Role (Were they a member or an leader of the community they posted in)
- Date of Community Creation (Used to compute when in the community lifecycle the post was made)

For each link, the following was captured:

- Source Community (Where it was posted)
- Source Tool (Which tool the link was posted in)
- Target Location (Internal or external to source community)
- Target Tool (The tool the link points to)
- Date (Time stamp when link was posted)
- Author ID (Unique identifier)
- Author Role (Were they a member or an leader of the community they posted in)

Creation was assessed as the total amount of content added to a community across different tools, i.e. the combined sum of forum posts and replies, blog posts and replies, wiki edits, and bookmarks. Following [43]**curation** was defined as the act of referencing already-created content and potentially annotating it for other community members.

From prior work, we know that links are used to reference external information sources that are relevant to community discussions and to organize content within the community [43, 75, 76]. The idea behind curation is related to the act of linking, where content from another source is being excerpted [55]. Previous work has examined curation communities that primarily serve as a repository for links to other social networks' content [56]. While more work has been focused on utilizing how links are being used across social media in order to improve content categorization [34] and identifying social curation behaviors [29].

Richer definitions of curation are difficult to accurately operationalize so linking is relied upon here. Prior work notes that community leaders are reluctant to copy and paste external content into their community, as this prevents it from being updated by its original leaders [43].

A small exploration of relationships to linking, roles, and community outcomes was conducted. A subset of communities were given a survey measuring member satisfaction, when examining those with linking present, controlling for other community metrics (number of members, leaders, contributors and contributor inequality i.e. Gini [22]) and removing outliers, 34 Communities within this survey had member linking present and 28 had leader linking present. Links from leaders were found to be positively related to member satisfaction (R^2. = 0.14, df = 26, $p < 0.05$); while links from members were not found to be related ($R^2 < 0.01$, df = 32, p = 0.61), further finding role differences in behaviors. While Wiki edits ($R^2 = 0.022$, df = 26, p = 0.44) and Bookmarks ($R^2 = 0.000$, df = 26, p = 0.88), both common curation style tools, showed no relationship to member satisfaction. These results suggest that leader linking behaviors possibly influence member perceived satisfaction from a community. This motivates further exploration of linking behaviors within the larger dataset of 2,010 communities.

I examined different types of social media tools including Blogs, Bookmarks, Forums (divided into initial posts and replies), and Wikis.

Bookmarks: Bookmarks are a critical curation tool [43, 49]; each bookmark posted within a community contains a link. Bookmarks were analyzed solely as a curation tool, as their aim is to signal relations between different content within a community.

Wikis: Wikis allow communities to post new content as well as to organize and summarize existing information [63, 69]. Unlike other social media tools, wikis can be edited after creation. Edits are considered as an act of creation because they add to community content.

Blogs: Participants post blogs to discuss their personal experience and knowledge on a given subject and prior work indicates that they can be used for both creation and curation [30]. In this work, blog posts are only considered to be creation, while a link within a blog is considered to be curation.

Forums: Unlike other tools, and based on prior research we distinguished between two types of forum post. *Initial posts* are analyzed separately from *replies*, as we wanted to capture critical differences between these post types.

In order to normalize for differences in date, a "Relative Date" variable was created for each post transforming the true date to be relative to the creation time stamp for each community, for example, month 1 indicates the behaviors of all communities at their age of 1 month. This removes the effects of outside events influencing aggregated behaviors across multiple communities, but also limits the amount of time we could examine lifecycle effects to 36 months due to community creation time disparities. I then took the simple approach of dividing time series data at different percentiles. For the majority of analyses, each 36-month timeline is split into two 18-month halves. This allows for a simple contrast between early emerging and late community behaviors. I experimented with 9-month and smaller splits as well, but these yielded similar statistical results as the 18-month split. Time series differences were analyzed using the Kolgormorov-Smirnov test. Most analyses were consistent across different split sizes, but I report where there were discrepancies between different split sizes. Some of the behaviors analyzed were relatively infrequent, occurring just a few times per month. However, it is important to note that all the communities that were analyzed were still active after 36 months, suggesting that low contribution activity does not necessarily mean community death.

3.1.2 Results

I compared the aggregated creation behaviors of members and leaders (see Fig. 5). Consistent with many studies documenting the importance of user-led content creation [30, 43, 52, 63, 69], members have higher overall creation rates ($D = 0.7838$, $p < 0.001$). This was true throughout the community time series. Members created more content than leaders in both the first half ($D = 0.7778$, $p < 0.001$) and the second half of the timespan ($D = 1$, $p < 0.001$).

Members also showed an *increase* in the number of posts/month over the lifetime of the sample ($D = 0.7281$, $p < 0.001$). This is consistent with many lifecycle models arguing that members take increasing responsibility for creating new content as the community matures. Leaders on the other hand, do not significantly increase their posting behavior ($D = 0.3655$, $p = 0.107$). From early on, members are the driving force in total content produced, and increase their production over time.

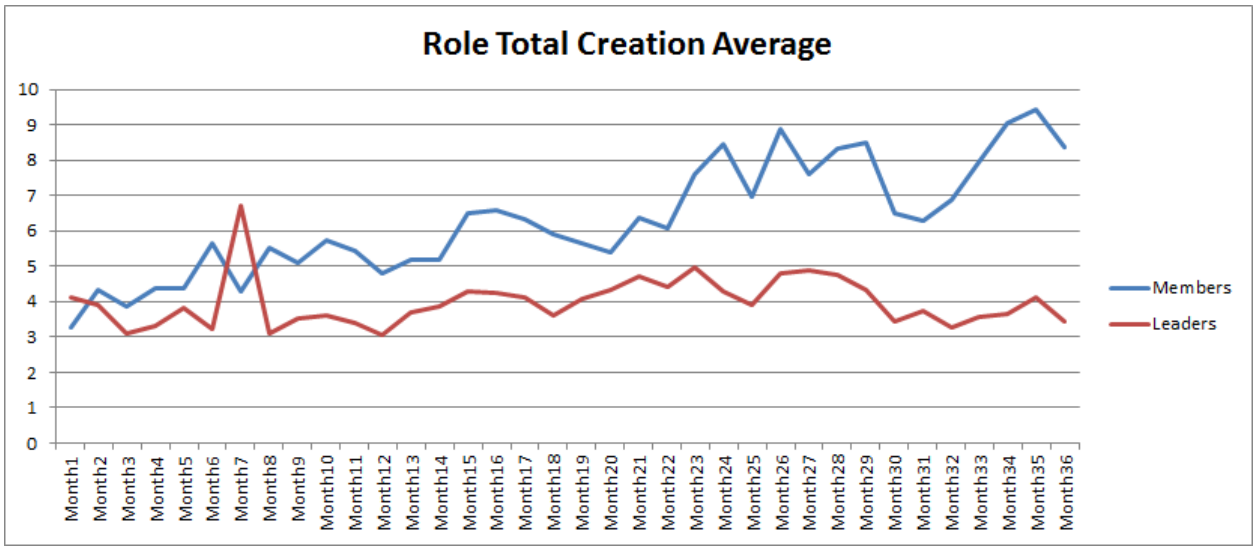

Figure 5 Creation rate by roles: members dominate creation

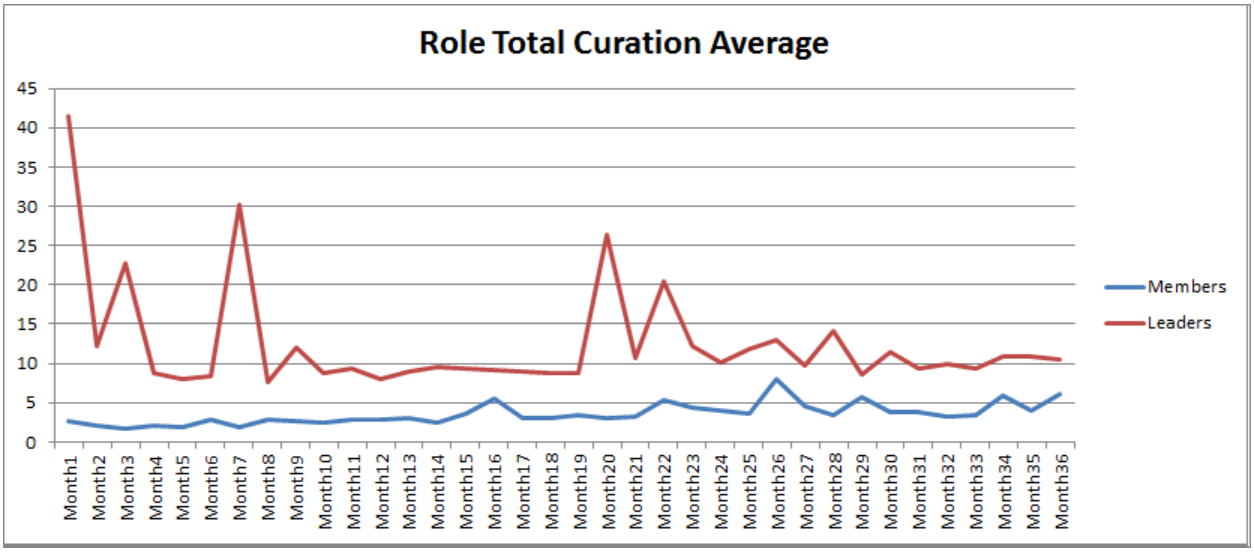

Figure 6. Curation rates by role: owners dominate curation, which decreases over time.

I next analyzed curation behaviors of leaders and members. In contrast to creation where members dominate, Fig. 6 shows that leaders are mainly responsible for curation ($D = 0.973$, $p < 0.001$). This difference was consistent throughout the first 36 months for both first and second half comparisons for leaders vs. members (Month 1-18: $D = 1$, $p < 0.001$, Month 19-36: $D = 1$, $p = 0.001$). However, there were differences within each role over time. Leaders' curation begins with an initial flurry of activity that drops over time, with the two halves of the timespan being significantly different ($D = 0.5117$, $p < 0.05$). Members, in contrast, show a slight rise in curation behavior ($D = 0.8363$, $p < 0.001$), but their levels of curation remain below that of leaders throughout. This greater rate of leader curation is striking given that there are an average of 850 members and just 9 leaders in each community.

These curation results are surprising in the context of lifecycle models [28, 35, 88]. Leaders dominated curation throughout, and contrary to those models, members did not take over curation as the community matured.

Next each type of tool was analyzed while here only reporting the most interesting findings in differences in how roles use Wikis and Forums. Overall, equal numbers of wikis were created in the first and second halves of the time series ($D = 0.3129$, $p = 0.2284$) despite the apparent spike in leader creation in month 7, which may have been the result of an leader offsite, data shown in Fig. 7. Leaders created significantly more wikis than members and this difference was enormous ($D = 0.8649$, $p < 0.001$). Leaders created more wikis than members in both the first half ($D = 0.9444$, $p < 0.001$) and the second half ($D = 0.7895$, $p < 0.001$) of the timespan. Wiki creation rates did not change over the two halves in the timespan for either members ($D = 0.1784$, $p = 0.8696$) or leaders ($D = 0.3041$, $p = 0.2733$). These differences contradict lifecycle models in two ways; wiki creation rates decreased over time and members failed to take increased responsibility.

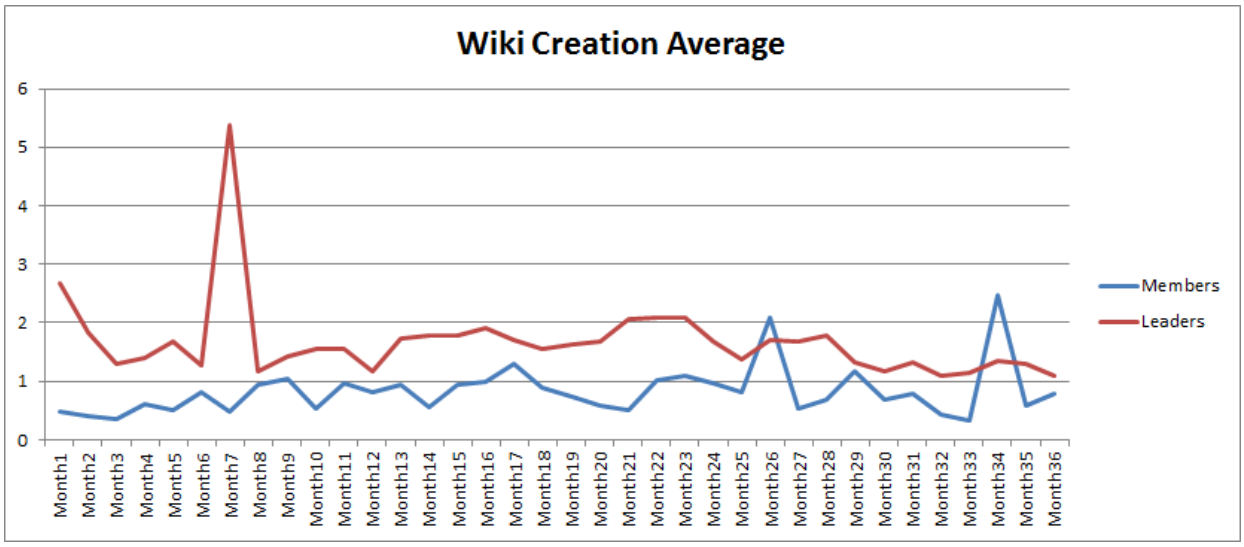

Figure 7. Wikis creation: owners dominate

Leaders also showed massively greater curation than members using wikis ($D = 0.973$, $p < 0.001$), shown in Fig. 8. Surprisingly, there was no change in leader curation rates in wikis between the two halves of the timespan ($D = 0.3509$, $p = 0.1588$). When examining this in a 4-way split, the behavior was similar to that of overall curation. The 3rd quarter (Months 19-27) was greater than the 2nd ($D = 0.8$, $p < 0.01$) and 4th quarter ($D = 0.5889$, $p < 0.05$) again suggesting a 'spring cleaning' activity with wikis. Members in contrast show a slight increase in wiki curation over time ($D = 0.4181$, $p < 0.05$). However, leaders consistently show significantly greater wiki curation rates than members (1st half: $D = 1$, $p < 0.001$, 2nd half: $D = 0.9474$, $p < 0.001$). Again this contradicts lifecycle models in that overall curation rates did not increase and members never exceeded leaders, although consistent with those models, members took on more wiki curation over time.

Initial forum posts (coded in figures as Forum Initial) have some of the lowest tool usage. Members created more initial forum posts ($D = 0.8919$, $p < 0.001$). This was true for both halves of the timespan: first ($D = 0.8333$, $p < 0.001$), second ($D = 0.9474$, $p < 0.001$). Lifecycle models predict that members should initiate more topics over

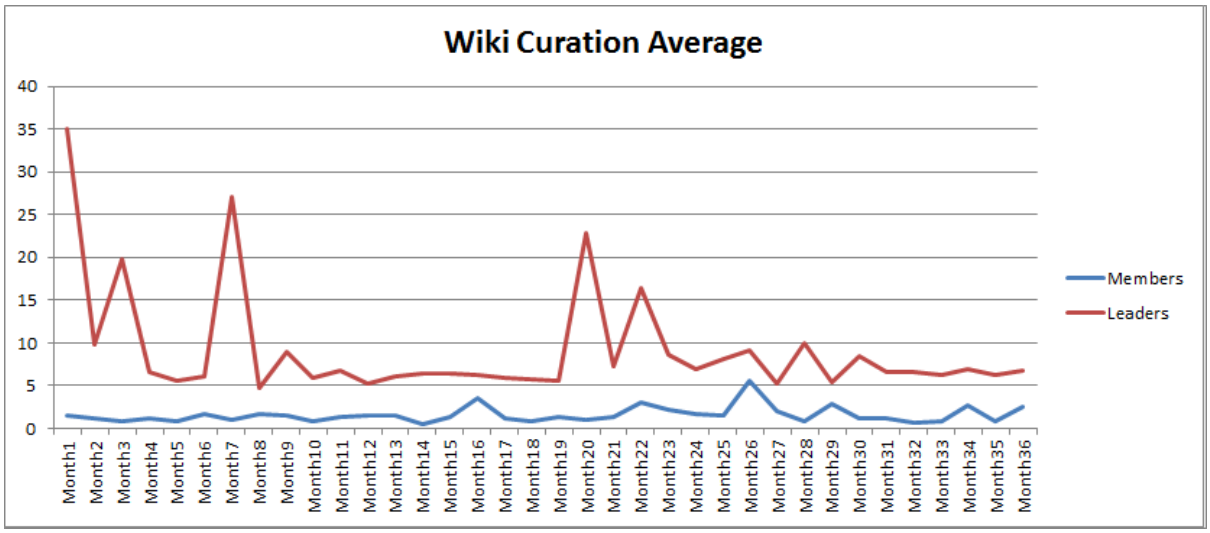

Figure 8. Wiki curation: owners dominate.

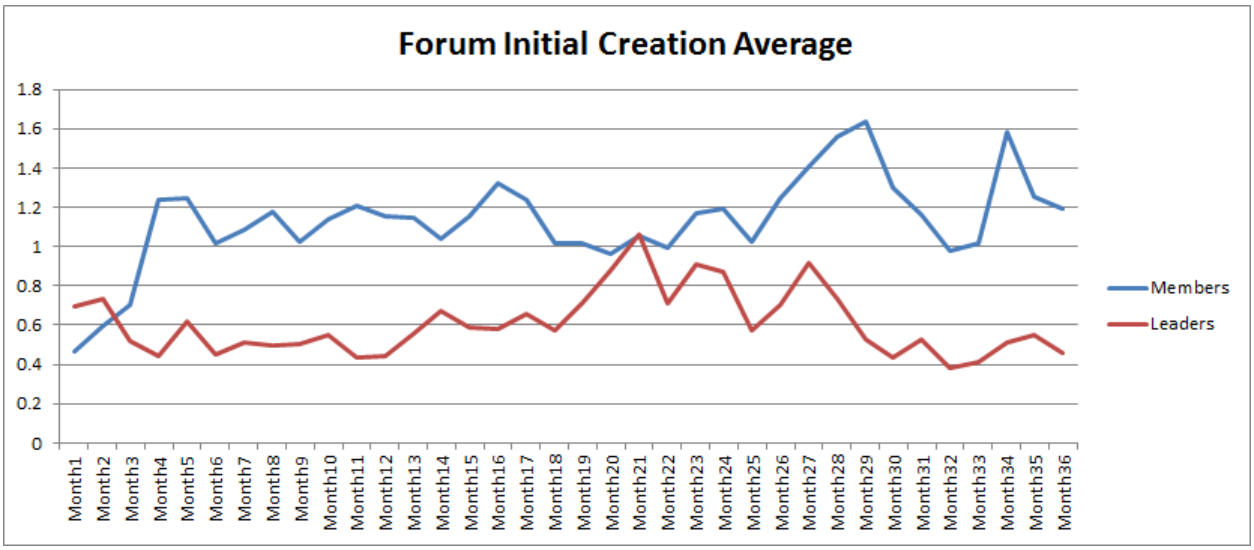

Figure 9. Members dominate initial forum creation.

time, however our data contradict this; members showed no increases over time ($D = 0.3129$, $p = 0.2284$).

For leaders there was an increase ($D = 0.4181$, $p < 0.05$). However, overall there was no increase in initial forum posts ($D = 0.3041$, $p = 0.2733$). Again these results partially contradict the predictions of lifecycle models. Although members were mainly responsible for topic initiation, we saw no evidence that as a community matures, members increase the rate at which they initiate topics.

In contrast for initial forum curation, leaders dominated ($D = 0.5405$, $p < 0.001$). Overall there is a slight increase in curation over time ($D = 0.4474$, $p < 0.05$). However there was a convergence of behavior toward the last few months. This convergence seems to arise from a decrease in leaders' curation behavior. Consistent with lifecycle models, members showed an increase in initial forum curation over time ($D = 0.7251$, $p < 0.001$), but this was not the case for leaders ($D = 0.3538$, $p = 0.1475$).

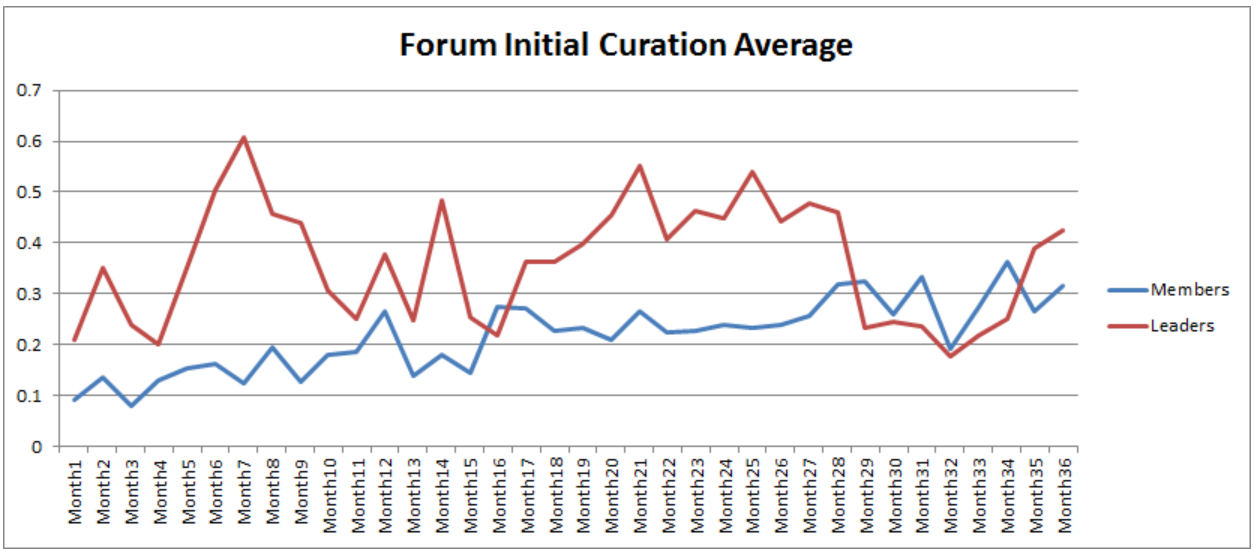

Figure 10. Members dominate initial forum creation.

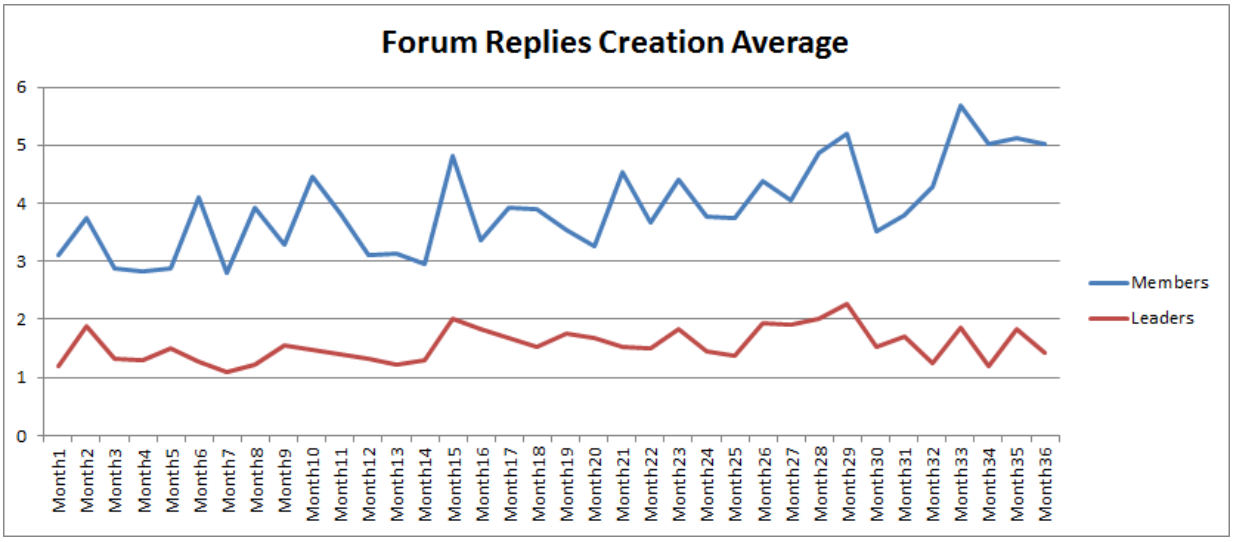

Figure 11. Members dominate forum reply creation.

Forum replies had more creation activity than any other tool, with activity increasing over time ($D = 0.4444$, $p < 0.05$). Members dominated creation of forum replies ($D = 1$, $p < 0.001$). Members also showed a slight increase over time ($D = 0.5029$, $p < 0.05$) whereas leaders were more consistent ($D = 0.3977$, $p = 0.08693$). Members dominated creation in the first ($D = 1$, $p < 0.001$) and second halves ($D = 1$, $p < 0.001$) of the timespan. Forum replies are the only tool where members showed more active curation than leaders ($D = 0.6757$, $p < 0.001$). Both members ($D = 0.731$, $p < 0.001$) and leaders ($D = 0.5146$, $p < 0.01$) also showed increased linking over time, but members increase at a higher rate than leaders. Members dominated initially ($D = 0.7778$, $p < 0.001$) and maintain that dominant status in the second half ($D = 0.9474$, $p < 0.001$). Forum replies overall increase for curation, just as with creation ($D = 0.7281$, $p < 0.001$). Results for forum creation and curation are generally consistent with lifecycle models.

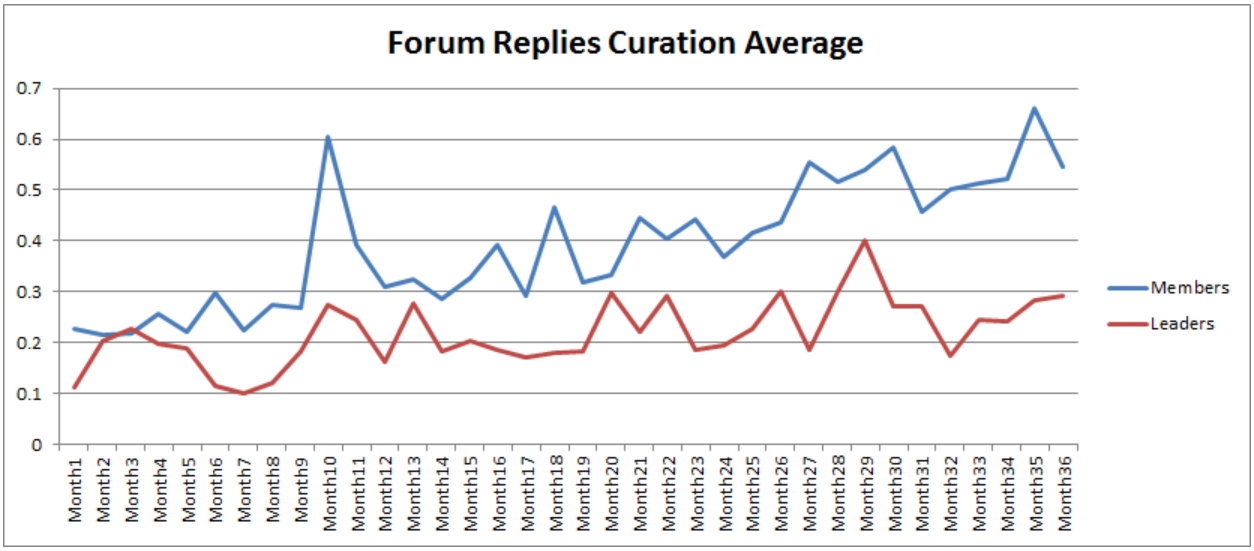

Figure 12. Members dominate forum reply curation

3.1.3 Conclusions

It was found that tool use patterns changed substantially over time. Most prior work has focused on forums and content creation. I can confirm work showing that forum replies increase over time, and that members are responsible for making more forum replies than leaders as a community matures.

As expected, members show increased linking, indicating curation, over time for forum replies, wikis, blogs, and forum initial postings. However in all cases except forum replies, members' activity levels are below that of leaders, suggesting that members never assume full responsibility for curation using these tools.

There are various limitations to these exploratory analyses. First, the curation measures may be conservative in focusing solely on links, as there are other ways that curation can take place. While this linking approach has been applied in prior research [29, 34, 43, 56, 75, 76], focusing on linking alone may miss other, more implicit methods of curation such as summarizing, quoting, or cross-posting. Future work might explore content-based approaches that determine when quoting or summarizations have taken place[68].

3.2 Linguistic Models of Online Communities

This work examines the role of emotional versus factual communication in contributing to perceived success in enterprise communities. This work is more methodological in application to how to study multiple contexts of online communities as well as examine how models can generalize across datasets. I adapted known models from a corpus of online debates[80] to develop an algorithm to detect the relative prevalence of emotional versus factual content in posts. I then applied to algorithm to an enterprise setting. It is predicted:

- The overall informational goals of enterprise communities should mean that an emphasis on factual communication leads to improved member satisfaction,
- Role differences in factual communication will vary in that leadership roles will be more factual.

This algorithmic approach allows the isolation and quantification of emotional versus factual language. The contributions of this work are to: (1) extend the understanding of what contributes to online community success by analyzing emotionality, (2) create a language style model that generalizes across multiple domains of social media, and (3) demonstrate how the effects of emotionality depend on community roles.

3.2.1 Methods and Modeling

I adapted previous work [1, 2, 83] in developing an algorithm that allowed me to distinguish emotional vs factual posts. This involved the following steps:

1. Gather a data set that contains distinct annotated examples of factual and emotional language,
2. Find a set of explanatory features relating to factual and emotional language,
3. Construct a model using said features,
4. Validate this model's output within the domain of interest (enterprise communities).

To accomplish the first step, I used the Mechanical Turk annotated 10,000 post-response pairs from the IAC corpus of online forum debates [80] about important societal issues such as abortion, religion, immigration, gay marriage and so on. The societal significance of these issues leads to engaged debate in which both factual and emotional language are overt and prevalent. The corpus annotates Factual vs. Emotional language for each post response on a scale ranging from -5 (Emotional) to +5 (Factual). Using this corpus allowed me to develop a model derived from multiple different types and valences of emotional and factual interaction. I modeled the extent to which a response to a post was emotional versus factual, which will be referred to as emotionality. To match this definition of emotionality, the valance of the scale for an emotional rating was inverted to then indicate an emotional post as a positive value and a factual post as a negative value.

The next step was to identify a set of explanatory linguistic features in the forum responses that would predict the Turkers' emotionality judgments. I explored both Lexical and Syntactic Features. Previous modeling work [1, 2, 83] derived lexical features from three sources: LIWC (Linguistic Inquiry Word Count)[74], EmoLex[51] and Subjectivity Lexicons[90]. I use the same lexical sources. Lexicon based approaches have limitations. They use a simple "bag of words" which assumes that social and psychological meaning can be derived from individual words alone. This ignores syntax, punctuation, conversational structure, and other relational features of text. I therefore used a part of speech (POS) tagger to count the relative frequencies of nouns, verbs, adjectives and adverbs, use of questions as well as tense and aspect information [79].

For modeling I used Scikit-Learn [61] a machine learning toolkit to build a regression model in order to match the continuous nature of the Turkers' responses. The dataset was split into a 85-15 training-test set. Within the training set, 5-fold cross validation was used to develop the best model, and then tested on the held-out test set. Evaluation of the model's performance was based on the Adjusted R^2 and Root Mean Squared Error (RMSE) on the test set. I used Adjusted R^2 to eliminate spurious variance increases arising with Unadjusted R^2.

3.2.2 Results

The best emotionality model was a linear regression model which had an Adjusted R^2 of 0.1968 and a RMSE of 1.38 for predicting the level of emotionality for forum responses. This model is highly significant ($p < 2.2e\text{-}16$). The level of RMSE shows that the model is varying around 13% in its predictions (given there were 11 possible values for the Turkers' to choose). In comparison, human annotators had a standard deviation of 2.08 for all posts, thus the model is varying in a way that is comparable with the overall judgements of a group of annotators.

To determine whether the emotionality model developed for online debate forums generalized to enterprise communities I first tested the model's ability to predict emotional judgements within Communities. I created a direct test set of annotated posts from the enterprise communities. Using the same procedure as for the IAC corpus annotations, I solicited judgments for 1000 enterprise posts selected at random from the communities, 7 posts had to be removed for not receiving enough annotations. Then I tested to see whether the model's predictions for each post agreed with the judges' emotionality ratings of that post. The model and judges' ratings were highly correlated, r =

0.54 (df = 991, $p < 0.001$) and had high agreement Kendall's Tau = 0.37 ($p < 0.001$). This shows that the emotionality model derived from debates generalizes to the enterprise community data.

	Model 1			Model 2		
	Control Model			Control + Emotion		
	Adj R^2	P		Adj R^2	P	
	0.09187	0.01381		0.1134	0.00329	
	Std Coef.	SE	P	Std Coef.	SE	P
Intercept	4.01	3.09E-01	***	-139	5.80E+01	.
Emotionality				-0.25084	1.87E+01	*
Type						
# of Leaders	-0.20334	4.43E-03	*	-0.18253	4.36E-03	.
# of Contributors	-0.19375	3.49E-04	.			
Gini	-0.17282	4.62E-01	.			
Word Count						
# Posts						
# Comments	0.28944	1.02E-04	*	0.23604	9.05E-05	*
# Views						

Table 6. Model 1 (Control) using the traditional measures for predicting member satisfaction. Model 2 (Control +Emotion) adds the Emotionality feature. ('*' indicates significance p<0.05, '.' p<0.10).

To evaluate the effects of emotionality within enterprise community success, I first created a Control Model containing the following (language independent) variables that have been proposed elsewhere as measures of community success [66]. The first model used these control factors to predict perceived user satisfaction. I next added emotionality to the Control Model to evaluate the prediction that greater emotional communication would decrease overall member satisfaction.

One limitation of the regression approach is that variables may be highly correlated or multi-collinear. I first tested for multi-collinearity using variance inflation factor (VIF). Following standard practice [27], variables with the highest VIF were removed until all variables were under a VIF threshold of 5. I then derived the Control model (Table 6, Model 1) using both-direction step-wise regression using AIC as a criterion, with the VIF filter applied. AIC is a common goodness-of-fit measure for linear regressions for model selection in step-wise regressions [27]. Using a both-direction step procedure is less biased than a one-way step. Stepwise selection led to the removal of Type, Members, Contributors, Gini, Word Count, and Views variables for the Control model.

Table 6 shows that the Control model (Model 1) has reasonable explanatory power (Adjusted R^2=0.091, AIC = 130.96) and is significant (p=0.013). # of Comments and Leaders are significant predictive factors of satisfaction. The Control + Emotional Model (Model 2 in Table 6) adds emotionality to the Control model to test whether emotional interaction increases satisfaction. Using the same feature selection procedure, I excluded highly collinear variables and again used both-direction stepwise regression. Adding the mean post emotionality of a community increases explanatory power (Adj R^2= 0.1134), decreases AIC in comparison to the Control Model (ΔAIC = -3.17), and the model is a significant predictor of member satisfaction (p=0.0032). The negative coefficient of the emotionality variable indicates that less emotional, i.e. more factual, content predicts satisfaction, confirming the prediction. This result contrasts with prior work on support forums where high amounts of emotional content benefit online interactions [83].

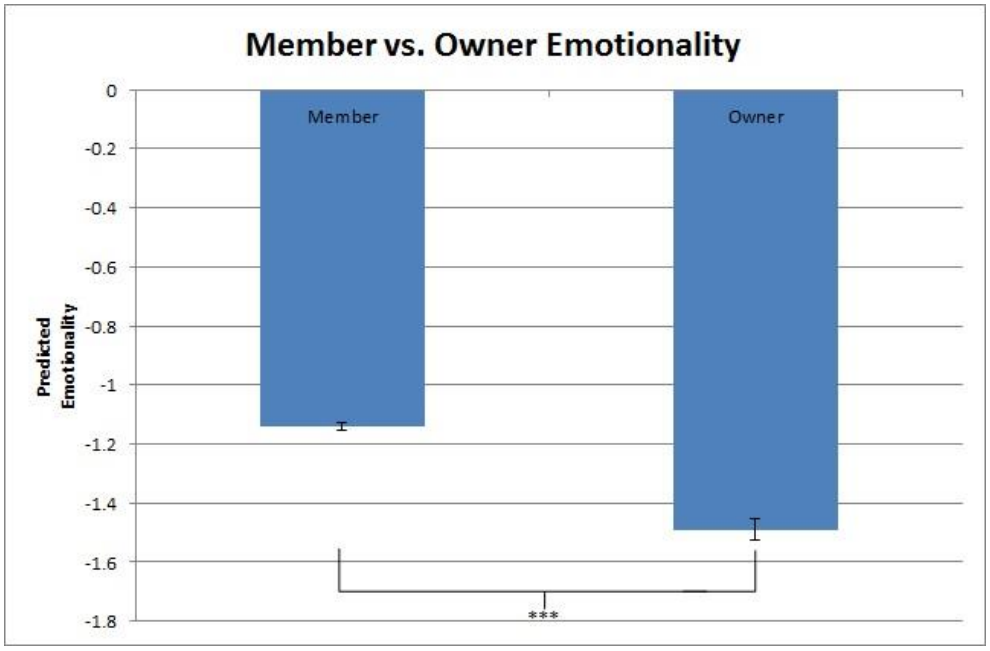

Figure 13. differences in emotionality predicted from members and owners. Owners are found to exert more factual language.

Further examination of emotionality across roles shows that leaders are expressing factual language more often than members, this was found to be significant in a two-tailed T-test ($t = 9.285$, $df = 633$, $p < 0.001$) , results shown in figure 13.

3.2.3 Conclusions

These results show a small but clear relationship between the use of emotional versus factual language in enterprise communities and member satisfaction, as well as show role differences in levels of emotional prediction. As predicted, it was found that factual language enhanced perceived satisfaction. These results also find more distinguishing differences in roles, particularly in a higher level (emotional) examination of linguistic differences and implying that leaders are more factual posters than members thus being more satisfactory to members. Future work can include an analysis in the interaction of role behaviors across type and social tools, as those were other factors found to have differences in emotional expression.

3.3 Leadership Style

In light of the results from Zhu et al. [94], where various leadership behaviors were observed in non-formal leadership roles, an exploration was conducted in applying the methods used from Zhu et al., towards enterprise online communities. Models were reproduced using the 500 posts that were provided by Zhu et al. Within Wikipedia it is found that leadership styles, as shown in table 7, can be identified using the linguistic features reported are as follows:

- Transactional leadership: the frequency of the word “barnstar” (a barnstar is a type of virtual reward in Wikipedia) and the frequency of the phrases “thanks for” and “thank you for”
- Aversive leadership: the frequency of a set of strong negative words, including “block,” “revert” and “remove”;
- Directive leadership: the frequency of “you” followed by modal words such as “should,” ”could,” “might,” and the frequency of the word “please” followed by a verb;
- Person-based leadership: the frequency of greeting words and smiley emoticons.

The modeling results are provided in table 6, reporting a binary classifier as to whether a post is expressing a style of leadership or not; a support vector machine classification model was created for each leadership style, using 5-fold cross validation.

Transactional Leadership	**Precision**	**Recall**	**F-Score**	**Support**
False	0.89	0.99	0.94	355
True	0.96	0.70	0.81	145
Avg / Total	0.91	0.91	0.90	500
ROC	0.88			
Aversive Leadership	**Precision**	**Recall**	**F-Score**	**Support**
False	0.86	1.00	0.93	396
True	0.98	0.39	0.56	104
Avg / Total	0.89	0.87	0.85	500
ROC	0.69			
Directive Leadership	**Precision**	**Recall**	**F-Score**	**Support**
False	0.85	0.92	0.88	291
True	0.88	0.77	0.82	209
Avg / Total	0.86	0.86	0.86	500
ROC	0.80			
Person-based Leadership	**Precision**	**Recall**	**F-Score**	**Support**
False	0.91	0.96	0.93	337
True	0.91	0.79	0.85	163
Avg / Total	0.91	0.91	0.91	500
ROC	0.91			

Table 7. Results from replicating the models reported in Zhu et al. [94]

These models were then applied to the enterprise online community dataset, where features were extracted from all text produced by individual members and leaders. The amount of individuals that were found to have leadership style was low, with only 718 showing transactional leadership, 49 showing aversive leadership, 1219 showing directive leadership, and 849 showing person-based leadership out of a total of 35238 users examined. These proportions aren't too far off from the linguistic classification model presented earlier classifying formal roles, from that model 9.08% of all individuals were classified as leaders, while in this work 8.04% are classified as showing a style of leadership.

When examining for differences in leadership style across roles, minimal differences were found, results shown in table 8. Differences in leadership were examined by the positive class probability of a leadership style,

using the likelihood of a leadership style given by the SVMs. For example, an individual with a model output of 0.6 in Transactional Leadership can be interpreted as having a 60% chance they are expressing transactional leadership.

LS-Features	Members vs. Owners		
Variable	98.75% CI (M-O)	98.75% Cohen's D	Conclusion
Transactional Leadership	[0.016; 0.020]	0.20 (negligible)	M>O
Aversive Leadership	[0.003; 0.004]	0.13 (negligible)	M>O
Directive Leadership	[0.014; 0.017]	0.17 (negligible)	M>O
Person-Based Leadership	[0.027; 0.030]	0.32 (small)	M>O

Table 8. T-Tests examining differences in leadership styles between members and leaders.

Such small results can be due to this simple analysis of looking for style difference across a vast number of low likelihood results (as seen by the Confidence Intervals reported in table 8). Also this can be due to the model never being tested for its ability to predict such leadership styles within enterprise online communities. Testing the model's generalizability within this separate context is the clear next step as these results (although weak) need to be verified. Such differences between the two contexts of Wikipedia communities and enterprise communities could be within the style of tool that is being used. Within Wikipedia, editors communicate through a chat log style tool, which can be rather different than the style of posts that exist within enterprise communities which are those are forums, blogs, and wikis.

Chapter 4

Proposed Work

This proposed work aims to extend our understanding of role, role dynamics, role ecology and their influence on online communities. My main goal is to *study the behaviors, interactions, influences, and changes that individuals of certain social roles go through in an online community* with an end goal of *modeling user role dynamics over time in order to evaluate proposed models of community and user life cycle stages.* This will allow us to connect community and user centric models.

To accomplish this goal, my research questions aim to evaluate and extend previous research. In particular I will examine if those prior observations and theories apply to enterprise and open source online community contexts. To reiterate the goals of this research, I will address the following questions:

1. ***Can we reliably distinguish member and leader roles?***
2. ***What is the ecology of roles?***
3. ***What is the effect of different behaviors on community success?***
4. ***How do roles change over time?***

4.1 Choice of Dataset

I am interested in studying online enterprise communities for theoretical and practical reasons. First, enterprise communities have been under-researched [43, 45, 46, 52, 95], so it is important to determine whether existing theory and results gathered for internet communities apply to enterprise contexts. Second, the enterprise dataset I have has important, unique properties. This dataset contains 2,010 enterprise communities who are considered to have a presence of active management as well as be updated within the last month of the time of collection (May, 2014). This dataset also includes a label of community types for the majority of communities. Each is labeled either a Community of Practice (821), Team (590), Technical focus (62), Idea Labs focusing on brainstorming around a set of questions or issues (23), or Recreational (8), though there are 506 communities with no label. Each community has access to various social tools such as Wikis, Blogs, Forums, and Bookmarks where each post collected has been labeled for which tool they were posted in. Furthermore, anonymized IDs are provided allowing for examining who is posting what content and what community they are posting in.

More importantly from a research perspective, the dataset has critical characteristics that are relevant to my research goals. Participants using communities declare their role in the community explicitly by stating whether they are a leader or a member providing a formal label on users and acting as a ground truth. This is an attractive property of the data. Typically, role researchers either need to conduct survey methods to identify roles [19] or examine properties of the data to infer roles sometimes algorithmically [13, 53]. In both cases, there is a possibility of role miss-assignment. Having data with labeled roles is a clear benefit.

Furthermore, I already have a good amount of experience working with this dataset and have developed methods to reduce its noise and complexity. For example, some posts contain different languages and needed to be filtered out since many linguistic methods only focus on English. Another example of noise is the aspect that post content typically include the post title and had to be removed in order to not be double counted by bag-of-word models. While there is noise that needs to be accounted for, this context for an online community is more ideal to linguistic methods as an enterprise context tends to have a traditional formal text speak, unlike other communities on the open internet which may have developed more net speak (i.e. Specific acronyms or L33t [6]). Critically for my third research question examining community success, this dataset also comes with a unique measure of community success, from a survey measure of member satisfaction from about 120 communities. While this sample is small compared to the entire dataset (roughly 5%), it does give an opportunity to discover important relationships relating to community success. Finally while the dataset is anonymized, it is also hashed allowing us to determine the same user's roles across multiple communities, for example allowing us to determine whether a specific user is a leader in one community but a member in another.

I will also compare my results to other communities outside the enterprise context. As has already been argued [28], community contexts directly influence individual behavior with important consequences for community success. I will therefore collect a second dataset outside of the enterprise context, focusing on Open Source Communities such as those present on GitHub. GitHub has been examined by community researchers [20, 39, 40]. There are important parallels between Open Source communities and team based enterprise communities as both focus on accomplishing task specific goals. It should also be possible to derive community success metrics for GitHub. GitHub communities typically work towards producing a specific product that is intended for public use [20]. While this additional work of gathering a new dataset will be challenging, it is important to replicate enterprise results on a dataset from open internet based communities.

4.1 Methods and Data

4.1.1 Distinguishing Member and Leader Roles

Keeping consistent with the format of the proposed sub questions within Chapter 1, I will now discuss how each sub question will be answered.

- *Do members and leaders differ in overall posting behaviors? Do leaders post more overall?*

This question can be answered through existing measures within the enterprise dataset. User posting behaviors already have been measured and can simply be compared across each role with standard statistical methods as T-tests and an anonva. While this is helpful in understand what these differences are, they can be included in existing predictive models which classify a user as either a member or a leader. Having the labeled roles within the enterprise community dataset provides an immediate high level distinction between a member and a leader as defined by the community itself. This gives a ground truth for testing methods of distinguishing these social roles as well as potentially informing exploration of other types of roles as defined in [53, 67].

- *Is the content of posts different across members and leaders? Are members more likely to post questions and leaders to reply to them?*

Content of posts have already been examined and explained to some extent within the completed work chapter. Models have been made based on bag-of-word lexicons examining the categories of words being used by members and leaders and found to have a strong predictive ability (accuracy of around 95%). One aspect that can be expanded upon is what the style of post or if the post is a type of question. Work has been done identifying the style of inquiries that can exist within online communication [89], and this work can be adapted to examine if roles are asking questions to coordinate events or ask favors (Anyone want to join me at <event>?), express opinions (Is there a way to accomplish this? Otherwise, what the hell?), promote information (Can't make it to <event> this week? Consider attending the live stream), or request information (Does anyone know where I can read about this?).

- *Do members and leaders vary in the community tools they use? Are leaders more likely to create wikis and members more likely to use forums?*

This question is a natural progression when examining role posts. Given the presence of different tools and the labels each post has as to what tool was used, this is an easy expansion on examining role differences. This can be accomplished by including a tool variable within exploratory statistics (ANOVAs) and also within predictive models which will give the added context of tools and thus can discover how roles are using them differently.

- *Do members and leaders reference third party content differently?*

Examining referenced content can be done by looking for the existence of links within a post. Links are found through scraping through the content of a post and seeing if a URL is present in the text. There are different types of links present within the enterprise communities; they can be self-referential links pointing to the same community it was posted in or a link referencing content outside of the community. Links contain the community id of which community content it is referencing within the enterprise context, to find outside references is to simply see if a community id pattern (that of a universal unique identification code) exists within the URL. This can be related to who posted the link to see if roles are posting different styles of links.

- *Do members and leaders vary in the social structure that they build for themselves? Are leaders more broadly networked across the entire community?*

Social network methods require that a graphical representation of the data can be created. Graphs have two distinctive components: nodes and edges. In order to build a graph of the enterprise data, nodes and edges must be defined. One definition can be nodes being the users within a community and edges will be the interactions between users. Interactions can be measured through many means; one way is through looking at replies within specific tools of each online community. If a user replies to a post from another user, then this is considered an interaction and thus will have an edge formed between the two; likewise for those users who reply to each other's replies. This can also go beyond just posting information and can look at community memberships where edges would be if user share common membership within a community.

With a graphical model in place, such social network analysis metrics can be applied. Some metrics which have already been applied in previous research and those of which will be explored in this work include assortativity

(tendency for a node in a network to be directly connected to other similar nodes) [9, 14], reciprocity (edges formed from previous existing edges) [82], behavior specific edges [40], community density [91], average in and out degree [91], and commonly found patterns within social networks (such as triangles) [91].

As theorized by previous work[67, 94], there can exist multiple other sub-roles beyond just these two labels. One method for examining possible sub roles is an unsupervised machine learning method called hierarchical clustering which has been used by Wang et al. [81]. Hierarchical clustering would allow me to examine whether members and leaders fulfill certain sub-roles within the higher level role distinctions by producing multiple levels of clusters containing common features between groups and subgroups. However, as pointed out previously, sub-role detection is not the main goal of this work. Subsequent parts of the thesis will build on the initial leader and member roles, and it is necessary that the distinguishing of identified roles have a high level of precision in order for modeling of network, success and temporal effects to have any sort of reliability. While interesting, the identification of sub-roles will be explored, but it isn't expected to be a main output of this research.

4.1.1.1 Machine Learning Methods

I will exploit these naturally defined labels using supervised classification machine learning methods. Many classification models are available like logistic regression, support vector machines, and random forests. Previous work has examined supervised methods to examine user behaviors [15, 70, 83]. I will follow standard machine learning practices to examine which are performing best in the experiment on hand. Following common practice, comparisons will be made through multiple standard evaluation metrics [13, 45, 46, 52, 95]. I will be using evaluation metrics such as Accuracy, Precision, Recall, F-Score, and the Receiver Operator Characteristic (ROC). Other evaluation metrics are necessary given modifications to sampling procedures, for instance if over or under sampling is used to remove the issue of skewed class distributions, then a balanced accuracy measure will be used [8]. The goal of this modeling will be to reproduce high precision (evaluation report of completed work reported in chapter 3: 95% accuracy, 0.98 precision, 0.98 recall, and 0.98 F-score) for classifying these two main roles (members and leaders) following the completed work chapter. While these initial thresholds are very high, it is encouraging that these roles were easily distinguished using only linguistic features.

I will also examine feature importance and selection for these models, using ranking methods [96], beta weights for regression models, and entropy measures like gini [73] for tree based models. These methods allow exploration as to what features distinguish the target classes, allowing machine learning models to be used for theory and exploration. I will compare my results with previous models [15, 28, 67, 70, 83] that formally define roles and the features that make them differentiable. This will extend prior models of social roles based on different datasets.

The feature space for examining role differences is rather large. Possible features that can be used based on previous work and completed work are:

- Term Frequency and Independent Document Frequency (TI-IDF) [53]

- Entropy/ Term Frequencies [15, 70],
- Topic models [83],
- Bag of words models [46, 83]
    - N-Grams
    - Linguistic Inquiry and Word Count [74]
    - Emotional Lexicon [51]
    - Subjectivity Lexicon [90]
    - Parts of Speech [79]
- Structural Network Features [5, 28, 50, 67, 87]
    - Assortativity [9, 14]
    - Reciprocity[82]
    - Behavior Specific Edges (Positive Feedback vs Negative Feedback) [40]
    - Community Density [91]
    - In and Out Degree [91]
    - Sub-group Membership (tri-ad groups) [91]
- Survey Report Data [19, 46, 52].

Such a large possible feature space makes it imperative that features and models are theoretically motivated. A systematic approach is necessary not only to control the amount of modeling that will be conducted, but also to understand how each feature influences the models. This work isn't only to see what the best models that can be produced are, but to understand how roles are different and provided theoretical grounding for future work to be done. To do this, hierarchical, incremental, or feature weighting techniques will be used in order to understand the relationships and effects of features.

4.1.1.2 Qualitative Methods

Qualitative methods are used in the social sciences to understand complex human behaviors, and there are many instances of such methods being deployed in online community research [26, 41, 45, 46, 52, 59, 62]. Surveys are a primary method for understanding participants' perspectives on the operation and success of their community [20, 46] and will be deployed for the Open Source Community dataset. For example, previous survey methods have asked users questions based on rating visit frequency and membership length [19, 52], how satisfied users are with the community [46], identifying leaders [19], accessing individual and community goals [45], and identifying cooperation and conflicting events [20]. I will adapt these surveys for collecting and understanding individual perceptions when I generalize the findings to Open Source Communities.

4.2 Relation to Community

Prior work examines subgroups within online communities [7], as well as membership overlaps between communities [52]. This is a start to the type of work I will conduct in understanding roles relation to communities, but I will use similar methods to first examine the population demographics of online communities and then to examine the influences of roles on each other.

- *Are a certain number of leaders or members necessary for a community?*

This is an exploratory question that first needs to examine what the distributions of leaders are within the sample of communities. A metric examining the ratio of leaders to members will be a good metric for this work as it gives a perspective that is normalized to the overall population of users. Examining if there is a level of necessary leadership within a community is a different question. If there exists communities with low or high ratios of leaders to members, this then gives the research the context necessary in order to examine how those communities are functioning differently. This question will be further examined in relation to success metrics and time effects.

- *Are members and leaders interacting within or between their respective roles?*

Measuring interactions between roles can be accomplished through social network analysis. As a graphical representation will have already of been built, this question can then be examined as to who each role is interacting with. Examining interaction networks is related to examining sub-groups and clustering methods can highlight these strong interactions. The graphical representation will need to include each node's role label.

- *Do community types influence the ecology of social roles?*

Since the enterprise dataset has a community type label present, this can be added on top of the previous questions described above and then see if leader to member ratios vary by community type as well as if interaction networks found by cluster methods also are influenced by community types.

- *Are leadership styles more interleaved within different types of communities?*

Leadership styles are the measure of the type of interaction going on within the content of a post. In order to examine if leadership styles are being used toward specific roles, a graphical representation will again be needed. Since leadership style expression is targeted toward a specific member type role[17], this introduces an interesting type of data representation where the label of a node in the graph becomes related to not only the observed characteristics of the node, but also the observed characteristics of the neighboring nodes and the labels of the neighbors to that node. Collective classification is a method that takes advantage of these types of dependencies, where traditional classification in machine learning would ignore them [72]. As role classification can take advantage of collective classification, another possible method of applying this is to classify the edges instead of the nodes, as this question is asking to predict the interactions taking place and not about predicting the role of a user.

Relationships between individuals within communities implies social network behaviors, where much work has been done on defining network attributes that indicate a social practice or behavior [84]. As network metrics will be applied in the previous section on distinguishing role behaviors, here they will be applied in understanding how each role potential affects each other. This expands on understanding what relationships are formed between roles, now understanding the influence in that role on the community.

4.3 Measuring Community Success

The research question to be assessed here is the extent to which role specific behaviors affect community success. As discussed previously there are many measures of community success, most of which focus on low level user behaviors, such as volume of members' posts, number of members, and quality of member relationships (extend of contact between members) [45]. However work has shown that such behaviors don't align well with participants' assessments of community success[45].

Many prior studies of online communities exploit observable traits within the data. However such traits are often specific to the domain of the community, such as tools like voting mechanisms [12]. The main focus of the current work is to study role behaviors. It will not therefore develop complex new ways to evaluate community success. Bearing in mind the complexity of evaluating online community success, survey measures of member satisfaction and simple observable measures such as activity and membership duration appear to be the best option for the goals in this work.

- *Are social role specific behaviors influential to success?*

I will use both behavioral metrics obtained by scraping community content as well as data collecting participants' subjective assessments of community success. My current enterprise dataset includes such a participant based evaluation of success, in which participants were asked how well the community is meeting their needs. This is a reliable metric that has already been validated as a significant predictor of community success [46]. Using these metrics and the previous results of distinguishing roles as well as in the context of role ecologies, I will then examine how such behaviors either predict or cause said metrics of success. Prediction can be accomplished through machine learning methods which work to model the dependent variable (success metrics) based on a set of independent variables (distinguished behaviors of members and leaders). Causality is a very different goal and as such will be hard to accomplish within a purely observational setting, however it is not unexamined. Toulis and Kao [78] introduced causal estimation procedures in order to account for social interference, complex response functions, and network uncertainty. When examining success factors, causal methods are more in favor as they give evidence to actual effects, while many typical prediction models tend to only show correlation relationships which can be influenced by many outside factors potential falsifying the results.

- *How do community ecologies relate to success?*

Ecological measures as discussed prior will also be examined for their relationship with member satisfaction. Simply examining how leader to member ratios related to perceived member satisfaction will address this question.

- *What interactions between roles are needed for a successful community?*

This question can be addressed by examining the quantity of interaction types (question and leadership styles) within a community and see if there is a relationship with those measures and member satisfaction.
Finally, as this work expands to other community contexts, surveyed success metrics like member satisfaction will need to be gathered for the Open Source Community. This does allow for an opportunity to examine other known success metrics such as goal achievement and leadership satisfaction [45].

4.4 Changes over Time

The research question I will be addressing here concerns changes in roles over time. Recall that theoretical models predict that certain members should shift roles to become more leader-like over time [28, 67]. As these theoretical stages of development are hard to operationalize, it is simpler to examine changes in role behaviors simply over time. A beginning method to do this is to examine the differences in early behaviors compared to later behaviors, where the split between the two is just at the mid-way point in the total timeline of contributions for a user.

- *Do users typically change as hypothesized in models, particularly in the early and late stages of said models* [67]*?*

This question can first be answered through such a time split. But it will need to be further explored beyond simply a time split, as users develop at different rates. Work by Rowe [70], adapts to these different rates in user development by bucketing user contribution based on their sequence rather than their distance in time. Rowe did a 4 post bucket, where the first 4 post by a user was considered the first time stamp, then the next 4 post were considered the second. I will adopt this bucketing strategy to examine more specific changes in roles over time. Other aspects of behaviors given by their proximity in time are sequential temporal analysis of behaviors. Some work has been done in cluster sequences in behaviors to find categories of roles [31, 81].

- *As members become more experienced, do they find themselves stagnate within a role or do they continue to gain responsibilities and become more leader-like* [47]*?*

This question will be examined by seeing how specific roles and users in those roles are changing. Specifically examining if there is a difference in the users that do change over time, for example testing if certain users are not under the influence of other users, therefore asking the question are leaders promoting leadership behaviors toward only certain members?

- *As the community ages and leadership and member roles are interchanged between users, is there ever a lacking of a role, in particular a lack of leadership, for a community or are roles constantly filled when there is a need* [28, 67]*?*

This question relates role dynamics to the changing needs of a community. Community and user lifecycles can be mapped together based on timestamps. Within these timestamps, it can be examined if there is a lacking of leadership presence in a community and are given members or leaders rising to the leadership void. Leadership presence can be measured by how much activity is either being classified as a leadership behavior or how many posts are being made by leadership roles.

- *Is there a relationship to collective user development over time that shifts the community norms or do community norms influence user development and influence whom* [15, 28, 42]*?*

This will be a difficult question to address as it is the combination of all previous work. Norms within a community can be measured by linguistic common shared behaviors [15, 70]. I can then measure how roles are either approaching the level of common linguistic behaviors or if the common linguistic behaviors are approaching levels

of key users. This can further take advantage by network methods seeing how much influence is being spread by simple measures of in and out degree of roles.

Comparisons of changing over time have many avenues available. From simple t-test comparison to vector regressions[97], time analysis can be rather complex as many of the variables are expected to be heavily dependent on time, therefore time is a rather encapsulating variable. I will use many methods to compare time differences, starting with examining individual social role changes with a generalized linear model, to then examining overall role differences with possible time series prediction models like ARIMA[85] and vectorized regression models[97]. Machine learning methods such as Bayesian regression have been used to examine time and will be used in this context when exploring overall role changes.

4.5 Timeline

		Fall '16	Winter '17	Spring '17	Summer '17	Fall '17	Winter '18	Spring '18
Distinguish Roles	*Do members and leaders differ in overall posting behaviors?*	X						
	Is the content of posts different across members and leaders?	X						
	Do members and leaders vary in the community tools they use?	X						
	Do members and leaders reference third party content differently?	X						
	Do members and leaders vary in the social structure that they build for themselves?		X					
Relation to Community	*Are a certain number of leaders or members necessary for a community?*		X					
	Are members and leaders interacting within or between their respective roles?		X					
	Do community types influence the ecology of social roles?			X				
	Are leadership styles more interleaved within different types of communities?			X				
Relation to Community Success	*Are social role specific behaviors influential to success?*	X	X					
	How do community ecologies relate to success?				X			
	What interactions between roles are needed for a successful community?				X			
Changes over Time	*Do users typically change as hypothesized in models, particularly in the early and late stages of said models*		X					
	As members become more experienced, do they find themselves stagnate within a role or do they continue to gain responsibilities and become more leader-like						X	X
	As the community ages and leadership and member roles are interchanged between users, is there ever a lacking of a role, in particular a lack of leadership, for a community or are roles constantly filled when there is a need						X	X
	Is there a relationship to collective user development over time that shifts the community norms or do community norms influence user development and influence whom						X	X
Open Source Communities	*Gathering GitHub data*			X				
	Explore GitHub success metrics available				X			
	Test overlaps and discrepancies between role models within enterprize online communities and open source communities					X		
	Examine if role relationships and network orientations are similar between the two online contexts						X	
	Explore temporal differences between the two online contexts							X

References

[1] Alm, C.O. et al. 2005. Emotions from text: machine learning for text-based emotion prediction. *Proc. of HLT/EMNLP*. (2005), 579–586.

[2] Aman, S. and Szpakowicz, S. 2007. Identifying expressions of emotion in text. *TSD*. (2007), 196–205.

[3] Arguello, J. et al. 2006. Talk to me: foundations for successful individual-group interactions in online communities. *Proc. of CHI*. (2006), 959–968.

[4] A working guide to boosted regression trees - Elith - 2008 - Journal of Animal Ecology - Wiley Online Library: *http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2656.2008.01390.x/full*. Accessed: 2016-05-19.

[5] Backstrom, L. et al. 2006. Group formation in large social networks: membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (2006), 44–54.

[6] Blashki, K. and Nichol, S. 2005. Game geek's goss: linguistic creativity in young males within an online university forum (94/$\backslash$/$\backslash$ 3 933k'5 9055oneone). *Australian Journal of Emerging Technologies and Society*. 3, 2 (2005), 71–80.

[7] Bos, N. et al. 2007. From shared databases to communities of practice: A taxonomy of collaboratories. *Journal of Computer-Mediated Communication*. 12, 2 (2007), 652–672.

[8] Brodersen, K.H. et al. 2010. The balanced accuracy and its posterior distribution. *Pattern recognition (ICPR), 2010 20th international conference on* (2010), 3121–3124.

[9] Buccafurri, F. et al. 2015. A new form of assortativity in online social networks. *International Journal of Human-Computer Studies*. 80, (2015), 56–65.

[10] Burke, M. and Kraut, R. 2008. Taking up the mop: identifying future wikipedia administrators. *CHI'08 extended abstracts on Human factors in computing systems* (2008), 3441–3446.

[11] Butler, B. and et al. '02. Community effort in online groups: Who does the work and why? *Leadership at a distance*. ('02).

[12] Cheng, J. et al. 2014. How community feedback shapes user behavior. *arXiv preprint arXiv:1405.1429*. (2014).

[13] Choobdar, S. et al. 2015. Dynamic inference of social roles in information cascades. *Data Mining and Knowledge Discovery*. 29, 5 (2015), 1152–1177.

[14] Chung, K.S.K. et al. 2012. Community evolution and engagement through assortative mixing in online social networks. *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)* (2012), 724–725.

[15] Danescu-Niculescu-Mizil, C. et al. 2013. No country for old members: User lifecycle and linguistic change in online communities. *Proceedings of the 22nd international conference on World Wide Web* (2013), 307–318.

[16] Denison, T. et al. 2002. Community networks: Identities, taxonomies and evaluations. *Electronic Networking 2002-Building Community (CCNR 03 July 2002 to 05 July 2002)* (2002), 1–17.

[17] D'Innocenzo, L. et al. 2014. A meta-analysis of different forms of shared leadership–team performance relations. *Journal of Management*. (2014), 0149206314525205.

[18] Ehls, D. and Herstatt, C. 2013. Open Source Participation Behavior-A Review and Introduction of a Participation Lifecycle Model. *35th DRUID Celebration Conference* (2013).

[19] Faraj, S. et al. 2011. Knowledge collaboration in online communities. *Organization science*. 22, 5 (2011), 1224–1239.

[20] Filippova, A. and Cho, H. 2016. The Effects and Antecedents of Conflict in Free and Open Source Software Development. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (2016), 705–716.

References

[21] Finin, T. et al. 2008. The information ecology of social media and online communities. *AI Magazine*. 29, 3 (2008), 77.

[22] Gastwirth, J.L. 1972. The Estimation of the Lorenz Curve and Gini Index. *The Rev. of Economics & Statistics*. 54, 3 (1972), 306.

[23] Gleave, E. et al. 2009. A conceptual and operational definition of'social role'in online community. *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on* (2009), 1–11.

[24] Global Internet User Survey 2012 | Internet Society: *http://www.internetsociety.org/internet/global-internet-user-survey-2012*. Accessed: 2016-05-15.

[25] Gong, N.Z. et al. 2012. Evolution of social-attribute networks: measurements, modeling, and implications using google+. *Proceedings of the 2012 ACM conference on Internet measurement conference* (2012), 131–144.

[26] Hou, H. 2015. What makes an online community of practice work? A situated study of Chinese student teachers' perceptions of online professional learning. *Teaching and Teacher Education*. 46, (Feb. 2015), 6–16.

[27] Hurvich, C.M. and Tsai, C.L. 1989. Regression and time series model selection in small samples. *Biometrika*. 76, 2 (1989), 297–307.

[28] Iriberri, A. and Leroy, G. 2009. A life-cycle perspective on online community success. *ACM Comput. Surv.* 41, 2 (2009), 1–29.

[29] Ishiguro, K. et al. 2012. Towards automatic image understanding and mining via social curation. *Data Mining (ICDM), 2012 IEEE 12th International Conference on* (2012), 906–911.

[30] Jackson, A. et al. 2007. Corporate Blogging: Building community through persistent digital talk. *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on* (2007), 80–80.

[31] Jeong, A.C. 2003. The sequential analysis of group interaction and critical thinking in online. *The American Journal of Distance Education*. 17, 1 (2003), 25–43.

[32] Johnson, C.M. 2001. A survey of current research on online communities of practice. *The internet and higher education*. 4, 1 (2001), 45–60.

[33] Kairam, S.R. et al. 2012. The life and death of online groups: Predicting group growth and longevity. *Proceedings of the fifth ACM international conference on Web search and data mining* (2012), 673–682.

[34] Kinsella, S. et al. 2011. Improving categorisation in social media using hyperlinks to structured data sources. *The Semanic Web: Research and Applications*. Springer. 390–404.

[35] Kraut, R.E. and Resnick, P. 2012. Building successful online communities: Evidence-based social design. *MIT*. (2012).

[36] Lazar, J. and Preece, J. 1998. Classification schema for online communities. *Classification schema for online communities*. (1998), 84–86.

[37] Leskovec, J. et al. 2008. Microscopic evolution of social networks. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (2008), 462–470.

[38] Liu, C.-C. 2011. Identifying the value types of virtual communities based on the Q method. *International Journal of Web Based Communities*. 7, 1 (2011), 52–65.

[39] Loyola, P. and Ko, I.-Y. 2014. Population dynamics in open source communities: an ecological approach applied to github. *Proceedings of the companion publication of the 23rd international conference on World wide web companion* (2014), 993–998.

[40] Luo, Z. et al. 2015. An Exploratory Research of GitHub Based on Graph Model. *Frontier of Computer Science and Technology (FCST), 2015 Ninth International Conference on* (2015), 96–103.

[41] Mačiulienė, M. and Skaržauskienė, A. 2015. Emergence of collective intelligence in online communities. *Journal of Business Research*. (2015).

[42] Malinen, S. 2015. Understanding user participation in online communities: A systematic literature review of empirical studies. *Computers in Human Behavior*. 46, (2015), 228–238.

References

[43] Matthews, T. et al. 2014. Beyond end user content to collaborative knowledge mapping: Interrelations among community social tools. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (2014), 900–910.

[44] Matthews, T. et al. 2013. Community insights: helping community leaders enhance the value of enterprise online communities. *Proc. of CHI*. (2013), 513–522.

[45] Matthews, T. et al. 2014. Goals and perceived success of online enterprise communities: what is important to leaders & members? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), 291–300.

[46] Matthews, T. and et al. 2015. They Said What? Exploring the Relationship Between Language Use and Member Satisfaction in Communities. *Proc. of CSCW*. (2015).

[47] McAuley, J.J. and Leskovec, J. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. *Proceedings of the 22nd international conference on World Wide Web* (2013), 897–908.

[48] McInnis, B.J. et al. 2016. One and Done: Factors affecting one-time contributors to ad-hoc online communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (2016), 609–623.

[49] Millen, D. et al. 2005. Social bookmarking in the enterprise. *Queue*. 3, 9 (2005), 28–35.

[50] Miritello, G. et al. 2013. Limited communication capacity unveils strategies for human interaction. *Scientific reports*. 3, (2013).

[51] Mohammad, S.M. and Turney, P.D. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to cre-ate an emotion lexicon. *Proc. of NAACL*. (2010), 26–34.

[52] Muller, M. et al. ’12. Diversity among enterprise online communities: collaborating, teaming, and innovating through social media. *Proc. of CHI*. (’12), 2815–24.

[53] Nolker, R.D. and Zhou, L. 2005. Social computing and weighting to identify member roles in online communities. *Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence* (2005), 87–93.

[54] Ortega, F. et al. 2008. On the inequality of contributions to Wikipedia. *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual* (2008), 304–304.

[55] Ovadia, S. 2013. Digital Content Curation and Why It Matters to Librarians. *Behavioral & Social Sciences Librarian*. 32, 1 (2013), 58–62.

[56] Padoa, C. et al. 2015. Investigating social curation websites: A crowd computing perspective. *Computer Supported Cooperative Work in Design (CSCWD), 2015 IEEE 19th International Conference on* (2015), 253–258.

[57] Panciera, K. et al. 2010. Lurking? cyclopaths?: a quantitative lifecycle analysis of user behavior in a geowiki. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), 1917–1926.

[58] Panciera, K. et al. 2009. Wikipedians are born, not made: a study of power editors on Wikipedia. *Proceedings of the ACM 2009 international conference on Supporting group work* (2009), 51–60.

[59] Panteli, N. 2016. On leaders’ presence: interactions and influences within online communities. *Behaviour & Information Technology*. (2016), 1–10.

[60] Panzarasa, P. et al. 2009. Patterns and dynamics of users’ behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology*. 60, 5 (2009), 911–932.

[61] Pedregosa and et al. 2011. Scikit-learn: Machine Learning in Python. *JMLR*. 12, (2011), 2825–2830.

[62] Pluempavarn, P. et al. 2011. Social roles in online communities: Relations and trajectories. *6th Mediterranean Conference on Information Systems, Nicosia, Cyprus. Retrieved October* (2011), 2012.

References

[63] Poole, E.S. and Grudin, J. 2010. A taxonomy of Wiki genres in enterprise settings. *Proceedings of the 6th international symposium on wikis and open collaboration* (2010), 14.

[64] Postmes, T. et al. 2001. Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin*. 27, 10 (2001), 1243–1254.

[65] Postmes, T. et al. 2000. The formation of group norms in computer-mediated communication. *Human communication research*. 26, 3 (2000), 341–371.

[66] Preece, J. 2001. Sociability and usability in online communities: Determining and measuring success. *Behaviour & Information Technology*. 20, 5 (2001), 347–356.

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

[68] Rambow, O. et al. 2004. Summarizing email threads. *Proceedings of HLT-NAACL 2004: Short Papers* (2004), 105–108.

[69] Richter, A. et al. 2013. Malleable End-User Software. *Business & Information Systems Engineering*. 5, 3 (2013), 195–197.

[70] Rowe, M. 2013. Mining user lifecycles from online community platforms and their application to churn prediction. *Data Mining (ICDM), 2013 IEEE 13th International Conference on* (2013), 637–646.

[71] Sassenberg, K. and Boos, M. 2003. Attitude change in computer-mediated communication: Effects of anonymity and category norms. *Group Processes & Intergroup Relations*. 6, 4 (2003), 405–422.

[72] Sen, P. et al. 2008. Collective classification in network data. *AI magazine*. 29, 3 (2008), 93.

[73] Strobl, C. et al. 2008. Conditional variable importance for random forests. *BMC bioinformatics*. 9, 1 (2008), 1.

[74] Tausczik, Y.. and Pennebaker, J.W. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*. 29, 1 (2010), 24–54.

[75] Terveen, L. et al. 1997. PHOAKS: A system for sharing recommendations. *Communications of the ACM*. 40, 3 (1997), 59–62.

[76] Terveen, L. and Hill, W. 2001. Beyond recommender systems: Helping people help each other. *HCI in the New Millennium*. 1, (2001), 487–509.

[77] Thorne, S.L. et al. 2012. The semiotic ecology and linguistic complexity of an online game world. *ReCALL*. 24, 03 (2012), 279–301.

[78] Toulis, P. and Kao, E. 2013. Estimation of causal peer influence effects. *Proceedings of The 30th International Conference on Machine Learning* (2013), 1489–1497.

[79] Toutanova, K. and et al. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proc. of NAACL*. (2003), 252–259.

[80] Walker, M.A. and et al. 2012. A Corpus for Research on Deliberation and Debate. *LREC*. (2012).

[81] Wang, G. et al. 2016. Unsupervised Clickstream Clustering for User Behavior Analysis. *SIGCHI Conference on Human Factors in Computing Systems* (2016).

[82] Wang, G.A. et al. 2015. An analytical framework for understanding knowledge-sharing processes in online Q&A communities. *ACM Transactions on Management Information Systems (TMIS)*. 5, 4 (2015), 18.

[83] Wang, Y. et al. 2012. To Stay or Leave? The Relationship of Emotional and Informational Support to Commitment in Online Health Support Groups. *Proc. of CSCW*. (2012).

[84] Wasserman, S. and Faust, K. 1994. *Social network analysis: Methods and applications*. Cambridge university press.

[85] Wei, W.W.-S. 1994. *Time series analysis*. Addison-Wesley publ Reading.

References

[86] Wellman, B. et al. 2002. The networked nature of community: Online and offline. *It & Society*. 1, 1 (2002), 151–165.

[87] Welser, H.T. et al. 2011. Finding social roles in Wikipedia. *Proceedings of the 2011 iConference* (2011), 122–129.

[88] Wenger, E. and et al. 2002. Cultivating communities of practice: A guide to managing knowledge. *Harvard Business*. (2002).

[89] Wen, X. and Lin, Y.-R. 2015. Information Seeking and Responding Networks in Physical Gatherings: A Case Study of Academic Conferences in Twitter. *Proceedings of the 2015 ACM on Conference on Online Social Networks* (2015), 197–208.

[90] Wilson, T. et al. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proc. of HLT/EMNLP*. (2005).

[91] Yang, J. and Leskovec, J. 2015. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*. 42, 1 (2015), 181–213.

[92] Zhang, J. et al. 2014. Role-aware conformity influence modeling and analysis in social networks. *AAAI* (2014), 958–965.

[93] Zhao, K. et al. 2014. Finding influential users of online health communities: a new metric based on sentiment influence. *Journal of the American Medical Informatics Association*. 21, e2 (2014), e212–e218.

[94] Zhu, H. et al. 2012. Effectiveness of shared leadership in online communities. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (2012), 407–416.

[95] Zhu, H. et al. 2014. Selecting an effective niche: an ecological view of the success of online communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), 301–310.

[96] Zien, A. et al. 2009. The feature importance ranking measure. *Machine Learning and Knowledge Discovery in Databases*. Springer. 694–709.

[97] Zivot, E. and Wang, J. 2003. Vector Autoregressive Models for Multivariate Time Series. *Modeling Financial Time Series with S-Plus®*. Springer. 369–413.