Title

Social Role Temporal Dynamics and Interactions in Online Communities: How are leaders and members different?

Author

Compton, Ryan

Publication Date

2019

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

SOCIAL ROLE TEMPORAL DYNAMICS AND INTERACTIONS IN ONLINE COMMUNITIES: HOW ARE LEADERS AND MEMBERS DIFFERENT?

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Ryan J. Compton

March 2019

The Dissertation of Ryan J. Compton
is approved:

Professor Steve Whittaker, Chair

Professor Marilyn Walker

Professor Lise Getoor

Lori Kletzer
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Social Role Temporal Dynamics and Interactions in Online Communities: How are leaders and members different?

by

Ryan J. Compton

Prior literature on online communities proposes an important function for social roles. Theoretical models argue that specific roles are critical for community success, and claim that individual roles shift in predictable ways over the lifetime of the community. However these models are currently hard to assess, as they do not provide systematic definitions of roles, nor empirical work evaluating the impact of roles on community success. Further they do not specify interrelations between different roles. This thesis addresses these questions in the context of enterprise online communities. Using gold-standard systematically-defined social roles, we explore role behaviors and their impacts using both quantitative and qualitative measures of community success. We find evidence contradicting prior theoretical claims about role shifts and the division of labor between roles. We also examine language styles within communities, finding a complex relationship in the types of support language that engender success. To develop new models of relations between roles, we utilize graphical methods to discover important community subgroups centering around leadership. These subgroups are important predictors of community success, significantly extending the explanatory power of existing network theories. Furthermore, subgroups serve as sources of significant topic-setting language. This work elucidates the workings of online communities at a user level, increasing theoretical understanding and suggesting how new community tools might be developed that promote community success.

To the love of my life,

Lisa Gatlin,

for all her support.

Acknowledgments

I would like to give recognition to all of those who supported me in making this work possible. First I would like to thank my advisor Steve Whittaker, he deserves a huge amount of credit for his guidance and mentorship in getting this work crafted and having me focused. I'd like to thank my committee members Marilyn Walker and Lise Getoor as their feedback and collaborations have been critical toward my PhD work.

Credit needs to be given to my lab mates in the HCI lab. Jeff Warshaw, Victoria Hollis, Aaron Springer, Lee Taber, Artie Konrad, Charlotte Massey, and Joel Schooler have been great colleagues and friends in providing advice, feedbacks, and always willing to lend a hand to someone in need.

Throughout my work, I've had many collaborators who exhibited amazing accomplishments and got us through tough challenges, which I want to thank them all. Hernan Badenes, in particular needs a strong recognition for his continued support through the years in data collection and storage.

My work couldn't have reached its goals without the amazing effort from my research assistants. Julius Kittler in his help replicating leadership models and beginning the ground work toward a fundamental aspect of testing role theories. Andrea Martinez in her dedicated effort in both building a coding book on role actions and help leading a team of researchers on qualitative coding. Marc de Giere in his excelling ambition to explore new methods in role research. Angela Ramirez who alone took on a substantial effort in exploration and building graphical representations. Mia Altieri for her supported efforts in coding and exploring natural

language processing algorithms. My work is built upon the support from all of these students.

I need to thank my friends and family for their endless support. My parents have provided me their limitless encouragement to always push myself. My brother being a influence and a wonderful friend. My friends never doubting me and always there to question when I get lost in a deep dive.

Lastly, I couldn’t be here today without the substantial support by Lisa Gatlin. While being my partner in life, she has gone above and beyond in helping through many hours of proof-reading and listening to my work

Chapter 1

Introduction

1.1 General Properties of Online Communities

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 daily with others online [2]. Online communities emerge from online connections through common interests or circumstances. These types of communities typically form from individuals who are unknown to each other offline. Over the past 20 years, online communities have been among the most popular applications on the internet [113, 129, 153, 160, 171, 173, 210]. Usenet, a worldwide distributed internet forum, had over 160,000 newsgroups active in 2006, and Yahoo claims to host over a million online groups [113], Wikipedia reportedly hosts over 35 million users with around 100,000 actively participating [6], and GitHub, a collaboration platform founded in 2008, has accumulated more than 31 million developers and 96 million repositories as of 2018 with about 8 million new users and nearly of third the total number of repositories being created over the past year [5].

There are many definitions of an online community [113, 129, 173], and though there are subtle differences, these definitions have a shared core. In this thesis we use the common definition of online communities adopted by Preece [173] and Kraut and Resnick [113], 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 derive a wide range of benefits that are similar to the benefits of offline communities [15, 23, 148, 174]. Participants within an online community have opportunities for information sharing and learning, companionship, social support, and entertainment [113, 133, 204]. These benefits can extend beyond the community to benefit non-community members by providing goods, such as open source software, product reviews, and encyclopedia pages [113]. These goods benefit from the diversity and range of contributions that an online community enables. As much prior work has argued [113, 119, 136, 173], the promise of online communities is that they break the barriers of time, space, and scale that limit offline interactions.

Online communities vary in size, ranging from just a few people to millions of users. These communities have also changed the way people interact by eliminating the obstacles that constrain offline interactions. While some online communities come from pre-established interpersonal connections, they typically connect networks of strangers around a common interest, topic, or goal [129, 173]. 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 and advice for those dealing with similar circumstances

[113, 204], curation groups on pinterest.com that organize ever-growing sets of domain-specific information [158], resource sharing platforms such as GitHub where people join teams to build software [74], and intranet enterprise communities allowing communication between workgroups [136]. With the rise of web-based social software, online communities also provide many different tools for collaboration. Historically, communities used simple communication tools such as email digests or Q/A forums [101] but newer social tools such as wikis, file repositories, and blogs have created diverse methods for communication and information sharing to occur [135].

There is extensive research studying various facets of online communities, including what makes communities successful [15, 99, 113, 171] and how communities change over time [18, 47, 99, 113, 121]. Communities also vary in type, from large groups with common interests or practices to smaller task-based groups with a shared goal for a particular project or function [146]. Researchers studying communities have defined community types in terms of their social attributes [119], supporting technology [134], relation to physical communities [120], functional characteristics [208], members' needs [125], and sponsoring organizations [60]. Some typical types found across multiple contexts are Communities of Practice and Teams. Communities of Practice are defined as groups of people who have a shared interest or practice, who share information and build social networks [146, 208]. Teams are defined as online groups working on a common goal, project, or function [146]. They typically work towards a well-defined "deliverable", and are common in contexts such as enterprise communities [146] and Open Source projects like those found on GitHub [126].

Other studies have examined communities at the user level, specifically the behaviors and formation of social roles [35, 129, 139, 166, 215, 219] and user dynamics over time [55, 68, 138, 141, 159, 160, 171, 181].

1.2 Social Roles in Online Communities

Most online communities are focused on collaboration and, as such, involve social interactions that have also been studied in offline settings [206]. One critical characteristic of communities is the concept of social role [81]. Within Usenet software, online communities were allowed to self organize and create their own policies outside of operating within a formal context. The content within these communities were generated by users themselves leading to a more democratic style of interaction and governance. During the advent of Web 2.0, they were among some of the first applications. This style of self governance and content creation 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, before joining the community participants do not usually know one another. This leads to questions about how community leaders incentivize participants to help others when they have no strong interpersonal ties to the community [35, 106, 113, 129, 178, 192]. 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 [99, 171, 184]? Current community models [99, 171] assume that participants are initially drawn to the community by an interest or specific question. Critical to these models is that communities provides multiple potential roles in an ecosystem that supports different levels of participation. These roles range from simply reading/lurking, making occasional posts, active contribution, through to active stewardship of the community. These community models [99, 171] 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.

A major binary distinction in research on social roles is between members 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 [33, 171]. Leaders also facilitate coordination among members [161]. Much research has been conducted on the benefits of leadership for a community [161], defining key leader behaviors [35, 69, 138, 219], as well as distinguishing leader and member behaviors [35, 133]. These theoretical distinctions between members and leaders [99, 171] are supported by some empirical work [35, 133, 161]. However other work suggests that the picture is more complex, finding that the same behaviors occur across roles [171, 219], with members sometimes enacting leadership behaviors. Certain behaviors that are typically associated with formal community leaders, such as offering directive, positive and negative feedback, and person-focused leadership styles, have been observed in both leaders and members indicating non-formal leadership practices [219]. Zhu et al. [219] also found that directive behaviors such as setting goals and actions occurred more frequently among regular community members than formal leaders. However, certain behaviors were more common in formal leaders who engaged in person-focused activities, including welcoming new members or simply posting positive emoticons.

While the existence of apparently overlapping behaviors between members and leaders makes it difficult to empirically distinguish categories of social roles, there are functional implications that can come from such work. A lot of what communities have to do is allocate volunteers to a specific task in the community, whether it's answering a post, seeding conversations, or moderating discussions [113, 136]. It really helps communities co-organize (especially leaders) if they can identify what skills/behaviors others have within the community (e.g. who has made successful posts in the past, who can moderate, etc.). Communities can also use roles to determine whether they are missing key skills needed for community success.

In addition to the functional motivations, there are theoretical arguments for distinguishing roles. Gleave et al. [81] argue that role clarity 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. Following this approach, social roles have served as a lens for social scientists to study the underlying structure of social interactions. For example, Nolker and Zhou [153] define a user to be a motivator if they exhibit a behavior of interest (keeping a conversation going) at a higher frequency than compared to all users thus making it a distinct role observation.

With this in mind, we are proposing to study the behaviors, interactions, influences, and changes that individuals enact in an online community with the goal of classifying those behaviors into formal categories we will call social roles. As already 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 Contributions

This section outlines the contributions of this thesis in each of the 5 studies; each study is described in an independent chapter.

1.3.1 Content Management through Referencing

Understanding content management is crucial as successful communities increasingly accumulate large amounts of content that need to be referenced for new or returning users. But many complex acts of content management are hard to operationalize across large datasets, because they require manual content analysis that relies on detailed domain knowledge [90, 115]. Our main goal in this study is to assess representative content management behaviors across a large dataset, so we chose to explore one specific quantifiable form of content management: hyperlink referencing. Hyperlinks are an efficient and pervasive general method to structure complex information [93], e.g. by referencing underlying content using a structured list or creating readable annotations [147]. We explore hyperlink usage across multiple communities over time, testing two predictions. We expected hyperlinking to increase over time, as community participants tried to organize the ever growing amounts of content within their community. Contrary to this expectation, linking decreased over time, despite accumulated content. Further hyperlink and content analysis suggested a possible reason for this, which is that participants are focused on recent content. We also examined predictions of lifecycle models which argue that community members will take increased responsibility for content management over time. Overall, links were mainly generated by leaders and there was no evidence that members linked more as the community matured.

1.3.2 Evaluation of Role Dynamics

There are conflicting accounts of how roles change within online communities. On one hand, the theoretical apprenticeship model proposed by Preece and Shneiderman [171] outlines a progression where participants adopt increasing responsibilities for community management over time. On the other, Panciera et al [159], argue that roles are static with participants showing consistent role behaviors during their entire time within a community. We test these contradicting models by examining role shifts during the online community life-cycle. We train accurate categorization and probabilistic models using formal gold standard data, then use these models to test the contradictory theories. If the apprenticeship model is correct, then we would expect performance of the categorization model to decrease with time as more members begin to behave like leaders. However, if the model performs equally over time, then this supports the claims of the static model. Likewise if apprenticeship occurs we would expect the probabilistic role model to show increased likelihood as the community mature that individual posts are generated by leaders. Both models provide evidence that roles tend to be static.

1.3.3 Factual versus Emotional Language and its relation to Community Success

We examine how the content of participants' conversations affects community success. A key question concerns the impact of emotional versus factual language. In some communities, emotional support for other members is critical [134, 204]. In contrast, other communities serve short-term informational needs, where the most common interaction is a simple information request from a first-time poster where a factual response is optimal [63]. This study examines the relationship between emotional versus factual communication on participants' perceived success of their community. We first develop an algorithm that accurately detects the prevalence of emotional versus factual content in posts. Contradicting earlier work for support communities, we find that more factual content is associated with higher levels of community success.

1.3.4 The Role of Subgroups in Online Community Interaction

Research on community roles has generally focused on individuals or characterized network relations for the entire community. Such prior models fail to explore subgroup structures and how different roles coexist within a community. The current study uses graphical methods to identify substructures within a larger network. These substructures are typically referred to as graphlets or motifs. We explore how these 4-node graphlet structures relate to individual roles and content production, and examine how these substructures predict metrics of online community success. Graphlets have considerable explanatory power improving network measures by 16% in predicting community success. Furthermore, graphlets are more likely to contain leaders who are in influential positions within the network.

1.3.5 Defining Subgroups through content

While the previous study finds that specific graphlets are predictive of online community success, such correlational approaches don't shed light on what makes these substructures important to communities. To explore this, we analyse how content production relates to graphlet type. First, machine learning models indicate that different graphlets produce different content types. Other analyses find that denser graphlets contain more content per post, and posts from denser graphlets also introduce anticipatory content that will later become more prominent within a community. Surprisingly, graphlets without leaders generate more content and anticipatory content.

Chapter 2

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 [81]. As we have seen, online communities provide a mechanism allowing for different modes of participation through different roles. One major distinction, and a key focus of this thesis, is between leaders and members. Members can contribute in many ways that have different effects on the community, the default contribution being a single post. Members can also be subcategorized, into categories like those discussed above in user lifecycle models [96, 99, 102, 113, 171, 207]: namely readers, contributors, chatters, collaborators, and motivators [99, 153, 171]. As described in lifecycle models [171], member participation follows a lifecycle apprenticeship model progressing through increasingly demanding roles: reading, contributing, collaborating, and leading. Figure 2.1 shows the expected evolution that a user might undertake as outline by Preece and Shneiderman [171].

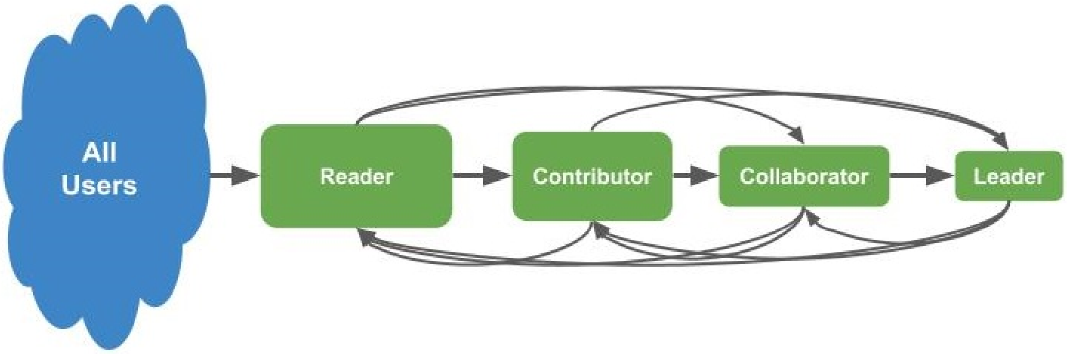


Figure 2.1: User Role Lifecycle Model as presented within Preece and Shneiderman [171]

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 do users become contributors [171]. Collaborators are those that go beyond just simple content contribution and focus more on developing relationships, working together, and setting goals. Finally, when users are actively promoting participation, mentoring novices, and setting and upholding policies are they considered leaders. Again, we should note that most such lifecycle models [113, 153, 171] are descriptive and provide many categories of role types with few studies actually operationalizing and measuring these stages. These roles models are hard to operationalise across varying community types and contexts however there there is additional work that makes very similar role categorizations [99, 113].

While those models are heavily theoretical, not all roles are defined by theoretical models. Some prior work takes a more empirical approach to identifying roles categories. Nolker and Zhou [153] used both network and information retrieval methods to capture role

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 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

Table 2.1: Attributes and measures of motivators and chatters gathered from Nolker and Zhou [153]

types within Usenet bulletin boards. Table 2.1 operationalizes the distinction between various role types as proposed by Nolker and Zhou [153]. Chatters are identified through conversation patterns that are not community supportive [153], such as influencing other users, but such conversations can be seen as social networking behaviors to build up relationships. Motivators are more prevalent roles when it comes to social relationships, and similar to the collaborator role as defined within [171] they aim to develop relationships.

In table 2.1, we see that motivators and chatters are differentiated through various behavioral and structural network measures like closeness (which is a measure of the average distance to all other members within a social network [205]), indicating that motivators interact with many others as opposed to chatters who are frequently interacting with only a few select individuals [153].

Additional work was conducted by Welser et al. [207], where they examined Wikipedia communities empirically analyzing common behaviors across users and qualitatively categorizing them based on context specific activities only found within Wikipedia. With one exception, the roles they found were similar to theorized roles. The four main roles they found were: Substantive experts, Technical editors, Counter-vandalism and Social networkers. Substantive experts are similar to experienced contributors, in that they provide substantive content to the community, either providing expert responses, or stimulating new discussions or topics. Technical editors correct the work of others to find small errors such as spelling, grammar, hyperlink format, or out of date facts. They contribute small but necessary aspects to the community. Preece and Shneiderman [171] describe such minor editing acts as ways that users can initiate their role as contributor. Counter vandalism is a specific role for Wikipedia, to discover vandalized articles, correct them, and sanction vandals. However these are similar to regular community leadership behaviors like enforcing community norms and policies [35, 99]. Lastly social networkers act 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. [207] further examined how each role relates to the community social network structure. The most noticeable differences between roles are that of networkers and substantive experts. Social networkers interact with a limited group of participants while experts show broader connections developing relationships with fellow experts beyond their immediate subgroup. This empirical work by [153, 207] 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 formally defined as the key role that promotes participation, mentoring, and setting and enforcing norms and policies [171, 219]. Previous theoretical work has identified a range of leadership behaviors: Transactional, Aversive, Directive, and Person-based leadership [219]. Table 2.2 shows these leadership styles as categorized by a machine learning model built by Zhu et al. [219]. Transactional leadership is when the interaction between the leader and member is considered a transaction or exchange, where the leader is providing praise or reward for the member ("Great job, thanks for the work!") and in some cases even withholding from punishment. Aversive leadership contrasts with 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 issuing 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 to members. Person-based leaders place emphasis on interpersonal relationships with members, and works through encouragement, inspiration, intellectual stimulation and empowering. These leaders focus on developing self-management skills of the member and team work.

These characteristics of leadership style contrast with member profiles allowing for the foundation to define various actions of leaders, thus allowing for stronger models of roles and their measurable influences on communities.

Machine Learning (ML) categories & Leadership type	**Sample messages**
ML category: Positive feedback **Leadership type**: Transactional leadership (Task-focused)	"I award this barnstar 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"
ML category: Negative feedback **Leadership type**: Aversive leadership (Task-focused)	"If you continue in this manner you will be blocked from editting without further warning."
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 tamplate..."
ML category: Social message **Leadership type**: Person-focused leadership	"Hi XX. Welcome to Wiki Project 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 :)..."

Table 2.2: Sample messages from Zhu et al. [219] from Wikipedia editors identifying correspondence between machine learning categories and leadership types.

2.1.2 More complex models of leaders.

Pluempavarn et al. [166] showed that formal roles of members and leaders vary across community contexts, while Zhu et al. [219] and D'Innocenzo et al. [61] 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 [61, 219] better explains instances of members being collaborators and motivators who also share the responsibility of promoting participation and mentoring typically associated with leaders. These subcategories of membership remain to be fully understood in their effect on communities.

Other work has examined different ways in which leaders' presence influences a community. Panteli [161] examined four different forms of leader presence. The four types of presence are interactive (leaders who interact with their members frequently, while 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 available to members, but does not interact with members on a frequent basis) as shown in table 2.3.

While the first three categories are similar to previously defined leadership behaviors, the fourth category, silent, is unique. Panteli [161] identified this by observing relatively infrequent and decreased posting by leaders, even though these limited posts attracted high member attention. This decrease in leader posting promoted higher member interactions, suggesting that the leader is allowing for more member interactions by interfering less. However, it is still necessary for the leader to show by occasional posts that they are still involved, suggesting that leaders are still actively reading others' content.

2.2 Beyond Individual Roles

To address the broader concern of how community roles interact we will need to the community as an ecosystem of social roles. Social role ecologies involve the interplay of roles within a community [81] and previous work examined multiple interacting communities as an ecosystem of communities addressing a common topic within a technology platform or organization [218]. An ecological perspective incorporates a perspective that is missing from lifecycle models, which typically define only dyad relationships, whereas there are much more complex relationships that can exist within a community. One form of ecological perspective can be based on organization ecological research [218] which is more community centric. Organization ecology research creates two ecosystem mechanisms: competition and complementarity.

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 toher **disappearance** has **not affected** the way I follow the site" (interviewee/BC member 10)

Table 2.3: Summary of Results from [161] indicating four leader types.

Competition is where organizations compete with others in the same ecosystem for common resources which is more intense if organizations demands similar resources. 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 modeled using structural network methods [18, 104, 121, 141, 162, 181].

Other work addresses population dynamics in Open Source Online Communities. Loyola and Ko [126] adapted biological models called Lotka-Volterra models, which are used for describing host-parasite interactions, to understand how contributions evolve within a GitHub community. This adaptation was able to model community dynamics over time. Other studies aim to understand the linguistic ecology of online gaming communities [195] and explore the information structures of hyperlinks to signal how community ecologies are organized [75].

Beyond some structural approaches [207], these complex role accounts do not specify how each role interact with the community as a whole. In contrast the approach we follow in Chapters 7 and 8 aims to characterize roles in relation to other community members, i.e. the community ecosystem, as well as how such interactions relate to community success metrics.

2.2.1 Role Interactions

If social roles are defined in relational terms, this means that certain roles cannot exist in the absence of other roles. Using an example from Gleave et al. [81], 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 [46, 153, 161, 215, 219]. Methods to find influential users include: examining conversations between users [15, 161, 219], identifying central nodes within a network using social network methods [46, 122, 153, 215], and examining difference between user behaviors and community norms [55, 181].

Work on role interactions takes two different approaches, either it examines the user-community interaction [15, 45, 55, 162, 181] or focuses on local dyadic interactions between two users [47, 215]. Both approaches contribute to our understanding of the interactions taking place within an online community. Effects on the overall community can happen by an aggregation or accumulation of common feedback that one receives from the contributions one makes, as found by Cheng et al. [45]. They found that community negative feedback leads a user to produce more content, but that content is of lower quality. Structural methods, applying models that take a social network perspective and measures of individuals relationships, influences, and position with the network, have also been used to study influences on users' joining, relationship forming, and communication behaviors [18, 104, 122, 162]. Other work shows first that structural features within a social network predict how likely a user is to conform to community norms [215] and second that linguistic deviations from community linguistic norms predict lower levels of user participation [55] or even lead the user to leave a community [181]. Individual effects are just as important as shown by Zhao et al.[217], finding that conversation sentiment can identify influential users who enact successful behaviors like community building and information retrieval. The work discussed previously by Zhu et al.[219] further shows the influence of individuals by examining which leadership styles promote contributions within

Wikipedia communities.

2.2.2 Network Approaches to Communities

Networks or graphs have been widely used to understand the connections and dynamics of online communities [77, 82, 122, 141, 153, 162, 216]. Typically networks use nodes to represent individuals, and edges to represent a connection between them [205]. While techniques evolved for offline behaviors, the emergence of online social networks has promoted widespread adoption of these methods. For example, Aumayr et al. [17] demonstrated how to convert forum based thread conversations into a relationship network representation. Similar network approaches have been used to analyze complex social phenomena, e.g. to identify key individuals and social roles within communities [16, 153, 216], relationships within threaded conversations [17], reciprocal relationships [24, 82], group dynamics over time [47, 82, 175, 183], as well as relationship formation and strength [84, 179].

Network representations provide a rich set of metrics that can help operationalize social theory. Rowe [181] used in-degree and out-degree distributions to measure social dynamics to model likelihood of individuals leaving a community, finding that lack of replies (in-degree) was a significant predictor of user churn likelihood [181]. More relationship-based measures extend these methods beyond dyadic one-to-one interactions (like degree) to instead model one-to-network interactions. Those interactions identify high connection, influential nodes within the graph, using measures like Centrality [20]. For example, Nolker and Zhou [153] used multiple relationship-based measures including degree, betweenness, and closeness to identify various role types in Usenet communities, finding more impactful roles (e.g. leaders and motivators) to have significantly higher network relationship measures [153]. Johnson et al. [102] also explored relationship-based measures in modeling role behaviors, finding leaders to be associated with k-core, a measure quantifying the network of nodes with at least k degree [102]. Sparrowe et al. [187] examined workplace relationships using a network representation to explore the relationship between centrality and task behaviors. Performance was positively related to centrality in cooperative networks with the opposite being true for uncooperative cases [187].

Other work has focused on different properties of network nodes. One key network topic is Assortative Mixing, or the tendency for nodes to interact with other similar nodes [150]. Chung et al. [47] explored assortative mixing in forums for government workers, and while they found mixed results they opened up new questions about community interactions in Web 2.0 tools [47]. Gong et al. [82] also observed early patterns of group node attributes within Google+ finding that social networks on this platform had lower social reciprocity and assortative mixing compared with other social networks [82]. More in-depth analyses found reciprocity related to common attributes of nodes within their networks, providing a more nuanced picture of assortative mixing.

Arnold et al. [16] recently explored a network-based approach to combine information about informal roles and their interactions over time. They were able to test whether overall performance of an online community depends on how individuals interact with one another. They found certain interactions, such as a chain with a Copy-Editor in Wikipedia connecting a newcomer, have a significant impact on the overall performance of the community. Their work points to a need for metrics targeting implicit coordination that brings together the right people [13, 16]. This thesis looks to extend that idea and suggests possible measures of implicit coordination.

Finally, network analysis has been extended beyond simple individual or network level metrics to explore different possible network configurations. Cummings and Cross [54] explored the relationship between network structure and performance in an online work community. They examined core-periphery and hierarchical network structures by measuring the structural holes within work groups, finding these to be associated with negative work performance, indicating a lack of connectedness between leaders and the rest of the network [54]. This work highlights a key point missing from many analyses of networks, that structure or network orientation is an important contributor to group success.

Prior work also has identified substructures and their relevance to social properties such as triadic tendency which was theorized from Social Exchange Theory [20, 54, 73]. Network science refers to these structures as graphlets or motifs. One example of a subnetwork type is a triad (involving three nodes) which has many different possible configurations such as connected (two edges) versus a clique (three edges). These structures have proved useful tools for evaluating social relationships within networks, i.e. transitivity of a network (being the proportion of three users forming a clique of three) has implications to the social ability and likelihood for connections to form [20, 24, 73]. These substructures may explain leader effectiveness, friendship formation and the community context as well as network evolution over time, [16, 20, 54, 62, 73, 163, 179]. Prior work has identified significant patterns involving triads [24, 73] and Faust [73] showed triad census (measuring the number of existing triads compared to those that could exists) exceed expectations from dyadic census, suggesting that substructures may provide information that is not captured by such lower-level measures. Dong et al. [62] also recently examined homophily through exploring various substructures across both Facebook and LinkedIn. Their findings are rather complex as they find the expectation of social structures (triads or cliques) being formed varies between different contexts (cliques were inconsistent between Facebook or LinkedIn). This work highlights a relationship in how individuals form connections and the motivations in which people join a given social network platform. We expand on their work by an examining the relationship between social roles and communication for different graphlet-based substructures.

Although people have recently begun to explore graphlets, they have traditionally not been assessed for computational reasons [8]. While counting the number of graphlets is possible [8, 122], efficient techniques for finding graphlets in networks have yet to be developed. As shown by Ahmed et. al [8], counting existing graphlets can be achieved through combinatorial means, but finding who participates in those graphlets needs a search procedure as all nodes need to be evaluated for the potential graphlets they could be a part of, hence making the search space of the number of possible 4 groups within all possible nodes. Gathering the sets of nodes that create graphlets is a very different task than measuring the number of existing graphlets within a network.

Using both graphlet counting and finding algorithms will be used within Chapters 7 and 8. Due to the computational complexity of graphlet finding (gather sets of nodes that create a graphlet), we limit this to only examining 4-node graphlets. This does go along with prior work on effective team sizes within work group, therefore the node size limitation is still within an interesting level of group size to explore [86].

Summary of the notation and properties for the graphlets of size 4. Note that ρ denotes density, Δ and $\overline{d}$ denote the max and mean degree, and assortativity is denoted by r. $|T|$ denotes the number of triangles, κ is the max k-core number, D denotes the diameter, B denotes the max betweenness, and $|C|$ denotes the number of components.

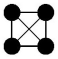
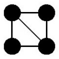
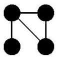
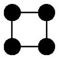
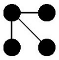
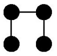
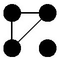
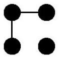
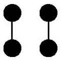
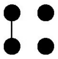
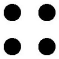
Graphlet	Description	ρ	Δ	$\overline{d}$	r	$\lvert T\rvert$	κ	D	B	$\lvert C\rvert$
Connected Graphlets										
	4-clique	1.00	3	3.0	1.00	4	3	1	0	1
	4-chordalcycle	0.83	3	2.5	−0.66	2	2	2	1	1
	4-tailedtriangle	0.67	3	2.0	−0.71	1	2	2	2	1
	4-cycle	0.67	2	2.0	1.00	0	2	2	1	1
	3-star	0.50	3	1.5	−1.00	0	1	2	3	1
	4-path	0.50	2	1.5	−0.50	0	1	3	2	1
Unconnected Graphlets										
	4-node-1-triangle	0.50	2	1.5	1.00	1	2	1	0	2
	4-node-2-star	0.33	2	1.0	−1.00	0	1	2	1	2
	4-node-2-edge	0.33	1	1.0	1.00	0	1	1	0	2
	4-node-1-edge	0.17	1	0.5	1.00	0	1	1	0	3
	4-node-independent	0.00	0	0.0	0.00	0	0	∞	0	4

Table 2.4: Summary table of Graphlets up to 4 nodes, retrieved from [8]. This table defines each possible graphlet type with 4-nodes. Along with a description (or label) for each type, network descriptive measures are provided to further differentiate between each type of graphlet.

2.3 Success

2.3.1 Success: Definitions and approaches

Much of the prior work on online communities aims to relate the phenomena of interest to significant outcomes within the community. Typically these outcomes are referred to as community success. Online community success has been assessed in various ways, from counting simple activities such as posts, voting, and editing [15, 45, 99, 173] to satisfying users' reasons for joining a community or user retention [15, 45, 99, 173]. While many studies use simple quantified metrics such as volume of posted messages, they have been criticized as ignoring the content and quality, of responses [99, 133]. For example, some messages in high volume threads may be spam or negative feedback. More precise measures targeting user perceptions such as measures of member satisfaction, reciprocity, and trustworthiness have been proposed instead [99, 133, 173].

2.3.2 Perspectives on Success

Many proposed success factors are interrelated, but focus on different community behaviors. Factors such as community growth, user retention, topic, interactions with other communities, and quality of information all take a community level perspective. In contrast, other measures focus on the individual level, such as response time for a user, quality of responses, and community feedback (e.g. through ratings of users' posts). No one measure seems adequate although many of these measures can be consistent. At the community level, growth may indicate how many users have an interest in the content present [99] or how much novel content is provided across multiple communities [218]. Content quality is intuitively related to user interest, i.e. for a community with unappealing content, users won't see the value and will ignore it [45]. While hard to measure directly, content quality can be theoretically representative by a measure of user satisfaction. A study by Matthews et al. [134], conducted a survey of users and asked "how well this community is meeting your needs" where users would answer on a Likert scale. This measure operationalizes some theories of community success [112, 172, 208]; however it is not an exhaustive measure of success as Preece [173] points out that factors relating to usability need to be taken into account as well as other social factors such as reciprocity and trustworthiness.

2.3.3 Who is successful? Variation in Goals of Social Roles

Member satisfaction addresses a key requirement brought up by Preece [173] and Matthews et al.[133], when asking about community success, who is being successful? This points to the fact that success for one individual does not necessarily indicate a successful community. Matthews et al. [133] found that members and leaders have different goals. For example leaders may be focused on community development; therefore a highly active community can be successful, while a member may be looking to for an answer to a specific question. These goals can also be in contention with each other, in the above example, the member's question may go unanswered in dense communities which can ignore non-important members.

Another complexity is that there are a variety of community types. Communities have different goals and needs which they aim to satisfy in different ways [99, 173]. Examples of different community types are those of Question/Answer communities where the primary actions are information seeking or Communities of Practice where user have shared interests, communicate, build relationships and share resources. These types can vary in their goals and actions, for instance Q/A communities are looking for quality answers, but communities of practice may be looking to build social connections around a given topic. Low level 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 variations in 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.

Another complexity in lifecycle models is noted by Iriberri and Leroy [99]. They hypothesize that success factors likely 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 than a community already focusing on existing members; for example a young community can be looking to define what need it is addressing (i.e. what makes this community unique) while older communities can be focused on more informational logistics trying to handle the vasts amount of content present in the community. Some of the general behavioral metrics proposed above work well for the life cycle degrees of success. Declines in many activity measures (e.g. number of posts, content quality, active contributors, response time) may indicate that a community is approaching death, while an increase in such measures can indicate robust growth. However even if time is included, Iriberri and Leroy still examine what the community and user's goals are, which points to the same conclusion that success can be evaluated as how well a community is satisfying its and the user base's goals. While this is an important factor to note for future work, this thesis will need to first address simpler operationalizations before moving into possible lifecycle stage success metrics.

The notion of success is central to all main chapters of this thesis. We will mainly rely on a survey measure of success, however we will combine this with many low level activity based metrics as they can be indicative of context specific success, i.e. Q/A community success can be measured by how well the community is responding to questions assessed by rate of responses and speed. This thesis focuses on success for the entire community but we will not examine multiple types of communities in depth, or role based success measures. Instead the focus on relationships to generalized community success, be that through qualitative metrics like Member Satisfaction or known quantitative metrics like quantity of content, reciprocity or language.

2.4 Role changes over time

2.4.1 Community dynamics

Models such as Preece and Shneiderman describe how individuals change roles over time. In contrast Iriberri and Leroy [99] propose a community lifecycle model which argues that overall interactions between members of the community undergo predictable shifts which they call stages: Inception, Creation, Growth, Maturity, and Death. Each stage is defined by member behaviors and needs for information, support, recreation, or relationships, as shown in figure 2.2.

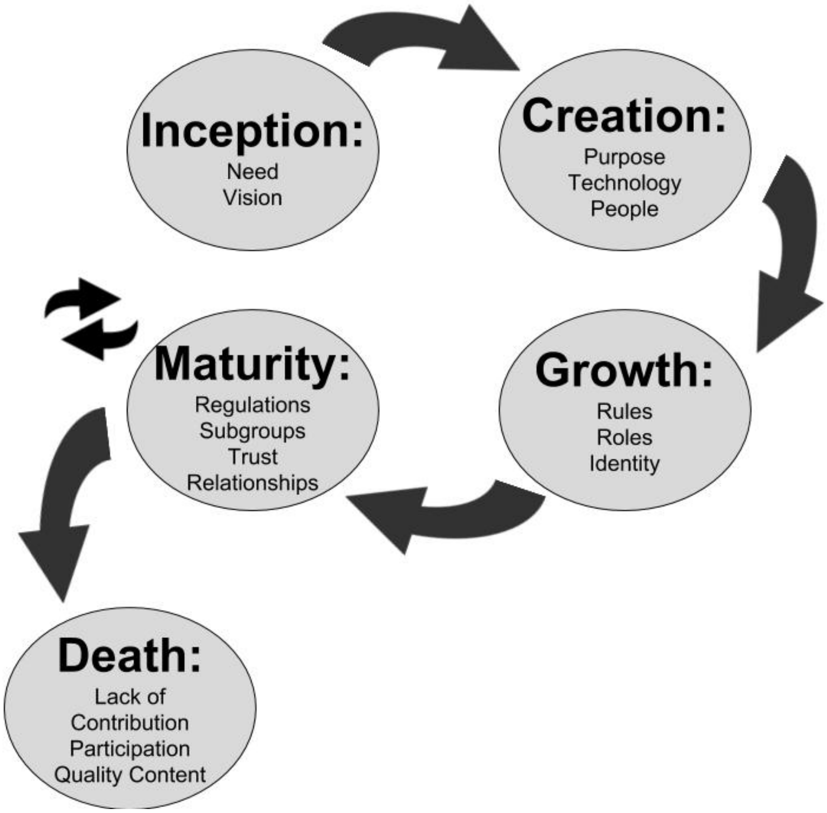


Figure 2.2: Stage Lifecycle model as proposed by Iriberri and Leroy [99]. Communities begin in the inception stage, work up to maturity where these stages can repeat, but end in death.

During inception, the initial idea for the online community arises from unspecific needs, and depending on the type of need, users may propose an intended goal for the community. With this goal in mind, participants may create simple policies to help maintain a focus. The creation stage begins once online technical components are in place, tools such as listservs, discussion forums, wikis, chats, and community blogs. The early group of users can now interact and further increase membership through recruitment. The growth stage is when the culture and identity of the community emerge. As more users join, they will start to emulate roles and identities within the community. Some more participative users will support others by leading discussions, while others may simply look for support and information. The maturity stage arises when a formal organization is required by the high amount of activity. The community therefore creates more regulations, contribution incentives, sub-communities, and wider range of discussion topics. While some early users may leave, new users join bringing new ideas for discussion and adopt community roles. Iriberri and Leroy [99] state that in maturity communities thrive for long periods being strengthened by trust and lasting relationships. Some communities may iterate through the lifecycle again as user interests shift. Alternatively, however they may approach the death stage when the community loses momentum and member interest leading to poor contributions and transient membership. Iriberri and Leroy [99] define each stage by characterising community needs or behaviors across multiple users. In contrast, user centric models offer an individualistic lens into how each user forms an identity within a community.

2.4.2 User Lifecycle

Preece and Shneiderman [171] generated a framework describing how users progress through various roles within an online community. Figure 2.1 demonstrates this progression of user activities and roles. Fewer users progress at each stage of the model, i.e. there are fewer leaders than any other role. Arrows show there may be a non-linear progression through roles, with many potential paths. However the model does not clearly describe what factors lead participants to transition between roles, but transitions can occur through repeat community visits that foster a growing sense of confidence and increased activity. These authors [171] further propose two paths of maturation for a user: users become more active within one stage or they move on to begin another stage. Users join a community by looking to satisfy their curiosity or needs. They begin reading, searching, and returning to the community if they feel they are satisfied. Users transition to 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 engage in more substantive acts such as tagging, reviewing, posting, or uploading to the community. When users establish mutual understanding or shared beliefs with a number of other participants, they may become collaborators. Collaboration in this context involves two or more contributors discussing or working together to create or share content. For example in Wikipedia, collaborators will share information correcting each other with the goal of producing a wiki of consumable information for other users. If users who are further motivated to improve the community may they become leaders. Leaders establish community norms and policies. Furthermore, leaders may promote participation and mentor other users.

Leaders typically contribute the largest number of comments and are the most active as can be seen from the observation that in many communities 90 percent of the comments come from less than 10 percent of the participants [157].

2.4.3 Deducing Lifecycle Behaviors

According to lifecycle models, user's behaviors change and user dynamics are a central aspect of community models [99, 171]. However there is little quantitative data supporting the precise evolution of user roles or stages in lifecycle models. Nor is there good evidence about the effects of role dynamics on community success. This thesis aims to connect these two important aspects of online community research.

Some quantitative research confirms change models showing that users' expertise in the content of a community increases as their time with the community gets longer [121, 138] supporting the aspects of user lifecycle models which state that users mature within a community gaining the skills to assist other users. It remains to be seen if such user dynamics are the same for all types of roles.

However other empirical studies make different claims, instead arguing that roles are relatively static, or that long-term participants become less active in communities over time. These role studies have used a combination of behavioral or social structural network analysis methods. One approach starts by operationalizing a social role; Panciera et al., [159] define a Wikipedian with over 250 edits to be a significant contributor within Wikipedia, then the authors identify users within a community that fit this role and then examine how those users' behaviors change over time. Overall Panciera et al. [160] found that significant contributors' behaviors did not change over time instead being consistent during their entire time in the community. Miritello et al.[141] also examined user dynamics assessing the number of social connections users maintain and the level of activity with other users. They found that users' social circles and interactions decreased with time. Danescu et al. [55] observed linguistic changes by modeling differences in term distributions in comparison to overall community norms. They found that users first adapt to community norms but then refrain from conforming. Other work has focused on the user-community interaction which takes the perspective of measure how users' behaviors differ from community norms. Rowe [181] 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. Such findings are contradictory to theorized lifecycle models which argue that overall interactions tend to increase over time, and become more consistent with community goals and practices.

These prior studies suggest some requirements for future work which are addressed in this thesis. First, recent quantitative studies show the need to refine stage models to characterize whether and how participants and communities change over time. A second major requirement concerns the development of new models that characterize subgroups within communities. Some work addresses roles in relation to others in the community [55, 153, 166, 181, 219] or the structural formation of online communities at the user level [47, 82]. However there is a lack of consensus about this.

Addressing these questions using quantitative methods would first update user centric theoretical models to accurately characterize role dynamics. It would also extend these models to encompass broader aspects of community ecology through subgroup formation [81]. Finally it would determine the contributions of different roles and subgroups on overall community success. This improved empirical and theoretical understanding should inform the development of new tools to enhance community that promote success [81, 171].

Chapter 3

Research Context and Methods

3.1 Research Context

This research was conducted in a global technical enterprise offering technology products and services to businesses. The company widely encouraged employee leadership of, and participation in, internal online communities by making easy-to-use commercial community technology available to all employees. For the remainder of this thesis, this enterprise community application will be referred to as "Communities". Communities is a pre-existing corporate product and was not developed for the purpose of the current study. A screenshot of the Communities landing page for a complex community working to launch a product is shown in fig. 3.1. The figure shows how that community synthesizes different resources around Web Marketing. Information is also shown about forum discussions, tags and community members, as well as bookmarks to related sites. All the communities we studied used the Communities application, which enabled participants to easily create a community space combining various social tools, e.g. forums, blogs, wikis, files, and bookmarks.

Across the thousands of participants and communities we sampled, skill sets varied widely as people were drawn from all across the company; skills ranged from highly technical (Software and Electrical Engineering) to less technical (Human Resources, Marketing and Management). Participants included both community owners and members, and we provide more details about these roles below. All participating users were employees of the same enterprise and were aware that this tool was being used for research purposes and proactively agreed in email to have their anonymized survey and logfile data used for analysis.

The software we studied was used for a very broad range of activities, from organizing social events in Communities of Practice, to goal focused activities like developing a marketing strategy for a range of new products across the company. The community activities we observed were very similar to those described in prior online communities research [35, 99, 171, 208]. Consistent with that other work, the communities we analyzed mainly focused on distributing knowledge, sharing expertise, answering questions and providing social support. Some communities appeared to be large communities of practice for members of a shared discipline (e.g. software engineers or marketers). Other communities appeared to be teams with executive leadership and more narrow goals specific to enterprise needs. Yet others were focused on specific recreational or technical problems over a shorter time-frame. To evaluate whether our communities functions overlapped with those identified in prior communities research, we surveyed community owners asking them to describe prevalent community activities and the type of the communities they managed. Owner responses were qualitatively clustered as follows: 41.1% Communities of Practice (many members, mainly expertise sharing

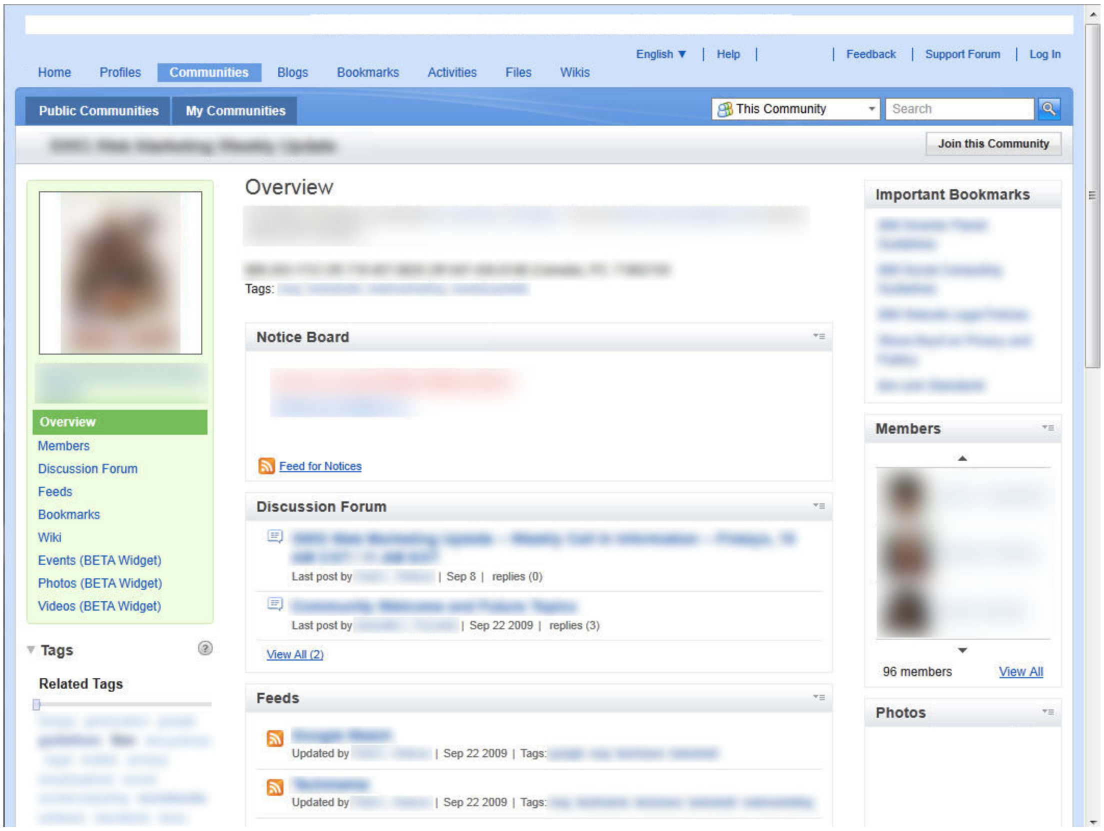


Figure 3.1: Screenshot of a ‘Communities’ landing page, where participants gather content, discuss strategy and compile data in relation to co-ordinating Web marketing. Personally identifiable information has been blurred out. Overview of the community’s media tools are provided in the left pane, recent discussions are within the center pane, and bookmarks, and community membership information are provided in the right panel.

and networking), 29.4% Teams (executive leadership, fewer members, goal oriented projects), 3% Technical Support (providing technical advice to end-users), 1.4% Others. These subtypes both match those described in other analyses of enterprise communities [146], as well as the literature more broadly [99, 113, 208]. We will present systematic analyses of different community types when differences are found to be present as contextual analyses [146] suggest such differences may exist. The median number of members per community was 850, although as in prior research [35, 113], there was considerable variability (95% CI [765.08. 934.27]). Many employees were members of multiple communities. In summary, the communities we studied involved varied participants and replicated many of the usage patterns that have been observed in other research on internet communities.

3.2 Social Roles

We have seen that prior work proposes different roles for community contributors, but without clear consensus about how roles are behaviorally defined [81, 159, 160, 166, 171]. Prior studies have also often employed inferential methods to distinguish leaders from members [102, 219]. However, our data has unique properties which allows us to take a different approach than inferential methods which have issues. In the Communities system, participants are officially designated to one of two roles: Owner or Member, each with different privileges. Members can view and post content with any tool, but may only edit their own content. They can also reference or link to others' content. Owners are considered leaders as they have members' rights but they can also edit any content, add/remove members, and configure tools. Owners are defined at community inception and do not change. These role definitions mean that our dataset has a gold standard for identifying social roles, removing the need for inferential methods and allowing direct measures to be made about role effects on specific behaviors of interest in relation to content management. Using a fixed definition of roles, we can examine role-specific owner and member management behaviors over time. While a fixed definition of roles may suggest limitations, prior empirical work shows peoples' online roles tend to be relatively static [55, 159]. Although our members lack certain privileges, in common with internet-based communities there is still considerable room for them to display important leader-like content management behaviors. For example, they can create links to manage forum, wiki or blog content.

To check for consistency between our system-designated owner and member and definitions of roles used in prior work, we compared typical role-specific behaviors identified in those prior studies, relating to networks, language use and identity behaviors. Consistent with that prior work, owners had larger communication networks [102], higher usage of social linguistic styles and adopted the community identity [102, 219], when compared with regular members.

3.3 Data Collection

Our criteria for including enterprise communities in our analysis were:

- **Active management**: Leaders had to sign up for Community Insights [136], a tool to help leaders enhance their community. A research goal is to make community design recommendations, so we wanted active leaders.

- **Active posting**: Updated in the last month. We wanted successful extant communities since our aim was to describe effective usage practices. We did not include communities that were inactive.

There were 2,010 communities that met our criteria of being active, generating a total of 428,476 posts. Recall that the Communities system was originally developed for the purpose of providing a platform to establish healthy enterprise communities that support employees and corporate processes. We were therefore able to collect log-files data on every user interaction, pages viewed, clicks on interactive widgets, from July 2007 to May 2014. This data was then linked to demographic Communities data and logged in a MySQL database. For each post, with participants' agreement we captured:

- **Community ID** (Where it was posted)
- **Author ID** (Anonymous Unique identifier)
- **Date** (Time stamp when post was made)
- **Tool** (Which tool the post was in: e.g., blog, wiki, etc)
- **Role** (Member vs Owner of community posted in)

3.3.1 Assessing Community Success

Many success metrics have been proposed for online communities. However these metrics are rarely validated and there is little agreement about which are most effective [30]. The most commonly proposed behavioral success metrics are: volume of members' posts [35, 67, 99, 146, 172], number of members [35, 67, 10], and quality of member relationships (e.g., measured as the extent of contact among members) [35, 67, 10]. Other common metrics include number of message threads [35], number of replies [35], threads with responses [35], and delay in response time [67]. Some researchers have developed algorithms combining multiple behavioral metrics to rate community content [91], community members [99], or the community itself [91].

One critique of these behavioral measures is that they are indirect. Other work therefore directly assesses participant perceptions, e.g. member satisfaction [10, 101], rather than inferring success from behaviors. It has been long known that successful online communities must meet member needs [113, 124, 208] and the relationship between behavioral measures and participant perceptions of their community's success in meeting its goals is explored in [49, 134]. That work uses member satisfaction as a measure of community success. One aim of the current thesis will be to re-examine how well these commonly proposed success metrics predict member satisfaction. Throughout all studies within this thesis, various types of quantitative success measure will be used along with a survey assessment of member satisfaction.

3.3.2 Survey Measures of Member Satisfaction

Workplace community members were surveyed as part of a larger research project [134]. This thesis involves a subset of the survey and communities from the larger study. Success was assessed using the most reliable survey question, the member satisfaction probe which asks community members "how well this community is meeting your needs", on a scale of 1=very poorly to 5=very well. We rely on this single question because it was highly correlated with other related questions, e.g. 'how successful is your community' as well as being predictive of other behavioral success measures [134].

A sample of actively managed communities was drawn from a pool of 666 communities whose leaders participated in an experimental deployment to help leaders enhance their community. These communities varied widely in terms of size, longevity, and purpose. The survey was sent to 20-26 members within each community. A stratified sampling method was used to balance the different types of community members. We next removed: communities with too few members (<20) and too few responses (<3) to yield a valid assessment of member satisfaction. We also removed 8 communities with incomplete data and 86 communities with <3000 words to ensure enough content to obtain accurate results from any content analyses. The word threshold was needed to reduce sampling error, i.e., the error across different lexical category frequencies when comparing small and large language samples from the same source [85]. Respondents represented a wide range of geographies, business divisions and roles.

The overall response rate was 19% for all participants surveyed, and an average of 5.9 members responded per community. We averaged member responses within each community, as a validity check showed good correlation coefficients (average ICC = 0.69) across respondents from the same community.

3.4 Classical Machine Learning

The analytical methods used through this thesis involve traditional statistical methods and classical machine learning modeling techniques. Conducting standard machine learning practices is set to help discern if meaningful relationships exist between various factors (i.e.

Features or independent variables) and targets of interest (i.e. Dependent variables).

Data splitting is conducted during each machine learning experiment producing separate train and test data sets. Each model is trained (or fit) on the train data set where the test set is used to evaluate a model's performance and generalizability. Within the training stage, k-fold cross-validation [111] is used to explore a more generalizable model. The parameter k for each procedure does vary depending on the amount of data available within the train set. Future sections will report such a parameter when used.

Various types of models are used throughout this thesis to explore which produces the best performance given the task. Typically models explored are those of Boosted Random Forests [44] and Support Vector Machines [42], in contrast to more traditional statistical models like Logistic and Linear Regression. The best performing model will be reported for relevant sections. Each model type has undergone a hyper parameter search to find the best performance. The most common method used throughout this work was through using the Grid Search Cross Validation method. The main library used for data splitting, cross validation, modeling algorithms, and parameter searching was the Python module Scikit-Learn [164].

3.5 Natural Language Processing

Some chapters will conduct a linguistic analysis on the content present within community posts. This is conducted through proven methods within the Natural Language Processing field. Primarily this work will be utilizing proven methods of feature extraction to find meaningful factors toward a given task. This incorporates finding features such as N-grams [41] or

Parts of Speech tagging [197].

Additional approaches will be utilizing lexical based approaches from existing libraries that have curated reliable signals of high order categories. The main library used is that of Linguistic Inquiry and Word Count (LIWC) [190]. This library is a text analysis tool that will extract which words are associated with psychologically-relevant categories. This approach is primarily a bag-of-words method in which words are measured out of context from the words surrounding them. In addition to this, two other lexical libraries are used, the Subjectivity Lexicon within the OpinionFinder tool set [212] and the NRC Emotion Lexicon [142]. Both of these are used in a similar fashion to the LIWC library.

3.6 Network Approaches

For the later chapters (7 and 8) graphical networks were built for each of the 2010 communities in our data using the NetworkX library within a Python 2.7 environment [87]. Each user in an online enterprise community was considered a node and edges were found through measuring a reciprocal relationship between nodes. Reciprocal relationships are defined as exchanges between two entities, which were operationalized as when a user responds to another user's post. We refer to this network as a reciprocal network due to the definition of an edge being a reciprocal action [24]. Reciprocity is already an indicator of online community success [173] and this type of representation allows for a more thorough examination into the group level aspects of reciprocity. Using replies to evaluate reciprocity meant that only community tools allowing for replies (Forums and Blogs) were included in this work. There can be multiple types of reciprocal relationships, one being a reply to an initial post and another type is a reply to a reply (nested thread structure). We were only interested in the pairwise interactions that come from post replies.

Utilizing these graphical networks, we were able to extract commonly used measures to explore the various community networks. Such measures were those of population metrics such as Degree (number of edges) and Nodes (number of nodes), while expanding to more complex metrics like Density, Bridges, Local Efficiency and K-Core [118]. More detail for each metric is provided within the methods section of Chapter 7. Additionally, intermediate level substructures within these networks were explored. Such substructures are commonly referred to as Graphlets (or motifs) [8]. A state of the art algorithm was used to discover graphlets within these network [8], and further detail is provided within Chapter 7.

Chapter 4

Posting and Linking Behaviors to Test Lifecycle Models

4.1 Introduction

Understanding the long-term practices of successful communities is critical to inform theory, the design of effective tools, develop success metrics and guidelines for effective online community building. In particular, a systematic understanding is needed for how long-term communities actively manage ever-growing amounts of content [99, 171]. In many cases, successful online communities have generated extensive shared resources and content. Effective content management is critical for retaining members and ultimate community success, as clear organization provides members with straightforward ways to access important community content [114, 127, 171, 192].

This chapter explores content management presenting an examination of one common technique for management, namely referencing via hyperlinks. We first qualitatively examine how links are used for content management and then quantitatively address two questions: (1) does content management via links increase over time as community content accumulates? (2) who takes responsibility for content referencing as communities evolve? It is well known that work is shared unequally in online communities [35, 159, 211], but little is known about how workloads shift among members in the long-term. Do the people who originally created the community remain responsible for content referencing or do newer members adopt responsibility as their level of participation increases [119]? Contrary to our expectations we find that referencing does not increase as content accumulates, which may be due to communities focus on recently created information. Furthermore, contradicting common lifecycle models of community success members do not assume responsibility for content management as the community evolves.

4.1.1 Content Management through Referencing

Creation involves generating new community content, and much is already known about creation practices [35, 106, 110, 113, 127]. This chapter defines content management as the active organization and annotation of pre-existing content created by the community or by others external to the community, in order to facilitate access to that content. With a few exceptions [116, 180], there has been little empirical research into long-term content management. Understanding management is crucial as successful communities increasingly accumulate content that needs to be referenced for new or returning users. Many complex acts of content management are hard to operationalize across large datasets, because they require manual content analysis that relies on detailed domain knowledge [90, 115]. The main goal in this chapter is to assess representative content management behaviors across a large dataset, focusing mainly on one specific quantifiable form of content management: hyperlink referencing. Hyperlinks are an efficient and pervasive general method to structure complex information [93], e.g. by referencing underlying content using a structured list or creating readable annotations [147]. Hyperlinks are used in many content management systems, such as wikis, as they promote straightforward access to complex content [59, 147]. Hyperlinks help manage content by creating navigational infrastructure [59], as well as supporting curation [89]. Because hyperlinks do not cover all aspects of content management, the organizational use of links is referred to as referencing and the discussion is limited in the results to address these specific referencing styles of content management. Although referencing is pervasive in online collaborative content management tools such as wikis, it is also used in many other tools. In this study we explore referencing via links across a range of social media tools, including wikis, forums and blogs, to determine whether it increases over time.

4.1.2 Changes in Roles

The second research question asks which community members take responsibility for online content management. Online community research has shown that a small active subset of users contribute the majority of work in online communities [35, 211], both in creating content and successfully coordinating the work of others [136, 219]. Models that describe long-term community lifecycles have suggested there are shifting responsibilities with leaders and members dominating at different points in the community's evolution [99, 171]. These lifecycle models stress the importance of early proactive leadership to seed interesting content, set policy, welcome newcomers, etc. [114, 136]. But these models also argue for the importance of apprenticeship so that as the community matures and accumulates content, some regular members gradually assume more responsibility [172, 208]. Such members begin as peripheral participants who simply read and lurk, but over time they take on increasing responsibility with a subset organizing and managing their community. Despite clear theoretical consensus of lifecycle models around apprenticeship and increasing member participation [99, 172], we are unaware of systematic quantitative analyses of how roles and responsibilities change for long-term content management.

This chapter therefore examines both how long-term communities actively manage content using referencing links as well as role changes in such management over time. To address these questions for enterprise communities, content management is first explored using referencing links at a post level, we then determine how accurately lifecycle theories predict role changes compared with actual practice. Using quantitative analyses, this chapter addresses these research questions in the context of mature enterprise communities that have access to a range of social media tools.

4.1.3 Contributions

We quantitatively characterize one important aspect of content management using reference links, where both long-term changes and role differences are examined. We contribute to communities existing literature by exploring the following questions: First, how does referencing using hyperlinks change over time and how does this relate to content creation? Second, as content accumulates, who takes responsibility for management: members or leaders? Do members assume more responsibility for referencing over time? Our exploratory findings are counterintuitive. First, active content referencing does not increase as content accumulates and second, contradicting lifecycle models, members never assume full responsibility for referencing. Content analysis suggests that recency bias is a possible reason for the absence of such referencing. We suggest new tools and community building practices that better support content management taking these findings into account.

4.1.4 Content Management and Hyperlinks

Various arguments have been made that collecting, organizing and actively maintaining content is critical to online communities [155, 172], with members being more likely to use communities that provide easily accessible information [135, 192, 208]. Well-organized content is also claimed to help retain members over the long-term [155, 172, 192] while disorganized content is argued to cause people to leave [71, 172, 192]. Tedjamulia et al. [192] argues that long-term participation depends on the community providing enough content, as well as the community's ability to leverage technology to provide access to that content.

Content management covers a range of different activities, including quoting already-existing content or summarizing prior useful content using FAQs [71, 90, 115]. In addition to these high-level management activities, there are also simple but pervasive methods such as referencing content through hyperlinks which are the focus of the current paper. Linking is commonly used for knowledge management across multiple tools including wikis, blogs, and forums [64, 200]. Hyperlinks were originally envisioned as a mechanism for both annotating individual documents and also indicating relationships between documents [149]. At the same time, links provide a straightforward way to navigate within and between document sets [28]. Much work has examined the uses and benefits of hyperlinks, showing that they serve to connect communities around similar content [64, 70, 200] and filter the abundance of content on the web [107, 108, 213].

Within wikis, hyperlinks are considered a fundamental aspect of content management as they connect topics and create context for those topics [7, 200, 201]. They also encourage cross-referencing, creating a navigable linked structure for networks of online resources, for example in educational contexts [64]. Within blogs and forums, hyperlinks are used as a resource for interlinking related ideas, typically associated with recommendation and summarization of said referenced content [28, 64, 108, 213]. While acknowledging the diversity of content management practices, in this chapter, hyperlinks are used as a measurable indicator of active content management behaviors. Additionally hyperlinks are profiled and analyzed based on different content sources to ensure measuring those used to actively manage content within and outside communities.

4.2 Methods

4.2.1 Models of Roles and Content Management

In this chapter, we test theoretical predictions about increasing member responsibility for content management [99, 171]. We assess this by observing whether members' target content management behaviors i.e. linking, increase with community age as hypothesized by lifecycle models. We also assess whether content management increases as the community ages.

4.2.2 Community Sampling and Data Collection

Out of the 2,010 communities within the communities sample, a total of 428,476 posts and 1,246,570 links exist. Recall that the Communities system was originally developed the purpose of providing a platform to establish healthy enterprise communities that support employees and corporate processes. We were therefore able to collect log-files data on every user interaction, pages viewed, clicks on interactive widgets, from July 2007 to May 2014. For each post, with participants' agreement we captured:

- **Community ID** (Where it was posted)
- **Author ID** (Anonymous Unique identifier)
- **Date** (Time stamp when post was made)
- **Tool** (Which tool the post was in: e.g., blog, wiki, etc)
- **Role** (Member vs Owner of community posted in)
- **Date of Community Creation** (To determine when in the community lifecycle the post was made)

Our focus was on links for content management, and for each link we 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** (Anonymous Unique identifier)
- **Author Role** (Member or owner)

4.2.3 Measures

4.2.3.1 Creation and Referencing

Creation was defined as new content added to a community. Content can be added in different ways using different tools. Tools available to the community were: forums, blogs, wikis, and bookmarks. Our measure of content created was the sum of the number of forum posts and replies, blog posts and replies, wiki edits, and bookmarks. Referencing is defined as the act of linking to already-created content and potentially annotating it for other community members through the use of a hyperlink in any of the social tools. From prior work, it is known that hyperlinks are used to reference external information sources that are relevant to community discussions and to organize content within the community [59, 135, 147, 194]. As noted earlier, referencing has been successfully used in prior work to assess content management, but it is not an exhaustive measure of content management, as it excludes behaviors such as quoting or summarizing prior content using methods such as FAQs. Nevertheless, it was chosen to measure content management via referencing, as it can be more reliably operationalized as an indicator of content management than those behaviors; hyperlinks are easily countable and extractable from posts. To validate links as a measure of content management, a qualitative analysis is reported showing that specific link types are reliably used for referencing.

Hyperlinks in the Communities data have different sources they are referencing. This work focuses on references that link to existing content within the community or another community. These hyperlinks were then labeled as Internal (hyperlinks referencing content within the community) or External (hyperlinks referencing content from another community but still within the enterprise intranet Communities app). Overall, 205,693 (16.5% of total hyperlinks) Internal references and 313,922 (25.1%) External references were found. Additional types of hyperlinks that exist were Enterprise intranet but outside of the Communities app (45.2%) and those referencing the Open Internet (12.8%). The remaining 0.4% of hyperlinks were unidentifiable. Below two analyses of link functions are presented. First, a machine learning analysis showing that links to the 4 different sources (Internal, External, Enterprise Intranet and Open Internet) involve distinctly different content. A second qualitative analysis examined the content management functions of these four different link sources, showing that only Internal and External links are actively invoked for content management.

4.2.3.2 Temporal Analysis and Lifecycle

The focus of this chapter is on whether and how community role behaviors change with respect to community lifecycle. Prior work has proposed different community lifecycle phases. However, these phases are extremely difficult to operationalize, e.g., how might we determine that a community has moved from inception to growth or from growth to maturity [99]? Communities may also develop at different rates, making it difficult to compare between them. Rather than proposing ad hoc behavioral indices for these phases, long-term data was collected over communities for a 36 month period in relation to their age. Not all communities analyzed were 36 months old but were included in the aggregated behaviors up to their age when the data was collected, for example a community that is only 15 months old at the end of our data collection would be included in the data for months 1-15 but not for months 16-36. The average community age was 24.2 months (95% CI [23.42, 24.98]).

We analyze time relative to the creation date of each community, for example, month 1 indicates the behaviors of all communities from their creation to age 1 month. This minimizes the possibility of outside events influencing aggregated behaviors across multiple communities. Outliers at each time step were filtered using a Median Absolute Deviation [123]. For each behavior, we first examine general trends by fitting a local polynomial regression to the time series of all communities. Polynomial methods are used because linear models offered poor fits and provide a stronger visualization of changes over time, although they don't allow for statistical comparisons. A nonparametric regression is preferred here since these time series analyses were found to be non-normal. 95% confidence intervals are plotted along the regression lines. To compare roles across time, for each target behavior we separate those that were conducted by owners and members for each month and fit another local poly regression for each role to visualize the differences. We then used a mixed model regression to evaluate statistical significance between roles.

Some of the behaviors we analyze are relatively infrequent, occurring a few times per month. However, it is important to note that all the communities analyzed are still active when the data was collected suggesting that even low levels of behavior are markers of long-term community survival.

4.3 Results

Before assessing the main research questions, an analysis of link usage to assess whether links reliably assess important aspects of content management is presented first. We begin with a machine learning analysis of different link sources, showing that sources are distinct, followed by a qualitative analysis indicating that only Internal and External links are being used for content management. Then both posting and referencing over time are explored. Finally, an evaluation of role changes by comparing leaders and members in their posting and referencing behaviors.

4.3.1 Referencing Link Sources For Content Management

4.3.1.1 Link Sources are Distinguishable by Machine Learning Classification

Recall that there are four sources of links based on whether they reference content that is Internal to the community, External, outside the Communities app but within the Enterprise Intranet, or from the Open Internet. A machine learning experiment was conducted on these 4 link sources to determine if links differ depending on the source content they reference, by evaluating the words used around each link. We extracted the content from the sentence before the posted link, as well as text that contained an embedded link. We then extracted a series of N-gram (1-2-3 gram) features from the text. We fitted a Support Vector Machine [145] to this data and used a 72-18-10 data split for training, validation, and test sets for modeling training. The validation set was generated using a 5-fold Stratified cross validation procedure. Stratified cross validation keeps the distribution of classes equal through the data splits, helping address

Link Source Classification			
	Precision	**Recall**	**F1-Score**
Internal	0.63	0.32	0.42
External	0.66	0.59	0.63
Enterprise Intranet	0.69	0.85	0.76
Open Internet	0.75	0.61	0.67

Table 4.1: Results for Classification of Link Types using a SVM model on N-gram feature set, showing that link sources reliably index different types of content. Enterprise Intranet has the best performance based on F1-Score, with Open Internet having the second best performance. Internal and External had the lowest performance but were still better than random.

unrepresentative data splits [111, 214]. To evaluate the model, we used Precision (True Positives divided by True Positives plus False Positives), Recall (True Positives divided by True Positives plus False Negatives), and F1-Score (a combination of precision and recall) on each link class [170]. Link class performance for the model is ranked based on F1-Score.

Table 4.1 shows the results. The model performed well with an overall F1-Score of 0.68, indicating a reliable difference in the type of words used in the sentence before the link, showing that referencing behaviors are reliably different based on the source of the referenced content. The model performed best in identifying link types in the Enterprise Intranet class (highest F1-Score), and Open Internet links were also classified well. Both Internal and External were fairly accurate in their classifications as both are still above baseline (F1-Score >0.25). Furthermore, confusion matrix analysis indicated that the majority of misclassifications involved Internal and External classes being confused for each other suggesting their functions overlapped. We return to this overlap in our qualitative analysis.

The machine learning analysis suggests that the 4 link source types are distinct, therefore we next went on to qualitatively analyze the functions of each link source to identify whether and how they were used for content management. Example posts of each link source category were explored, namely Internal, External, Enterprise Intranet and Open Internet links. Example posts show how different sources of referencing links managed content. Explicit URLs are indicated by "<HYPERLINK>" and embedded references by hyperlink tags bookending the text "<HYPERLINK>" "</HYPERLINK>". Personally identifiable information has been anonymized.

4.3.1.2 Internal Links To Reintroduce Existing Content.

Internal references were commonly used by experienced users to draw attention to prior material within the community that relates to a new post. Referencing was done for the benefit of newer users who seemed to be unaware of that existing content [113].

> Hi User 1, that's a very good idea. We actually have a ”XX” forum where everyone can access and trade their stuff. You should check it out too. Here's the link <HYPERLINK>

This post links to existing community content that the newcomer User 1 seems unaware of. The explanatory text labels the prior content via a short description, and the reference provides direct access to that content. Using the link serves two important content management functions. It avoids duplicating prior content and so reduces the accumulation of content within the community. Use of the link also explicitly signals relations between content in different parts of the community, in this case between the current post and the "XX" forum resource.

Other Internal references were similarly used to announce new content to the community, where that content is being posted in a pre-existing community resource.

> User 2 has announced that User 3 is the new crucial position... Read the announcement on the <HYPERLINK>YY Community Wiki</HYPERLINK>here: <HYPERLINK>

This post again serves multiple referencing functions. As in the prior example, the embedded reference to the ‘YY Community Wiki’, reminds readers about the existence of that local community wiki resource. The link also provides ready access to the content of that announcement. Again using the link reduces content accumulation, as content is not duplicated within the current post, but can be accessed from the Wiki by those interested in the announcement details. Other posts used internal references to promote group action, in this case a forum for community brainstorming.

> Foster the collaboration. If you have ideas on Tools saving, pls. put them here -><HYPERLINK>. If you have any further questions please let me know.

This post is proactive in encouraging new community posts as opposed to organizing existing information. Nevertheless, it uses the same content management approach as the prior examples; it directs the reader to existing internal community resources where they should post new content without redundantly duplicating a detailed prior description of those resources. Other referencing behaviors promoted content outside the community. The functions of these different types of outside links are now characterized.

4.3.1.3 Functions of External Links

Many External links overlap with functions already identified for Internal links. External links identify directly relevant resources in other enterprise communities that help the local community better organize their own information.

> Some more info i have [sic] digged up is a wiki with guidance: It is from the <HYPERLINK>Community Builders Wiki</HYPERLINK>: Community Leader tips from social science research. That page describes what you should to do in advance and after creating a community, to ensure it becomes successful. A lot of questions I have seen (like small or big) are answered by that material indirectly.

Here the poster promotes external information they believe is relevant to their own community. They briefly summarize the content of the reference, justifying why it is relevant. This use of referencing means that content is not duplicated or proliferated within the community. Other posts involving external references aim to organize material relevant to the local community, with less focus on explaining the referenced external content.

> Welcome to the BC Tools Focal Point Topic in the BC Focal Point Forum Some tools links:
> <HYPERLINK>
> <HYPERLINK>
> <HYPERLINK>

This post sets up a simple structure to reference external content while directly imposing an active organization on that material.

These examples of internal and external links indicate how referencing is used to achieve important aspects of content management. Each case involves a combination of linking to outside community content, while simultaneously summarizing or annotating the content. Use of the reference link avoids content duplication within the community reducing accumulation of content over the long-term. Most importantly links organize or impose structure on that internal or external content.

4.3.1.4 Enterprise Intranet and Open Internet Links.

Other forms of linking outside the community have different functions that less directly involve content management. Enterprise Intranet links reference the corporate intranet and can identify important resources such as programs or events. These resources can assist with tasks users are trying to execute, but tend not to actively relate to community content:

> Hi User 4, I suggest two panels being opened when viewing a document in this tool
> 1) All open docs for tool goes like this (example): <HYPERLINK>
> 2) Doc Journal which goes like this (example): <HYPERLINK>
> Obviously you'd need to replace the value, second value or number with the one at hand

A final class of links reference the Open Internet. Open Internet links do not point to existing organizational resources. Instead they usually identify additional external information, for example identifying the person introduced in the post with a link to their personal webpage. These Open Internet links usually do not reference organized existing information or provide follow-up actions for readers.

Internal and External links reference community content, providing active resources for organizing and sharing target community information. In contrast Intranet and Open Internet do not seem to actively address community organization the way that internal and external links do. Therefore, further quantitative analyses are limited to Internal and External links, i.e. posts that link within a community or to another related enterprise community.

4.3.2 Overall Lifecycle Trends in Creation and Referencing: Total Content Increases but Linking Decreases.

The qualitative analysis characterized referencing for Internal and External community links showing active content management. Now we examine these specific management behaviors over time using quantitative methods. In each of the following analyses, a time series trend is computed by fitting a local polynomial regression for each month. Analyses of absolute rates of posting and linking are reported, but similar analyses that normalize these rates by community show similar results. Different analyses were conducted exploring whether there are differences in creation and referencing for different types of communities and whether there are differences between different tools (blogs, wikis, forums, bookmarks). There were no noteworthy differences between community types, so in the interests of space we do not report these results. There were differences between tools. As anticipated, wikis were used more often for referencing, which we discuss later.

Following claims in prior work [99, 171, 208] we examine whether active referencing activities are a response to accumulating content.

4.3.2.1 Content Accumulates Over Time

The left plot in fig. 4.1 shows a local polynomial regression on the cumulative content for all communities by month. Over time, communities acquire ever-larger amounts of content. Overall, content shows a steady visual linear accumulation with a slight increase as communities approach 20 months, dropping slightly in months 32 and 33. This content accumulation would seem to demand greater organization; communities on average have about 3.5 times the amount of content at month 36, compared with month 10. As communities mature, we should therefore expect them to engage more actively referencing activities to manage this accumulated content.

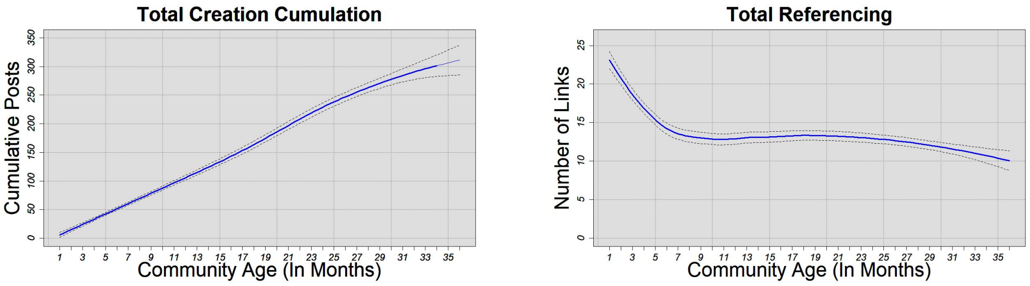


Figure 4.1: Communities accumulate ever-larger amounts of content over time. Local Poly Fitted Regression of cumulative posts/month (95% confidence intervals shown as upper and lower dotted lines). Referencing starts high but decreases until around 15 months, then remains steady. Figure shows local poly fit of referencing within communities over community age: (95% confidence intervals shown as upper and lower dotted lines).

We evaluated the relationship between total content and active organizational linking using a repeated measures regression with communities as subject. Somewhat unexpectedly, and contradicting lifecycle models, this analysis showed a very weak relationship between cumulative content and linking as the beta weight is extremely low, ($\beta = 0.008$, SE = 0.001). Even though this relationship is statistically significant ($p < 0.001$), the small coefficient indicates that cumulated content has little predictive power in explaining linking behavior. Next, we went on to examine linking behavior directly to explore why this is the case.

4.3.2.2 Referencing Rates Drop Over Time

To assess referencing visually, a local poly regression was plotted of the number of links/month over the age of the community in the right plot of figure 4.1. Surprisingly, linking rates decreased over time. Despite each community having many more posts to organize, linking rates dropped over time. The reason for this drop becomes clear when we examine trends in referencing. Most referencing occurs within the first month. We see a large negative rate of change in the first few months but this approaches 0 around the end of the first year. Consistent with prior qualitative work on communities [136], figure 4.1 suggests an intense referencing phase in the initial months. This start-up phase involves active referencing as participants seek to proactively organize early content to allow other members to effectively navigate that content.

After the startup phase, we expected the growing weight of accumulated content would demand active organization, leading to an increase in referencing over time. Instead it was found that referencing decreased after the first month, levelling off after 10 months, despite ever-accumulating amounts of content late in the community's lifespan. These results over 36 months therefore challenge one aspect of those models which assume that content management increases over the lifetime of communities. A possible explanation for this might be that communities tend to focus on content that is more recent and ignore older accumulated content thereby removing the need to organize that overlooked older content.

4.3.2.3 Referencing Has a Recency Bias

To test that possibility, posts that contained at least one link were explored to see if these posts are more focused on present rather than past content. Examining the content of a post that contains a link, we used the tense categories from the tool Linguistic Inquiry and Word Count (LIWC) [190] as a proxy for temporal focus within the post and examined tense use over time. Fig. 4.2 shows the average Past, Present, and Future language use over community age reported as a percentage of words in a post. Posts were far more likely to focus

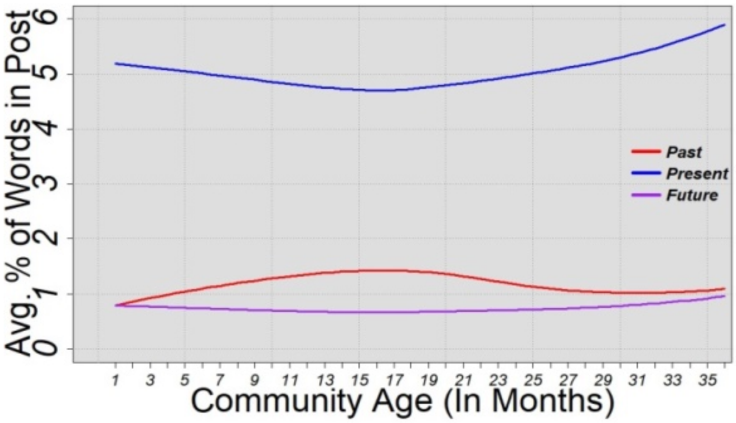


Figure 4.2: Results of analysis on link temporal focus. Left shows the results of examining language tense in posts containing at least one link. Posts contain more Present than Past and Future tense words.

on Present than both Past and Future tenses. More significantly we saw no increases in Past tense usage. If participants were actively referencing past archival content we might expect Past tense references to increase throughout the community lifespan as past content accumulates. Overall this is consistent with a recency based content focus.

While this tense analysis shows that link posts are more focused on the present, it doesn't directly examine the age of older content that is being referenced. To assess this, instances of links that were repeated were analyzed. The median time difference between repetitions of the same link was measured. In the last 6 months of the community lifecycle, the largest median time difference between repetitions was 27 days, a 2% time window within a community lifecycle of 36 months, indicating that communities are focused on recent content. Overall, tense and link repetition analyses argue that referencing content has a recency bias.

4.3.2.4 Role Differences: Members are increasingly responsible for content creation but not for referencing

Our second research objective was to examine Lifecycle models which claim that community responsibilities shift over time. Those models argue that, compared with community owners, members take an increasingly active role over time both creating and managing content. We therefore expected members to engage in higher rates of linking as the community matured. To establish baselines we first analyze content creation rates in members versus owners. Recall too that members and owners are formally defined in the communities that we are analyzing.

4.3.2.5 Members Dominate and Take Increasing Responsibility for Content Creation over Time

We compared the creation behaviors of members versus owners (see figure 4.3). The figure indicates that owners show higher content creation rates only at the community's inception. As expected, the average number of member posts increases over time, whereas owner posting rates remain steady. This is consistent with theoretical lifecycle accounts arguing that members take increasing responsibility for creating new content as the community matures. After very early stages, members drive content production, increasing production faster than owners, but both members and owners decrease production rates as the community matures after 30 months.

To test the significance of this difference, we used a zero-inflated negative binomial mixed-model regression. Poisson and Binomial regressions are typically used for fitting count data and due to the unequal values of the mean and variance for each time point in the data,

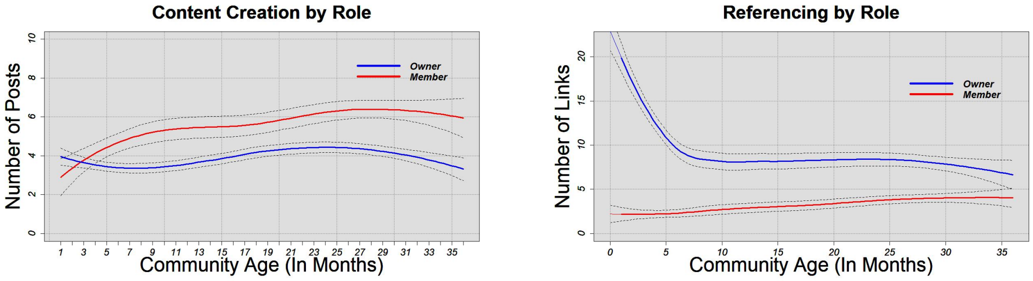


Figure 4.3: (Left) Local poly fit for role creation over time. Members dominate content creation after first few weeks. (Right) Local Poly fit for role referencing over time. Owners dominate referencing throughout, but quickly decrease after a few months and then have a gradual decline. Members referencing remains relatively similar throughout.

binomial distributions produced a better fit (p <0.0001) and were not over dispersed. A zero-inflated model was used, as roughly 55% of the data at each timestamp are zeros. This type of model accounts for high proportions of zeros by considering them in a separate process. A mixed model was used to account for individual community variance. The libraries lme4 and glmmADMB in R were used for modeling [1, 22, 25], the zero-generation process was handled by the glmmADMB library which treated the zero-count outcome as a mixture of structural and sampling zeros. The model applies this as follows: given the outcome of the model (Y) and the probability of the outcome being equal to zero being p, a proportion of Y of size p comes from extra zeros and a proportion of Y of size 1-p comes from the Binomial distribution. The density of this model can therefore be characterized following Bhning et al. [37] as:

$$f(y;p,\mu) = p * BN(y,0) + (1-p) * BN(y,\mu) \tag{1}$$

Equation 1 shows the model is a mixture of two classes, the first class having a fixed value of 0 and the second being the Binomial distribution (indicated by BN in Function 1) [1, 37]. The p for this model would be 0.55, given the sample. Such a model has a high computational

ZF Negative Binomial Mixed Model Regression Creation Results			
	Estimate	**Std. Error**	**P-value**
Role	0.578	0.107	6.4e-8***
Time	0.044	0.011	8.2e-5***
Role*Time	-0.015	0.006	0.012*

Table 4.2: Results for Mixed Model for Community Roles For Creation

ZF Negative Binomial Mixed Model Regression Referencing Results			
	Estimate	**Std. Error**	**P-value**
Role	1.751	0.121	<2e-16***
Time	-0.026	0.013	0.047*
Role*Time	-0.044	0.007	5e-10***

Table 4.3: Results for Mixed Model for Community Roles For Referencing

demand, the model was restricted to run on a random sample of 200 communities. The results of the model are in shown in table 4.2.

The model shows differences in both Role and Time, with an interaction between the two. This is consistent with the left plot in figure 4.3, showing overall differences between roles with members creating more content and time as overall differences in content created increased by month. The role by time interaction follows from members overtaking owner's initial posting rates after month 2.

4.3.2.6 Members Don't Increase Referencing Over Time

Next referencing behaviors comparing different roles were analyzed. Following lifecycle models, we anticipate that members' referencing would increase as they assume greater responsibility for content management compared with owners. However, this was not confirmed. In contrast to creation, where members increasingly dominate, the right plot in figure 4.3 shows owners show greater referencing and this is maintained throughout. Using the same modeling procedure as before, table 4.3 again shows main effects of role and time and an interaction between role and time. Consulting the right plot in figure 4.3, it can be seen that differences between Roles arose because owners created more links than members. Effects of time arise because overall linking decreased over time, again contradicting lifecycle models. The role by time interaction results from the decrease in owner linking after the first few months, as owners shift from intensive early link creation, whereas member linking is relatively stable throughout the community lifespan.

4.3.2.7 Analysis of Wiki Usage

To confirm the referencing results, a second exploratory tool-centric analysis was conducted. Recall that the collected data was from several social media tools. As noted earlier in this chapter, there are multiple methods to support content management, with wikis being a tool that is commonly used for this purpose [135, 200]. To check and extend the link analysis, we therefore assessed whether wiki tool usage was consistent with the link referencing behavior we had observed across all tools. It was found that the rate of wiki creation remained constant across the community lifespan and did not increase over time. There were no differences in the mean number of wikis created when comparing between the first and second 18 months of the communities' lifespans, using a Kolgormorov-Smirnov test (D = 0.3129, p = 0.2284). Turning to role behaviors for wikis, it was found that owners on average created more wikis than members, with owners creating 10.8 wikis and members creating 0.3 wikis. With respect to time, again it was found that owners consistently created more wikis than members throughout the community lifespan (D = 0.8649, $p < 0.0001$). Overall, then, this analysis of wikis confirms our link data. Similar to the results for overall linking, there is little increase in use of wikis overall; owners dominate wiki referencing with no significant changes in member wiki creation over time.

4.3.2.8 Limitations

There are limitations to this analysis. As we noted, we restricted quantitative analyses to one operationalizable aspect of content management using hyperlinks. Although we present other consistent exploratory data from wiki usage, it could be that using these link measures leads us to underestimate other content management behaviors and tool usage. Furthermore, communities may turn to external tools to manage content that we were unable to measure. We also rely on simple binary distinctions between owners and members and we are aware that theoretical work proposes more subtle differences in community roles [99, 171, 208, 219]. However, these decisions were motivated by the need to gather reliable quantitative data to make comparisons, acknowledging the difficulty of accurately operationalizing prior qualitative definitions of complex community behavior. Furthermore, our analysis is limited to the first 36 months of interactions, and it could be that referencing only becomes critical later in a community's lifespan, although lifecycle models would have to be modified to incorporate this claim. Finally, this work explores referencing in enterprise settings. Although past work shows strong overlaps in community behaviors across contexts, future work should explore whether our results generalize to internet communities.

4.4 Discussion

While this study is exploratory, the findings are nevertheless counterintuitive. First, we expected to find increases in referencing over time predicted by influential community lifecycle models [99, 171, 208]. One possible explanation is that after an early burst when a community is initiated, referencing ignores older 'stale' content and instead focuses on recent active content rather than trying to organize the entire set of community content. Data from tense usage and link repetitions are consistent with this recency bias. An alternative explanation could be that older references are handled through other content management methods, e.g. external storage systems (version control, file storage or synchronization services like Google Drive). However it seems unlikely that communities would link to external Wikis when they have the more straightforward alternative of using their own community wiki for content management. Alternate content management methods were tested by exploring community wikis, and the exploratory results confirmed the referencing analysis. Content management using external methods (e.g. versioned repositories, external wikis), could of course have occurred and be the cause for this drop in referencing. Future work is needed to examine more specific open internet referencing using external management tools.

A second research question concerned division of labor between members and owner roles and how this changes over time. Consistent with lifecycle models, members were increasingly responsible for content creation over time. But contrary to those models, referencing was managed by owners throughout the life of the community. This discrepancy in management is notable given that there were almost 100 times as many community members as owners, and that content has increased many times over during community's lifetime.

Another important question is the extent to which the results generalize to other online community contexts. Overall results which seem to indicate that our work is representative of other well-studied online communities. First the data for content referencing replicate known power law effects with the majority of referencing effort being contributed by relatively few [159, 211]. Second the data includes many commonly occurring community types observed elsewhere, such as communities of practice, teams and small work groups [39]. Furthermore, although the data used fixed as opposed to flexible participant roles, an analysis of owner behaviors in our sample revealed strong overlaps with prior reported norms for leader behaviors [102]. Finally the analysis of hyperlinks matched use cases detailed in prior work [59, 200]. Together these observations suggest clear consistencies with other online communities' research, giving confidence that results will generalize elsewhere. The results on reduced referencing and recency bias are important for theory and practice. These phenomena may not have been observed before because there have been relatively few long-term analyses of long-term community behaviors and little focus on referencing. Theories need to incorporate our results, explaining why communities apparently fail to organize and refer to accumulated prior content. It could be that communities are inherently biased to focus on present discussions rather than extensive past content. This bias may reflect short-term participation with members joining the community but leaving after relatively short periods of active contribution. Such 'churn' would make it hard for communities to build a shared long-term perspective on their content. It may also be that members find it hard to use tools such as wikis that promote referencing and we return to this point below.

The other results extend prior literature on distribution of labor within communities. Referencing follows the well-known power law distribution [159, 211] where only a select few users (owners) conduct the majority of work. However, this was not the case for content creation as the much larger group of members enacted the majority of work. It may be that referencing follows a pattern similar to Panciera et al. [159], where owners' initial commitment to the community leads them to be relatively more active than members throughout the community lifespan. However it is also clear that owners' levels of referencing drop as the community ages, which may reflect burnout, or an unwillingness to organize content created by others. Again these are interesting questions for future research.

The results have implications for theory as well. First, they suggest a need to refine lifecycle models to include a greater emphasis on actual community practices, in particular to incorporate these findings about decreased referencing over time. Second, owners were mainly responsible for content referencing over the life of the community; this suggests the need to educate and encourage members not only to create content but also to reference content. Third, while owners' referencing is consistent with standard power law accounts [211], this wasn't the case for creation where members posted more content than owners. This suggests a need to refine power law accounts to include the specific tasks involved in the community. Overall, the findings on temporal characteristics generate new questions about lifecycle theories, suggesting the design of new tools and metrics to assess community success.

Finally, these results inform the design of new tools and metrics for community building. In particular, organizational tools such as wikis and bookmarks might be designed to encourage more active participation by members. Prior work shows that content management activities play an important usability role for community members [135] and that content management tools, i.e. wikis and bookmarks, are currently challenging for members to deploy [155]. Automatic text processing methods could also assist in referencing, for example by summarizing existing content. It may also turn out that community's recency bias means that content management tools only need to focus on newer information. An alternative design approach may be to modify community members' recency bias by designing new interfaces that draw their attention to interesting older content. One solution might be to model the approach taken by Facebook's "On this Day" [4] or Google Photos "Rediscover this day" [3], which both focus on re-presenting older content relevant to recent activity or content that previously received extensive active feedback. For example, communities might resurface older posts based on their direct relevance to recent posts, or more simply because the re-presented posts promoted highly involved discussion in the past. Of course such representations would need to be carefully designed so that users are aware of the motivations for resurfacing older posts. Another design approach might encourage community leaders to flag interesting content for later resurfacing. More automated solutions could involve detecting overlap between a new contribution and a successful prior contribution, leading the application to resurface the prior solution. This is similar to work done on automatic answer detection in question oriented forums [51].

Chapter 5

Role Dynamics: Stability versus Change Over Time?

5.1 Introduction

5.1.1 Contrasting theories of community roles: evolving apprenticeship versus static roles

As outlined in Chapters 1 and 2, online community theoretical models [99, 171] predict changes in online community roles, arguing that such role shifts are needed to meet the additional tasks confronting the community as it evolves. These ‘apprenticeship’ models claim that a subset of individuals who begin as regular community members take on increasingly greater responsibilities, changing their behaviors and starting to act like leaders as the community evolves. For example, Preece and Shneiderman [171] propose the ‘reader to leader’ model in which community participants initially engage in ‘lurker’ behaviors, simply reading content that others have produced. As their exposure to the community increases they engage in ‘peripheral participation’, beginning by contributing simple edits or clarifications to work created by others. As time progresses, some of these editors adopt greater responsibility, and end up leading activities within the community, for example by seeding new discussions, setting community policy or moderating others’ conversations.

A key element of this apprenticeship model is that some participants take advantage of opportunities in the community to shift roles from regular members to responsible leaders. In contrast, Panciera et al. [159] studied Wikipedia finding contradictory evidence; they saw no major role shifts as the community evolves. Instead, certain contributors act as power users engaging in leader-like behaviors, but counter to apprenticeship accounts these contributors do not alter their work styles over time. On entering the community, these power users immediately and consistently engage in leader-like behaviors. We call this the static role model.

This chapter contrasts these conflicting theoretical accounts using quantitative modeling to evaluate whether role change occurs in online enterprise communities. We first develop reliable machine learning models to distinguish community leaders from members, and then test contrasting theoretical predictions about how these models will perform over the community lifespan. If, following the apprenticeship account, some members increasingly engage in leader-style behaviors, over time we would expect models to: (a) make more errors as some members start to engage in leader-like behaviors; (b) predict a greater likelihood that a given post came from a leader. Results show that these predictions aren’t supported; over the community lifetime we neither find an increase in categorization errors nor an increased prediction that posts came from leaders. Therefore, this work supports the static model of Panciera et al.

[159], that roles do not shift over time.

5.1.2 New Models Distinguishing Social Roles

To test these predictions, we first need to develop accurate models that distinguish leader versus member roles. Recent work has built reliable statistical models mapping potential roles to an individual's actions and network relationships within a community [96, 102, 219]. This statistical work has often relied on measures like number of posts, number of replies, tenure in community, and social networking measures [96, 102]. Other work utilizes quantitative content analysis techniques assessing high level linguistic features to understand role contributions [102, 219]. Such features include emotional language [96] or content uniqueness [102]. A different approach aims to understand broader leadership style themes and patterns of language use [219]. Overall these different statistical approaches show it is possible to develop reliable models that distinguish online leaders from members. Leaders are found to have significantly more posts, credibility, network centrality, and use of language expressing affect, assertiveness and linguistic diversity [96, 102].

However this prior work has generally constructed aggregate models that combine every post a participant makes to the community. Aggregating data in this way makes it hard to model changes in individual behavior over time. We therefore develop two sets of linguistic models, the first replicating prior work based on an aggregation of all content produced by an individual and the second using individual participant posts. The first model establishes that we can reliably distinguish role differences in a context not previously explored, while the second model allows us to test the dynamic versus static role models we described earlier.

We also extend the set of linguistic features used in these models. Prior work focuses on generalizable characteristics of leaders across many different types of online communities [96, 102, 219]. We extend that work by first replicating known results that leaders are distinguishable from members using lexical (LIWC [190]) measures, but then expand on this using a linguistic modeling approach involving more context sensitive (N-gram [41]) features to focus on a post level instead of aggregate content.

5.1.3 Contributions

This work advances social role theory in online communities. We test contrasting model predictions by examining role shifts during the online community life-cycle. We train accurate categorization and probabilistic models using formal gold standard data, then use these model to test these contrasting accounts. If the apprenticeship account is correct, we would expect performance of the role categorization model to decrease with time as more members begin to behave like leaders. Likewise as the community matures, we would expect the probabilistic role model to show increased likelihood that individual posts are generated by leaders. However, neither prediction is supported indicating that roles are static. This work also develops new models of community roles that extend the set of linguistic features used, as well as building post-level rather than aggregate models. Overall the work provides new quantitative evaluations of descriptive community theory, and extends the static role account from wikis into the domain of enterprise systems.

5.2 Online Activity over Time

Recapping the literature discussed in Chapter 2, much prior work discusses how online communities change over time, and different long-term models have been proposed. Wenger et al. [208] describe a sequence of stages: potential, coalescing, maturing, stewardship, and transformation. Communities typically start as loose social networks with the potential for becoming connected. As connections form they coalesce into a community. This matures as members adopt active stewardship to manage increased content and emerging practices. Finally, a community transforms as it becomes irrelevant or members go on to other activities. Iriberri and Leroy [99] review multiple prior models of community development proposing a similar 5-stage model to [208]. They also argue for distinct stages: inception, creation, growth, maturity, and death. Kraut and Resnick [113] also recognize that communities have different needs at different times. They do not explicitly outline a long-term stage model but like [99, 208] characterize problems confronting communities at different stages of their lifespan. Problems include: startup, attracting and socializing newcomers, encouraging commitment, encouraging contribution, and regulating behavior.

A common theme in lifecycle models is that communities have an increasing demand for work. Within these models, one goal for leaders is to mentor future community leaders, and one characterization of this process is the apprenticeship model. This model outlines how community leaders encourage ever growing responsibilities in committed members (those who have a vested interest in the community and take on more administrative work than others). Leaders help by guiding apprentices about how to conduct the needed work.

Other research has examined long-term community behaviors that are unrelated to roles. One study explored a specific quantified aspect of community behavior by looking at linguistic changes over time. Danescu-Nisculescu-Mizil et al. [55] researched how individuals follow community linguistic norms in two large beer review communities finding that users would first align to the language of the community, but would then be resistant to follow how linguistic norms change and would be considered stuck in the past. Rowe [181] further examined user-lifecycle behaviors through linguistic methods as well as social network analysis, and found that user churn prediction could be predicted given linguistic deviation from community norms.

The current chapter takes a linguistic machine learning approach to first develop reliable models of social roles in online enterprise communities. We then use these models to examine role change over time to evaluate different life cycle models.

5.3 Methods

5.3.1 Data

Recall that one critical feature of the Communities software is that it provides objective definitions of roles which are formally allocated within the system, and these do not have to be inferred or handcoded. The two roles of interest here, namely owners and members, have different permissions in the software, with owners being able to edit the posts of others. Furthermore prior work [50] shows that behaviors in each the community roles of leader or member are consistent with the research literature: owners tend to have larger communication networks [96, 102], higher usage of social linguistic styles and adopted the community identity [96, 102, 219] when compared to regular members.

The Communities software also provides multiple social media tools, including Forums, Wikis, and Blogs. In Forums and Blogs, commenting features facilitate questions and answers, interpersonal interactions, and commentaries. Our focus in this work is on social media tools that promote more discourse focused content. We therefore only examine Forums and Blog posts in this chapter.

5.3.2 Modeling Procedure

5.3.2.1 Model granularity

We developed two models based on the sources of the text gathered. Both models attempt to accomplish the same task, namely: given text produced by a community participant, determine whether this text was produced by an owner or a member. However the models had one major difference:

Aggregate Models: Aggregate models were based the aggregation of the entire content each individual contributed to the community. We collected this content by extracting all forum and blog posts by an individual and measured psychologically meaningful lexical based features indicating high-level use cases. More detail on such a feature set is described below.

Post-level Models: The second model characterizes participant behavior at the level of individual posts. We collect and extract the same lexical features as the aggregate model. However given to the limitations of such approaches [190], we also develop other models that employ N-grams [41].

The aggregate model served to replicate prior research and to serve as a performance baseline. However the post-level model was used for our main analysis as critically this allows us to look at an individual's contributions over time. For each model, the dataset was split into 90-10 train-test splits and a 5-fold cross validation was conducted within the training set to ensure more generalizability. Additionally to account for the large difference in members and owners, class weights were added within the modeling parameters to allow for class balancing. This is a common procedure shown to remove modeling bias in predicting one class over another [43]. For each model, we conducted a grid parameter search [188] on Logistic Regression models to find the best Cost, Tolerance, and optimization for the feature space. We used logistic regression, as it performs well for textual features and generates a probabilistic output, which will be used in our later analyses.

5.3.2.2 Model Features

We explored two types of language based features, first a high-level categorical based approach using the Linguistic Inquiry and Word Count (LIWC) tool [190] and then a more context specific approach by creating various N-gram [41] related features to the text samples. We explored from Unigrams up to Trigrams and a combination of all N-gram features. We limited our exploration up to Trigrams as the frequency of unique higher level N-grams is extremely low [103], leading the model to overfit to more individual cases. We introduce the N-gram feature set in the post-level model as this is more robust than LIWC for smaller text samples [190], and we report results of the performance with LIWC features in comparison.

5.4 Results

5.4.1 Roles are distinguishable

Our first goal was to determine if roles can be distinguished based on participant posts, in other words could we develop models that accurately categorize leaders versus members based on their aggregate posts? Using the LIWC measures as features, we modeled author's roles and found a good level of performance with an overall F1-Score of 0.79. This shows that our systematically defined roles are distinguishable and replicates prior work [96, 102]. These results are further used as a baseline in comparison to the post-level model's performance.

Next we began modeling at the post-level where we explored 5 different linguistic models that aim to distinguish roles; the first model used LIWC to derive its lexical features and we subsequently developed four different N-gram models. We explored N-grams as posts are often short, and LIWC is less reliable for small numbers of input words [190]. An N-gram feature set in contrast can handle varying amounts of content.

As the models are intended to examine changes in individuals' behaviors over time, a special type of data selection procedure was implemented. Typically within machine learning experiments, the train and test set split is conducted randomly. In other words, data can be selected for either the train or test set, regardless of any metadata concerning instances in the main data set. However we were concerned that communities could show shifts in language behavior over time which could compromise our training and test approach. We did not want for example to be assessing models trained on mature community behaviors, against test data derived from early stage community behaviors. We therefore incorporated time in the train-test splitting, to ensure that the training and test sets were drawing on a consistent time period for each model. As the model was trained on a stratified amount of data from each time point, this controls for possible shifts in language in a community lifecycle. The procedure was as follows: Categorize each post by the month in which it was posted, for example if a post was made in the first month after a community's creation, that post is considered to be of age 1. Next iterate through each possible post age (ages ranged from 1 month to 80 months) and conduct a randomized 90-10 train-test split within that age. This provides a time-stratified set of data within the training and test sets over time, allowing us to test the performance of the model each time point.

Using this procedure, we built 5 models: LIWC, Unigram, Bigram, Trigram and a combination (Uni + Bi + Tri-grams). The highest performing model was the post-level model trained on a combination of Uni, Bi, and Tri gram features. Table 5.1 shows the results of our experiments. The N-gram combination model outperformed all other sets for AUC, F1-Score, LogLoss metrics, and performance was generally good. As expected all N-gram models outperformed LIWC.

Given that we have demonstrated a reliable model based on a combination of N-gram features, we now explored how this best model performed over time to test theoretical predictions of the different role theories.

Figure 5.1 shows how the best post-level model (using all N-gram types) performs over time. As the model was trained on a proportionally equal amount of data from each time point, this controls for possible temporal language shifts. Following apprenticeship theories [99, 171], we expected the model to decrease in accuracy as the community ages. According

Post Level Performance on Modeling Roles			
Feature Set	**AUC**	**F1-Score**	**LogLoss**
Unigram	0.69	0.725	9.568
Bigram	0.69	0.747	8.472
Trigram	0.66	0.721	9.337
N-Gram (Uni + Bi + Tri)	0.71	0.758	8.221
LIWC	0.63	0.638	12.483

Table 5.1: Modeling results for different feature sets. There is a performance drop in comparison to aggregated content, but the combination of Uni, Bi, and Tri grams produces a reliable enough model in distinguishing roles at the post level. LIWC, as expected, is showing the weakest performance.

to these theories, more members will begin to act like leaders over time, thus reducing the model's ability to correctly distinguish between members and leaders. However, fig. 5.1 shows no downward trend in modeling performance and in fact shows a slight upward slope. This contradicts apprenticeship theory, instead social roles appear to become more entrenched over time, with both leaders and members behaving more consistently over time. This supports the Panciera et al. [159] hypothesis that individuals have static roles that don't shift over time.

There are two limitations of these results however. The first is that this is showing data derived from the entire community, whereas lifecycle models strictly predict changes over time in an individual's behavior. A second, related point is that there is a great deal of churn in community membership, making it possible that across all participants, could outweigh changes in behaviors of long-term members. For example a large influx of late joining members engaging in role-consistent behaviors (i.e. member-like behaviors), could make imperceptible the behaviors of a smaller number of members who have switched to leader-like behaviors.

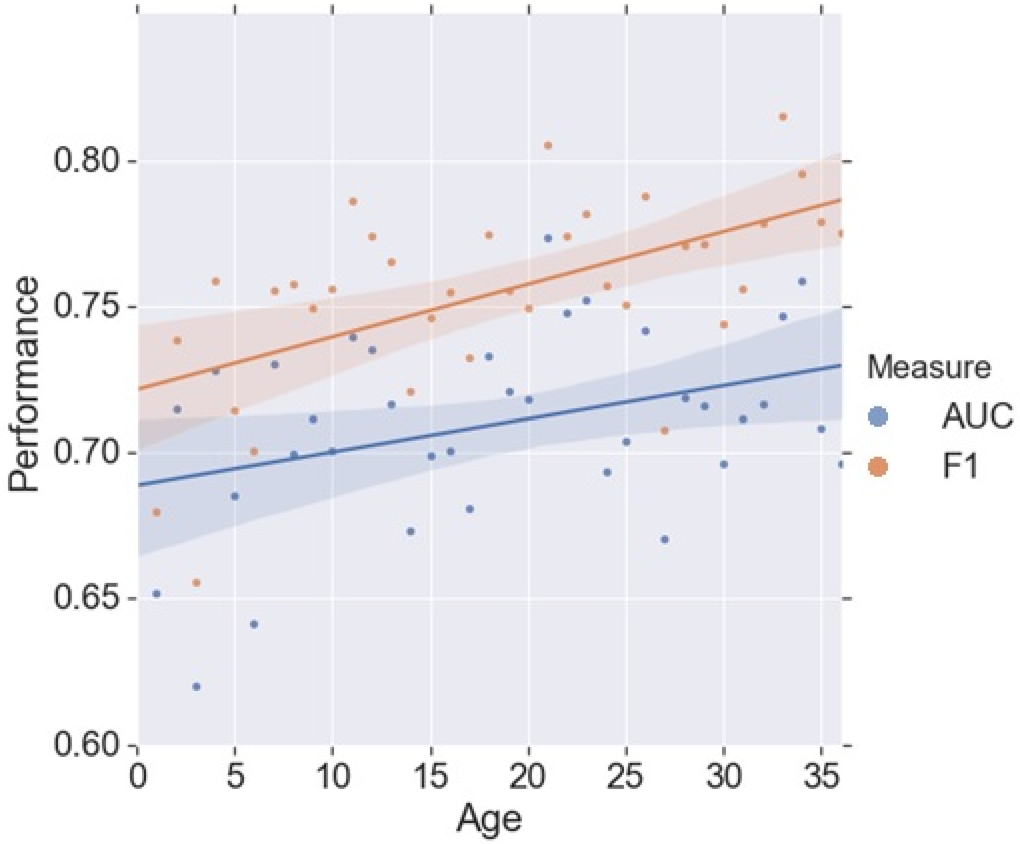


Figure 5.1: N-gram combination model's performance over time. Model performance is good and trend lines show no decrease in model accuracy as the community ages.

Linear Mixed Model Regression			
	Estimate	**Std. Error**	**P-value**
Time	2.598e-5	1.81e-5	0.1336
Role	7.001e-1	1.523e-3	<2e-16***
Time*Role	2.138e-5	1.523e-3	0.2301

Table 5.2: Results of the Linear Mixed Model Regression with leadership likelihood as the dependent variable. Role and members having their last post as a leader were considered separately from members consistently predicted as leaders. However time was found to not be significant. The only interaction effect that was found to be significant was within the group of members with their last post being predicted as a leader.

5.4.2 Roles are static over time

To further test the apprenticeship hypothesis, we used the post-level model trained to produce a probabilistic output for how likely the model evaluated a given post to be authored by an owner, represented as a continuous value between 0-1. Using this continuous output, we can model how this likelihood changes over time for each individual. We use a Mixed Linear Regression to control for individual random effects and included Time, Role, and an interaction effect between Time and Role. As individuals contribute differently within communities [181], our time variable is the order in which the posts occurred in the sequence of all posts by an individual, hence a first post by an individual would be considered time 1, with the second post being time 2 and so on. This controls for individual differences in how far apart in minutes, hours, or days they frequently post and is more focused on each individuals life-cycle in the community.

Apprenticeship models predict an interaction effect between Role and Time. This interaction can be explained as follows: we should see time differences only within the member case as members begin to post like leaders over time, but leaders roles in contrast should be constant throughout. Table 5.2 shows the result of the mixed regression. As expected, role differences were significant, with the difference between roles being around 0.70 indicating that posts by owners were 70% more likely to be considered from an owner. However, time was not significant overall. Turning to the interaction, Time by Role was also not found to be significant, again supporting Panciera et al.'s [159] claim that individuals tend to consistently post according to their starting role.

However, this model also predicted a few members to be owners. Overall the model predicted a tiny 1.7% of members whose last posts were evaluated as being generated by an owner. Could these be apprentices shifting their role? To test this, we conducted an additional Mixed Model Regression with a dummy variable determined by whether members had their last post predicted as to be from an owner. Significant differences were found with this dummy variable indicating these participants were indeed acting more as leaders. However this model did not support shifting theory. When we examining the interaction of between group with time we found the interaction coefficient to be negative, indicating that this group actually became less leader-like over time. This is an interesting result, but not entirely unexpected as prior work reports that some members are retrograde in their use of language, becoming less like their community the longer they are members [55]. Future work is needed to explore how member language changes with respect to community norms.

5.5 Discussion

Overall we provide empirical evidence for static roles [159], disconfirming the apprenticeship theory [171]. We developed two sets of predictive ML role models (aggregate and post-level), but found little evidence for the distinct role shifts argued for by apprenticeship theory. Our post models did not show predicted increases in error rates over time predicted by the apprenticeship account. Nor did they yield the predicted increases in leader-like posts generated by members. These are important findings because they generalize static theory from initial results gathered in the context of Wikipedia [159] to our very different enterprise setting.

This work has limitations. First, it models simple binary owner vs member distinctions. Other work has proposed a broader set of community roles [99, 171] that future work might explore, although empirical evidence for such varied roles remains scanty. Future studies might also explore models that incorporate richer feature sets, for example to include network or graph-theoretic measures [150]. We also restricted our approach to examining participant behaviors within a single community, and did not explore whether the same participant enacted different formal roles across different communities [219]. Future work could explore whether participants with different formal roles across communities show different overall behaviors from those with a consistent role across communities. Furthermore, our data were gathered in a specific enterprise community context. However our static role findings are consistent with results arising in a very different domain focused on peer-based content production [159]. Consistent results across very different settings provide a case for generalization.

Turning to implications, our results suggests a paradox; we know that successful communities have an increased need for more demanding work as they grow [113]. Nevertheless, members do not seem to shift roles to pick up the slack by engaging in more leader-like activities. How then do communities manage this additional leadership workload? One possibility is that long-term leaders continue to handle demanding tasks [105]. Or it may be following Panciera et al. [159], that a small number of members who are consistent power users assist in management tasks despite their official designation and system privileges as members. Or it could be that members occasionally step up to assist with leader tasks, but do so for specific short-term goals, without then increasing the long-term balance of leaders within the community. An alternative viewpoint is that communities may finesse complex issues concerning long-term content management by focusing on recent posts rather than prior accumulated content [50]. This further raises the possibility that theoretical models may overestimate the need for additional leadership work.

Our results also give rise to further intriguing theoretical questions. We have seen that apprenticeship models outline a process to how these individuals learn to be a leader. However along with Panciera et al. [159], we find that a few members enter a community immediately ready to act as a leader. But how do static models explain how these individuals know what to do? And how do they acquire the skills to act like a leader? One possibility is that these power users learn appropriate leader behaviors through lurking before they contribute [154]. Another is that individuals learn from prior experiences as formal leaders in other communities and then transfer that experience into struggling communities [220]. Future work is needed to test these hypotheses.

Our findings also have technical implications for the design of new enterprise community tools. We have seen that members are unlikely to take over critical community management tasks. How then might members be encouraged to take responsibility and how might these tasks be supported? One possibility is that automatic posting tools, such as bots, may provide assistance to community newcomers or help with task allocation. Prior work [52] supports task routing within Wikipedia, shows it is possible to encourage new members to target pressing community needs and address outstanding community tasks. Additionally, within online forum groups like Reddit.com, we see auto-moderation tools that inform new members about community norms before they first post [132]. Other tools may support skill crafting to enhance member behaviors to focus on leadership responsibilities [76, 209]. Such tools may incentivize members by offering extrinsic rewards [97], instead of relying on intrinsic motivations [99, 171].

In summary, this work challenges apprenticeship theory [99, 171]. We did not observe shifts in community roles over time, with participants instead enacting static roles. These findings confirm earlier static results obtained for Wikipedia. Nevertheless we found instances of many enterprise communities that were able to succeed in the long-term. They did this by building on the efforts of a small number active leaders supplemented by a few power users who consistently engaged in leader-like behaviors despite being officially designated as members. Future work should extend these findings, building new tools to help these active community members carry out their demanding work over time.

Chapter 6

Community Language and Success

6.1 Introduction

Recent work has begun to explore community interaction examining how the content of members' conversation affects success [63, 151, 191, 202]. A key question concerns the impact of emotional versus factual language. In some communities, emotional support for other members is critical [134, 204]. In these communities it may be vital to respond to the affective content of a post rather than the factual information that the poster is nominally requesting. This interpersonal focus may then promote long-term relations between members [137, 146, 208]. In contrast, other communities serve short-term informational needs, where the most common interaction is a simple information request from a first-time poster where a factual response is optimal [63, 113, 151, 191, 202].

This chapter examines the relationship between emotional versus factual communication and the perceived success of enterprise communities. We adapt known models from a corpus of online debates to develop an algorithm to detect the relative prevalence of emotional versus factual content in posts. We then apply the algorithm to the under-researched context of enterprise communities. However one challenge in examining content relationships is that there is a huge spectrum of online communities that are often used very differently. We therefore also explore how emotional versus factual communication differs as a function of the community type and social media tool used.

6.1.1 Community Types and Community Tools

Prior work has noted key differences between community types that engender very different forms of communication [67, 91, 120]. For example, Communities of Practice [208] or CoPs are oriented to relationships and support, leading to more emotional language use [177, 204, 208]. In contrast, project based teams focus more on factual information because of their predominantly instrumental goals [113, 146, 172]. We therefore will analyze community type.

While much of the research on communities has focused on discussion forums, there are now multiple tools available, including Wikis and Blogs. Tool type may affect content posted. For example, Wikis are often used to post descriptive information as a resource for new members [136]. And while other tools such as Forums feature an initial post that nominally elicits information or assistance, the resulting discussion responses may be discursive and evaluative [91]. Blogs in turn may aim to be more evaluative in their tone. While prior work has explored content usage for each of these tools in isolation, it has not explored how tools are used interdependently. Another goal of this chapter is to therefore examine differences in emotional content between different tools when community members have access to multiple social media tools and can choose which tool they want to post their information to. Increased understanding of the role of emotions should improve the design of community tools and practices across different community types.

6.1.2 Hypotheses

The predominantly informational goals of enterprise communities should mean that an emphasis on factual communication leads to improved member satisfaction.

Conflicts can arise between emotional and factual goals, fostering disagreements between members who respond to the poster's emotional needs, and others who view such emotional responses as being 'off-topic' [172, 177]. Furthermore, research on social media communities shows that for controversial political topics, using emotional rather than factual language is counterproductive, leading to deadlock in negotiating legislation [15]. The expectation is therefore that factual rather than emotional language will promote success within enterprise communities.

The role of emotion is moderated by community type, with emotional communication being more important than factual content in communities that have greater focus on relational goals such as CoPs.

Community goals align with community type [67, 91, 120]. The enterprise context includes not only communities that aim to promote large scale social interactions around common topics of interest, but also smaller team-based projects. We expect that these larger Communities of Practice with their social goals will rely on emotional language. In contrast teams with instrumental goals should have more factual content.

There will be differences in use of emotional language both between and within community tools. We expect Wikis will be more factual than other tools, and that initial posts will be more factual than responses.

Tools have stated purposes and intended uses. Wikis tend to be used for descriptive purposes while Blogs and Forums are more social and conversational. However it is also the case that deeper threads within a conversation seem to engender more emotional communication [221]. We therefore expected to find more factual content in Wikis and the initial posts of Blogs and Forums to be more factual than the comments and replies.

This algorithmic approach allows me to isolate and quantify the contribution of emotional versus factual language and test these predictions of enterprise communities. The contributions of this chapter are to: (1) extend 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 relationship of emotionality depends on community type and tool.

6.1.3 Algorithmic Analysis of Emotions

We use automatic linguistic and statistical methods to model emotional and factual language use in online communities. Prior work has used such methods to: identify emotional and mood expression in social media [11, 12, 19, 58, 144, 156]; understand linguistic predictors of member satisfaction in communities [48, 134]; and predict member retention following exposure to emotional language [204].

6.1.4 Sentiment and Emotion Detection

Early work on sentiment detection made bimodal judgments about whether a given text expresses a positive versus negative evaluation [152, 176]. Sentiment detection uses a mixture of machine learning, lexicon based, and hybrid approaches [176]. Areas of research within sentiment detection have focused on Polarity determination [80, 143, 199], Multi-lingual and cross-lingual analysis [21], and cross-domain classification [88]. More recent work has extended bimodal positive vs negative distinctions, instead aiming to identify the presence vs absence of emotions in text [39]. For example [12] classifies whether a sentence is emotional or not, at 73% accuracy compared with human judges. Similarly [11] developed and used both semantic and lexical machine learning features to obtain an accuracy of 69% in classifying neutral vs. emotional sentences. Other work develops methods to recognize different types of emotions in text [19, 38, 58, 168].

The aim here is different from both sentiment and emotion recognition. Like [11, 12, 204], rather than determining the valence or type of an emotion, instead we want to determine the degree to which a community post is emotionally versus factually focused. We therefore adapt these methods to automatically rate texts on a scale that ranks posts from highly emotional to highly factual. Furthermore, like some previous work [26, 182], We are looking to train within a well annotated domain and expand that predictive power onto a different but related domain.

6.1.5 Assessing Community Success

Many success metrics have been proposed for online communities. However these metrics are rarely validated and there is little agreement about which are most effective [30]. The most commonly proposed behavioral success metrics are: volume of members' posts [35, 67, 99, 146, 172], number of members [67, 10], and quality of member relationships (e.g., measured as the extent of contact among members) [67, 10]. Other common metrics include number of message threads [35], number of replies [35], threads with responses [35], and delay in response time [67]. Some researchers have developed algorithms combining multiple behavioral metrics to rate community content [91], community members [99], or the community itself [91].

One critique of these behavioral measures is that they are indirect. Other work therefore directly assesses participant perceptions, e.g. member satisfaction [10, 101], rather than inferring success from behaviors. It has been long known that successful online communities must meet member needs [113, 124, 208] and the relationship between behavioral measures and participant perceptions of their community's success in meeting its goals is explored in [134]. That paper uses member satisfaction as a measure of community success. One aim of the current paper will be to re-examine how well these commonly proposed success metrics predict member satisfaction.

To further add to the complexity of defining success, there may be differences between types of online communities. Some claim that different types of communities have different goals with corresponding different success metrics [67, 91, 120]. Porter [169] argues that source of initiative is a key factor, leading to different goals in organization- vs. member-initiated communities. Muller et al. [146] compared different types of communities, e.g. Communities of Practice (CoPs), collaborative teams, and technical support communities, showing measurable differences in behavior. One goal of this paper is to determine whether there are differences in the relationships of emotional versus factual language use on perceived success depending on community type.

6.1.6 Language Use in Communities

A number of researchers have explored the effects of language use on online community members' behaviors [134, 142, 146, 151, 204]. Matthews et al. [134] used the Linguistic Inquiry and Word Count tool (LIWC) [190] to understand what community linguistic behaviors predict member satisfaction in enterprise communities. They found simple linguistic predictors of member satisfaction that included the use of inclusive language, low anger and increased anxiety.

Two other types of language use have received theoretical and empirical attention emotional support and factual support. The majority of this work has not examined enterprise communities however. Participants in online support groups experience different forms of emotional support. This can be either direct, such as messages of caring and concern, or indirect, e.g. comparisons with others with similar experiences. Cancer patients often claim that emotional support is the most helpful type of support they receive and seek [66]. Prior work suggests that peer discussion towards emotional support enhances cancer patients' psychological adjustment [92].

Wang et al. [204] explored the role of emotions in health support communities. They developed two machine learning models to automatically identify messages which contain (a) emotional vs. (b) factual support. With these models they found greater emotional support lowered the risk of dropout whereas factual support had the opposite effect. The authors speculate that emotional support enhances member's relationships with one another, whereas more factual responses may simply satisfy simple information needs. However, factual information is a key function of online communities [15, 32, 40, 63, 151]. Participants in health support groups also obtain factual support, e.g. about the course of their disease, treatments, side effects, communication with physicians, and financial problems and other burdens.

Currently there is no consensus about the relative benefits of emotional versus factual communication for enterprise community success. To explore these relationships quantitatively we used a similar approach to [204]. Building on [11] and [12], we developed a single machine learning metric to determine whether community post contains more emotional versus factual content. This single emotionality measure allows me to quantify the relationships of emotional vs. factual focus within communities.

6.2 Method

6.2.1 Adapting the Emotionality Algorithm

We adapted previous work [11, 12, 204] in developing an algorithm that allowed me to distinguish emotional vs factual posts. This involved the following steps:

1. Find a set of explanatory features relating to factual and emotional language, given a data set that contains distinct annotated examples

2. Construct a model using said features

3. Validate this model's output within the domain of interest (enterprise communities)

6.2.2 Developing Explanatory Features

To accomplish the first step, we used the 10,000 post-response pairs from the IAC corpus of online forum debates [203] 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 to +5. Each forum response was annotated by 5-7 annotators. To ensure reliability of judgments, Turker judgments were filtered based on two criteria, 1) each response had >4 ratings and 2) the standard deviation for ratings on each response <3.0. Following prior work [186], these criteria optimized the number of judgments used, while maintaining reliability. Using this corpus allowed me to develop a model derived from multiple different types and valences of emotional and factual interaction. We modeled the extent to which a response to a post was emotional versus factual, which we refer to as emotionality. We wanted to identify a set of explanatory linguistic features in the forum responses that would predict the Turkers' emotionality judgments. We explored both Lexical and Syntactic Features (table 6.1).

Previous modeling work [11, 12, 204] derived lexical features from three sources: LIWC (Linguistic Inquiry Word Count), EmoLex and Subjectivity Lexicons. We use the same lexical sources. Each of these lexicons classifies words into a parent category (e.g. 'anger', 'annoyance' belong to the negative emotion parent category). The lexicon is used to find the relative frequency of words in the target text that correspond to each category.

Lexical Features: LIWC v2007 [190] is a lexicon that provides frequency counts of words that signify important psychological constructs, as well as some relevant topics (e.g. Leisure, Work). LIWC is widely used and reliable compared with human judges [134, 190, 204]. The LIWC dictionary defines 81 word categories, each containing multiple words. It indexes categories such as pronouns ('I', 'you'), as well as words with psychological relevance, e.g. that express positive and negative emotion or verbs of cognition. Categories are not exclusive; so words can belong to multiple categories.

The Emotion lexicon (EmoLex) [142] is specifically focused on emotional terms. It contains 14182 words classified into 10 emotional categories: Anger, Anticipation, Disgust, Fear, Joy, Negative, Positive, Sadness, Surprise, and Trust.

The final lexicon was less directly concerned with emotions. Instead it was focused on whether words expressed positive or negative sentiment. The Subjectivity Lexicon is part of OpinionFinder [212]. It consists of 8222 stemmed and un-stemmed words annotated by a group of trained annotators as either strongly or weakly subjective. Subjectivity has been found as a useful lexicon for analyzing sentiment [27, 176].

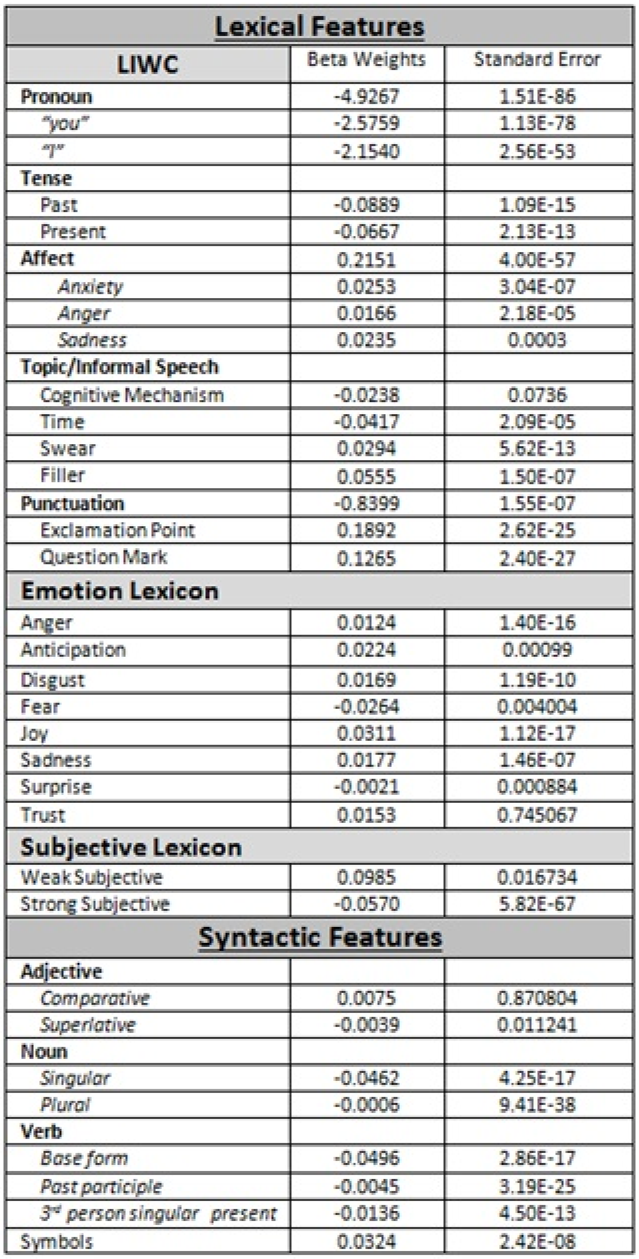
Syntactic Features: Lexicon based approaches have limitations. They use a simple "bag of words" which assumes that social and psychological meaning can be derived from individual words alone. This ignores syntax, punctuation, conversational structure, and other relational features of text. We therefore also included structural features of language use, such as use of questions and grammatical tense that might also signal emotional or factual expression. Syntactic choices show an emphasis on concepts (nouns) versus actions (verbs) [165]. Syntax also indicates a focus on past, present or future. We 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 [197].

6.2.3 Creating the Model for Detecting Emotionality

We wanted to model the relationship between these features and the emotionality judgments generated by Turkers for the debate corpus. Our first approach was to combine judgments into a binary classification of Emotional vs Factual responses. However actual Turker responses ranged from extremely to mildly emotional responses and conflating these would result in data loss. Furthermore, the majority of the data was distributed close to the mean (zero), i.e. 45% of posts were judged as mildly emotional (0 to -1) or mildly factual (0 to +1), so treating a score of -0.1 as similar to a score of -4.9 and categorically different from a score of +0.1, is likely to reduce model reliability. One solution to this is excluding intermediate data and focusing on clear-cut cases, however this again results in data loss. Using scores of <-2 and >+2 results in only 39% of data being used and <-1 and >+1 results in only 55% of data being used.

We therefore abandoned the binary classification approach and instead developed a regression model which outputs a scalar emotionality evaluation, which better represents the given input feature representation for each debate response, resulting in less data loss. This model assesses the extent a given text is factual versus emotional.

We used Scikit-Learn [164] a machine learning toolkit to build a regression model. The dataset was split into a 85-15 training-test set. Within the training set, 5-fold cross valida-



Lexical Features		
LIWC	Beta Weights	Standard Error
Pronoun	-4.9267	1.51E-86
"you"	-2.5759	1.13E-78
"I"	-2.1540	2.56E-53
Tense		
Past	-0.0889	1.09E-15
Present	-0.0667	2.13E-13
Affect	0.2151	4.00E-57
Anxiety	0.0253	3.04E-07
Anger	0.0166	2.18E-05
Sadness	0.0235	0.0003
Topic/Informal Speech		
Cognitive Mechanism	-0.0238	0.0736
Time	-0.0417	2.09E-05
Swear	0.0294	5.62E-13
Filler	0.0555	1.50E-07
Punctuation	-0.8399	1.55E-07
Exclamation Point	0.1892	2.62E-25
Question Mark	0.1265	2.40E-27
Emotion Lexicon		
Anger	0.0124	1.40E-16
Anticipation	0.0224	0.00099
Disgust	0.0169	1.19E-10
Fear	-0.0264	0.004004
Joy	0.0311	1.12E-17
Sadness	0.0177	1.46E-07
Surprise	-0.0021	0.000884
Trust	0.0153	0.745067
Subjective Lexicon		
Weak Subjective	0.0985	0.016734
Strong Subjective	-0.0570	5.82E-67
Syntactic Features		
Adjective		
Comparative	0.0075	0.870804
Superlative	-0.0039	0.011241
Noun		
Singular	-0.0462	4.25E-17
Plural	-0.0006	9.41E-38
Verb		
Base form	-0.0496	2.86E-17
Past participle	-0.0045	3.19E-25
3rd person singular present	-0.0136	4.50E-13
Symbols	0.0324	2.42E-08

Table 6.1: Example of Features along with their model weights. Emotional predictors have positive weights; Factual predictors have negative weights.

tion 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. We used Adjusted R^2 to eliminate spurious variance increases arising with Unadjusted R^2. Unlike Unadjusted R^2, Adjusted R^2 only increases the explained variance if a newly added variable explains more of the variance than would be expected by chance. RMSE allows for a comparative evaluation of the model's variation to the variation of the original Turker annotations.

To ease interpretation, we transformed the valence of the emotionality score so that a higher positive number indicates a higher level of emotion, and a lower value is more factual. 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 statistically 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. The most significant features and their standardized coefficients (beta weights) for the regression model are shown in table 6.1. The table shows a mixture of lexical and syntactic features predict emotionality.

Examining table 6.1 suggests that the features that predict emotional judgments seem to have face validity. Predictive features include swear words and specific punctuation, e.g. exclamation or question marks (!!!!, ????), that are commonly used to express emotions in other contexts, while all forms of punctuation was found to be factual. Other features that predict factual evaluations, pronouns overall and specific personal pronouns such as “you” and “I” were important predictors, as were a number of specific syntactic features. We can also contrast factual versus emotional predictors. With the exception of pronouns these were distinct, with emotional features drawing heavily on specific types of punctuation whereas factual predictors were more often syntactic.

As an initial quantitative check of face validity, we determined if the model’s outputs corresponded with the pre-existing emotion lexicons. We correlated the emotionality score generated by our model with the frequencies of emotional categories in LIWC and EmoLex. Responses that the algorithm classified as emotional had higher frequencies of emotional terms in LIWC and EmoLex: Emotional posts typically contained 10.54% emotional terms while Factual posts just 4.97% ($t = 18.87$, $d = 0.46$, $df = 4786$, $p < 0.001$).

Having developed the emotionality model using debate forums we next evaluated whether the model generalized to other contexts by validating it on online enterprise community data.

6.2.4 Linguistic Data

For each of the Communities, we collected all the content posted to their discussion forums, blogs and wikis over the community’s life, including original posts and comments/replies. We analyzed wiki, forum, and blog posts, as these were the three tools in Communities that generated significant amounts of created content. There were 428,476 posts overall. While all world continents were represented, this work examined only English posts.

6.2.5 Behavioral Data

We also collected behavioral data for each of the Communities. We collected the most common behavioral measures of community success used in prior literature. While there are other potential behavioral metrics, the following have commonly been proposed to explain community success and we wanted to include these as control variables [169, 172, 208].

- Membership: # of leaders, # of members, # of contributors
- Contribution: We collected the number of posts across a range of community tools, including # of wiki, blog and forum posts, # of blog and forum comments
- Equality: From these data we also computed the gini measure of equality of contribution [79]
- Consumption: # of wiki and blog views

6.3 Results

6.3.1 Cross Data Set Validation

To determine whether the emotionality model developed for online debate forums generalized to Communities we first tested the model's ability to predict emotional judgements within Communities. We created a direct test set of annotated posts from Communities. Using the same procedure as for the IAC corpus annotations, we solicited judgments for 1000 Communities posts selected at random from the communities we had surveyed, 7 posts had to be removed for not receiving enough annotations. Then we tested to see whether the model's predictions for each post agreed with the judges' emotionality ratings of that post. Model and judges' ratings were highly correlated, r = 0.54 (df = 991, p <0.001) and Kendall's Tau = 0.37 (p <0.001). This shows that the emotionality model derived from debates generalizes to the community data.

Next we explore the role of emotional versus factual communication in online communities, to evaluate the relationship of emotional communication on perceived user satisfaction.

6.3.2 Predicting User Satisfaction in Enterprise Communities

Again we used regression methods, where multiple models' performance will be sequentially compared using Adjusted R^2. To evaluate the relationship of emotionality on community success, we first created a Control Model containing the following (language independent) structural variables that have been proposed elsewhere as measures of community success. Our first model used these structural factors to predict perceived user satisfaction. We next added emotionality to the Control Model to evaluate our prediction that greater emotional communication would predict lower overall member satisfaction.

Control variables were:

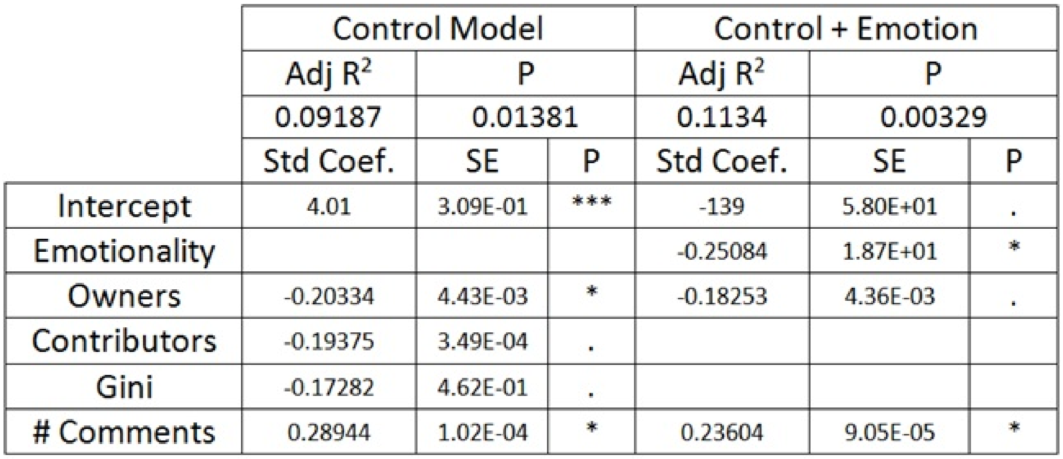
- Community Type: (CoP, Teams, Recreational, etc.) Membership: # of leaders, # of members, # of contributors
- Gini: From these data we also computed the gini measure of equality of contribution.
- Contributions: # of words, # of posts (wiki pages and edits, forum and blog posts, bookmarks and file uploads), # of comments (blog comments, forum replies)
- Consumption: # of views (wiki and blog views, file downloads)

For the 93 communities, the average and standard deviation values for the control variables were as follows: # members (1729, 2860), # contributors (116, 159), total # posts of all types (605, 660), total wiki+blog+file views (37239, 62537), total comments (372, 553). All data was centered and the resulting distributions were normal.

One limitation of the regression approach is that variables may be highly correlated or multi-collinear. We first tested for multi-collinearity using variance inflation factor (VIF). Following standard practice [98], variables with the highest VIF were removed until all variables were under a VIF threshold of 5.

We then derived a Control model (Table 6.2, 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 [98]. Using a both-direction step procedure is less biased than a one-way step. Stepwise selection led to the removal of Type, Members, Contributors, #Posts, Gini, Word Count, and Views variables for the Control model.

Table 6.2 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. However our main interest was in exploring the role of emotional communication. A simple one way correlation between emotionality score and satisfaction (r=-0.219, df=91, p=0.034) suggests that emotionality may contribute to satisfaction.



	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	*
Owners	-0.20334	4.43E-03	*	-0.18253	4.36E-03	.
Contributors	-0.19375	3.49E-04	.			
Gini	-0.17282	4.62E-01	.			
# Comments	0.28944	1.02E-04	*	0.23604	9.05E-05	*

Table 6.2: Model 1 (Control) using the traditional measures for predicting member satisfaction. Model 2 (Control +Emotion) adds the Emotionality feature which improves predictive power and shows a negative relationship (‘*’ indicates significance p<0.05, ‘.’ p<0.10).

6.3.3 Facts Not Emotions Predict User Satisfaction

The Control + Emotional Model (Model 2 in table 6.2) adds emotionality to the Control model to test whether emotional interaction increases satisfaction. Using the same feature selection procedure, we 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 our prediction. It is important to note that this relationship depends on the degree of emotionality rather than the valence of emotions expressed; independent analyses exploring positive and negative emotions revealed no significant predictors. Nor did sentiment predict community success.

This result contrasts with prior work on support forums where high amounts of emotional content benefit online interactions [204]. However the result may reflect the overall goals of the communities studied. To explore community goals further we next examined relationships between structural factors (including community type) and satisfaction.

6.3.4 Emotional Language Has A Negative Relationship With Satisfaction In Communities Of Practice

We first examined whether the relationship of emotional language and member satisfaction depended on the type of community. Emotional language in a community of practice (CoP) should improve satisfaction more than using emotional language in a team [169]. We initially checked for any interactions in the regression analysis but the results were borderline. To examine further we simplified the Emotionality measure into a median split of High Emotion and Low Emotion.

Figure 6.1 shows how emotionality interacts with Community Type to influence satisfaction. It contrasts satisfaction in CoPs with other types of communities. The figure suggests that highly emotional language in CoPs has a negative relationship with satisfaction, a relationship that is less pronounced in other Types of community. Using Hedge's G to calculate the effect size between the two groups shows that within CoPs (g = 0.714, CI = [0.617, 0.810]) there is a strong difference, while within other communities this difference is weaker (g = 0.314, CI = [0.176, 0.453]).

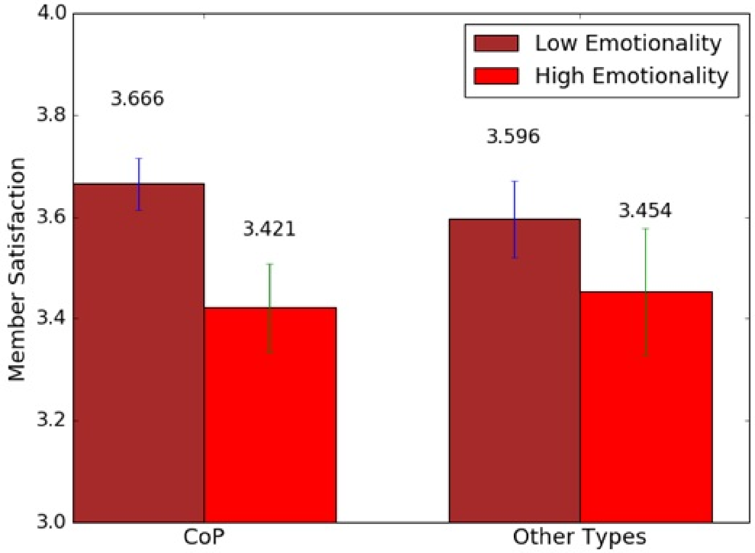


Figure 6.1: Emotional communication has lower Satisfaction, with this relationship being more marked in Communities of Practice. Satisfaction for COP vs other community types contrasting communities with low and high emotional language use.

6.4 Limitations

We studied one company and the results may not generalize outside this context. Among other things, our study demonstrates that the relationship between emotional versus factual language and member satisfaction depends on community type. We suggest future studies explore this relationship in different contexts and for different community types. A second potential limitation of this work is the member survey method. Respondents were asked to respond from their own personal experience. It is possible that the members who responded to the survey were not representative of the community membership as a whole. However, we believe this concern is limited because our member respondents agreed with each other per community (see the high intra-class correlation coefficients noted above). A third issue concerns our cross-sectional analysis; while the model suggests a relationship between emotionality and member satisfaction, it does not indicate the causal relationship between them. Finally this work is purely quantitative, and qualitative analyses of emotional communication in enterprise communities would extend and add nuance to the results presented here.

6.5 Discussion

We first developed an emotionality detection algorithm and then used it to evaluate perceived community success. The results show a small but clear relationship between the use of emotional versus factual language in enterprise communities and member satisfaction. As predicted, we found that factual language enhanced perceived satisfaction. Furthermore, counter-intuitively, emotional language use reduced satisfaction in CoPs where social relations and emotional support are thought to be important. This relationship relates to the presence of emotional language rather than to the valence of emotions expressed.

Increased understanding of the role of emotions should improve the design of community tools and practices across different community types. These results contrast with work on health support communities where greater use of emotional language is associated with member retention. But this discrepancy may result from the different goals of enterprise and support communities. Successful enterprise CoPs may rely on factual language, with emotional language signaling a breakdown of communication within the community. These negative consequences of emotional discussions echo recent work on political discussions where use of emotional language has been shown to lead to deadlock and lack of legislative progress [58]. We plan to explore this further in future work to determine whether satisfaction is reduced specifically by negative emotional discussion or by emotional interaction in general.

There are important practical implications to these findings. Community leaders might directly apply the results by introducing policy guidelines concerning the use of emotional language or attempt to moderate posts that are 'overemotional'. For example, typical factual posts in the communities with the highest user satisfaction stated: "Team, I am starting this forum to track any issues found during Production Pre-deployment on 27th and 28th of April Please list those issue individually on the following forum." This post is indicative of a directive and goal oriented style of post. In contrast typical emotional posts found in low user satisfied communities stated, "Thank you .... Happy Christmas holidays to you, your family and all Blue Community members!". This style of post may not directly relate to enterprise community goals.

Theoretically the results are also important, in showing that the relationship of emotional language is not always intuitive instead depending on the precise context in which that language is used. Methodologically this work extends prior analyses of emotional community language use. Prior work relying on modeling using LIWC showed effects of emotional language but across a large array of implicit and explicit lexical features. By developing a single emotional classifier We was able to isolate and quantify the predictive power of emotions, and explore how emotional language interacted with other variables such as community type and tools. Furthermore, by validating the model's ability to generalize outside its training domain, this opens up new possibilities of exploring emotional language in educational, therapeutic or political settings.

Chapter 7

New Methods to Identify Subgroups in Online Communities

7.1 Introduction

Online communities provide support via weak ties that are often unavailable through people's existing strong tie networks [83]. Such communities also help maintain complex social relations both pairwise between individuals, as well as between subgroups of three or more participants. Early communities work documented the dyadic processes by which individual community members pose questions and receive answers from domain experts [91, 172, 193]. However, it is apparent that there are other important multiway relationships existing above these dyadic connections. For example, a three-way triadic relationship frequently arises when a community leader recommends that a new member with a topical question consult with a third person who is an expert on that topic [56, 57, 73, 94]. Or in a peer production context, a leader might negotiate with multiple volunteers who form a small team adopting differing roles to address a community related task [100, 117, 130, 177]. Such interactions give rise to complex small group networks and these relationships remain to be thoroughly explored quantitatively. This chapter proposes new methods to analyze and understand the effects of higher level subgroups on communities. We also examine the implications of different small group configurations for community success.

A common approach to the analysis of community relationships is Social Network Analysis, which utilizes the mathematical field of graph theory. This approach aims to characterize the many different types and dynamics of relationships by formulating an overall network of relationships. Graph theory has many advantages for examining social phenomenon as it treats the existence and description of relationships between individuals as interesting in their own right [78]. It has produced powerful general metrics such as range, centrality, core periphery, density, and strength of ties for understanding individuals' relations within a network. However, social network analysis does not yet provide systematic fine-grained methods for identifying and analyzing small sized groups [78].

Recently granular graphical methods have been proposed to identify substructures within a larger network. These substructures are typically referred to as graphlets or motifs. One type of substructure is the triangle, a 3-node connection between three community members. But despite the intuition that these substructures capture meaningful social relations, it has nevertheless been argued that substructures do not add additional explanatory information because standard higher level network measures of density or core periphery are a byproduct of these lower level organizations [14, 36, 73].

However Faust [73] has challenged that argument by showing that 3-node substructures supplement information available from lower order graph features (dyads). Faust found that triadic structural tendencies (i.e. social tendencies to form groups) align dyads into triads in ways that depart from expected census. Here we take a different approach than Faust [73], exploring the utility of graphlets in assessing community health. The current chapter utilizes her finding as a motivation for exploring more complex 4-node graphlets and examine whether identifying and characterizing such substructures has implications for the important issue of online community success.

This chapter examines 4-node graphlet substructures in Enterprise Online Communities. We explore groups of size 4 as it is both computationally viable and has been found to be just under the optimal group size found within prior work on certain decision-making tasks [86]. We explore how these different 4-node structures relate to different social roles and dominant content producers, and examine how these substructures predict metrics of online community success.

Specifically, this chapter addresses the following questions:

- Do subgroups relate to online community success?
- Does the identification of subgroups have explanatory power compared with more standard global network metrics?
- What are common subgroups in online enterprise communities?

While triadic tendencies in social networks have been thoroughly examined [73, 128], higher level structures in online communities have not. This work explores 4 person groups in enterprise communities. We expect that denser groups are likely to be less common than sparsely connected groups.

How might different subgroup types promote community success? We also examined whether subgroup structures existence predict other adaptive community behaviors associated with success, such as fast response times to posts and reciprocity. We expected denser subgroups to facilitate cooperation and hence success [77, 216], in contrast to sparser subgroups where there are more holes in the network [20, 34, 77].

Are key community roles, such as leaders, common within subgroups? In general, assortativity (also referred to as homophily) is prevalent in communities with people tending to talk to others like themselves [53, 62]. However this bias undermines the potential weak-tie value of heterogenous subgroups. We therefore explore whether subgroups are made up of heterogeneous networks containing a mix of leaders and members, or instead whether they favor assortativity.

7.1.1 Contribution

This chapter highlights the importance of online subgroups identified by graphlets in explaining community success. We further examine how the structural properties of these subgroups contribute to community success through response time, as well as supporting heterogeneous interactions involving leaders and members.

7.2 Assessing Social Phenomena Using Network Methods

As discussed within Chapter 2, networks or graphs have been widely used to understand the connections and dynamics of online communities [77, 82, 121, 141, 153, 162, 216]. Typically networks use nodes to represent individuals and edges to represent a connection between them [205].

Network representations provide a rich set of metrics that can help operationalize social theory. Rowe [181] used in-degree and out-degree distributions to measure social dynamics to model likelihood of individuals leaving a community, finding that number of replies (in-degree) was a significant predictor [181]. More relationship-based measures focus not on dyadic one-to-one interactions (like degree) but instead model one-to-network interactions. Those interactions identify high connection, influential nodes within the graph, and one such measure is Centrality [20]. Nolker and Zhou [153] used multiple relationship-based measures including degree, betweenness, and closeness to identify various role types in Usenet communities, finding more impactful roles (e.g. leaders and motivators) to have significantly higher network relationship measures [153]. Johnson et al. [102] also explored relationship-based measures in modeling role behaviors, finding leaders to be associated with k-core, a measure quantifying the network of nodes with at least k degree. Sparrowe et al. [187] examined workplace relationships using a network representation to explore the relationship between centrality and task behaviors. Performance was positively related to centrality in cooperative networks with the opposite being true for uncooperative cases.

Other work has focused on different properties of network nodes. One key topic is

Assortative Mixing, or the tendency for nodes to interact with other similar nodes [150]. Chung et al. [47] explored assortative mixing of forums for government workers, while they found mixed results they opened new questions to community interactions in Web 2.0 tools. Gong et al. [82] also observed patterns of group node attributes early within the Google+ deployment and that social networks were found to be unique in this platform as they had lower social reciprocity and assortative mixing compared to other social networks. They later go more in-depth and find reciprocity to be related to common attributes of nodes within their networks, providing a more nuanced picture of assortative mixing.

Finally, network analysis has extended beyond simple individual or network level metrics to explore different possible network configurations. Cummings and Cross [54] explored the relationship between network structure and performance in an online work community. They examined core-periphery and hierarchical network structures by measuring the structural holes within work group, finding these to be associated with negative work performance, indicating a lack of connectedness between leaders and the rest of the network [54]. This work highlights a key point missing from many analyses of networks, the structure or network orientation is an important contributor to group success.

Prior work also has identified substructures and their relevance towards social properties such as triadic tendency which was theorized from Social Exchange Theory [20, 54, 73]. Network science tends to refer to these structures as graphlets or motifs. One example of a subnetwork type is a triad (involving three nodes) and their many different possible configurations like connected (two edges) versus a clique (three edges). The existence of such has proven valuable to evaluate the social relationships within networks, i.e. transitivity of a network being the proportion of triads forming cliques has implications to the social ability and likelihood for connections to form [20, 24, 73]. These substructures can explain leader effectiveness, friendship formation and the community context as well as network evolution over time, [16, 20, 54, 62, 73, 163, 179]. Prior work has identified significant patterns involving triads [24, 73] and Faust [73] showed triad census exceed expectations from dyadic census, allowing for the argument that substructures may provide information that is not captured by such lower-level measures. Dong et al. [62] also recently examined homophily through exploring various substructures. Their findings are rather complex in that existence of two nodes within a similar substructures varies in its implications if those two nodes share a connection with one-another. They find in certain contexts the expectation of shared structures to be positive with connects but negative in other implying different network context properties. Either way, that work highlights a relationship in how individuals form connections. We expand on that work by incorporating an examination of social roles.

Graphlets, while relatively recently are being explored, have traditionally not been measured for computational reasons [8]. While counting the number of graphlets has proved tractable [8, 121], efficient techniques for finding graphlets in networks have yet to be developed. Finding graphlets is a very different task than measuring the population of graphlets. As shown by Ahmed et. al [8], counting existing graphlets can be achieved through combinatorial means, but finding who is within those graphlets still needs a search procedure as all nodes need to be evaluated for the potential graphlets they could be a part of, hence making the search space of the number of possible 4 groups within all possible nodes.

This chapter uses graphlets to identify network substructures. We show that these substructures help explain the success of online enterprise communities. Further exploration of graphlets indicates why: community leaders are highly active within these substructures, and denser substructures promoting efficient community interactions.

7.3 Methods

7.3.1 Community Sampling and Data Collection

Out of the 2,010 communities within the communities sample, a total of 428,476 posts exist. Recall that the Communities system was originally developed the purpose of providing a platform to establish healthy enterprise communities that support employees and corporate processes. We were therefore able to collect log-files data on every user interaction, pages viewed, clicks on interactive widgets, from July 2007 to May 2014. For each post, with participants' agreement we captured:

- Community ID (Where it was posted)
- Author ID (Anonymous Unique identifier)
- Date (Time stamp when post was made)
- Tool (Which tool the post was in: e.g., blog, wiki, etc)
- Role (Member vs Owner of community posted in)
- Date of Community Creation (To determine when in the community lifecycle the post was made)

7.3.2 Owners and Members

To reiterate what we know about roles and networks, prior work has found and highlighted various roles and activities for community contributors, however there is no clear consensus about how roles are behaviorally defined [35, 102, 153]. Prior studies mainly employ inferential methods to distinguish these various roles [102, 153, 215]. Again, our data has unique properties that allow us to avoid the problems that come with inferential methods. In Communities, participants are systematically designated to be either an owner or member, each with different privileges. These role definitions supply us with a gold standard for identifying social roles, removing the need for inferential methods and provide more direct measures of owners and members. While a fixed definition of roles may suggest limitations, prior empirical work shows peoples' online roles tend to be relatively static [55, 159]. Of direct interest in this Chapter is the extent to which these fixed role designations map onto substructures detected in graphlets. We also assess how these substructures relate to standard network phenomena such as assortativity (homophily).

7.3.3 Survey Measures of Member Satisfaction

Reviewing our measures of success, workplace community members were surveyed as part of a larger research project [134]. This chapter, just as with chapter 6, involves a subset of the survey and communities from a larger study. Success was assessed using the most reliable survey question, the member satisfaction probe which asks community members "how well this community is meeting your needs", on a scale of 1=very poorly to 5=very well. We rely on this single question because it was highly correlated with other related questions, e.g. 'how successful is your community' as well as being predictive of other behavioral success measures [134].

A sample of actively managed communities was drawn from a pool of 666 communities whose leaders participated in an experimental deployment to help leaders enhance their community. These communities varied widely in terms of size, longevity, and purpose. The survey was sent to 20-26 members within each community and the response rate was 19% for all participants surveyed. A stratified sampling method was used to balance the different types of community members. We next removed: communities with too few members (<20) and too few responses (<3) to yield a valid assessment of member satisfaction. We also removed 8 communities with incomplete data; and 86 communities with <3000 words to ensure enough content to obtain accurate results from lexical features. The word threshold was needed to reduce sampling error, i.e., the error across different lexical category frequencies when comparing small and large language samples from the same source [85]. We arrived at this threshold after piloting, where we tried to maximize the number of communities we included while reducing this error. This left a total of 93 communities for analysis. Respondents represented a wide range of geographies, business divisions and roles.

The overall response rate was 19% for all participants surveyed, and an average of 5.9 members responded per community. We averaged member responses within each community, as a validity check showed good correlation coefficients (average ICC = 0.69) across respondents from the same community.

	Owner	**Member**	**Hedge's G**
Degree	0.0075	0.0010	0.5442 (Moderate)
Triangles	2.1114	0.4526	0.1927 (Weak)
K-Core	0.6857	0.5503	0.1337 (Weak)
Clustering	0.0494	0.0515	0.0107 (Weak)
Centrality	0.0017	8.25e-5	0.3179 (Moderate)

Table 7.1: Average network measures for roles.

7.3.4 Network Definition

7.3.4.1 Online Communities as Networks

Networks were built for each of the 2000 communities using the NetworkX library within a Python 2.7 environment [87]. Each user in the online enterprise community was considered a node and edges were found through measuring a reciprocal relationship between nodes. Reciprocal relationships are defined as exchanges between two entities, which we operationalized as when a user responds to another user's post. We refer to this network as a reciprocal network due to our definition of an edge being a reciprocal action [24]. Reciprocity is already an indicator of online community success [173] and this type of representation allows for a more thorough examination into the group level aspects of reciprocity. Using replies to evaluate reciprocity meant that only community tools allowing for replies (Forums and Blogs) were included in this work. There can be multiple types of reciprocal relationships, one being a reply to an initial post and an additional type is that of a reply to a reply (nested thread structure). We are only interested in the pairwise interactions that come from post replies.

7.3.4.2 Traditional Network Measures

We capture many traditional network measures used in prior work on social roles [102, 207]. Network measures included simple population and activity metrics like:

- **Degree**: Number of edges
- **Nodes**: Number of nodes

But these are the basis for more complex metrics such as those examining Small-World Networks [118] for example:

- **Density**: Number of actual edges divided by number of possible edges.
- **Bridges**: A bridge within a graph is an edge in which if it is removed then no path exists between those two nodes.
- **Avg. K-Core**: A sub-graph in which all nodes within this graph have at least a degree of k, average k-core is the arithmetic average of the max k-core for all nodes in the network [102].
- **Triangles**: A group of three nodes in which all three nodes share an edge between each other [73].
- **Local Efficiency**: The efficiency of two nodes is the inverse of the shortest path between those two nodes. The local efficiency of a node is the arithmetic average of the efficiency of the subgraph induced by the neighbors of that node. To find the local efficiency of a network, just find the average efficiency of all nodes [118].

Previous work show that such network measures predict social roles [102, 207], consistent with the view that higher impact roles (leaders and social networkers) have strong effects on the overall network. Table 7.1 shows these differences for node level network measures of the roles examined here, along with distributional effect size measures (Hedge's G). As expected, owners have higher values for all network metrics, suggesting owners are more connected within the community than members. This in addition to other comparison made to typical role-specific behaviors identified in prior studies [35, 102], the roles within Communities were consistent with expectations and hence owners are similar to prior work definitions of leaders. Here we explore what group level activities different roles participate in and look at graphlets as potential structures associated with community leadership.

These measures were used to compare the frequency of 4-node graphlets as a potential predictor of success as they have foundations for being influential towards community success [102, 113, 173, 207]. Triangles were included as a comparison set for 4-node graphlets. The number of edges within each community heavily influences some of these measures (Degree and Bridges), therefore these measures were divided by the number of nodes in the network to normalize across communities. All measures were then z-normed before modeling

7.3.5 Graphlet Algorithms

Prior work has assessed methods for counting graphlets of various sizes within networks [8, 29, 185]. We use a proven method for counting graphlets [8] and calculate the percentage of all 4-node graphlet types within a community. However, graphlet frequencies and percentages do not allow the examination of the underlying makeup and interactions of graphlet types. Since we wanted to examine the effects of different graphlet types as well as distributions of roles within such graphlets, we needed to find the nodes that form graphlets.

Increasing the number of nodes within graphlets leads to combinatorial explosion; the number of possible graphlets increases factorial as nodes increase [8]. Therefore, we limited the graphlet finding algorithm to only Connected 4-node graphlets. We limited the finding algorithms to exclude unconnected 4-node graphlets due to computational restraints at scale, for instance, the number of Unconnected graphlets within a network of 400 nodes can range into the billions. Fortunately, connected graphlets are of more interest to the research questions as network connections have more theoretical associations with community success we wish to test here [102, 207] as well as being a more common size of teams within enterprises [86]. Additionally we limit our finding to only 4-node graphlets as 5-nodes also lead to combinatorial explosion and theoretical work also argues that higher level structures (e.g. 5 node and above) have weaker effects on community behaviors [65, 167]. Figure 7.1 shows 4-node graphlet types distinguishing Connected (all nodes have a path to each other) from Unconnected (at least one node does not share a path to another). Additionally fig. 7.1 shows the density and density category for only those Connected types, as this is relevant for later analyses. Graphlets will be analyzed through these bins unless they deviate from one-another which they will be reported separately.

We identified connected graphlets using the NetworkX graph objects and functions [87]. To find all graphlet types of interest, we first iterated through all edges in a graph, for each edge finding the adjacency lists for the nodes linked with that edge. This allows for detecting all connected graphlets except 3-Star, which is the only 4-node connected graphlet without an

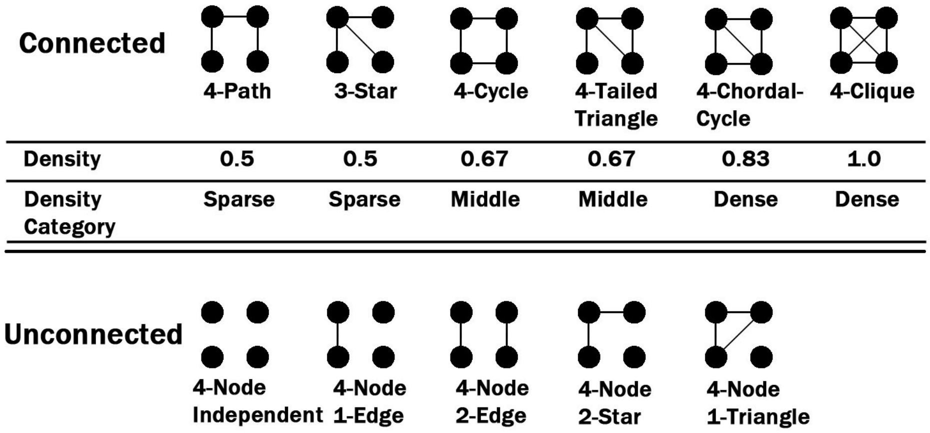


Figure 7.1: Configurations of all possible 4-Node graphlets broken into two groups, Connected and Unconnected. Graphlets are oriented from sparsest (missing edges) on the left to densest on the right. Only Connected Graphlets are binned into density categories, as further analyses cannot explore Unconnected.

internal 4-Path. More specifically, given an edge E with nodes n1 and n2, find a node n3 from the adjacency list of n1 and find a node n4 from the adjacency lists of n2. From here, gather all edges between those four nodes and classify the subgraph based on how many edges exist within the subgraph of all 4-nodes. For finding 3-Star graphlets, we found the subset of nodes from n1's adjacency lists that are not within n2's adjacency list, and found all pairs from that subset as they form a 3-Star graphlet.

Only unique sets of individuals were stored for each graphlet and in their densest possible structure. This is to avoid double counting as each 4-node graphlet exists within the denser form (i.e. all 4-node graphlets can exist within a 4-clique). Hence, we gathered groups of 4 that make a single graphlet type and there exists no overlap between the sets of graphlets we found.

7.4 Results

7.4.1 Graphlet counts are a significant predictor of Member Satisfaction

Using the percentage measure of frequency counts for each 4-node graphlet type in a community, we conducted a hierarchical modeling procedure with Member Satisfaction as the dependent measure. Recall that this percentage measure for all 4-node graphlets is the percentage of a specific graphlet type (4-Clique for example) in relation to the total number of 4-node graphlets within that community. Additionally, recall that we have Member Satisfaction data only for a subset of 93 communities.

We used a Support Vector Machine to capture non-linear relationships and found the radial basis function kernel to be the best fit compared to both a linear and polynomial kernel. Table 7.2 shows the results. Consistent with prior work [102], traditional network measures were a moderately good fit with an adjusted R^2 of 0.4034. Keep in mind that included with our network measures is the frequency of triangles within community networks, to compare 4-node graphlet proportions predictive power against size node size graphlet frequencies. However, the 4-node Graphlet percentages alone had a stronger fit than the Network measures (Adj. R^2 = 0.4717). To explore any possible redundancies between these apparently independent variables, Graphlet proportions were added to the Network measures. Adding Graphlets radically improved the model, suggesting too that variables are providing independent types of information.

While these models show Graphlets are a strong success predictor, they don't inform us about the influence each graphlet type has on Member Satisfaction. Therefore, we next explore connected graphlets in more detail, to get a better sense for the directionality of the

SVM Fit for Member Satisfaction	
Variable Set	**Adj. R^2**
Network Measures	0.4034
Graphlet Percentages	0.4717
Network + Graphlet	0.8276

Table 7.2: Hierarchical modeling of Member Satisfaction using traditional network measures, graphlet frequencies, and combined feature sets with a support vector regression and a radial basis function. Graphlets add a great deal of explanatory power to the network measures.

relationship between the connected graphlet frequency and member satisfaction, fig. 7.2 shows the median split of communities with high or low member satisfaction and their average proportional frequency for only connected graphlets. We binned graphlets by density (Dense $\geq$ 0.8 density, Sparse = 0.5 density), due to the differences between the middle density graphlets we keep them separate.

Density seems to relate to Member Satisfaction; more satisfied communities tend to have around 5 times a higher frequency of Denser connected graphlets (Note that fig. 7.2 has a log scaled y-axis). Less satisfied communities in contrast have 1.7 times more Sparse graphlets.

These results confirm and extend Faust [73] in demonstrating the value of substructure modeling as adding graphlet frequency proportions to the model increases predictive power. This is consistent with other approaches [10, 35, 67], in suggesting subgroup interactions contribute to member satisfaction. Caution should be exercised however with this interpretation as these are correlational not causal modeling analyses.

Having demonstrated the importance of these subgroups for success, we now explore how they might be having positive effects. To do this we: 1) explore a larger sample of communities to assess how graphlets relate to a different measures of community success, response

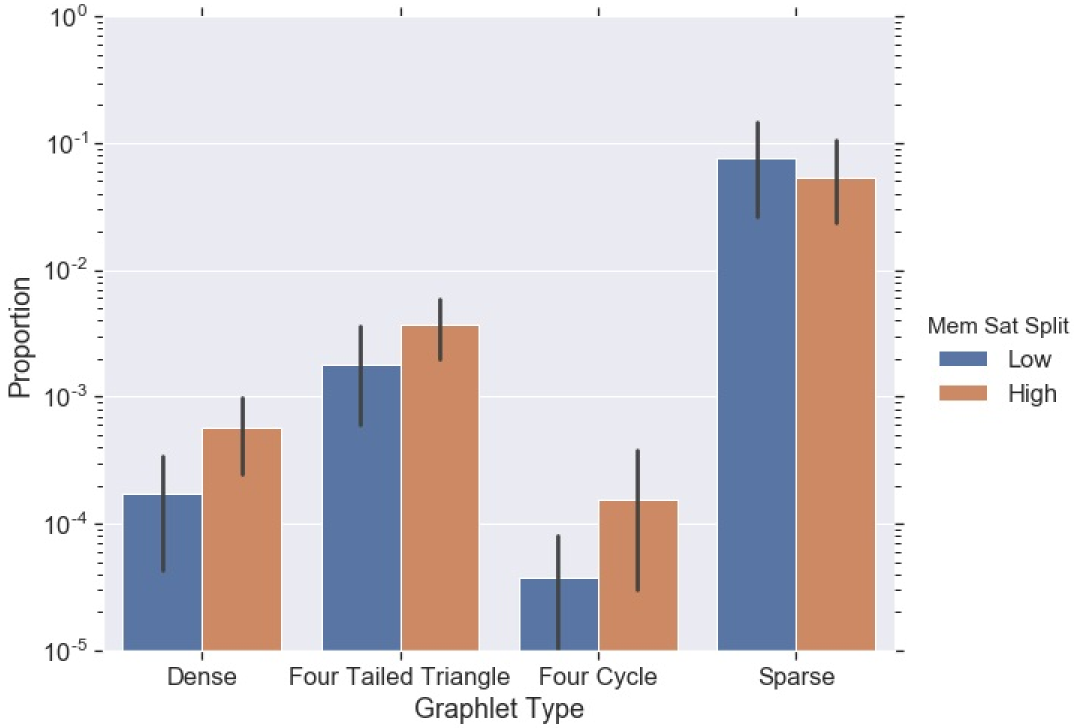


Figure 7.2: Avg. Proportion of Connected Graphlets for Low and High Member satisfaction communities, y-axis is log scaled; Error bars show 95% Confidence Intervals. Communities with High member satisfaction trended to have higher proportions of denser connected graphlet types. Low member satisfaction communities only had a slightly higher average proportion of sparser graphlets.

time [35, 67] and 2) examine the roles that participate in them.

As noted previously, finding graphlets in networks is computational expensive. For this reason, the analyses that follow are restricted to examining connected graphlets. Future work will be needed to explore unconnected graphlets further.

7.4.2 Responses are Quicker in Dense Graphlets

The previous analysis suggests relations between graphlet types and one measure of community success. The results are encouraging, although limited by the scope of the survey (which involved just 90 communities). To extend this, we used a different independent success factor. It is well known known that communities who respond to posts in a timely manner have higher levels of success [35, 67, 173]. Since each edge within the network is a reciprocal relationship (Post-Reply pair), we can measure the response time of the reply within that edge. We explore whether responsiveness relates to participants' connectivity within a graphlet across our whole sample of 2,000 communities. We use the density categorization shown in figure 7.2.

Fig. 7.3 shows the median response time for different graphlet density. Using an ANOVA we found significant differences across graphlet densities ($p < 0.0001$) indicating as expected that densely connected graphlets have faster response times. This suggests why communities with higher proportions of denser connected graphlets were considered more successful. Not only are they involved with more communication, the individuals involved respond on average 25% percent faster than individuals within a sparse graphlet.

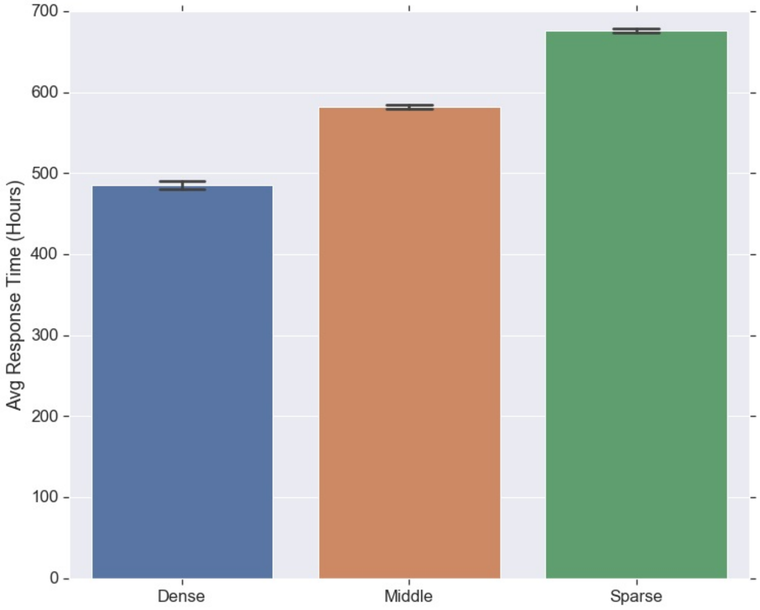


Figure 7.3: Median response time plotted over Connected 4-Node Graphlets based on graphlet density. Error bars are 95% confidence intervals.

7.4.3 Graphlets tend to include leaders

Recall that participants within the dataset have pre-specified community roles (namely Owner vs Member). We therefore explored how often community owners were present within graphlets. Since a graphlet can contain a range of 0-4 possible owners, fig. 7.4 plots this distribution of owners found within graphlets, showing graphlets tend to contain owners (On average 25% of graphlets contained no owners).

If we compare the ratio of owners to members across communities (On avg. 90 members to 1 owner), the finding that owners are involved in the majority of graphlets suggests a functional role for graphlets. Taking into account that owners have more posts this result makes sense. However this greater involvement may simply result from owners' higher connectivity and not leadership in which case it should also be observed in highly connected members. We

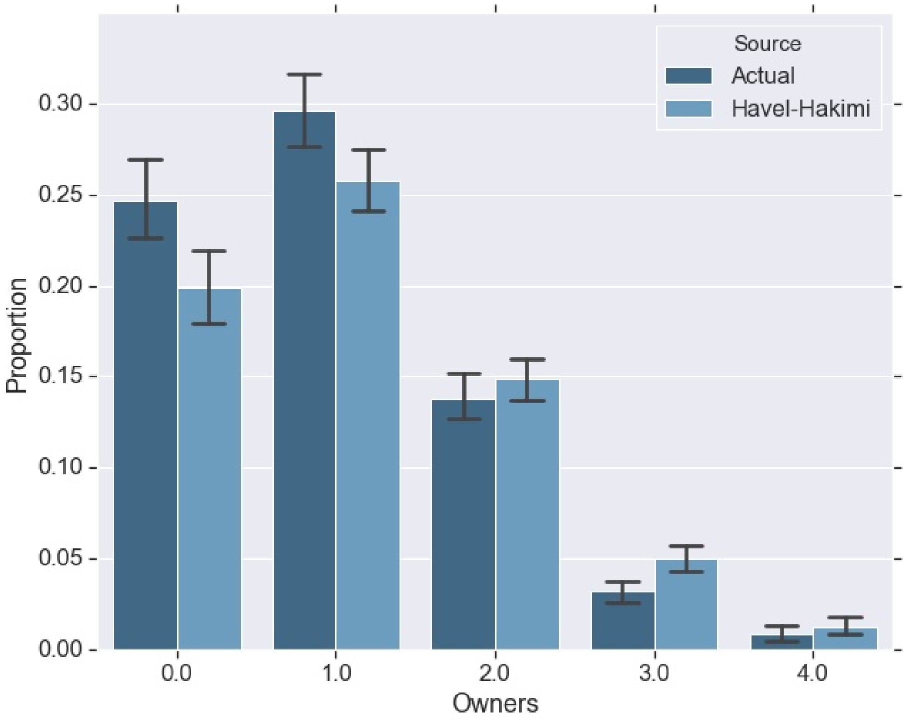


Figure 7.4: Average proportion of graphlets with a given number of owners (ranging from 0 to 4) compared to Havel-Hakimi random graph generated results. Error bars represent 95% confidence intervals across all communities; communities with zero graphlets were included thus all bars don't add to 1. Actual shows similar pattern to the Havel-Hakimi graphs as the majority of graphlets have at least 1 owner. Actual results show a higher frequency of zero and single owner graphlets and lower frequency of three and four owner graphlets.

therefore generated random graphs for all communities for members with a similar degree of connectedness to those of owners. This process was conducted through the random graph algorithms within the NetworkX library [87]. Two generation algorithms were chosen, first, a baseline model where graphs were generated through the algorithm presented by Miller and Hagberg [140]. This creates a graph with a similar node degree distribution to the graph provided as an input, meaning that this graph will contain nodes that have an equal degree (edges) to those within a graph representing natural relationships. The second used the Havel-Hakimi algorithm [87, 109] where nodes of highest degree are connected with other nodes of highest degree, thus introducing an assortative feature into the generated graphs. Assortativity is a common feature of online communities [47, 62, 118]. Assortativity argues that network subgroups should include similar roles, e.g. separate subgroups containing solely members or leaders but not a mix of the two roles. These two algorithms were chosen as they both allow for generation of graphs with matching degree distributions and allow us to test for possible assortative effects in high degreed individuals when comparing the baseline and Havel-Hakimi generated graphs to the actual online community graphs.

The baseline model was found to poorly fit the owner degree distributions of both our Actual data and the Havel-Hakimi generated data. Therefore, fig. 7.4 shows only in addition, the distribution of “owner”-like nodes within the Havel-Hakimi generated graphs. The the Havel-Hakimi graphs showed a similar trend to that of our actual data, indicating it is more common for there to be at least one “owner”-like node within each graphlet and indicating that this owner distribution is partially explained by Assortativity. However there was a significant interaction between the Actual and Havel-Hakimi results in terms of the number of owners when compared through an ANOVA (p <0.0001). This can be seen in fig. 7.4 where the Actual data shows higher occurrences of graphlets with zero and one owners, but fewer occurrences of graphlets with 2 or greater owners. This implies that owners are connecting more with members than other owners compared to the expectations from assortativity.

7.5 Discussion

We make both methodological and theoretical contributions to online communities research. We demonstrate that graphlets offer significant insight into subgroup structure, as well as explaining global aspects of community interaction. Methodologically, we show that graphlets offer one approach to identifying functional subgroups in communities. These subgroups were significantly related to measures of online community success, suggesting their explanatory validity. At a theoretical level, these results contradict prior claims [14, 36, 72] arguing that global social network measures alone are sufficient to explain community behaviors. Including graphlet frequencies significantly improved models of member satisfaction when compared to standard social network measures, with graphlets explaining an additional 40% of the variance. This suggests subgroups are significant determinants of success, which is consistent with theories of group behavior arguing that overall community accomplishments arise from combinations of small group interactions [167]. Furthermore, while social theorists have argued for the importance of subgroups in explaining community behaviors [167], prior work has lacked analytic methods to reliably identify and explore such structures.

More supporting evidence for the utility of graphlets is provided by analyses showing that graphlet structures directly predict other measures of community behavior, including response time. Finally, contradicting prior work on assortativity, most graphlets contain exactly one leader, suggesting that graphlets offer a way for community leaders to influence and interact with a small group of significant members. Assortativity is a common feature of online communities [47, 62, 118]. Assortativity argues that network subgroups should include similar roles, e.g. separate subgroups containing solely members or leaders but not a mix of the two roles. Instead we found that graphlets tended to included mixes of both leaders and member roles. Why might this be the case? It may be that participating in small heterogeneous subgroups allows leaders to create the weak-tie relationships with regular community members that are known to be critical for community success [134]. This is consistent with other work showing that communities are successful if leaders are well networked to many community members)[20, 99, 102]. Future work should explore the content of the interactions within these heterogeneous graphlets to assess the nature of these weak-tie interactions.

Overall this work suggests that graphlets capture important network subgroups. The results suggest important questions for future work relating to the structural properties of graphlets; could it be that sparser graphlet types represent low priority communications? Could denser graphlets indicate stronger bonds as expected from past work on cohesion? Furthermore are different graphlet structures associated with different types of community language behaviors, and how do graphlet structures evolve over time?

Design Implications: Other research demonstrates the utility of community dashboards in allowing leaders to diagnose and steer overall community interactions [136, 189]. We have shown that specific graphlet structures are associated with key community behaviors, e.g. faster response times. New tools that identify such productive subgroups could be incorporated into community dashboards, allowing leaders to identify opportunities to modify and optimize community behaviors around those subgroups. For example, tools identifying that sparse subgroups are prevalent in a given community, would allow a leader to encourage community behaviors that foster denser subgroup connections, that we have shown to predict faster response times.

These results also have important practical ramifications informing community leader practices. Based on our findings, leaders might intentionally instigate community interaction policies that promote certain community subgroups that are associated with productive behaviors. Or leaders could aim to modify cases where they communities are engaging in less adaptive subgroup interactions.

In summary, this work successfully applies a method for exploring substructures in networks, showing that these have considerable value in explaining social phenomena in online communities. We develop new theoretically motivated network measures that take into account the complexity of subgroup interactions at a level rarely explored in online community research, suggesting new methods and questions for future online community research.

Chapter 8

Interactions in Community Subgroups

8.1 Introduction

Online communities provide support via weak ties that is often unavailable through existing strong tie networks [83]. Early communities work documented the dyadic processes by which individual community members pose questions and receive answers from domain experts [91, 172, 193]. However, it is apparent that there are other important multiway relationships existing above these dyadic connections. For example, a three-way triadic relationship frequently arises when a community leader recommends that a new member with a topical question consult a third person who is an expert on that topic [56, 57, 73, 94]. Or in a peer production context, a leader might negotiate with multiple volunteers who form a small team adopting differing roles to address a community related task [100, 117, 130, 177]. Such interactions give rise to complex small group networks and these relationships remain to be thoroughly explored quantitatively.

A great deal of research has examined network representations of online communities

[31, 47, 73, 96, 181, 196, 216]. However there has been little work exploring intermediate subgroups within communities. The network sciences has developed methods for identifying and contextualizing such structures, known as Graphlets [8, 9, 29, 122, 185, 198]. Chapter 7 showed a strong relationship between connected 4-node graphlets and known measures of online community success was shown. This chapter extends this result by exploring whether there are differences in content between different graphlet types and whether specific graphlet conversations may predict community success.

Recent work has explored community interaction examining how the content of members' conversation affects success [63, 151, 191, 202]. In some communities, emotional support for other members is critical while others benefit from group supportive language [134, 204]. Interpersonal and group focused conversations may both promote long-term relationships between members [137, 146, 208]. More recent work has looked into how individual's posting content changes over time compared to their community. It shows that tenured members are more likely to leave if they begin to deviate from community linguistic norms [55]. Chapter 7 showed that graphlets were important in predicting community success, making it vital to understand underlying mechanisms by which graphlet subgroups promote success. Examining the content shared across different graphlet types may shed light on these mechanisms. This chapter therefore looks to explore connected 4-node graphlets in using standard content analytic methods, and success metrics. We conduct three related evaluations looking at linguistic content, future-orientation and aimed to categorize graphlet types.

First, we conduct a Machine Learning experiment using the content of graphlets aiming to classify different graphlets types. This model has good performance indicating graphlets to vary in their content. Interestingly, the main features are found to be those of content LIWC categories, while more generic or context relevant features (Punctuation or Work words) are least predictive.

One aspect of online community success is the generation of content which draws people into a community. We explore this through simply looking at amount of content within a post. Prior work has identified successful communities based on the population generating content [173]. While word count is rather simplistic to imply success, as there is a potential that many words can indicate negative communication (flame war), this is a first step in identifying potential differences that may be why such subgroups (graphlets) are more successful than others. Additionally, as this is an enterprise context where users are known to each other, the level of malicious behavior is low. Other more linguistic markers of content, identified using LIWC [190] are associated with community success [134], for example communities with more first person plural and assent (‘agree’, ‘yes’, ‘Ok’, etc) uses show higher member satisfaction [134]. This may result from active leadership behaviors to develop a stronger community identity [102]. Or more inclusive word use might help maintain and build relationships [113].

We also explore content through a temporal lens. Communities change over time [99, 113] and successful individuals are known to post more relevant content and drive community topics [55, 181]. Prior work [55], developed a new measure evaluating the content of a post to be either Progressive or Conservative, by examining the similarity between the content of the post and subsequent or preceding posts. This measure assumes that content that is more similar to future posts is more relevant and driving the community forwards [55, 113].

8.1.1 Contributions

This work explores community subgroups and their functions by examining the content associated with different graphlet types. We also explore relations between graphlet types and community success metrics. Linguistic analyses show that denser graphlets are associated with more verbose posts containing larger word counts. Furthermore connected graphlets were found to contain more progressive (i.e. future-focused content), suggesting these graphlets are driving the community topics of conversation. Counterintuitively having more leaders in a graphlet led to less inclusive language and were more present focused, but as expected more leaders led to more group oriented language. This may be an indication of conversations being held by leader-to-leader or member-to-member post-reply pairs. Leader-to-leader pairs may be more focused on the present to address current community needs while member-to-member pairs are leading community topics and building relationships.

8.1.2 Content Exploration of Online Interactions

Network representations provide a rich set of metrics that can help operationalize social theory. Rowe [181] used in-degree and out-degree distributions to measure social dynamics to model likelihood of individuals leaving a community, finding that number of replies (in-degree) was a significant predictor [181]. More relationship-based measures focus not on dyadic one-to-one interactions (like degree) but instead model one-to-network interactions. Those interactions identify high connection, influential nodes within the graph, and one such measure is Centrality [20]. Nolker and Zhou [153] used multiple relationship-based measures including degree, betweenness, and closeness to identify various role types in Usenet communities, finding more impactful roles (e.g. leaders and motivators) to have significantly higher network relationship measures. Johnson et al. [102] also explored relationship-based measures in modeling role behaviors, finding leaders to be associated with k-core, a measure quantifying the network of nodes with at least k degree. Sparrowe et al. [187] examined workplace relationships using a network representation to explore the relationship between centrality and task behaviors. Performance was positively related to centrality in cooperative networks with the opposite being true for uncooperative cases.

A number of researchers have explored the effects of language use on online community members' behaviors [134, 142, 146, 152, 204]. Matthews et al. [134] used the Linguistic Inquiry and Word Count tool (LIWC) [190] to understand what community linguistic behaviors predict member satisfaction in enterprise communities. They found simple linguistic predictors of member satisfaction that included the use of inclusive language, low anger and increased anxiety.

Some prior work has explored how content measures and social relationships relate to online leadership, but we are unaware of any work that explores content within intermediate structures of online community networks. Prior work has highlighted the importance of all aspects of reciprocity, number of words, number of replies, and social connectedness. However this chapter looks to expand on this by providing a more detailed understanding of exactly how small groups communicate and are influenced by leadership presence.

Prior work has also identified substructures and their relevance towards social properties such as triadic tendency which was detailed in Social Exchange Theory [20, 54, 73]. Network science tends to refer to these structures as graphlets or motifs. One example of a subnetwork type is a triad (involving three nodes) and their many different possible configurations like connected (two edges) versus a clique (three edges). These substructures have proven valuable in evaluating social relationships within networks, i.e. transitivity of a network being the proportion of triads forming cliques has implications to the social ability and likelihood for connections to form [20, 24, 73]). These substructures can also explain leader effectiveness, friendship formation and the community context as well as network evolution over time, [16, 20, 54, 62, 73, 163, 179]. Prior work has also identified significant patterns involving triads [24, 73] and Faust [73] showed triad census (frequency measure within a network of all the existing configurations of a group of three, for example three nodes having three edges is one possible configuration of a triad) exceeds expectations from dyadic census (frequency of possible dyad configurations in directed network), indicating that substructures provide information that is not captured by such lower-level measures. Dong et al. [62] also recently examined homophily by exploring various substructures. Their findings are rather complex in that existence of two nodes within a similar substructures varies in its implications if those two nodes share a connection with one-another. They find in certain contexts the presence of shared structures is positively associated with connects but negative in others implying different network context properties. Either way, that work highlights a relationship in how individuals form connections. We expand on that work by more deeply examining social roles (i.e community leaders versus members).

Other work has looked at how content evolves over time. Danescu et al. [55], examined how the content of individuals' posts change in relation to the rest of the community, primarily finding that older members find themselves making progressively less relevant posts until they leave the community entirely. To characterize this, they introduce new linguistic measures of content; Linguistic Progressiveness/Conservativeness. Both compare a given post with the content of subsequent and antecedent posts in the community. For a post to be Progressive, it must contain more similar words to subsequent content, whereas Conservative posts are more similar to antecedent content. While this measure can be categorized into two groups, it is primarily a continuous measure on a scale of -12 to 12, negative indicated Conservative and positive indicates Progressive. This metric provides a method to examine how the post content potentially impacts a community, with Progressive posting indicating that the poster is setting the content agenda for the community. We use this metric to explore content differences between graphlet types.

8.2 Methods

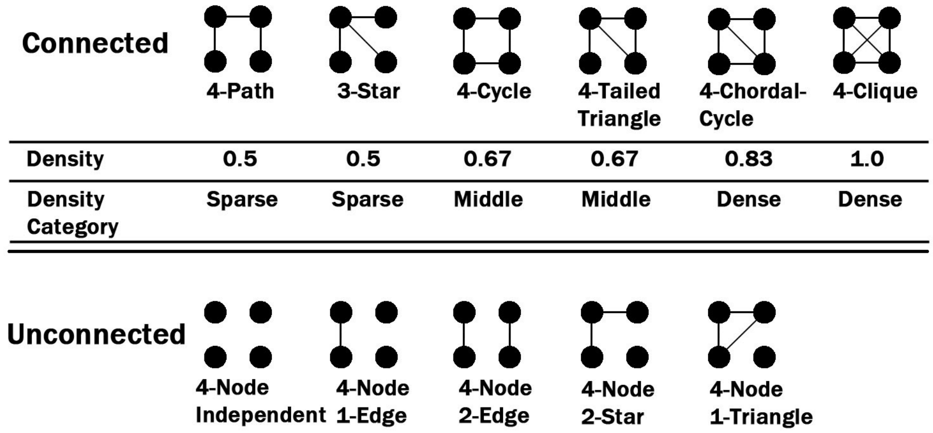
8.2.1 Online Communities as Networks

Networks metrics were derived for each of the 2000 communities using the NetworkX library within a Python 2.7 environment [87]. Each user in the online enterprise community was considered a node and edges were found through measuring a reciprocal relationship between nodes. Reciprocal relationships are defined as exchanges between two entities, which we operationalized as when a user responds to another user's post. We refer to this network as a reciprocal network due to our definition of an edge being a reciprocal action [24]. Reciprocity is already an indicator of online community success [173] and this type of representation allows for a more thorough examination of group level aspects of reciprocity. Using replies to evaluate reciprocity meant that only community tools allowing for replies (Forums and Blogs) were included in this work. There are potentially multiple types of reciprocal relationships, in addition to a reply to an initial post, there are more complex relations, e.g. reply to a reply, or a reply to a reply to a reply (nested thread structures). To simplify our analysis here we focus on simple pairwise interactions (whether a given post replied to a prior one) and do not include these more complex reciprocal structures.

8.2.2 Graphlet Algorithms

Prior work has developed methods for counting graphlets of various sizes within networks [8, 9, 29, 185]. We use a proven method for counting graphlets [8] to calculate the percentage of all 4-node graphlet types within a community. However, simple overall measures of graphlet frequencies and percentages do not allow us to examine the underlying structure and content within graphlet types. Since we wanted to examine the effects of different graphlet types as well as distributions of roles within such graphlets, we needed to identify specific graphlet instances.

While there are methods to identify 3-node graphlets, Increasing the number of nodes leads to combinatorial explosion; the number of possible graphlets increases factorially as nodes increase [8]. Therefore, we limited the graphlet finding algorithm to Connected 4-node graphlets. We therefore excluded unconnected 4-node graphlets as the number of Unconnected graphlets in a network of 400 nodes can potentially range into the billions. Theoretically however, connected graphlets are of more interest to our research questions as denser connections are argued to promote community success [102, 207]. Additionally, limiting to 4-node is still



	4-Path	3-Star	4-Cycle	4-Tailed Triangle	4-Chordal-Cycle	4-Clique
Density	0.5	0.5	0.67	0.67	0.83	1.0
Density Category	Sparse	Sparse	Middle	Middle	Dense	Dense

Figure 8.1: Configurations of all possible 4-Node graphlets broken into two groups, Connected and Unconnected. Graphlets are oriented from sparsest (missing edges) on the left to densest on the right. Only Connected Graphlets are binned into density categories, as further analyses cannot explore Unconnected.

theoretically interesting as prior work has reported this being a more common size of teams within work groups [86]. For the same computational reasons we also limit our finding to 4-node graphlets as 5-nodes lead to further combinatorial explosion. In addition, theoretical work argues that higher level structures (e.g. 5 node and above) have weaker effects on community behaviors [65, 167]. Figure 8.1 shows 4-node graphlet types distinguishing Connected (all nodes have a path to each other) from Unconnected (at least one node does not share a path to another). Additionally fig. 8.1 shows the density and density category for only those Connected types, as this is relevant for later analyses. Graphlets will be analyzed through these bins unless they deviate from one-another which they will be reported separately.

We identified connected graphlets using the NetworkX graph objects and functions [87]. To find all graphlet types of interest, we first iterated through all edges in a graph, for each edge finding the adjacency lists for the nodes linked with that edge. This allows us to detect all connected graphlets except 3-Star, which is the only 4-node connected graphlet without an internal 4-Path. More specifically, given an edge E with nodes n1 and n2, find a node n3 from the adjacency list of n1 and find a node n4 from the adjacency lists of n2. From here, we gather all edges between those four nodes and classify the subgraph based on how many edges exist within the subgraph of all 4-nodes. For finding 3-Star graphlets, we found the subset of nodes from n1's adjacency lists that are not within n2's adjacency list, and found all pairs from that subset as they form a 3-Star graphlet.

Only unique sets of individuals were stored for each graphlet and in their densest possible structure. This is to avoid double counting as each 4-node graphlet exists within the denser form (i.e. all 4-node graphlets can exist within a 4-clique). Hence, we identified groups of 4 that make a single graphlet type without overlap between the different graphlet types.

8.2.3 Content Analysis

Using this graphlets discovery algorithm, we extracted the content associated with each edge. Content can be rather complex within the edges of the reciprocal network. As edges are established as a post-reply format, each edge has at least 2 content contributions. As there is no limit to how many times a pair can reply to each other, this may include many post-reply pairs. For each linguistic measure in this Chapter, we gathered measures on each piece of text (Post and reply, which there can be more than one post and reply) within an edge and average each measure to summarize that edge. Hence this has a normalizing effect so any difference in edges won't be based on the effects of repetitive post-reply pairs and instead focuses on the content within post-reply pairs.

We first conduct a machine learning experiment exploring whether graphlet types are distinguishable in their content. This is similar to prior machine learning experiments described in this thesis. We create a train-test split, cross validation is used for training more generalizable models, and parameter searching to find the best models. We use an XGBoosted Random Forest [44] as a model to explore which features are most important to distinguishing graphlet types.

We explore differences in graphlets through various linguistic measures here. We begin with a simple measure of word count per post to see if there is a relationship between graphlet types and a known (but approximate) measure of online community success [173]. We follow up with purely lexical measures using LIWC which offer more psychological assessments of content associated with graphlet types. Lastly, we examine the measure introduced by Danescu-Niculescu-Mizil et al. [55] called Linguistic Progressiveness . This measure explores how similar content is to antecedent and subsequent posts within the community through the cosine similarity score [95]. This measure aggregates the cosine similarity of pairs of posts across time. Linguistic Progressiveness indicates that the post is more similar to subsequent posts, while Conservativeness indicates the post is more similar to its antecedents. We use this measure as an indicator of forward thinking to determine which individuals are anticipating the future directions of discussions within the community. As mentioned before, this is a continuous measure on a scale of -12 to 12. For clarity, we will refer to a measure in the positive end to be Future and a measure in the negative end to be Past.

8.3 Results

8.3.1 Graphlet types are distinguishable by content

To see if graphlets were distinguishable from each other, we conducted a machine learning experiment using the LIWC categories as a feature set. For our train-test splits, we wanted to be sure no possible information is being shared between the splits, therefore we conducted a split over a sample of 574 communities using an 80-20 train test split. This type of split ensures that no graphlet can share similar content with each other as the train set is within different communities than the test set. Additionally since we are looking at content across communities, each LIWC measure was z-normed with respect to the community they were within. Overall we identified 61,152 graphlets in this experiment.

Fitting an XGBoosted Random Forest to the train set, we obtained a good F1-score of 0.81, which was found to be consistent across all graphlet types. This shows that each graphlet type is distinguishable based on contents amounts. The result is generalizable too as the test set involved communities not seen in training.

Table 8.1 shows the top and bottom 10 features ranking according to the feature importance measure within XGBoosted Forests. Interestingly, content based categories of LIWC appear to give the model the most information (left hand side of table 8.1). Graphlets seem to be mainly discernible through the content they are discussing rather than function words they used, as more common function categories such as verbs, tense, and prepositions aren't very helpful (right hand side of table 8.1). The one outlier is the content category work which is a poor predictor of graphlet type. However it isn't surprising that this has such a low signal as

Top 10 Features		Bottom 10 Features	
Feature	Relative Importance	**Feature**	Relative Importance
Swear Words	0.0567	**All Punctuation**	0.0
Religion	0.0425	**Common Verbs**	0.0030
Assent	0.0348	**Present Tense**	0.0031
Death	0.0319	**Prepositions**	0.0035
Family	0.0307	**Work**	0.0041
Sexual	0.0292	**Quote**	0.0042
Ingestion	0.0224	**Comma**	0.0046
Anger	0.0220	**Exclusive**	0.0047
Leisure	0.0219	**Past Tense**	0.0048
Exclamation Mark	0.0194	**Relativity**	0.0049

Table 8.1: Feature Importance rankings for the features used within the XGBoosted Random Forest. Left shows the Top 10 features as ranking by relative importance while the right shows the Bottom 10 features. Relative Importance indicates the amount of information a feature provides in determining differences in the classes the model attempts to predict.

these are enterprise focused communities therefore many of them are discussing work matters. Overall the strength of linguistic content features is consistent with section 8.4.1's findings that graphlets contain high degrees of community content.

8.3.2 Denser graphlets contain more words and more progressive content

We explored content using multiple measures. First we examined the raw amount of content in posts, to examine whether different graphlet types generated different amounts of content. Content was normalized by community to allow for the fact that different communities posted differing amounts of content overall. To better understand relationships between content and graphlet types, we first categorized each graphlet into one of three different groups based on the density of edges within the graphlet, similar to the density classification described in Chapter 7.

Fig. 8.2 shows the average normalized word count per post for each graphlet density.

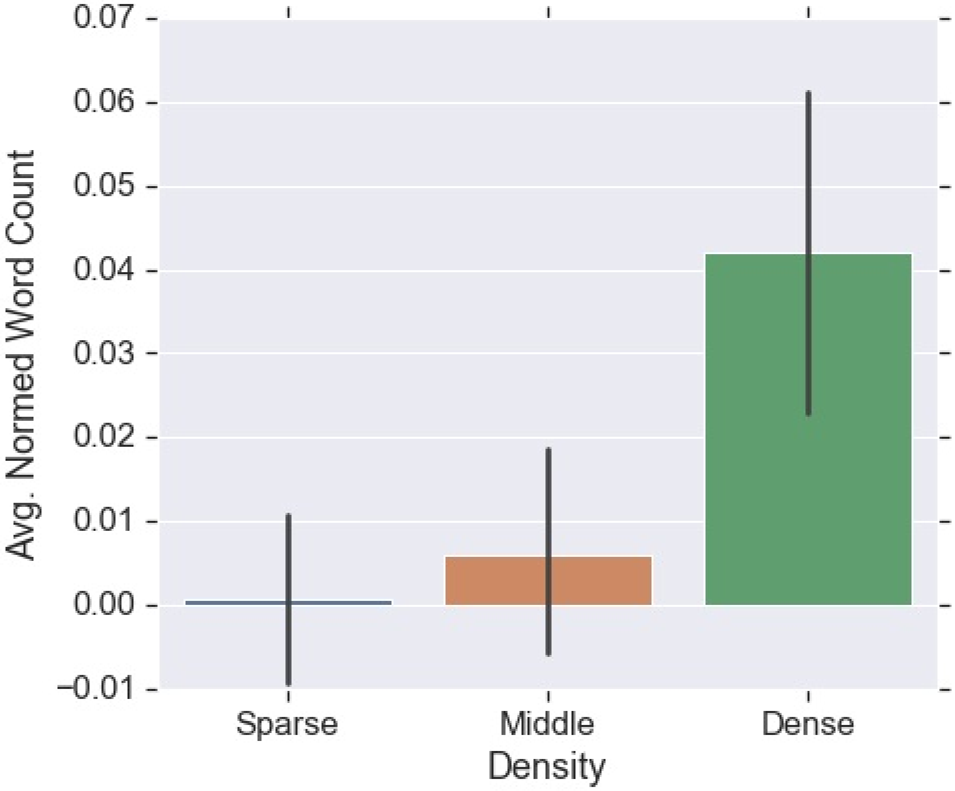


Figure 8.2: Normalized Word Count by graphlet densities. Denser graphlets tend to have greater word count per post. Error bars indicate 95% confidence intervals.

This figure shows a clear positive relationship between content volume and graphlet density, which was found to be statistically significant, ANOVA (F = 7.98, p <0.001). This confirms our expectation that denser graphlets should produce more discussion.

Although volume has been proposed as a measure of online community success [173], but it is clearly a highly imprecise assessment as highly volumes can arise for multiple reasons [210]. We therefore explore a more nuanced measure of content in communities, the Linguistic Progressiveness metric [55], to explore Future and Past oriented posts.

Fig. 8.3 shows Linguistically Progressiveness for graphlet densities. Overall all com-

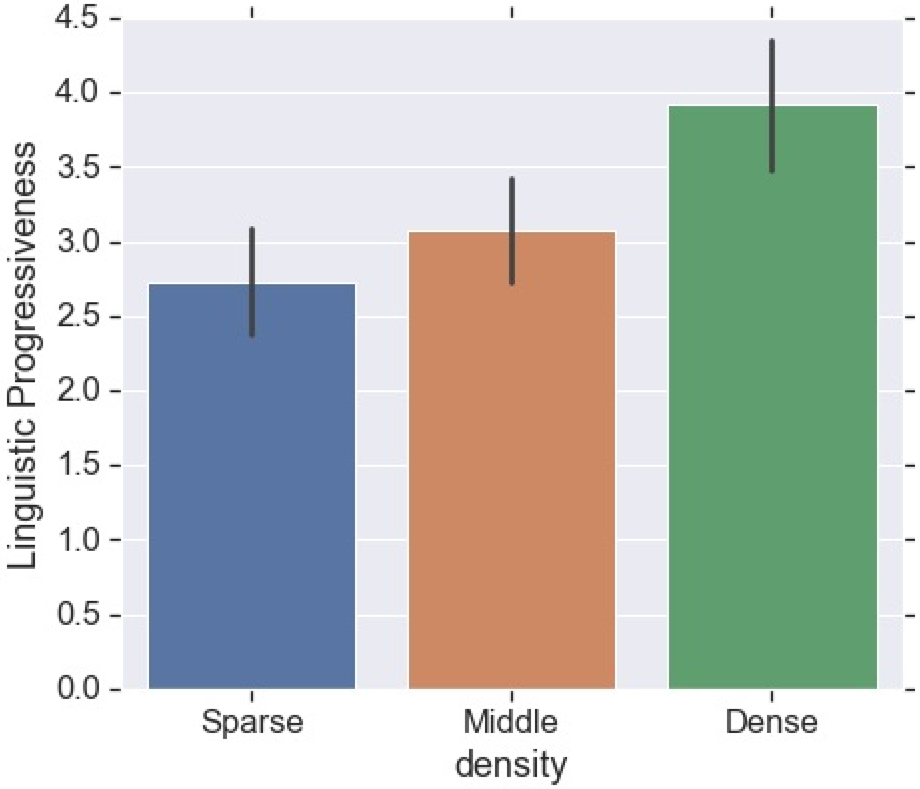


Figure 8.3: Linguistic Progressiveness by graphlet density. All graphlet types show conversations are future rather than past oriented. However denser graphlet types have higher future orientations, indicating that denser graphlets are driving the community agenda

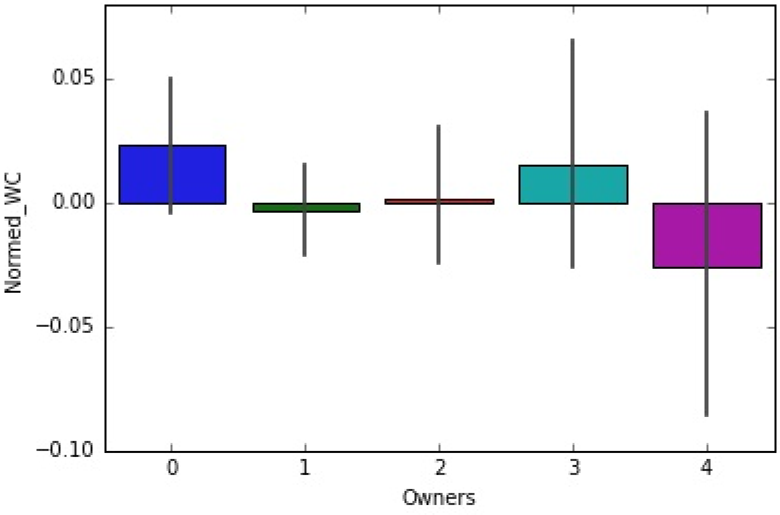


Figure 8.4: Normed Word Count for each post plotted against number of owners in a graphlet. No relationship was found between the number of owners and normed word count. While there is a suggestion that graphlet with 4 owners post less content, this was not statistically significant.

munities are future focused but as with word count (fig. 8.2), denser graphlets are more future oriented ANOVA (F = 8.34, p <0.001). This suggests that denser graphlets drive may online communities being more likely than sparse graphlets to introduce relevant content.

8.3.3 Having more Owners reduces content volume, and induces less future oriented thinking

We now explore the effects of owner presence on word count. We again categorize graphlets by the number of owners that are present. As we are only looking at graphlets of size 4, there can be a range of 0 to 4 possible owners within a graphlet. First we explored normed word count per post, in relation to owner presence. We found almost no association between number of owners and word count. The figure suggests that graphlets consisting solely of owners (4 owners) actually have less content in their posts with respect to the community.

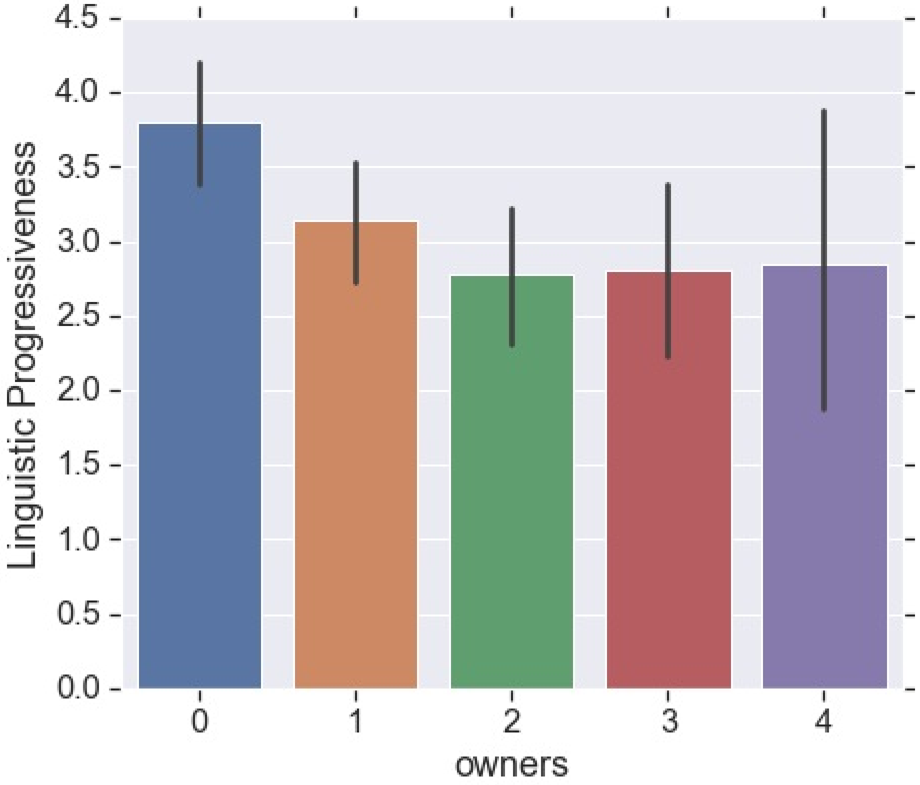


Figure 8.5: Linguistic Progressiveness by number of owners present within graphlet. An early negative trend is found with the amount of linguistic progressiveness and the number of owners being present. The most progressive graphlet was that which contained no owners.

However this was not significant in an ANOVA examining effects of number of owners (F = 1.04, p = 0.38).

Next we explored the relationship between Linguistic Progressiveness and number of owners in a graphlet (see fig. 8.5). Counter to expectations, having more owners appears to reduce the Linguistic Progressiveness of graphlets (F = 3.34, p = 0.009). This implies that having more owners in a graphlet reduces the future orientedness of posts, although posts overall tend to be future oriented.

8.3.4 Having more Owners increases Inclusive Language

Prior work by Matthews et al. [134] showed relationship between certain Linguistic Inquiry and Word Count categories and member satisfaction in online communities. We choose to explore if there were any associations with these linguistic markers and graphlets to further explore why graphlets might predict success, as suggested in Chapter 7.

Matthews et al. [134] showed First Person Plural and Singular word usage are closely correlated with community member satisfaction, which in turn was highly related to overall community success. More usage of first person plurals (‘we’, ‘our’, ‘ours’) was correlated with high member satisfaction and first person plurals (‘I’, ‘my’, ‘mine’) with low member satisfaction. To evaluate such usage we developed a combined metric by taking the percentage of Plural words minus the percentage of Singular words divided by the total use of Plural and Singular. As Matthews et al. [134] showed opposite effects of Plural vs. Singular first person use in relation to Member Satisfaction, this acts as an aggregated metric as the valence of the metric (Positive or Negative) indicates its relation to Member Satisfaction. This is also a practical single metric because speakers are usually making a direct choice between one expression or the other.

Fig. 8.6 shows a discrete differences between when a leader is present and not present. When no leaders are present, there is a stronger use of First Person Singular. One a single leader is presence within a graphlet though, the content is more First Person Plural focused, which stays around the same level regardless of the amount of leaders. An ANOVA showed these differences to be significant (F = 15.54, $p < 2e\text{-}12$).

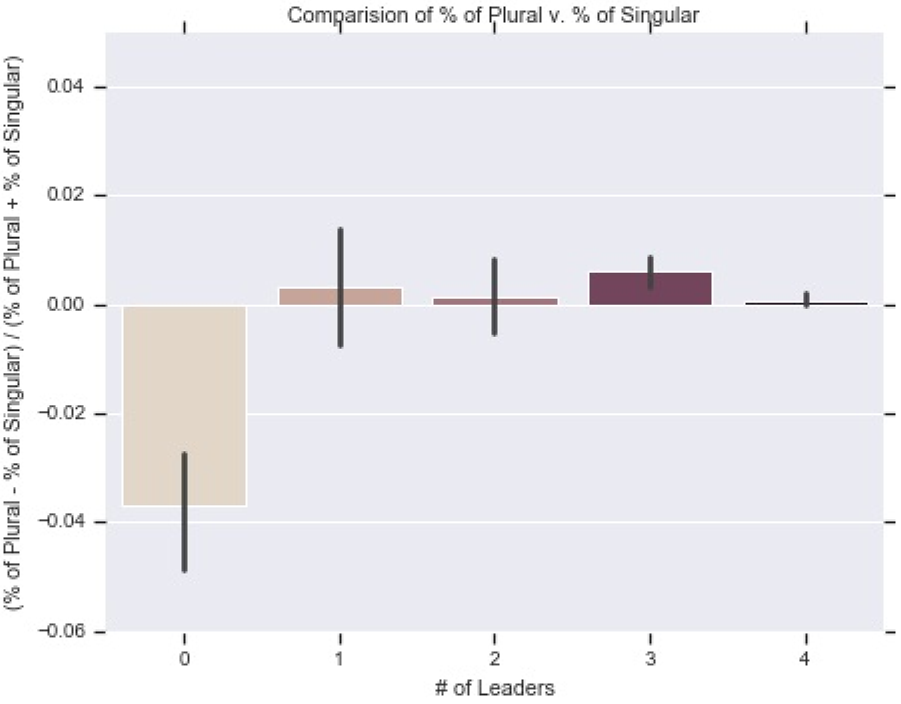


Figure 8.6: First Person Plural-Singular word use plotted against the number of leaders in a graphlet. Graphlets with 0 or 1 leader are more likely to use singular pronouns with equal numbers of singular and plural pronouns in graphlets with more leaders.

8.4 Discussion

This chapter expands on the methodological contributions of Chapter 7. It identifies differences between graphlet structures and the types of content associated with them. Dense graphlets have higher word counts per post and more progressive language [55, 173]. This suggests denser subgroups may contribute to community success, which is consistent with theories of group behavior arguing that overall community accomplishments arise from combinations of small group interactions [167]. Furthermore, while social theorists have argued for the importance of subgroups in explaining community behaviors [167], prior work has lacked analytic methods to reliably identify and explore such structures.

More supporting evidence for the utility of graphlets is provided by analyses showing that graphlet content has direct relationships with the presence of leadership roles. We found contrary to our expectations, that fewer leaders involved in subgroups was associated with higher levels of inclusive language. As leaders are expected to be ambassadors of a community, by helping onboard newcomers [113], we would expect more inclusive language to be when they were involved. Why then might this be the case? It may be that within enterprise communities we see responsibility have dispersed as team cohesion is important to team success [131]. However, we confirmed that more leaders were associated with greater usage of plural first person pronouns, consistent with the view that leaders help instantiate a stronger community identity [171]. Prior work has also found such language to be associated with more Member satisfaction [134]. Interestingly, as this was seen most in posts from heavily leader dominated graphlets, perhaps such leader-to-leader discussion is more beneficial to the community as a whole? Future work is needed to understand potential contextual differences within such reciprocal forum threads.

Lastly, we were able to show that graphlets are distinguishable from each other. Using LIWC as a feature set, we found a good model to distinguish graphlet types. Surprisingly, we found that LIWC content categories were the most informative, implying graphlets vary significantly in how much focus to a given subject. Not surprisingly, we see more common features like punctuation types and the Work LIWC category as the least informative. This may reflect the enterprise context in this sample. One future research topic might be to identify subgroup interactions that are associated with community failure and propose interventions to redress such problems.

This work has additional technical implications for online community tools. Participants in sparser graphlets might be encouraged to focus on more future focused content or leaders could be encouraged to reply to such a subgroup.

Chapter 9

Discussion and Future Work

9.1 Summary

This thesis extends our existing understanding of social roles in online communities to the novel domain of enterprise communities. In the context of existing community theories, we examine long term content management, temporal changes in social roles, develop new metrics identifying community subgroups and show that such subgroups contribute to community success. Each study offers empirical quantitative evidence advancing our knowledge. We first summarize major contributions and then discussion limitations.

9.1.1 Content Management and Linking

This chapter quantitatively characterizes one important aspect of content management using reference links, examining both long-term changes and role differences. It contributes to existing literature by exploring the following questions: First, how does referencing using hyperlinks change over time and how does this relate to content creation? Are links more prevalent over time as content builds up and does accumulation of content lead to increased referencing? Second, as content accumulates, who takes responsibility for management: members or leaders? Following current community lifecycle models, do members assume more responsibility for content referencing over time? Our findings are counterintuitive. First, active content referencing does not increase as content accumulates and second, contradicting lifecycle models, members never assume full responsibility for referencing. Content analysis suggests that recency bias is a possible reason for the absence of such referencing. We suggest new tools and community building practices that better support content management taking these findings into account.

9.1.2 Evaluating Role Models

This chapter develops new statistical models for evaluating role shifts over community lifespans. Our findings contradict apprenticeship models [171] which argue that some community members change their roles over time to adopt more responsibility for community management. Instead following Panciera et al. [159], we found that most community members retain their initial roles. We discuss the theoretical implications of these findings.

9.1.3 Emotional Language Use

We develop an emotionality detection algorithm and then use it to evaluate how community language use affects perceived community success. The results show a small but clear relationship between the use of emotional versus factual language in enterprise communities and member satisfaction. As predicted, factual language enhanced perceived satisfaction. Furthermore, counter-intuitively, emotional language use reduced satisfaction in communities of practice where social relations and emotional support are thought to be important.

9.1.4 Subgroups: Identification and Impact

Research on community roles has generally focused on individuals and dyads or characterized network relations for the entire community. Such prior models fail to explore complex subgroup structures and how different roles interact within such structures. We incorporate graphical methods to identify substructures within a larger network. We explore how these 4-node graphlet structures relate to individual roles and content production, and examine whether these substructures predict metrics of online community success. Graphlets have considerable explanatory power improving network measures by 16% in predicting community success. Furthermore, graphlets are more likely to contain leaders in influential positions within the network.

9.1.5 Graphlet Content Analysis

The previous chapter finds that specific graphlets are predictive of online community success, yet these correlational approaches don't shed light on what makes these substructures important to communities. We analyse how content production relates to graphlet type. First, machine learning models indicate that different graphlets produce different content types. Other analyses find that denser graphlets contain more content per post, and posts from denser graphlets also introduce anticipatory content that will later become more prominent within a community. Surprisingly however graphlets without leaders generate more content and anticipatory content.

9.2 Future Work

We now discussion overall limitations of the thesis, using these to motivate future work. One limitation is that the thesis is reliant on large scale data analyses using statistical methods. However qualitative approaches could extend our data by indicating important underlying community mechanisms implied by quantitative methods, as well as suggesting hypotheses for future quantitative evaluation. Our qualitative content management analysis showed the value of this approach in identifying functional differences between link types, allowing our quantitative analysis to target key referencing behaviors. With the exception of that linking analysis, we did not generate such data, and future work might extend our findings by conducting such analyses.

A second general question concerns the extent to which our results generalize outside the enterprise context. All of our analyses were conducted on a specific class of data emerging from enterprise communities which are not yet well explored. We have reason to be optimistic about generalization, however, as many of our results have also been found on the open internet, e.g. inequality of posting content, consistent role behaviors for leaders and members and few observed role shifts. Nevertheless future research should explore whether findings about content management and subgroup behaviors are also observed in open internet contexts such as Reddit or Wikipedia, where these topics have not yet been broadly researched.

Another limitation of this work is that for each of our empirical findings, we make implementation recommendations about new technologies that might aid communities in managing content, allocating work to qualified individuals or identifying key subgroups within a community. For example we make recommendations about how communities might be reminded about prior valuable content (Chapter 4 - referencing), how outstanding group tasks might be allocated to qualified community members (Chapter 5 - roles), how leaders might track content to make adaptive adjustments (Chapter 6 - emotions) or monitor specific subgroups to encourage active contributions (Chapters 7, 8 - graphlets). Future systems work should implement these suggestions and observe whether their introduction led to anticipated changes within communities. If successful these tools could help address the quite significant problems experienced by online communities in retaining active members and generating long-term activity as the community ages. Detecting the changes resulting from successful tool interventions would also confirm the validity of our empirical findings as well as contributing to the development of future theory.

Turning now to community theory, such tool based interventions could confirm the empirical observations emerging from our analyses. For example, we could evaluate whether a tool that recommended valued prior content to community members could spark community interactions and promote long term community success. Or we could determine whether informing leaders about sparse subgroups within their community might motivate them to intervene with these subgroups to promote rich long-term interactions.

In addition our results suggest that current community theories are in need of significant overhaul. In particular apprenticeship and peripheral participation accounts need to be radically modified to incorporate our results. We found for example that communities did not manage long-term content as predicted, nor did members assume additional management responsibilities over time. Theory should also develop more nuanced accounts about which types of communities benefit from emotional interactions to explain the findings of Chapter 6. Additionally our results contradict network approaches arguing that subgroup structures add little explanatory power beyond standard network metrics. Again existing theory needs to be extended to account for the role and value that these subgroups bring to the community and flesh out our suggestive findings about how subgroups aid community interactions. We also found quite significant changes in community behaviors over time, e.g. with respect to unmanaged content growth. These need to be better explained by current theories which should also be extended to cover very long term communities that are beginning to emerge. Furthermore, there is a significant theoretical gap between communities research that tends to focus on individuals and dyads, and network approaches that analyze either egocentric individuals or entire networks. How then might we extend existing theorising to include these subgroups which seem to support important community functions? What different types of reliable subgroups exist and what community functions do they serve? What do our findings about small groups say about key assortativity results? And how do these structural accounts mesh with work from organizational theory about the importance of small groups?

Another important theoretical contribution of this work has been to provide replicable operationalizations of existing theoretical constructs, where one important weakness of prior theorising was that key theory concepts were not well defined. Exploiting the unique properties of our enterprise data allowed us to develop accurate statistical models of specific roles (owners vs members), as well as critical content management constructs (referencing via links). Having well defined constructs allowed us to systematically test theoretical claims. However while these constructs are well defined, they do not cover more ambiguous cases and future work should be extended to cover such cases. How for example might we operationalize and test findings about complex role types where there are often very subtle differences between roles (e.g. collaborator vs leader vs contributor)? And how might we define and test complex forms of content management involving FAQ creation or moderation?

Another significant methodological contribution of the work has been the application of graphical techniques to the community context. As far as we are aware this is the first work to exploit graphical methods to observe significant explanatory effects for intermediate community substructures. While there are significant computational barriers to identifying complex substructures, future work could explore further synergies between the development of new graph analytic methods and community research and theorising. Our work also exploited a member-generated metric of success elicited by a survey of active community participants, and one methodological weakness of prior work has been to rely on strictly behavioral success metrics, e.g. increased posting, fast responses to posts. Future work needs to use more nuanced methods to evaluate community success as opposed to relying on easy-to-capture, but ambiguous behavioral metrics.

Bibliography

[1] The glmmADMB package.

[2] Global Internet User Survey 2012 | Internet Society. http://www.internetsociety.org/internet/global-internet-user-survey-2012.

[3] Google Photos - All your photos organized and easy to find.

[4] Introducing On This Day: A New Way to Look Back at Photos and Memories on Facebook | Facebook Newsroom.

[5] The State of the Octoverse. https://octoverse.github.com/.

[6] Wikipedia:Statistics, January 2019. Page Version ID: 880210525.

[7] Sisay Fissaha Adafre and Maarten de Rijke. Discovering missing links in wikipedia. In *Proceedings of the 3rd international workshop on Link discovery*, pages 90–97. ACM, 2005.

[8] Nesreen K Ahmed, Jennifer Neville, Ryan A Rossi, and Nick Duffield. Efficient graphlet counting for large networks. In *2015 IEEE International Conference on Data Mining*, pages 1–10. IEEE, 2015.

[9] Nesreen K Ahmed, Jennifer Neville, Ryan A Rossi, Nick G Duffield, and Theodore L Willke. Graphlet decomposition: Framework, algorithms, and applications. *Knowledge and Information Systems*, 50(3):689–722, 2017.

[10] Nader Ale Ebrahim, Shamsuddin Ahmed, and Zahari Taha. Virtual r&d teams in small and medium enterprises: A literature review. *Scientific Research and Essays*, 4(13):1575–1590, 2009.

[11] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics, 2005.

[12] Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer, 2007.

[13] Aris Anagnostopoulos, Luca Becchetti, Carlos Castillo, Aristides Gionis, and Stefano Leonardi. Online team formation in social networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 839–848. ACM, 2012.

[14] Brigham S. Anderson, Carter Butts, and Kathleen Carley. The interaction of size and density with graph-level indices. *Social networks*, 21(3):239–267, 1999.

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

[16] Thomas Arnold, Johannes Daxenberger, Iryna Gurevych, and Karsten Weihe. Is Interaction More Important than Individual Performance?: A Study of Motifs in Wikia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1609–1617. International World Wide Web Conferences Steering Committee, 2017.

[17] Erik Aumayr, Jeffrey Chan, and Conor Hayes. Reconstruction of threaded conversations in online discussion forums. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[18] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM, 2006.

[19] Alexandra Balahur, Jesús M Hermida, and Andrés Montoyo. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4):742–753, 2012.

[20] Prasad Balkundi and Martin Kilduff. The ties that lead: A social network approach to leadership. *The Leadership Quarterly*, 17(4):419–439, 2006.

[21] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. Sense-level subjectivity in a multilingual setting. *Computer Speech & Language*, 28(1):7–19, 2014.

[22] Douglas M. Bates. lme4: Mixed-effects modeling with R, 2010.

[23] Gerard Beenen, Kimberly Ling, Xiaoqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, and Robert E. Kraut. Using social psychology to motivate contributions to online communities. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 212–221. ACM, 2004.

[24] Per Block. Reciprocity, transitivity, and the mysterious three-cycle. *Social Networks*, 40:163–173, 2015.

[25] Ben Bolker, Hans Skaug, Arni Magnusson, and Anders Nielsen. Getting started with the glmmadmb package. *Available at glmmadmb.r-forge.r-project.org/glmmADMB.pdf*, 2012.

[26] Danushka Bollegala, David Weir, and John Carroll. Cross-domain sentiment classification using a sentiment sensitive thesaurus. *Knowledge and Data Engineering, IEEE Transactions on*, 25(8):1719–1731, 2013.

[27] Johan Bollen, Mao Huina, and Zeng Xiaojun. Twitter mood predicts the stock market. *Journal of Computational Science*, 2.1:1–8, 2011.

[28] Leanne Bowler, Daqing He, and Wan Yin Hong. Who is referring teens to health information on the web?: hyperlinks between blogs and health web sites for teens. In *Proceedings of the 2011 iConference*, pages 238–243. ACM, 2011.

[29] Marco Bressan, Flavio Chierichetti, Ravi Kumar, Stefano Leucci, and Alessandro Panconesi. Counting graphlets: Space vs time. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 557–566. ACM, 2017.

[30] S.I. Brown, A. Tilton, and D.M. Woodside. The case for online communities. *The McKinsey Quarterly*, 1, 2002.

[31] Francesco Buccafurri, Gianluca Lax, and Antonino Nocera. A new form of assortativity in online social networks. *International Journal of Human-Computer Studies*, 80:56–65, 2015.

[32] M. Burke, E. Joyce, T. Kim, and et al. Introductions and Requests: Rhetorical Strategies That Elicit Response in Online Communities. In *Communities and Technologies*, pages 21–39. Springer, 2007.

[33] Moira Burke and Robert Kraut. Taking up the mop: identifying future wikipedia administrators. In *CHI'08 extended abstracts on Human factors in computing systems*, pages 3441–3446. ACM, 2008.

[34] Ronald S. Burt. Structural holes versus network closure as social capital. In *Social capital: Theory and research*, pages 31–56. 2001.

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

[36] Carter T. Butts. Exact bounds for degree centralization. *Social Networks*, 28(4):283–296, 2006.

[37] Dankmar Bhning, Ekkehart Dietz, Peter Schlattmann, Lisette Mendonca, and Ursula Kirchner. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(2):195–209, 1999.

[38] Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

[39] Erik Cambria, Paolo Gastaldo, Federica Bisio, and Rodolfo Zunino. An ELM-based model for affective analogical reasoning. *Neurocomputing*, 149:443–455, 2015.

[40] Justine Cassell, David Huffaker, Dona Tversky, and Kim Ferriman. The language of online leadership: Gender and youth engagement on the internet. *Developmental Psychology*, 42(3):436, 2006.

[41] William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175. Citeseer, 1994.

[42] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[43] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[44] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[45] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[46] Sarvenaz Choobdar, Pedro Ribeiro, Srinivasan Parthasarathy, and Fernando Silva. Dynamic inference of social roles in information cascades. *Data mining and knowledge discovery*, 29(5):1152–1177, 2015.

[47] Kon Shing Kenneth Chung, Mahendra Piraveenan, and Shahadat Uddin. Community evolution and engagement through assortative mixing in online social networks. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 724–725. IEEE Computer Society, 2012.

[48] Ryan Compton. Theory driven community analytics and influence on community success. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, pages 135–138. ACM, 2016.

[49] Ryan Compton, Jilin Chen, Eben Haber, Hernan Badenes, and Steve Whittaker. Just the Facts: Exploring the Relationship between Emotional Language and Member Satisfaction in Enterprise Online Communities. *Proceedings of the Eleventh International Conference on Web and Social Media*, pages 500–504, 2017.

[50] Ryan J Compton, Jeff Warshaw, Hernan Badenes, Barton Smith, and Steve Whittaker. Living in the present: Understanding long-term content referencing in enterprise online communities. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):39, 2018.

[51] Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. Finding question-answer pairs from online forums. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 467–474. ACM, 2008.

[52] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. Suggestbot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 32–41. ACM, 2007.

[53] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168. ACM, 2008.

[54] Jonathon N Cummings and Rob Cross. Structural properties of work groups and their consequences for performance. *Social networks*, 25(3):197–210, 2003.

[55] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318. International World Wide Web Conferences Steering Committee, 2013.

[56] James A. Davis. Clustering and structural balance in graphs. *Human relations*, 20(2):181–187, 1967.

[57] James A. Davis and Samuel Leinhardt. The structure of positive interpersonal relations in small groups. 1967.

[58] M. De Choudhury, M. Gamon, and S. Counts. Happy, nervous or surprised? classification of human affective states in social media. *Proc. of ICWSM*, 2012.

[59] Juliette De Maeyer. Towards a hyperlinked society: A critical review of link studies. *New Media & Society*, 15(5):737–751, 2013.

[60] Tom Denison, Gary Hardy, Graeme Johanson, Larry Stillman, and Don Schauder. Community networks: Identities, taxonomies and evaluations. In *Electronic Networking 2002-Building Community (CCNR 03 July 2002 to 05 July 2002)*, pages 1–17. Centre for Community Networking Research, 2002.

[61] Lauren D‘Innocenzo, John E. Mathieu, and Michael R. Kukenberger. A meta-analysis of different forms of shared leadershipteam performance relations. *Journal of Management*, 42(7):1964–1991, 2016.

[62] Yuxiao Dong, Reid A Johnson, Jian Xu, and Nitesh V Chawla. Structural diversity and homophily: A study across more than one hundred big networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 807–816. ACM, 2017.

[63] James A Dove, Dawn L Eubanks, Niki Panteli, Leon A Watts, and Adam N Joinson. Making an entrance 2.0: The linguistics of introductory success in virtual communities. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2011.

[64] Peter D. Duffy and Axel Bruns. The use of blogs, wikis and RSS in education: A conversation of possibilities. *Proceedings of the Online Learning and Teaching Conference*, 2006.

[65] Robin IM Dunbar, N. D. C. Duncan, and Daniel Nettle. Size and structure of freely forming conversational groups. *Human nature*, 6(1):67–78, 1995.

[66] C Dunkel-Schetter, L.G. Feinstein, S.E. Taylor, and R.L. Falke. Patterns of coping with cancer. *Health Psychology*, 11(2):79–87, 1992.

[67] Kathleen T Durant, Alexa T McCray, and Charles Safran. Modeling the temporal evolution of an online cancer forum. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 356–365. ACM, 2010.

[68] Daniel Ehls and Cornelius Herstatt. Open Source Participation Behavior-A Review and Introduction of a Participation Lifecycle Model. In *35th DRUID Celebration Conference*, 2013.

[69] Samer Faraj, Sirkka L. Jarvenpaa, and Ann Majchrzak. Knowledge collaboration in online communities. *Organization science*, 22(5):1224–1239, 2011.

[70] Henry Farrell and Daniel W. Drezner. The power and politics of blogs. *Public choice*, 134(1-2):15–30, 2008.

[71] Robert Farrell. Summarizing electronic discourse. *Intelligent Systems in Accounting, Finance & Management*, 11(1):23–38, 2002.

[72] Katherine Faust. Triadic configurations in limited choice sociometric networks: Empirical and theoretical results. *Social Networks*, 30(4):273–282, 2008.

[73] Katherine Faust. A puzzle concerning triads in social networks: Graph constraints and the triad census. *Social Networks*, 32(3):221–233, 2010.

[74] Anna Filippova and Hichang Cho. The Effects and Antecedents of Conflict in Free and Open Source Software Development. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 705–716. ACM, 2016.

[75] Tim Finin, Anupam Joshi, Pranam Kolari, Akshay Java, Anubhav Kale, and Amit Karandikar. The information ecology of social media and online communities. *AI Magazine*, 29(3):77, 2008.

[76] Snehal Neil Gaikwad, Durim Morina, Rohit Nistala, Megha Agarwal, Alison Cossette, Radhika Bhanu, Saiph Savage, Vishwajeet Narwal, Karan Rajpal, Jeff Regino, et al. Daemo: A self-governed crowdsourcing marketplace. In *Adjunct proceedings of the 28th annual ACM symposium on user interface software & technology*, pages 101–102. ACM, 2015.

[77] Dale Ganley and Cliff Lampe. The ties that bind: Social network principles in online communities. *Decision Support Systems*, 47(3):266–274, 2009.

[78] Laura Garton, Caroline Haythornthwaite, and Barry Wellman. Studying online social networks. *Journal of computer-mediated communication*, 3(1):JCMC313, 1997.

[79] Joseph L Gastwirth. The estimation of the lorenz curve and gini index. *The review of economics and statistics*, pages 306–316, 1972.

[80] M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282, 2013.

[81] Eric Gleave, Howard T. Welser, Thomas M. Lento, and Michael A. Smith. A conceptual and operational definition of'social role'in online community. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pages 1–11. IEEE, 2009.

[82] Neil Zhenqiang Gong, Wenchang Xu, Ling Huang, Prateek Mittal, Emil Stefanov, Vyas Sekar, and Dawn Song. Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 131–144. ACM, 2012.

[83] Mark S. Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

[84] Yupeng Gu, Yizhou Sun, Yanen Li, and Yang Yang. RaRE: Social Rank Regulated Large-scale Network Embedding. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 359–368. International World Wide Web Conferences Steering Committee, 2018.

[85] Eben M. Haber. On the Stability of Online Language Features: How Much Text do you Need to know a Person? *arXiv preprint arXiv:1504.06391*, 2015.

[86] J Richard Hackman and Neil Vidmar. Effects of size and task type on group performance and member reactions. *Sociometry*, pages 37–54, 1970.

[87] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[88] Zhen Hai, Kuiyu Chang, Jung-Jae Kim, and Christopher C. Yang. Identifying features in opinion mining via intrinsic and extrinsic domain relevance. *Knowledge and Data Engineering, IEEE Transactions on*, 26(3):623–634, 2014.

[89] Catherine Hall and Michael Zarro. Social curation on the website Pinterest.com. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–9, January 2012.

[90] Derek L. Hansen. Knowledge sharing, maintenance, and use in online support communities. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, pages 1751–1754. ACM, 2006.

[91] F Maxwell Harper, Daniel Moy, and Joseph A Konstan. Facts or friends?: distinguishing informational and conversational questions in social q&a sites. pages 759–768, 2009.

[92] Vicki S Helgeson, Sheldon Cohen, Richard Schulz, and Joyce Yasko. Group support interventions for women with breast cancer: who benefits from what? *Health psychology*, 19(2):107, 2000.

[93] Itai Himelboim. The international network structure of news media: An analysis of hyperlinks usage in news web sites. *Journal of Broadcasting & Electronic Media*, 54(3):373–390, 2010.

[94] Paul W. Holland and Samuel Leinhardt. A method for detecting structure in sociometric data. In *Social Networks*, pages 411–432. Elsevier, 1977.

[95] Anna Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56, 2008.

[96] David Huffaker. Dimensions of leadership and social influence in online communities. *Human Communication Research*, 36(4):593–617, 2010.

[97] Shin-Yuan Hung, Alexandra Durcikova, Hui-Min Lai, and Wan-Mei Lin. The influence of intrinsic and extrinsic motivation on individuals' knowledge sharing behavior. *International Journal of Human-Computer Studies*, 69(6):415–427, 2011.

[98] Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[99] Alicia Iriberri and Gondy Leroy. A life-cycle perspective on online community success. *ACM Computing Surveys (CSUR)*, 41(2):11, 2009.

[100] Oskar Jarczyk, Szymon Jaroszewicz, Adam Wierzbicki, Kamil Pawlak, and Michal Jankowski-Lorek. Surgical teams on GitHub: Modeling performance of GitHub project development processes. *Information and Software Technology*, 100:32–46, 2018.

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

[102] Steven L. Johnson, Hani Safadi, and Samer Faraj. The emergence of online community leadership. *Information Systems Research*, 26(1):165–187, 2015.

[103] Dan Jurafsky and James H. Martin. *Speech and language processing*, volume 3. Pearson London, 2014.

[104] Sanjay Ram Kairam, Dan J. Wang, and Jure Leskovec. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 673–682. ACM, 2012.

[105] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. Surviving an eternal september: How an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1152–1156. ACM, 2016.

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

[107] Sheila Kinsella, Mengjiao Wang, John G. Breslin, and Conor Hayes. Improving categorisation in social media using hyperlinks to structured data sources. In *The Semanic Web: Research and Applications*, pages 390–404. Springer, 2011.

[108] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[109] Daniel J. Kleitman and Da-Lun Wang. Algorithms for constructing graphs and digraphs with given valences and factors. *Discrete Mathematics*, 6(1):79–88, 1973.

[110] Joon Koh, Young-Gul Kim, Brian Butler, and Gee-Woo Bock. Encouraging participation in virtual communities. *Communications of the ACM*, 50(2):68–73, 2007.

[111] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA, 1995.

[112] A.D.I. Kramer, S.R. Fussell, and L.D. Setlock. Text analysis as a tool for analyzing conversation in online support groups. *CHI Ext. Abstracts*, pages 1485–88, 2004.

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

[114] Robert E. Kraut and Andrew T. Fiore. The role of founders in building online groups. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 722–732. ACM, 2014.

[115] Yu-Sheng Lai, Kuao-Ann Fung, and Chung-Hsien Wu. Faq mining via list detection. In *proceedings of the 2002 conference on multilingual summarization and question answering-Volume 19*, pages 1–7. Association for Computational Linguistics, 2002.

[116] Shyong (Tony) K. Lam. Collaborative curation in social production communities. August 2012.

[117] Cliff Lampe and Erik Johnston. Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 11–20. ACM, 2005.

[118] Vito Latora and Massimo Marchiori. Efficient behavior of small-world networks. *Physical review letters*, 87(19):198701, 2001.

[119] Jean Lave and Etienne Wenger. Legitimate peripheral participation in communities of practice. *Supporting lifelong learning*, 1:111–126, 2002.

[120] Jonathan Lazar and Jenny Preece. Classification schema for online communities. In *AMCIS 1998 Proceedings*, pages 84–86, 1998.

[121] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470. ACM, 2008.

[122] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on World Wide Web*, pages 695–704. ACM, 2008.

[123] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, July 2013.

[124] Hui Lin, Weiguo Fan, Linda Wallace, and Zhongju Zhang. An empirical study of web-based knowledge community success. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 178c–178c. IEEE, 2007.

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

[126] Pablo Loyola and In-Young Ko. Population dynamics in open source communities: an ecological approach applied to github. In *Proceedings of the companion publication*

of the 23rd international conference on World wide web companion, pages 993–998. International World Wide Web Conferences Steering Committee, 2014.

[127] Pamela J. Ludford, Dan Cosley, Dan Frankowski, and Loren Terveen. Think different: increasing online community participation using uniqueness and group dissimilarity. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 631–638. ACM, 2004.

[128] Ravindranath Madhavan, Devi R. Gnyawali, and Jinyu He. Two's company, three's a crowd? Triads in cooperative-competitive networks. *Academy of Management Journal*, 47(6):918–927, 2004.

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

[130] Jennifer Marlow, Laura Dabbish, and Jim Herbsleb. Impression formation in online peer production: activity traces and personal profiles in github. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 117–128. ACM, 2013.

[131] Rainer Martens and James A Peterson. Group cohesiveness as a determinant of success and member satisfaction in team performance. *International review of sport sociology*, 6(1):49–61, 1971.

[132] Adrienne L Massanari. Contested play: The culture and politics of reddit bots. In *Socialbots and Their Friends*, pages 126–143. Routledge, 2016.

[133] Tara Matthews, Jilin Chen, Steve Whittaker, Aditya Pal, Haiyi Zhu, Hernan Badenes, and Barton Smith. Goals and perceived success of online enterprise communities: what is important to leaders & members? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 291–300. ACM, 2014.

[134] Tara Matthews, Jalal U. Mahmud, Jilin Chen, Michael Muller, Eben Haber, and Hernan Badenes. They Said What? Exploring the Relationship Between Language Use and Member Satisfaction in Communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work Social Computing*, pages 819–825. ACM, 2015.

[135] Tara Matthews, Steve Whittaker, Hernan Badenes, and Barton Smith. Beyond end user content to collaborative knowledge mapping: Interrelations among community social tools. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 900–910. ACM, 2014.

[136] Tara Matthews, Steve Whittaker, Hernan Badenes, Barton A. Smith, Michael Muller, Kate Ehrlich, Michelle X. Zhou, and Tessa Lau. Community insights: helping community leaders enhance the value of enterprise online communities. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 513–522. ACM, 2013.

[137] Tara Matthews, Steve Whittaker, Thomas P Moran, Sandra Y Helsley, and Tejinder K Judge. Productive interrelationships between collaborative groups ease the challenges of dynamic and multi-teaming. volume 21, pages 371–396. Springer, 2012.

[138] Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908. International World Wide Web Conferences Steering Committee, 2013.

[139] Brian James McInnis, Elizabeth Lindley Murnane, Dmitry Epstein, Dan Cosley, and Gilly Leshed. One and Done: Factors affecting one-time contributors to ad-hoc online communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 609–623. ACM, 2016.

[140] Joel C. Miller and Aric Hagberg. Efficient generation of networks with given expected degrees. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 115–126. Springer, 2011.

[141] Giovanna Miritello, Rubn Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3, 2013.

[142] Saif M. Mohammad and Peter D. Turney. Emotions evoked by common words and phrases: Using Mechanical Turk to cre-ate an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34, 2010.

[143] Rodrigo Moraes, Joao Francisco Valiati, and Wilson P. Gavio Neto. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, 40(2):621–633, 2013.

[144] Mohamed M. Mostafa. More than words: Social networks text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241–4251, 2013.

[145] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 412–418, 2004.

[146] Michael Muller, Kate Ehrlich, Tara Matthews, Adam Perer, Inbal Ronen, and Ido Guy. Diversity among enterprise online communities: collaborating, teaming, and innovating through social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2815–2824. ACM, 2012.

[147] San Murugesan. Understanding Web 2.0. *IT Professional*, 9(4):34–41, July 2007.

[148] Lisa Neal, Kate Oakley, Gitte Lindgaard, David Kaufman, Jan Marco Leimeister, and Ted Selker. Online health communities. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems*, pages 2129–2132. ACM, 2007.

[149] Ted Nelson. A file structure for the complex, the changing, and the indeterminate. *ACM Annual Conference. Proc. of the 1965 20th National Conference*, pages 84–100, 1965.

[150] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[151] Dong Nguyen and Carolyn P Rosé. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Languages in Social Media*, pages 76–85. Association for Computational Linguistics, 2011.

[152] Thin Nguyen, Dinh Phung, Brett Adams, and Svetha Venkatesh. A sentiment-aware approach to community formation in social media. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[153] Robert D. Nolker and Lina Zhou. Social computing and weighting to identify member roles in online communities. In *Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence*, pages 87–93. IEEE Computer Society, 2005.

[154] Blair Nonnecke and Jennifer Preece. Shedding light on lurkers in online communities. *Ethnographic Studies in Real and Virtual Environments: Inhabited Information Spaces and Connected Communities, Edinburgh*, 123128, 1999.

[155] Blair Nonnecke and Jenny Preece. Why lurkers lurk. *AMCIS 2001 Proceedings*, page 294, 2001.

[156] Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. *arXiv preprint arXiv:1709.05295*, 2017.

[157] Felipe Ortega, Jesus M. Gonzalez-Barahona, and Gregorio Robles. On the inequality of contributions to Wikipedia. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pages 304–304. IEEE, 2008.

[158] Carlos Padoa, Daniel Schneider, Jano Moreira de Souza, Palma J. Medeiros, and others. Investigating social curation websites: A crowd computing perspective. In *Computer Supported Cooperative Work in Design (CSCWD), 2015 IEEE 19th International Conference on*, pages 253–258. IEEE, 2015.

[159] Katherine Panciera, Aaron Halfaker, and Loren Terveen. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 51–60. ACM, 2009.

[160] Katherine Panciera, Reid Priedhorsky, Thomas Erickson, and Loren Terveen. Lurking? cyclopaths?: a quantitative lifecycle analysis of user behavior in a geowiki. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1917–1926. ACM, 2010.

[161] Niki Panteli. On leaders presence: interactions and influences within online communities. *Behaviour & Information Technology*, pages 1–10, 2016.

[162] Pietro Panzarasa, Tore Opsahl, and Kathleen M. Carley. 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):911–932, 2009.

[163] Ashwin Paranjape, Austin R. Benson, and Jure Leskovec. Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 601–610. ACM, 2017.

[164] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[165] James W Pennebaker. *The secret life of pronouns*, volume 211. Elsevier, 2011.

[166] Patchareeporn Pluempavarn, Niki Panteli, Adam Joinson, Dawn Eubanks, Leon Watts, and James Dove. Social roles in online communities: Relations and trajectories. In *6th Mediterranean Conference on Information Systems, Nicosia, Cyprus. Retrieved October*, volume 4, page 2012, 2011.

[167] Marshall Scott Poole and Andrea B. Hollingshead. *Theories of small groups: Interdisciplinary perspectives*. Sage Publications, 2004.

[168] Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, and Sivaji Bandyopadhyay. Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, (2):31–38, 2013.

[169] Constance Elise Porter. A typology of virtual communities: A multi-disciplinary foundation for future research. *Journal of computer-mediated communication*, 10(1):JCMC1011, 2004.

[170] David Martin Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies.*, 2011.

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

[172] Jenny Preece. *Online communities: Designing usability and supporting socialbilty*. John Wiley & Sons, Inc., 2000.

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

[174] Jenny Preece and Diane Maloney-Krichmar. Online communities: focusing on sociability and usability. *Handbook of human-computer interaction*, pages 596–620, 2003.

[175] Jiezhong Qiu, Yixuan Li, Jie Tang, Zheng Lu, Hao Ye, Bo Chen, Qiang Yang, and John E. Hopcroft. The Lifecycle and Cascade of WeChat Social Messaging Groups. In *Proceedings of the 25th International Conference on World Wide Web*, pages 311–320. International World Wide Web Conferences Steering Committee, 2016.

[176] Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, (89):14–46, 2015.

[177] Yuqing Ren, Robert Kraut, and Sara Kiesler. Applying common identity and bond theory to design of online communities. *Organization studies*, 28(3):377–408, 2007.

[178] Catherine Ridings, David Gefen, and Bay Arinze. Psychological barriers: Lurker and poster motivation and behavior in online communities. *Communications of the Association for Information Systems*, 18(1):16, 2006.

[179] Rahmtin Rotabi, Krishna Kamath, Jon Kleinberg, and Aneesh Sharma. Detecting strong ties using network motifs. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 983–992. International World Wide Web Conferences Steering Committee, 2017.

[180] Dana Rotman, Kezia Procita, Derek Hansen, Cynthia Sims Parr, and Jennifer Preece. Supporting content curation communities: The case of the Encyclopedia of Life. *Journal of the Association for Information Science and Technology*, 63(6):1092–1107, 2012.

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

[182] M. Rushdi Saleh, M. T. Martn-Valdivia, A. Montejo-Rez, and L. A. Urea-Lpez. Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications*, 38(12):14799–14804, November 2011.

[183] Tiago Santos, Simon Walk, and Denis Helic. Nonlinear Characterization of Activity Dynamics in Online Collaboration Websites. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1567–1572. International World Wide Web Conferences Steering Committee, 2017.

[184] Mark Sharratt and Abel Usoro. Understanding knowledge-sharing in online communities of practice. *Electronic Journal on Knowledge Management*, 1(2):187–196, 2003.

[185] Xiutao Shi, Liqiang Wang, Shijun Liu, Yafang Wang, Li Pan, and Lei Wu. Investigating Microstructure Patterns of Enterprise Network in Perspective of Ego Network. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*, pages 444–459. Springer, 2017.

[186] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[187] Raymond T. Sparrowe, Robert C. Liden, Sandy J. Wayne, and Maria L. Kraimer. Social networks and the performance of individuals and groups. *Academy of management journal*, 44(2):316–325, 2001.

[188] Carl Staelin. Parameter selection for support vector machines. *Hewlett-Packard Company, Tech. Rep. HPL-2002-354R1*, 2003.

[189] Bongwon Suh, Ed H. Chi, Aniket Kittur, and Bryan A. Pendleton. Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1037–1040. ACM, 2008.

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

[191] Yla R Tausczik and James W Pennebaker. Participation in an online mathematics community: differentiating motivations to add. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 207–216. ACM, 2012.

[192] Steven JJ Tedjamulia, Douglas L. Dean, David R. Olsen, and Conan C. Albrecht. Motivating content contributions to online communities: Toward a more comprehensive theory. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 193b–193b. IEEE, 2005.

[193] Loren Terveen and Will Hill. Beyond recommender systems: Helping people help each other. *HCI in the New Millennium*, 1:487–509, 2001.

[194] Loren Terveen, Will Hill, Brian Amento, David McDonald, and Josh Creter. PHOAKS: A system for sharing recommendations. *Communications of the ACM*, 40(3):59–62, 1997.

[195] Steven L. Thorne, Ingrid Fischer, and Xiaofei Lu. The semiotic ecology and linguistic complexity of an online game world. *ReCALL*, 24(03):279–301, 2012.

[196] Sergio L. Toral, Mara del Roco Martnez-Torres, and Federico Barrero. Analysis of virtual communities supporting OSS projects using social network analysis. *Information and Software Technology*, 52(3):296–303, 2010.

[197] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[198] Ngoc Hieu Tran, Kwok Pui Choi, and Louxin Zhang. Counting motifs in the human interactome. *Nature communications*, 4:2241, 2013.

[199] Angela Charng-Rurng Tsai, Chi-En Wu, Richard Tzong-Han Tsai, and Jane Yung-jen Hsu. Building a concept-level sentiment dictionary based on commonsense knowledge. *IEEE Intelligent Systems*, (2):22–30, 2013.

[200] Christian Wagner. Wiki: A technology for conversational knowledge management and group collaboration. *The Communications of the Association for Information Systems*, 13(1):58, 2004.

[201] Christian Wagner. Breaking the knowledge acquisition bottleneck through conversational knowledge management. *Information Resources Management Journal (IRMJ)*, 19(1):70–83, 2006.

[202] Claudia Wagner, Matthew Rowe, Markus Strohmaier, and Harith Alani. What catches your attention? an empirical study of attention patterns in community forums. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[203] Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *LREC*, pages 812–817. Istanbul, 2012.

[204] Yi-Chia Wang, Robert Kraut, and John M Levine. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 833–842. ACM, 2012.

[205] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[206] Barry Wellman, Jeffrey Boase, and Wenhong Chen. The networked nature of community: Online and offline. *It & Society*, 1(1):151–165, 2002.

[207] Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. Finding social roles in Wikipedia. In *Proceedings of the 2011 iConference*, pages 122–129. ACM, 2011.

[208] Etienne Wenger, Richard Arnold McDermott, and William Snyder. *Cultivating communities of practice: A guide to managing knowledge*. Harvard Business Press, 2002.

[209] Mark E Whiting, Dilrukshi Gamage, Snehalkumar Neil S Gaikwad, Aaron Gilbee, Shirish Goyal, Alipta Ballav, Dinesh Majeti, Nalin Chhibber, Angela Richmond-Fuller, Freddie Vargus, et al. Crowd guilds: Worker-led reputation and feedback on crowdsourcing platforms. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1902–1913. ACM, 2017.

[210] Steve Whittaker, Loen Terveen, Will Hill, and Lynn Cherny. The dynamics of mass interaction. In *From Usenet to CoWebs*, pages 79–91. Springer, 2003.

[211] Dennis M. Wilkinson. Strong regularities in online peer production. In *Proceedings of the 9th ACM conference on Electronic commerce*, pages 302–309. ACM, 2008.

[212] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

[213] Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. Social context summarization. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 255–264. ACM, 2011.

[214] Xinchuan Zeng and Tony R. Martinez. Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1):1–12, 2000.

[215] Jing Zhang, Jie Tang, Honglei Zhuang, Cane Wing-Ki Leung, and Juanzi Li. Role-aware conformity influence modeling and analysis in social networks. In *AAAI*, volume 14, pages 958–965, 2014.

[216] Jun Zhang, Mark S. Ackerman, and Lada Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230. ACM, 2007.

[217] Kang Zhao, John Yen, Greta Greer, Baojun Qiu, Prasenjit Mitra, and Kenneth Portier. Finding influential users of online health communities: a new metric based on sentiment influence. *Journal of the American Medical Informatics Association*, 21(e2):e212–e218, 2014.

[218] Haiyi Zhu, Jilin Chen, Tara Matthews, Aditya Pal, Hernan Badenes, and Robert E. Kraut. Selecting an effective niche: an ecological view of the success of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 301–310. ACM, 2014.

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

[220] Haiyi Zhu, Robert E. Kraut, and Aniket Kittur. The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 281–290. ACM, 2014.

[221] Fabiana Zollo, Petra Kralj Novak, Michela Del Vicario, Alessandro Bessi, Igor Mozeti, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Emotional dynamics in the age of misinformation. *PloS one*, 10(9):e0138740, 2015.