The Emergence of Online Community Leadership

Pre-Publication draft of work published in *Information Systems Research*
See also: http://dx.doi.org/10.1287/isre.2014.0562

Steven L. Johnson, Hani Safadi and Samer Faraj

Abstract

Compared to traditional organizations, online community leadership processes and how leaders emerge are not well studied. Previous studies of online leadership have often identified leaders as those who administer forums or have high network centrality scores. Although communication in online communities occurs almost exclusively through written words, little research has addressed how the comparative use of language shapes community dynamics. Using participant surveys to identify leading online community members, this study analyzes a year of communication network history and message content to assess whether language use differentiates leaders from other core community participants. We contribute a novel use of textual analysis to develop a model of language use to evaluate the utterances of all participants in the community. We find that beyond communication network position--in terms of formal role, centrality, membership in the core, and boundary spanning-- those viewed as leaders by other participants, post a large number of positive, concise posts with simple language familiar to other participants. This research contributes a language model to study online language use and by pointing to the emergent and shared nature of online community leadership.

Keywords: online communities, leadership, natural language processing, knowledge management, network analysis, computer-mediated communication and collaboration.

The Emergence of Online Community Leadership

"The key to successful leadership is influence, not authority." – Kenneth H. Blanchard

Introduction

Supported by the widespread usage of social media, online communities have rapidly emerged as essential new forms of organizing (Benkler 2006, Kraut and Resnick 2011, Preece 2000). Online communities are large collectivities where members with shared goals and interests interact primarily via the Internet (Sproull and Arriaga 2007). They bring together thousands of strangers across national, geographic, time zone, and organizational boundaries. Some communities focus on sustaining social ties and friendship (e.g., Facebook). Others serve as platforms for knowledge integration (e.g., Wikipedia), for sharing creative output (e.g., YouTube), for open source software development (e.g., Linux) or for answering questions (e.g., Quora.com). There is literally an online community to support every kind of interest, self-identified group, or creative endeavor.

In spite of the rapid growth of this new organizational form, research has been slow to examine the points of commonality and difference between traditional organizations and online communities. Principally, little is known about the rich diversity of forms of online collaboration, how they are structured, and how they sustain themselves (Faraj, et al. 2011). Many of these communities are characterized by a core-periphery structure suggestive of interactions typical of communities of practice (Collier and Kraut 2012, Wasko, et al. 2009). Members' decisions to participate may be due to a variety of intrinsic and extrinsic motivations (Kankanhalli, et al. 2005, Lakhani and von Hippel 2003). Their level of engagement is affected by the strength of their identification with the group and the kind of interpersonal bonds they develop (Ren, et al. 2012). Further, in production or expertise based communities, continued participation is linked to the depth of embeddedness in the social practice encompassing the communal activity (von Krogh, et al. 2012a, Wasko and Faraj 2005). Online communities are often characterized by high turnover, fluid boundaries, expertise-based authority, and emergent roles (Faraj, et al. 2011, Ren, et al. 2007).

In this paper, we focus on leadership processes in online communities. While thousands of published works have enriched the understanding of organizational leadership, much less is known as to what constitutes effective leadership online. Given the lack of face-to-face communication, the mediated nature of interactions, and the primacy of text-based asynchronous exchanges, online leadership is bound to differ in some substantial ways from more familiar in-person and synchronous settings. Early findings indicate that leadership roles are more informal and emergent (Butler, et al. 2007, Collier and Kraut 2012). Network position at the center of the exchanges seems to matter greatly (Sutanto, et al. 2011). Leaders are heavily involved in the social practice of the community, its core mission, and core activity whether it is developing code in open source software or answering questions in an expertise based community (Dahlander and Frederiksen 2012, von Krogh, et al. 2012a, Wasko and Faraj 2005). What makes someone a leader online, given the relative weakness of hierarchy and bottom up governance structure, remains an open research question (O'Mahony and Ferraro 2007, von Krogh, et al. 2012b, Yoo and Alavi 2004).

As there is no one true definition for leadership, definitions should be made consistent with a study's substantive and methodological approach (Bass and Bass 2008). We define an online community leader as a participant recognized by other participants as influential in what the community does or how it does it (Yukl 2010). As such, online community leadership is not a stable global designation. Formally occupying a defined role of authority is neither a necessary nor sufficient condition for demonstrating online community leadership. Although interactions in the form of message posts are visible to all participants, different participants read different content and interpret content differently. Online communities are characterized by fluid structures, and shifting membership, and are sustained through the voluntary contribution of members. Their structure is dependent on active posting and is therefore constituted by the interactions. Thus, a premise of this paper is that online community leadership is a local designation both distinct from formal roles and emerging from observable interactions.

We are motivated by the goal of understanding what leaders actually do in online communities. We consider both the network position resulting from a leader's interactions as well as the characteristics of a leader's written communications. Therefore, we first examine how communication network position--

in terms of formal role, centrality, membership in the core, and boundary spanning--affects the likelihood of being seen as a leader. Then we contribute a novel use of textual analysis to develop a language model of utterances in the community to evaluate how convergent or divergent leader language is compared to the community as a whole. Our findings suggest that the most influential participants of any online community, those viewed as leaders by other participants, are not just among the most central but also post a large number of positive, concise posts with simple language familiar to other participants.

Three important innovations strengthen our results. First, the leaders in our study are identified by community members rather than deduced based on structural position or behavior, as has been the common practice in the majority of online leadership studies. Second, we contribute methodologically by comparing the identified leaders to a comparable set of participants that post an equivalent number of messages rather than to attempting to compare to an average participant of the community-- something futile given that an "average" participant is non-representative in online communities characterized by power law distribution of participation (Faraj and Johnson 2011, Newman 2003). Finally, our language model offers a sophisticated set of semantic and syntactic tools for analysis of community discourse, again an advance over previous research models based on frequencies of often pre-identified words.

Conceptualizations of Leadership in Online Communities

We argue that a synthesis of organizational leadership theories is required for a deeper understanding of leadership processes in online communities. In drawing on theories of leadership that emphasize behaviors associated with leadership, we identify four as particularly relevant to leadership in online communities: functional leadership, leader-member exchange, shared leadership, and communication as constitutive of organizing. Next, we discuss each theory and how it can be applied to online settings.

Functional leadership theory identifies behaviors that distinguish successful leaders and looks for associations between effective leadership and the functions performed (Burke, et al. 2006). Leadership is not considered a personal characteristic but, rather, can be identified as a set of behaviors that contribute to a group's goals and operation. The theory focuses on general leader behaviors and elaborates how they influence team processes and outcomes (Hackman and Walton 1986, Morgeson, et al. 2010). Like functional leadership theory, we are also interested in what leaders do in online communities. Nonetheless, two specific limitations require adaptation of leadership theory to the online communities. First, given that online communities lack the stable structure and visible leadership that characterize traditional organizations and teams, it is not clear that a focal leader can be identified a priori (Butler, et al. 2007). Second, both the functions of leadership (Butler, et al. 2007) and outcomes of successful leadership (Huffaker 2010, Zhu, et al. 2012) are different in online communities.

Leader-member exchange theory (LMX) focuses on the dyadic relationship between leader and team member. LMX assumes that the characteristics of the interactions between a leader and each of their team members is correlated with leadership processes and organizational outcomes (Gerstner and Day 1997). Given that team members have diverse abilities, commitments, roles and responsibilities, then a leader can improve team function by engaging in customized interactions with each team member based on their potential contribution to the team task (Graen and Uhl-Bien 1995). We share this view that online leadership behaviors are contingent and situated. Yet, the ability to direct apply LMX is constrained by large differences in group size between task-oriented workgroups and online communities. Specifically, given that online communities typically contain thousands of members, direct leadership relationships are bound to be tenuous and lacking in direct influence when compared to smaller, organizationally embedded teams in traditional face-to-face settings (Kiesler, et al. 2012). Indeed, communication in online communities tends to include not only patterns of direct dyadic reciprocation, but also generalized exchange patterns of indirect reciprocation (Faraj and Johnson 2011). Furthermore, because communication in online communities is typically open--all participants can read all communication--the ability for a leader to engage in differentiated direct exchange is diminished.

Shared leadership theory emphasizes the need for members to co-lead each other. Also known by labels such as horizontal, distributed or collective leadership, shared leadership theory views leadership as a set of actions, rather than a designated role. It is "leadership that emanates from members of teams, and not simply from the appointed leader" (Pearce and Sims 2000, p. 115). Shared leadership reflects a web of mutual influences and shared responsibility and is associated with enhanced outcomes in a variety of settings including work groups, virtual teams, and virtual collaborations (Hoch and Kozlowski 2014, Perry, et al. 1999, Sutanto, et al. 2011, Wang, et al. 2014). Likewise, we argue that leadership in online communities also emerges through interactions. Distinctively, though, the combination of open, voluntary participation and paucity of formal leadership roles in online communities means that leadership is inherently shared. Whereas in formal organizations the relative concentration or distribution of leadership may be considered a strategic choice, we argue that shared leadership is an intrinsic property of online communities.

Finally, we consider the Communication as Constitutive of Organizing (CCO) theory as a pertinent perspective to understand online community leadership (Cooren, et al. 2011, Taylor and Van Every 1999), This theory emphasizes the dynamic processes of communication in organizations and how these communication flows enact the social structure via interactions. Organizations are both a network of conversations and the symbolic dimension to interpret these conversations. These communicative interactions act as a structuring process for organizational processes and reinforce organizational processes (Robichaud and Cooren 2013). From this perspective, collaboration or even leadership cannot be conceived as independent of the text that forms the base of organizational conversations. These conversations build on the textual corpus to transcend the text to move to the realm of action and interactions. For example, when confronted to a text produced elsewhere in the organization, people evaluate it for relevance to their own context. They interpret it based on their own experience and their reactions are shaped by norms within their specific community of practice (Taylor and Van Every 2010). This approach has been applied to online settings, to identify communication patterns of online leaders (Huffaker 2010, Zhu, et al. 2012). Although the CCO framework appears most pertinent to the structured world of within-organization communication, its emphasis on explaining how conversations cycles support networking and social structuring makes it relevant to examine the utterances of online community leaders.

As our theoretical review of the four leadership theories indicates, significant differences exist between theories developed for explaining team and organizational leadership and the setting of online communities. We share the emphasis of functional leadership theory on the functions of leadership rather than the behavior of formally designed leaders. We draw on leader-member exchange theory to stress that leadership is contingent and situated. In agreement with theories of shared and distributed leadership, we recognize that leadership is not restricted to designated leaders. Finally, we draw on the communication as constitutive of organizing framework to emphasize the role of online interactions in sustaining leadership.

Applying Leadership Theory to Online Communities

Given that no single theory of leadership seems uniquely suited to online communities, we propose that multiple theories can be productively applied to this setting. Three major attributes of online communities necessitate adaptation of existing organizational leadership theories. First, like other voluntary collectives there are few participants with formal power. Rather than formal roles and responsibilities dictating interaction and communication norms, efforts are predominantly performed in informal voluntarily roles defined by behavior (Butler, et al. 2007, Collier and Kraut 2012). When they exist, positions of formal power (such as moderation) appear to possess a limited range of rewards and sanctions. There are no tangible resources to distribute and few formal sanctions short of removing content or members. Thus, governance and leadership structures are emergent and highly situated to each community's setting (O'Mahony and Ferraro 2007). Second, compared to formal organizations, online communities are dominated by bottom-up emergent processes rather than top-down centralized interventions. They are fluid as they morph and change their boundaries, yet retain their shape and basic characteristics (Faraj, et al. 2011). Finally, asynchronous written communication in online communities is intrinsically limited compared to face-to-face interactions. Participants lack the broad range of verbal nuances, non-verbal cues, and physical status characteristics that enrich other forms of communication. Yet, the online space is a social field where participants select distinct strategies of participation, produce and evaluate each other's content, and pursue distinction and marks of status. (Levina and Arrigara

Forthcoming). Thus, online community members are constantly engaged in contribution strategies that positively differentiate them from others.

Given these unique characteristics of online community dynamics and membership, we suggest that any theorizing of online community leadership will require a contextualized synthesis of the four leadership theories describe in the previous section. First, we must build on the functional leadership theory in order evaluate the specific behaviors that differentiate leaders from non-leaders. Second, the online setting allows us to explore specific ties and interactions between leaders and non-leaders and thus offers a unique opportunity often unavailable in face-to-face settings. Third, given the size of the community and fluid membership, indications are that leadership is broadly distributed and thus would be shared. Finally, the CCO theory with its emphasis on how communication flows enact the social structure is highly relevant for understanding online communities where by definition one only "exists" if they post. The balance of this section reviews existing empirical research on online community leadership to evaluate whether certain theoretical subtleties and empirical findings can further enhance our theorizing.

In a discussion of critical online community behaviors, Butler, et al. (2007) identify four distinctive categories of maintaining infrastructure, social control and encouragement, external promotion, and content provision and consumption. Both infrastructure maintenance and social control require formal powers. Only authorized users can configure supporting communication infrastructure or remove unwanted content or members. These activities provide leadership through "a process of originating and maintaining the role structure" (Bass and Bass 2008, p.18). In the moderated forums in our study, these activities are performed primarily by the designated roles of administrators and moderators. Other influential roles in building community do not require formal powers. Any community member can provide social encouragement, promote the community externally, or create and read content. Nonetheless, some individuals will be more influential than others as they perform these activities. Applying the typology of leadership definitions described by Bass and Bass (2008), these participants can be said to emerge as leaders through influence processes resulting in recognition of informal leadership by others.

Recent research on leadership in online collectives has investigated the adjacent settings of open source software development, Wikipedia, and online communities. In a study of the governance structures and leadership in open source software development, O'Mahony and Ferraro (2007) collected interviews, secondary data, and project documentation to understand phases of governance over a 13-year period. Through an analysis of 815 participants during a time-period of stabilizing governance they identified behaviors and characteristics that increased the likelihood of being assigned to a formal leadership role. They found that tenure, the quality of contributions, and degree centrality all predicted leadership team membership. Likewise, Fleming and Waguespack (2007) performed longitudinal analysis on 16-years of membership in the Internet Engineering Task Force to identify the types of human and social capital associated with moving into formal leadership roles in this voluntary community. They also concluded that network position (boundary spanning) and quality of technical contributions are associated with formal leadership roles. Although the communities in both studies have substantial in-person interactions, they nonetheless support the proposition that network position predicts leadership.

Multiple studies have also considered leadership characteristics in the production community of Wikipedia. Collier and Kraut (2012) analyzed 2,442 candidates under consideration for the formal leadership role of Administrator in Wikipedia. The data, spanning six years of leadership deliberations, supports the importance of network position to formal leadership advancement. In turn, Zhu, et al. (2012) used a novel machine learning technique to evaluate the effectiveness of the leadership styles. Through the analysis of 1.6 million messages by 31,676 unique Wiki editors, they found both that leadership styles vary in effectiveness based on roles and that those in formal roles are more influential than other leaders. Together, these studies demonstrate the importance of formal roles in production-oriented communities as well as the value of considering structure and linguistics in relationship to leadership in online collectives.

Huffaker (2010) took a linguistic analysis approach to develop a comprehensive look at leadership behaviors in online communities. Focusing on the leadership role of generating interaction, he analyzed the structural and linguistic characteristics of the participants whose posts have the most impact on community discussions. Based on 2-years of communication in 16 Google Groups, his study encompasses over 600,000 messages from over 33,000 participants. Controlling for communication frequency, he found strong support for the importance of multiple structural and linguistic measures in triggering replies, creating conversations, and diffusing group-specific language. Specifically, "online leaders influence others through high communication activity, credibility, network centrality, and the use of affective, assertive, and linguistic diversity in their online messages" (Huffaker 2010, pg. 593).

In summary, our review of prior empirical findings lead us to conclude that leadership is influence based and involves aspects of all four theories described above. Thus, we propose an integrated model for leadership in online communities and adopt a popular definition of leadership (Yukl 2010): a leader is someone who is viewed by other participants as influencing what the online community does or how it does it. We further draw inspiration from a guiding idea of these studies: leadership is associated with participant roles and structural position. In the next section we describe a model of online community leadership that also incorporates the importance of language usage.

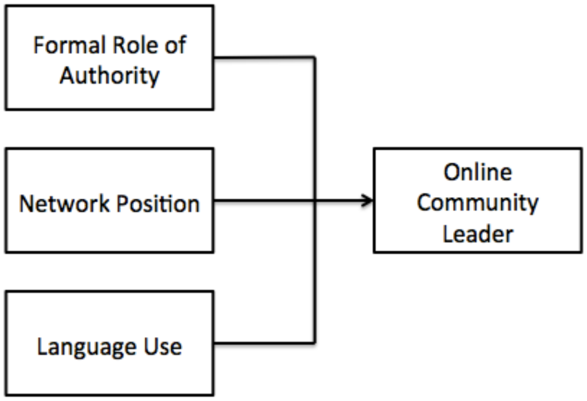

Figure 1: Model of Leadership in Online Communities

Model of Online Community Leadership

Our study builds on these previous empirical findings by identifying leaders through peer nominations and adopting a robust set of structural and linguistic measures. We investigate the characteristics of online community participants associated with exhibiting leadership. As shown in

Figure 1, three characteristics of leadership stand out as most relevant to online communities and other open communication networks. First, participants in formally designated roles are more likely to be viewed as leaders. Second, filling an informal leadership role is not a single static designation but, rather, an emergent role based on the structure of repeated interactions. Third, not only is the behavior of regular interaction important but also communication qualities of those interactions matters.

Formal Roles of Authority

Many online communities, including those studied in this paper, grant formal authority to designated administrators and moderators. Typically, these participants have community-recognizable handles or identifying markers in their signatures, and as a result are likely to "have disproportionate influence, through possession of consensual prestige or the exercise of power, or both, over the attitudes, behaviors, and destiny of group members" (Hogg 2001, p.188). There are three distinct processes that suggest those filling these formal roles are will also be online community leaders. These relate to their recruitment, status characteristics, and role behaviors.

First, administrators and moderators are typically recruited from the most active and engaged participants. For an online community to grow from a handful of active users to hundreds, thousands, or more, the number of participants providing leadership must also grow (Butler, et al. 2007). The most likely candidates for filling formal roles of authority are those who have demonstrated leadership qualities such as the interest and aptitude in helping to shape community dynamics. Given the importance of repeated interaction in establishing one’s status online, active participation can be regarded as a sine qua non condition of being considered a leader.

Second, in the online communities in this study the formally designated roles of administrator and moderator are prominently displayed next to all content posted by those users. With traditionally significant status characteristics (e.g., age, gender, ethnicity, personal appearance) largely hidden online, those that remain are even more salient (Hogg 2001). Finally, although these roles enjoy few formal capabilities, administrators and moderators play a major role in structuring interactions in online communities. Structuring of participant interactions is a leadership behavior (Reicher, et al. 2005).

Moderators and administrators can move and remove content, ban members, and use the threat of such to coerce desired behaviors. In summary, we propose:

> *Proposition 1: Occupying a formal role of administrator or moderator is positively associated with online community leadership.*

Communication Network Position

A central activity of online communities is written communication visible to all participants. Participant posting creates a communication network that is amenable to social network analysis. This approach has been applied both to the larger question of the structural role of leadership as well as to leadership in online communities. The emergent consensus is that leaders score highly in various centrality measures and also play a boundary spanning role in order to acquire information or resources (Balkundi and Kilduff 2006, Barge 1994).

In online settings, given the lack of face to face connection, communication network position is primarily based on where online contributions are made, how new ties are formed, and how those ties influence others' impressions (Dahlander and Frederiksen 2012, Donath 2007). Empirical studies indicate that online leaders tend to be longer term participants of the group, entertain more ties with different others, and post frequently (O'Mahony and Ferraro 2007). Yet, in communities based on knowledge sharing, online leaders are not necessarily more "chatty" than others. In a study of a legal community, Wasko and Faraj (2005) found that experts, while being more central, were also suspicious of the validity of content provided by non-experts and engaged in exchanges with little expectation of reciprocity. Taken together these studies indicate the importance of a holistic view of leadership in communication networks.

For example, in online communities there is support for leadership being associated with network centrality, though results diverge on the type of centrality. A study of 16 Google Groups (discussion forums) by Huffaker (2010) found that expansiveness (out-degree centrality) was associated with leadership behaviors but brokering (betweenness centrality) was not. Looking at virtual collaboration supported by Second Life and in text-based chat rooms, Sutanto, et al. (2011) found that both degree and betweenness centrality were associated with emergent leadership, but closeness centrality was not.

Closely related to centrality, the concept of core/periphery provides a complementary understanding of the structure of a communication network (Borgatti and Everett 2000). Compared to continuous measures of centrality, core/periphery suggests that there are distinct sub-groups of participants with jointly occupied, structurally equivalent positions. Core/periphery structures have been identified in smoking cessation (Cobb, et al. 2010) and video-blogging online communities (Warmbrodt, et al. 2008). Membership in the core is associated with leadership in open source software developer communication networks (Crowston and Howison 2005) and in Wikipedia (Collier and Kraut 2012).

The online communities studied in this paper are supported by asynchronous discussion boards and are organized with participation structures of threads and forums. Some participants may focus their participation within a single topic, while others may have participation spanning the topic boundaries of threads and forums. The former have low boundary spanning and the latter have high boundary spanning. Although high boundary spanning has been associated with leadership characteristics in multiple domains including knowledge-intensive work (Levina and Vaast 2005) and open innovation communities (Fleming and Waguespack 2007), overall evidence of an association is mixed (Reagans and Zuckerman 2001). The primary value of boundary spanning is derived through information brokering (Burt 1995). In online communities where posts are visible to all members, the ability to broker information is reduced. Further, our sample is of complex knowledge-rich topics where no single individual can be an expert in all areas. This further reduces the ability to broker information and favors participants who have deep, rather than broad, knowledge to share. Thus, we argue that participants who are central in the communication network, are part of the communication network core, or have low boundary spanning are more likely to be online community leaders than others. In summary, we propose:

> *Proposition 2: Communication network position (high centrality, a core position, and low boundary spanning) is positively associated with online community leadership.*

Language and Leadership

Both where and how communication occurs are salient to organizational processes. Indeed, many scholars have recognized that "communication is the medium through which leadership occurs" (Barge

1994, p. 29). For example, discursive leadership theory focuses on communication to understand behaviors consistent with leadership (Fairhurst 2007). Barge (1994) stresses that linguistics is integral to phrasing persuasive messages yet should also be tailored to individual context. Barrett (2008) notes the importance of language as a way for leaders to influence others and recommends use of concise positive messages. In regards to leadership in online communities, these works suggest that it is not just a matter of which participants communicate with each other but also the characteristics of that communication.

Looking more closely at communication online, multiple studies find an association between language usage and demonstrating leadership. Yoo and Alavi (2004) analyzed communication among team members of seven executive student project teams. They found that the team members that emerged as leaders wrote longer and more frequent emails than other team members. Wickham and Walther (2007) analyzed discussions of 18 small groups working on a decision-making task. They also found that higher levels of communication activity were consistent with being identified as a group member exhibiting leadership. In a review of leadership perceptions in both online and offline small group settings, Hollingshead (2011) notes that quantity of participation is highly correlated with leadership. However, her review also notes that in knowledge-oriented forums the quality of participation is more closely associated with leadership than merely quantity of participation. Frequency alone is not enough to be a leader, the quality of communication also matters. Additional studies analyze behaviors in terms of leadership styles and language characteristics. For example, Huffaker (2010) found that participants identified as online community leaders used more affective and assertive language than others. Finally, in a study of influence behaviors seen in Wikipedia, Zhu, et al. (2012) identified different types of language used as associated with different leadership styles. Together, these studies support the general idea that leadership is associated from differences in language use in online settings but provide limited guidance regarding specifically how those differences manifest themselves.

We propose a linguistic analysis model to shed further light on language use consistent with being considered a leader in online communities. Drawing on work in computer science, artificial intelligence and linguistics, Natural Language Processing (NLP) offers a promising approach to enable computers to derive meaning from human utterances, but more crucially, to derive a multidimensional understanding of bodies of text (Clark, et al. 2010, Mitkov 2005). The NLP approach splits language analysis into theoretical (often hierarchal) levels (Mitkov 2005). Our operationalization of the NLP algorithm relies on generating data along these four major dimensions of language: morphology, lexicography, syntax and semantics.

Several organizational leadership theories support the relevance of language to leadership. Empirical evidence from written electronic communication finds that leaders use language differently than non-leaders. NLP provides a systematic approach to identifying those differences. We propose five core linguistic features, each mapping to one of the four major dimensions of language, as consistent with online community leadership: readability (morphology), vocabulary richness and external linking (lexicography), proto-typicality of vocabulary (syntax), and positive sentiment (semantics).

First, we propose that text readability is positively associated with leadership communication. Brevity and succinctness are characteristic of effective communication (c.f., Zinsser 2006) and conciseness is recommended to achieve a leadership purpose (Barrett 2008). Holding all else equal, an online participant who can express their ideas simply (with improved readability) is more likely to influence others than one who expresses their ideas in a difficult to read manner.

Readability and vocabulary are related, yet distinct, linguistic features. Whereas readability is based on characteristics of single words (e.g., length, number of syllables) and sentences (e.g., number of words) vocabulary richness reflects how many different words someone uses. A body of text containing a large number and variety of short simple words is more readable with more vocabulary richness than text with a small number of long, complex words. All else equal, someone who commands a larger vocabulary has more tools available to word and reword ideas, thus to better influence others. Huffaker (2010) in his study of 16 Google Groups measured online leadership as having influence on the communication behaviors of other group members. Participants with increased linguistic diversity had more influence than those with less linguistic diversity.

Providing direct access to online resources (via hyperlinks) is another aspect of online written communication. Providing URLs in online text is likely to increase online influence for multiple distinct reasons. First, providing a link can serve as verifiable evidence for arguments made in online rhetoric. Second, a link may provide a resource of value to the online community, be it news, information, or entertainment. Third, a link may directly address a question or concern of others. These are all pathways to influence and demonstration of leadership.

Another linguistic feature related to vocabulary is how distinctive word choices are in relationship to the frequency of words used by others. In an online community this takes the form of comparing all of the words used by a single individual to all of the words used by the rest of the community. The closer an individual's word usage is to the collectives', the more they represent an average or proto-typical vocabulary use for that collective. In social identity theory prototypicality is both a process leading to, as well as an outcome of, social influence and leadership (Ashforth and Mael 1989, Hogg 2001). Thus, we expect that the more prototypically a participant uses the vocabulary of an online community, the more likely they are to be identified as an online community leader.

Finally, we consider the role of sentiment in communication. Sentiment is commonly associated with leadership. Leader mood is contagious (Sy, et al. 2005). Positive emotions create affective bridges that serve as channels of influence and some scholars view that "shared affect could be more salient basis for group formation than shared cognition" (Weick 1969, p. 14). Leadership emerges from positive sentiment and micro-effective events (Johnson and Dasborough 2008). In summary, we propose:

> *Proposition 3: Unique patterns of language use (readability, vocabulary richness, external linking, proto-typicality of vocabulary, and positive sentiment) are positively associated with online community leadership.*

Research Method

Research Design

Testing the propositions requires four different types of data. First, survey data is used to measure the dependent variable of online community leadership; participants in three communities focused on technical topics were asked to identify other participants who they regarded as most influential in what that online community did or how that online community did it. Second, the participants in formal roles of authority (administrator or moderator) were collected from public lists posted at each of the three online communities. Third, the structure of the communication network was gathered through automated collection that identified how participants interacted through threaded discussions. Finally, the full text of all of the parsed posts was collected to document the content of interactions. This text forms a corpus of full text messages analyzed with natural language processing (NLP) algorithms. Because the focus of this study is to compare participants (online community leaders vs. other participants), measurements are aggregated to the participant level.

Table 1 provides an overview of the targeted communities, all of which focus on technical topics and use vBulletin, an open, asynchronous, web-based message board technology. Community discussions are organized by message thread, with each message thread belonging in a single higher-level topic forum. These communities were chosen from a sample of several dozen online communities surveyed in spring, 2008, for a broader-based study of online participation (citation blinded). These three were randomly selected from those with at least a dozen survey-nominated leaders and a year of pre-survey full message-level communication available.

Table 1: Online Community Descriptive Statistics

Online Community	**Blender Artists**	**Gearbox Software**	**Northern Sounds**
Tagline	Community of artists using Blender, a 3D creation tool	The official community of Gearbox games	Northern Sounds software
Collection URL	blenderartists.org/forum	gbxforums.gearboxsoftware.com	northernsounds.com/forum
Inception	14 October 2001	12 July 2002	1 February 2003
Members	10,264	1,644	2,488
Forums	28	27	37
Threads	32,656	5,383	8,472
Posts	308,682	118,924	51,472
Words	17,088,714	4,673,512	4,357,698
Survey Responses	19	16	21
Participants in Formal Roles	8	11	8
Online Community Leaders	23	21	15

Identifying Online Community Leaders

To test the propositions we needed to identify participants who are online community leaders. Most existing empirical studies of leadership behaviors outside of formal positions of authority focus on small work teams in educational or organizational settings. As such, they predominately operationalize leadership through work team peer ratings where each participant rates the leadership qualities of all other team members (Pfeffer and Cialdini 1998, Walter, et al. 2012). This approach is not practicable in online communities with thousands of active participants. Instead, online community leaders were identified by asking survey respondents to name up to three participants who had the most influence on what the communities does or how it does it; wording directly applied from Yukl's (2010) definition of leadership. The terms "leader" and "leadership" were intentionally avoided in the prompt so that respondents would focus on leadership outcomes rather than formally designated roles. Consistent with our theoretical stance that online community leadership is a local temporary designation, we consider anyone identified by a fellow participant as demonstrating leadership to indeed be an online community leader. As such, the dependent variable in our analysis is an ordinal value reflecting if any other participants have nominated the focal participant as a leader.

Table 2: Relationship Between Identified Online Community Leaders and Participants in Formal Roles

	Blender Artists	Gearbox Software	Northern Sounds	Total
(A) Identified Online Community Leaders	23	21	15	59
(B) Participants in Formal Roles of Authority	8	11	8	27
(C) Identified Online Community Leaders also in a Formal Role of Authority	4	5	3	12
% of Participants in Formal Roles (B) who are also Online Community Leaders (C)	50% (4 of 8)	45% (5 of 11)	38% (3 of 8)	44% (12 of 27)
% of Online Community Leaders (A) who are also in a Formal Role (C)	17% (4 of 23)	25% (5 of 21)	20% (3 of 15)	20% (12 of 59)

Identifying participants occupying formal roles of authority is a straightforward process. The three online communities in our sample each display a public list of moderators and administrators. This list was captured immediately prior to the survey response period. As demonstrated in **Table 2** participants identified as online community leaders and participants with formal authority are related, yet distinct categories. Consistent with the first proposition (above), 44% of those in the latter role are also in the former (50%, 45%, and 38% respectively in the three communities). The validity of the online community leadership construct is strengthened by the much smaller overlap between the two categories. Only an average of 20% of identified online community leaders are moderators or administrators (17%, 25%, and 20% in the three communities). The two measures have a 0.33 correlation (p<0.001) in the analyzed sample. (Correlations are provided in Appendix A for all of the study measures.)

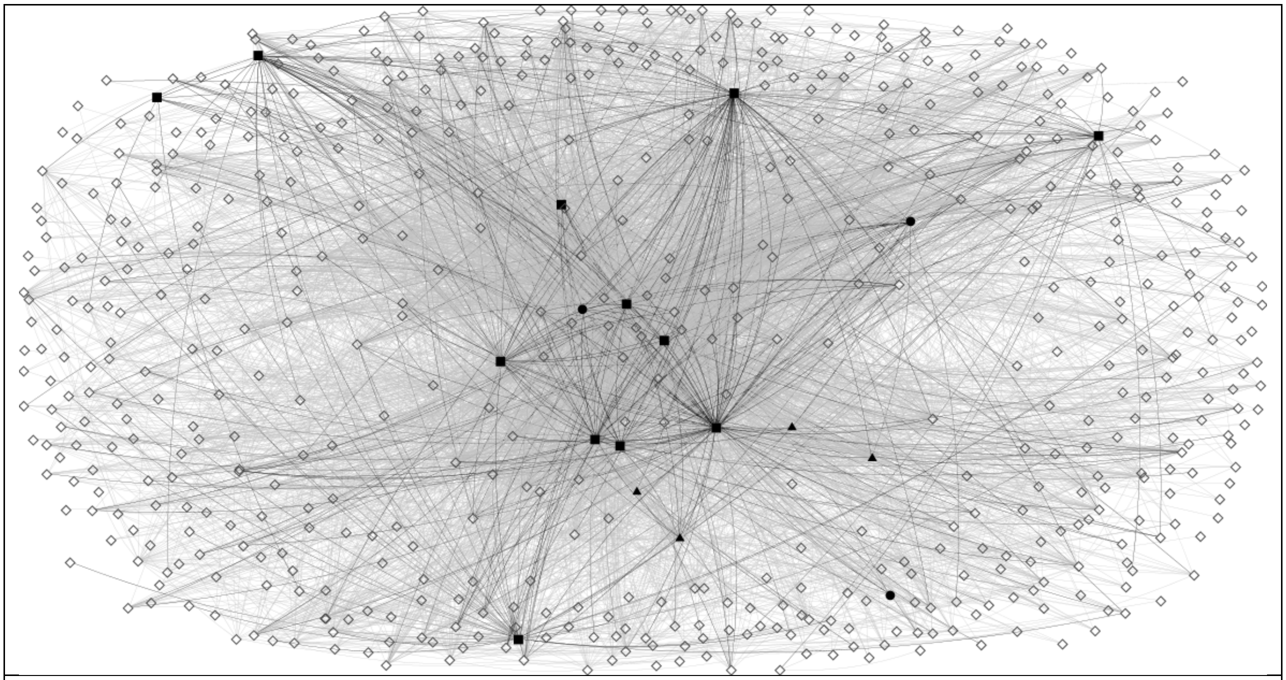

Figure 2: Communication among active participants of Northern Sounds Community

Structural Model

Archival data was used to calculate the communication network structure measures of centrality, core position, and boundary spanning. Communication in the studied communities was modeled as affiliation networks with two node types: participants and threads (Borgatti and Halgin 2011). A communication network link (i.e. an edge in the graph) exists between a participant node and a topic node when the participant posts to a message thread. The network link carries the attribute of the communication such as the text, the date of the post, and the order of the post in the discussion. Figure 2 shows a graph representing the relationships among active participants (top 20% frequent participants). Black circles represent online community leaders who are also in a formal role of administrator or moderator. The black squares are other online community leaders. The black triangles are administrators and moderators not identified as leaders. White diamonds represent all other participants. Dark edges represent communication originating from community leaders while light edges represent all other communication in the community.

Several items of note can be seen in this representation of the active core of this community. First, a central network position is neither necessary nor sufficient for being an online community leader. Leadership is a local designation; within the core leaders occupy both central and peripheral positions. Second, there is no discernable relationship between serving in a formal role of authority and network position. Although a relatively high percentage of those in formal roles are also identified as online community leaders, network position does not suggest which ones. Third, although a small fraction of the community, leaders communicate with many other participants regardless of their formal designation or network position. This is visually evident by the spread of darker edges over the community.

Modeling the community as a communication network also allows for the investigation of network characteristics associated with nodes and to numerically ground the observations drawn from examining the graphical representation (Knoke and Yang 2008). Several measures exist to describe centrality in a network. First, degree centrality, the fraction of nodes in the graph connected to the node under consideration. Second, betweenness centrality, which is the sum of the ratio of all pair's shortest paths that pass through a focal node. Third, closeness centrality (measured as the reciprocal of the normalized average distance to other nodes), a measure of how long it takes to sequentially disseminate a message to all other nodes in the network. As these three network measures were highly correlated in our sample, only one was retained for the measurement model. Betweenness centrality was chosen as the most theoretically meaningful of the three measures as it takes into account both the local connections of a node as well as its global position in the network (Mehra, et al. 2001). Supporting this approach, post-hoc analysis using alternative measures of centrality did not affect our results.

We also use core-periphery measures in order to account for the possibility that the network had an active core of highly involved participants primarily engaged with each other and a larger periphery of less involved (peripheral) participants. Indeed, studies of online communities show that the behavior of participants is impacted by their position vis-à-vis the core (Dahlander and Frederiksen 2012, Liu 2011). The k-core number is used to divide the network into layers of cores. The cores are sub-networks with k connectivity. For example, it is possible to fragment the 1-core by deleting one edge, while at least two edges need to be deleted in order to fragment the 2-core. A higher k-core number occurs for nodes that are located in a densely connected parts of a network (Seidman 1983).

Boundary spanning is operationalized as a ratio of number of messages to the breadth of areas those messages appear in. Specifically, boundary spanning is measured as the ratio of the number of unique threads a participant posted messages in, divided by their total number of posts. Lower values of this measure occur when a participant concentrates their posts into a smaller number of threads. Higher values occur when a participant posts messages in many different threads. Thus, this measure reflects the degree of specialization of topics. Together, centrality, core position and boundary spanning provide a robust structural assessment of position in the communication network.

Model of Language Usage

Advances in both online data availability and computing processing speed have opened up new opportunities for automated communication analysis. The application of both natural language processing (NLP) and computation linguistics (CL) have recently flourished (Clark, et al. 2010, Jurafsky and Martin 2008) and are well suited to help assess how subsets of a group, such as leaders and others, differ in language usage. NLP and CL are used in real life applications including speech recognition, text translation, and question answering that exist in a wide variety of platforms ranging from mobile devices such as the Siri personal assistant on the iPhone (Aron 2011) to supercomputers such as IBM Watson, the world champion of Jeopardy (Ferrucci 2010). The ultimate goal of NLP is to mathematically model the understanding and the generation of human language.

In this paper we apply NLP algorithms to better understand how language usage is associated with online community leadership. Table 3 provides a representative sample of posts from the Northern Sounds community by different participant types. Examining these posts gives a rudimentary idea of how leaders may differ from non-leaders in terms of language use. Online community leaders tend to use simple positive language with high readability. Other active participants tend to use less familiar language with more complex sentences. We expand on previous research identifying the importance of language usage (Hollingshead 2011, Huffaker 2010, Zhu, et al. 2012) by applying a systematic computational approach to quantify online community participant messages.

Table 3: Example Posts in Northern Sounds Online Community by Participant Type

Participant Type	**Representative Post Text**	**Characteristics of Text**
Online Community Leader In Formal Role of Authority	*"Nieves, …" Excellent rhythms going on in this piece. Excellent percussion writing and full of energy. I agree with Reegs that Drumlines would like this. I can't wait to see what you can do with the Marching Band library. Keep on doing what you are doing.*	Simple positive language with high readability
Online Community Leader not In Formal Role of Authority	*Thank you very much you all good people, who made this course possible. It was a great experience.*	Simple positive language with high readability
Active Participant In Formal Role of Authority	*I have been a member of this community for a while now but this is the first time I have posted a work in the Listening Room. http://www.michaelsroom.co.uk/Handel - Organ Concerto in F (The Cuckoo and the Nightingale) 2nd Movement.mp3 For those who are unfamiliar with this, the nickname (Cuckoo and Nightingale) comes from the bird song motifs to be heard in this movement. Coincidentally, this work was completed by Handel in April (2nd) 1739.*	Less typical language for the community.
Active Participant not In Formal Role of Authority	*In my opinion, wait a few months and see how GS4 turns out and let the bugs get fixed, then its time to get 1 machine that is as bad to the bone as you can afford (quad core, 8 gigs ram?) with a 64 bit OS (xp64 or vista64) and upgrade to GS4. I think once the kinks are worked out, and knowing tascam there will be some lol, that it will be quite awesome to go with a beefy GS4 machine*	Lower readability.

Given the complexity of human language in action, the NLP approach splits language analysis into theoretical (often hierarchal) levels (Mitkov 2005). Our operationalization of the NLP algorithm relies on generating data along these four major dimensions of language: semantics, syntax, lexicography, and morphology. Together, these measurements allow the analysis of how language is used in online communities using multiple and complementary linguistic perspectives (a prism model). Like a prism breaking light into its full spectrum, NLP analytical techniques (Bird, et al. 2009) break text into multiple components (Figure 3). Identifying a full spectrum of linguistic characteristics provides a robust method to compare leaders and other online community members based on their expressed language corpus. The starting point for our analysis is morphology (the sub-word level) and continues to semantics (the meaning of text). The goal is not to completely model the use of language at each intervening level, but rather to identify representative indicators in order to compare participants of online communities. Each of the four levels (morphology, lexicography, syntax, and semantics) is described further below.

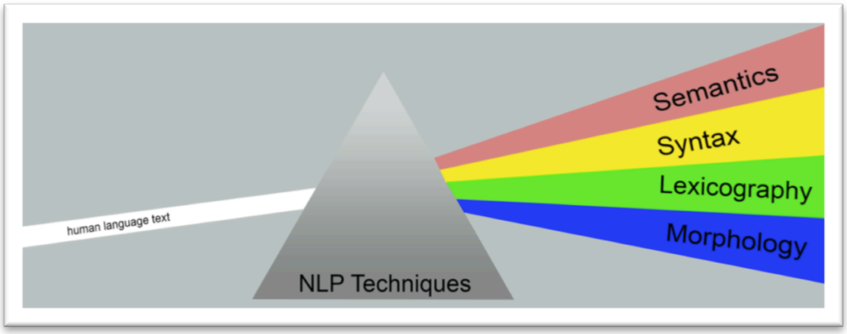

Figure 3: The Natural Language Processing Analysis Prism

Morphological Analysis. Morphology studies how words are formed in natural language. More precisely, it is how the words are segmented into components that form those words via concatenation (Goldsmith 2001). Two main types of such decomposition exist: morpho-phonology, in which the subcomponents correspond to spoken syllables, and morpho-syntax, in which the subcomponents are syntactical (such as prefixes and suffixes). An example measure is the number of syllables per word. Words with more syllables are considered more complex; their usage indicates a higher command of language (Gunning 1969). At the morphological level there are three well-known indices of readability: the Automated Readability Index (ARI), the Flesch-Kincaid Reading Ease (Kincaid, et al. 1975), and the Gunning-Fog Readability Index (Gunning 1969). Because all three measures are highly correlated in our data set, we choose the simplest of the three, the ARI, for analysis. The ARI takes into account the number of characters, words, and sentences in a post. It yields higher score for longer words and longer sentences. It is computed with the following equation:

$$ARI = 4.71 * \frac{\#characters}{\#words} + 0.5 * \frac{\#words}{\#sentences} - 21.43$$

Lexical Analysis. Whereas morphological analysis focuses at the sub-word level, lexical analysis focuses on characteristics of participants' vocabulary at the level of words. For example, the number of words used by an online community participant can be assessed as well as the qualities of those words (e.g., by matching them against precompiled dictionaries). First, a dictionary is compiled for each participant. The dictionary contains the unique words that the participant used in their posts. Next, the size of participants' dictionary is normalized based on their number of posts. The resulting indicator (vocabulary richness) measures the richness of vocabulary. In addition, the use of hyperlinks is considered as a special vocabulary. The average number of links per post is also calculated for each participant.

Syntactic Analysis. The syntactical level examines how words are combined and used to form sentences and posts. This paper adopts an NLP technique called Statistical Language Modeling (Jurafsky and Martin 2000) that assigns a probability to a sentence in a textual corpus given the likelihood of that sentence based on the rest of the textual content of the corpus. This probability is an indicator of whether that sentence conforms in its syntax with the rest of the sentences in the corpus. For example, if we take the Bible as a corpus, a sentence like "Computers are used to process words." will have a very low probability, whereas a sentence like "The word was with God." will have a high probability. The latter sentence is thus more proto-typical of the Bible than the former.

Proto-typicality is measured as the inverse of entropy, a sophisticated approach to comparing individual and group language usage. Statistical language models (Jurafsky et al., 2000) allow for calculating the likelihood that a word, a sentence, or a set of sentences is representative of the language used in a bigger collection of text, also called the text corpus. As such, it can be used to compute the probability that a post was authored by a participant of the community given what all other participants wrote.

Building a model that estimates the exact sentence probability is computationally challenging because of the large number of potential word combinations in sentences of varying length. The solution is to approximate the computation of the sentence's probability by considering word sequences of fixed length in the sentence. Those sequences are called N-grams. A sequence of one word is called a unigram, a sequence of two words is called a bigram, and a sequence of three words is called a trigram. In most cases a trigram model (n=3) provides a strong approximation and has been considered the standard statistical language model for more than 30 years (Clark, et al. 2010). A trigram statistical language model is used. For ease of exposition a bigram model representation is shown in Table 4.

Table 4: Computing the Probability of a Sentence W Composed of N Words using a Bigram Model

(A) Probability of a bigram	$\mathrm{P}(\mathrm{w_n} \mid \mathrm{w_{n-1}}) = \mathrm{Count}(\mathrm{w_{n-1}w_n})/\mathrm{Count}(\mathrm{w_{n-1}})$
(B) Probability of a word in a sentence is based on (A)	$P(wn \mid w) \approx P(wn \mid wn\text{-}1)$
(C) Probability of a sentence is based on (B)	$P(W) = P(w_1 w_2 \dots w_n) = \mathrm{P}(\mathrm{w_2} \mid \mathrm{w_1}) * \mathrm{P}(\mathrm{w_3} \mid \mathrm{w_2}) * \dots * \mathrm{P}(\mathrm{w_n} \mid \mathrm{w_{n-1}})$
(D) Word entropy of a sentence is based on (C)	$Entropy(W) = -\log(P(W))/N$
(E) Prototypicality is based on (D)	*Entropy * -1*

The probability of a two-word sequence (i.e. a bigram) is estimated by how many times the two words occur together in the text corpus divided over the frequency of the first word. The bigram and unigram probability are estimated from the text corpus. Those probabilities are used then for assessing complete sentences that can come from the text or can be new unseen sentences. In a bigram model, the probability of a word in a sentence is approximated by the bigram probability, i.e., the bigram model looks back one word only to assign a probability to a word in a context. Next, the probability of a sentence is the multiplication of the probability of the sequence of bigrams in the sentence. Because the probability of a sentence is a very small number and because this number depends on the length of the sentence, the entropy of a sentence is defined as the negative log of its probability divided over the number of words in the sentence.

The entropy number can be used to compare two sentences in light of the training corpus. The sentence with higher entropy has lower probability of occurrence and is (probabilistically) more unique than the first one, while that of lower entropy has a higher probability of occurrence and is (probabilistically) more of an average sentence than the first one. Bringing this measurement to the level of the participant, participants whose posts have on average higher entropy are contributing unique posts that deviate from what other participants are writing and vice versa. Thus, participants whose contributions are characterized by high levels of entropy are not prototypical of the conversation of the community and are possibly offering novel information. As such, prototypicality is measured as the additive inverse of entropy.

Semantic Analysis. At the level of semantics, the goal is to go beyond the structure of words and sentences to identify the meaning of what is written. As such the semantic analysis is the first step of natural language understanding—a step that is considered the most challenging in linguistic analysis (Shahaf and Amir 2007). A simple form of semantic analysis is assessing the sentiment of posts in terms of their polarity (i.e. positive vs. negative) (Pang and Lee 2008). For our study, the sentiments expressed in posts were assessed in terms of their negative vs. positive polarities. A word-based classifier of sentiments based on a dictionary of emotionally-rated English words, AFINN (Nielsen 2011), was used. The dictionary used is a customized version of the package ANEW (Bradley and Lang 1999), which provides a set of normative emotional scores for a large set of English words. AFINN customized the ANEW dictionary to tailor it to the Internet language of web logs, discussion forums and tweets (Nielsen 2011). In addition, AFINN associates a score to each post based on the emotions expressed within the words of that post. A negative score implies negative emotions or polarity in the post and vice versa for a positive score. A score of zero implies a neutral tone.

A number of additional pre-processing and data cleaning steps were required to facilitate linguistic analysis of participants' postings. First, all of the collected posts were pre-processed to remove HTML formatting (e.g., bold, italics). Second, the special content of web links and formal quotes of other participants was identified and removed. These filtering steps are important (a) to focus the language modeling toward what a participant says rather than the text that is being repeated from a previous post and (b) because many of the linguistic measures are poorly suited to marked-up text. Finally, the linguistic characteristic of each post was assessed with measurements aggregated per online community participant. As described above, the four dimensions of language process focused on are morphology, lexicography, syntax, and semantics.

Model Testing

Because online community leadership is a binary categorical variable, logistic regression (Long and Freese 2006) is a natural choice to model its relationship with the independent variables (summarized in Table 5). Since participants are nested within communities, we use a random-effect logistic regression model where the effect of community membership on emergent leadership role is a random-effect coefficient. Three potential concerns arise regarding this analysis approach: first, the dependent variable is sparse (i.e. very few of the participants were nominated as leaders). Second, most of the structural independent variables are not normally distributed. For example, the centrality and core variables follow a power-law that is commonly found in network data (Faraj, et al. 2008). Third, because of the power-law distribution and the large sample size many outliers are found in most variables. The three concerns are interrelated and are indeed typical characteristics of large networks.

Table 5: Participant Measures

Measurement	**Type**	**Description**
Online Community Leader	Leadership	Identified as a leader by a fellow participant
Formal Role of Authority	Leadership	In a formal role of authority (administrator, moderator)
Centrality	Structural	Betweenness centrality of participant (larger value is more central)
Coreness	Structural	k-core number of participant node (larger value is more in the core)
Boundary Spanning	Structural	Ratio of number of unique message threads posted in divided by total number of posts.
Readability	Linguistic (Morphology)	Automated Readability Index
Vocabulary Richness	Linguistic (Lexicography)	Vocabulary Richness (average number of unique words per post)
External Linking	Linguistic (Lexicography)	Average number of web links
Prototypicality	Linguistic (Syntax)	Prototypicality of participant language use when compared to other participants of the same community
Positive Sentiment	Linguistic (Semantic)	Average sentiment polarity score

The sparseness of data could affect the power of the statistical analysis but this is addressed by the large sample size (thus providing greater statistical power). However, the large sample size poses its own concerns. The large data set we obtained (14,396 participant observations across the three communities) can be criticized from a theoretical and a practical perspective. First, from a theoretical perspective, communication networks exhibit a power-law distribution structure (Faraj, et al. 2008). Most participants contribute little while a few contribute a lot to the community. Indeed, 40% of participants in the three communities contributed only one or two posts. Therefore, it is problematic to compare leaders to everyone else knowing that most participants contribute few posts. A more appropriate comparison group is community participants who are also frequent contributors, but were not nominated as leaders. Second, from a practical perspective, the large sample size leads to statistically significant results. The ability to harvest large data sets from the Internet may be problematic when practical interpretation of significant coefficients is difficult (Royall 1986).

To address all three issues in a rigorous way, we focus on the most theoretically and computationally relevant sub-set of the available sample. Because we are interested in comparing leaders to participants who are equally engaged and contribute to the community but were not identified as leaders, we use the number of messages a participant posts to the community as a threshold variable. This variable follows a power-law distribution. We keep observations of participants with their number of messages in the top 20% percentile. This corresponds to more than 18, 64, and 14 messages in Blender Artists, Gearbox Software and Northern Sounds respectively. The new sample size is reduced to 2,947 observations from the original 14,396 observations. Finally, we used robust standard error estimation to deal with the misspecification of the normal distribution in the independent variables (variables with power-law distribution remain so in reduced sample size as power-law distribution is scale free). Descriptive statistics and correlation tables are provided in Appendix A.

In summary, we use a hierarchical analysis technique to test the extent of association between our hypothesized variables and online community leadership. The first model analyzes the association between leadership and formal roles of authority. The second model analyzes the impact of network measures in combination with formal roles. The final model adds linguistic variables. All three models are random-effect logistic regressions. Finally, in order to more directly interpret results, we have standardized all of the research variables except the two (binary) leadership variables.

Two types of evaluations are employed to compare the three models. First, we look at the independent variables' coefficients and goodness of fit indices in the three models to evaluate the effect of these variables on leadership in the studied communities. Second, we perform two model difference tests comparing the nested models in order to judge the added value of linguistic variables in determining leadership. We also perform post-estimation tests for the full model (C) in order to ensure the validity of the results. We test for multicollinearity using variance inflation factors and we test for overfitting using cross validation. Results of these additional validation steps are reported in Appendix A.

Results

The analysis results are provided in Table 6. Examining the regression coefficients of participant-level variables, formal role is the most important predictor for online community leadership. Holding everything else constant a participant who is in a formal role of administrator or moderator is 35 times more likely to be viewed as a leader than other active participants. Structural and linguistic variables are also important. For example, increasing centrality by one standard deviation increases the odds of leadership by 150%. Similar increments in coreness, readability, prototypicality and positive sentiment all enhance the chance of being identified as a leader. However, an opposite effect is found for boundary spanning and vocabulary richness: a one standard deviation increment in boundary spanning or vocabulary richness almost halves the odds of being identified as a leader. Only external linking turns out to be non-significant in predicting online community leadership.

The intragroup intraclass correlation reflects the effect of group membership on being identified as a leader. Because identified leaders are equally distributed among groups (Table 2), the intragroup correlation is on the low side (9% in model C). This is an important indicator because it suggests that group membership does not overshadow participant-level variables in predicting leadership. In order to evaluate the added value of structural variables and linguistic variables we perform two Chi^2 difference tests comparing models A and B, and B and C. The two tests are significant indicating that both structural (Δ Chi^2 = 99.89***, Δ df = 3) and linguistic characteristics of participants (Δ Chi^2 = 19.34**, Δ df = 5) are important predictors of online community leadership.

In summary, both structural and linguistic participant-level variables are associated with leadership in addition to formal roles of authority. On the structural side, a more central position toward the core of the community increases the odds of a participant being identified as a leader. The opposite is true for boundary spanning; a participant who spreads their messages across different threads and topics is less likely to be identified as a leader. On the linguistic side, a participant with language that is more readable, with a simpler vocabulary, more prototypical, and of more positive sentiment is more likely to be identified as an online community leader.

Table 6: Hierarchical Logistic Regression Model with Dependent Variable of Online Community Leadership

Measure	**Model A**	**Model B**	**Model C**
Group-level Coefficients			
Intragroup correlation	0.11	0.37	0.096
Participant-level Coefficients as Odds Ratios, Standard Errors in Parenthesis			
Formal Role of Authority (P1)	58.98*** (28.00)	37.45*** (20.16)	35.20*** (19.86)
Structural Measures (P2)			
Centrality		1.58*** (0.12)	1.54*** (0.14)
Coreness		3.90*** (1.44)	2.59** (0.87)
Boundary Spanning		0.50*** (0.10)	0.57** (0.12)
Linguistic Measures (P3)			
Readability			1.33* (0.19)
Vocabulary Richness			0.45* (0.15)
External Linking			1.21 (0.16)
Prototypicality			1.66* (0.35)
Positive Sentiment			1.51* (0.28)
Goodness of Fit Indices			
Log likelihood	-248.5	-198.6	-188.9
Chi2	73.75	115.4	118.6
AIC	503.0	409.2	399.8
BIC	521.0	445.1	465.7
Comparison with previous model (Δ chi2)		**99.89*** **	**19.34** **

N= 2947; Odds ratios; Standard errors in parentheses; * p<0.05, ** p<0.01, *** p<0.001

Model Validation

In order to ensure the validity of the analysis we employed several post estimation tests. Multicollinearity is of concern because of the potential overlap among related measures used to evaluate structural and linguistic characteristics. Although standardizing the independent variables can alleviate multicollinearity in data (Barry 2011), we also test for this posthoc. We computed the Variance Inflated Factors (VIF) of the regression variables after estimation (results in Appendix A). All factors are below 2 with an average of 1.31 indicating the absence of multicollinearity concerns.

Next we address the concern that our model overfits the research data and could have little validity outside the research setting. This is a typical issue in machine learning and classification tasks because of the existence of noise in the training set making it difficult to judge whether the model parameters had fitted the real data or the noise (Hart, et al. 2001). We have partially addressed this issue by reducing our dataset to 20% of the original sample size and focusing only on participants with large number of messages comparable to those of leaders.

As further validation of the model's robustness, we have conducted an area under the Receiver Operating Characteristic Curve (ROC) analysis and we computed a 10-fold cross validation analysis (Hosmer and Lemeshow 2005, Kohavi 1995). The Area under the ROC curve analysis is most suitable for testing the ability of two-class classifiers to detect the true signal and separate it from noise (Hosmer and Lemeshow 2005). The area under the curve ranges from zero to one with any value above 90% indicating an excellent discriminative ability. Our model achieved an area of 91.81%.

Next, we check external validity by conducting a 10-fold cross validation test (Kohavi 1995). The measurement sample was split into 10 randomly selected sub-samples and the following procedure is repeated for each of those 10 samples: (1) the measurement model is run for a sub-sample; (2) the value of the dependent variable in the remaining 9/10 of the data is predicted from the regression coefficients calculated in the sub-sample analysis; (3) the root mean square errors (RMSE) is calculated as a measure of the difference between the predicted values and the actual values (for the 9/10 sub-sample). As noted in Appendix A, the average RMSE of these 10 analyses is 12%, indicating good external validity. Because of the two-class setup, this RMSE value corresponds to 90% accuracy of predicting the classes of instances outside of its training set correctly (Alpaydin 2004). Although other learning algorithms (such as a naïve classifier) may achieve a better RMSE, taking into account that other goodness of fit indices were also good (Table 6), the model indicates good external validity (Wolpert and Macready 1997).

Sensitivity Analysis

We performed several additional tests both to assess the sensitivity of results to the analysis sample selection and also to further explore how different subsets of online community leaders compare to one another. Additional logistic regressions are provided in Appendix A for (a) the full data set and (b) the analysis data set with all participants in formal roles of authority removed. Support is found for all three propositions in both of these tests. This strengthens the conclusion that formal role of authority, structural characteristics, and linguistic characteristics are all associated with being identified as an online community leader.

The sensitivity analysis also provides more nuance regarding which individual structural and linguistic characteristics. Analysis of variance tests are reported for (a) a comparison of online community leaders with and without formal roles of authority and (b) a comparison of online community leaders identified by one participant to those identified by two or more participants. No significant differences in found in either structural or linguistic variables between online community leaders with and without formal roles of authority. Differences are identified, though, between leaders identified by a single participant and those identified by two or more participants. In terms of structural variables, online community leaders identified by two or more other participants post more and have higher centrality compared to those identified a single time. Online community leaders identified by two or more other participants have lower vocabulary richness, and higher prototypicality compared to those identified a single time. Using simpler language that is most familiar to the participants is consistent with the highest likelihood of leadership identification.

Discussion

The goal of our research was to investigate whether, beyond network position, online community leaders had distinctive written communication patterns. Using participant surveys to identify leading online community members, this study analyzes a year of communication network history and message content to identify if leader contributions have unique qualities compared to the utterances of other core community participants. We first examine how communication network position--in terms of formal role, centrality, membership in the core, and boundary spanning--affects the likelihood of being seen as a leader. Then we contribute a novel use of textual analysis to develop a language model of utterances in the community to evaluate how convergent or divergent leader language is compared to the community as a whole. Our findings suggest that the most influential participants of any online community, those viewed as leaders by other participants, are not just among the most central but also post a large number of positive, concise posts with simple language familiar to other participants. Thus, leadership is not merely filling an assigned role nor occupying a communication network position. Online community leadership is multi-faceted, enacted through unique language patters, and based on the perception of others.

Theoretical Implications

Our paper makes four major contributions to the understanding of leadership in online communities. First, our findings lend support to a multi-faceted approach for understanding leadership in online communities. We integrate four sets of empirics (formal leadership roles, peer nominations, network position, and content of utterances) to offer a deeper understanding of leadership processes in online settings. Our findings build on and augment previous studies that had prioritized network position as proxies for leadership (Huffaker 2010, Sutanto, et al. 2011) by delineating the relative importance of network position compared to other correlates of leadership. In addition, by comparing leaders to other participants of equivalent rank, we were able to establish the ways by which leaders demarcate themselves from other members at the core of the community. Thus, we offer a more fine-grained evaluation of the activities of those active in the core and thus extend earlier findings regarding the core-periphery perspective on online communities (Cobb, et al. 2010, Collier and Kraut 2012, Crowston and Howison 2005, Warmbrodt, et al. 2008).

Second, our findings align with recent theorizing in leadership theory regarding the importance of shared leadership in knowledge and team work (see Pearce and Sims 2002; Carson et al. 2007). Our findings align with this trend and show that shared and emergent leadership is strong in online communities focused on knowledge exchange. Just as there are several complementary leadership perspectives on leadership in organizations, there is a need to recognize a similar, if not richer, diversity in online settings where individuals generally do not know each other, have ambiguous identities, and are limited in their communications to text-based exchanges. The shared nature of online community leadership is not yet directly recognized in the literature but is in line with findings about emergent roles, fluid boundaries, and the seeking of position of influence in online communities (Butler, et al. 2007, Faraj, et al. 2011, Levina and Arrigara Forthcoming).

Third, this study has implications for how researchers study online communities. By asking directly community participants to nominate those they consider leaders we expand on previous studies of online leadership. For example, Collier and Kraut (2012), O'Mahony and Ferraro (2007) emphasis formal leadership roles. Huffaker (2010) defines leaders as those who generate the most responses or whose language is most frequently adopted by other participants. Finally, Zhu, et al. (2012) focuses on leadership behaviors that any community member can perform. Our study offers a direct identification of leaders where peers nominate leaders and thus offers a stronger identification approach than those derived from leader activities or network position. We endorse the view that any participant in an online community may demonstrate a leadership behavior (c.f., Zhu, et al. 2012), but in considering the most influential participants, we focus on the most active participants. Given that the distribution of online participation follows a scale-free power law (Faraj, et al. 2008, Newman 2003), it becomes crucial to carefully select an appropriate sub-sample for a characteristic of interest as was done here by comparing participants at similar levels of participation.

Finally, this study's use of Natural Language Processing (NLP) provides a set of robust and rich measures to differentiate language use among online community leaders and other participants. With advances in computational linguistics, it is increasingly feasible to collect and analyze large samples of written text. Whereas future research is needed to determine how our specific findings (e.g., positive, concise posts with familiar language) generalize to other settings, our findings support the application of an NLP prism model utilizing multiple levels of linguistic analysis.

Nonetheless, the exact nature of these patterns, both in terms of communication network structures and linguistic characteristics, may well vary dependent upon the context. For example, a closed community handling a crisis situation may have very different patterns of influential communication than a group of hobbyists. Different communities have different values, purpose, and social context that each shape and reinforce the behaviors associated with leadership. Indeed, an open question is to what extent it is possible to theorize about online communities as a unitary phenomenon. Given recent suggestions that online communities differ greatly in terms of regulative behaviors (Kiesler, et al. 2012), role taking (Faraj, et al. 2011) and social stratification (Levina and Arrigara Forthcoming) it is probable that future research would benefit from approaches that emphasize a richer and more detailed data collection strategy, one that respects the embeddedness and situatedness of the social dynamics in online communities.

Practical Implications

Our results can inform participants and community managers regarding influence and leadership in online communities. First, individual participants can apply the findings of this study. Although occupying a formal leadership role is consistent with being viewed as influential, it is neither a necessary nor a sufficient condition. A participant seeking to become one of the most influential participants of a community should be a highly active participant in many messages threads concentrated on closely related topics and should communicate with positive familiar language.

Second, for those who manage or sponsor online communities the findings demonstrate the utility of seeking peer nominations to identify the most influential participants of a community. Whereas the ability to perform a robust linguistic analysis of participant posts is beyond the resources of a typical community manager, participant surveys are quite feasible. Our results demonstrate that, compared to the thousands of participants in the studied communities, even a relatively small number of survey responses can generate a valid list of peer-nominated influential participants. Finally, the findings also demonstrate that when a community manager seeks to reward, incentivize, or for any reason identify the most influential participants, they should not limit themselves to only those already in formal leadership roles.

Limitations and Directions for Future Research

The study design and findings point to multiple avenues for future research. First, more research is warranted to identify additional leadership behaviors and leadership styles associated with online community leadership. These include supporting and leading by example; transaction vs. transformative styles; and directive vs. empowering communication (O'Donnell, et al. 2012, Sims Jr, et al. 2009, Yukl 1999). Second, given the wide variety of online communities (Kietzmann, et al. 2011), it is quite possible that some traits associated with online community leadership are universal while others are idiosyncratic. Also, future research is needed to identify interactions between formal roles, network configuration or position, and the linguistic characteristics associated with leadership.

Second, while the majority of our propositions are indeed supported two unexpected findings merit further research. We had expected leaders to use a more sophisticated vocabulary than the average community member because they possibly needed a richer vocabulary to fully answer questions, provide deeper explanation, and to delve in discussions about complex knowledge topics. Instead, our measure of vocabulary richness was statistically significant but in the opposite direction than theorized. The results suggest that simpler language, particularly if it prototypical of community utterance, may be important for effective communication to a wider audience. Second, of the linguistic measures, we find that providing useful resources in the form of web links did not increase the likelihood of being identified as a leader. We had argued that web links represent a resource of value to other participants and thus would be perceived as a particularly helpful contribution. The negative finding indicates that web links may serve as poor proxies for actual resources, may be already known to the recipient, or may be perceived as a throwaway pointer reflecting a lack of engagement in the conversation. It is possible that providing resources is indeed valued but that web links serve are a poor proxy of such. In fact, as a measure solely of quantity without regard to quality, posted links could be off-topic, self-promotional, or otherwise not of general value.

Finally, this study supports the perspective that leadership in online communities emerges from both the structural and linguistic characteristics of participant communication but does not attempt to identify consequences of online community leadership. No doubt, not all online community leaders are equally effective. In addition, the compatibility of individual attributes such as personality, spatial and time separation (Espinosa, et al. 2012) and identification with community (Ren, et al. 2007) are all likely to impact online community leadership processes. More work is needed to gain a greater understanding of leadership effectiveness in online communities.

Conclusion

This study integrates four perspectives of online leadership--formal leadership roles, peer nominations, network position, and language use--to provide a richer understanding of leadership processes in online settings. Compared to traditional hierarchical organizations, online communities are heavily influenced by emergent leadership processes. To investigate how network structure and language use leads to influence we analyzed a combination of participant surveys, communication history, and user profiles. Our findings indicate that the participants viewed as leaders not only occupy central, core network positions, but generate distinctive written communication patterns of positive posts using language familiar to other participants. Thus, being an online community leader is associated with both where and how participants post; quantity, position, and quality all matter.

References

Alpaydin, Ethem, *Introduction to machine learning*, MIT Press, Cambridge, Mass., 2004.

Aron, J., "How innovative is Apple's new voice assistant, Siri?," *The New Scientist*, 212, 2836, (2011), 24.

Ashforth, Blake E and Fred Mael, "Social Identity Theory and the Organization," *Academy of Management Review*, 14, 1, (1989), 20-39.

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

Barge, J.K., *Leadership: Communication skills for organizations and groups*, St. Martin's Press, New York, 1994.

Barrett, Deborah J., *Leadership communication*, McGraw-Hill Irwin, Boston, 2008.

Barry, John, "Doing Bayesian Data Analysis: A Tutorial with R and BUGS," *Europe's Journal of Psychology*, 7, 4, (2011), 778-779.

Bass, Bernard M and Ruth Bass, *The Bass handbook of leadership: Theory, research, and managerial applications*, Free Press, New York, NY, 2008.

Benkler, Yochai, *The wealth of networks: how social production transforms markets and freedom*, Yale University Press, New Haven, 2006.

Bird, Steven, Ewan Klein and Edward Loper, *Natural language processing with Python*, O'Reilly Media, Inc., Sebastopol, CA, 2009.

Borgatti, Stephen P and Daniel S Halgin, "Analyzing affiliation networks," *The Sage handbook of social network analysis*, (2011), 417-433.

Borgatti, Stephen P. and Martin G. Everett, "Models of core/periphery structures," *Social Networks*, 21, 4, (2000), 375-395.

Bradley, Margaret M. and Peter J. Lang, *Affective norms for English words (ANEW): Instruction manual and affective ratings*, The Center for Research in Psychophysiology, University of Florida, 1999.

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

Burt, Ronald S, *Structural holes: The social structure of competition*, Harvard University Press, Cambridge, Mass., 1995.

Butler, Brian S., Lee Sproull, Sara Kiesler and Robert Kraut, "Community effort in online groups: Who does the work and why?," In *Leadership at a Distance: Interdisciplinary Perspectives*, S. Weisband (Ed.), Lawrence Erlbaum Associates, Mahway, N.J., 2007, 171-193.

Clark, Alexander, Chris Fox and Shalom Lappin, *The handbook of computational linguistics and natural language processing*, Wiley-Blackwell, Chichester, 2010.

Cobb, N.K., A.L. Graham and D.B. Abrams, "Social network structure of a large online community for smoking cessation," *American Journal of Public Health*, 100, 7, (2010), 1282-1289.

Collier, Benjamin and Robert Kraut, "Leading the Collective: Social Capital and the Development of Leaders in Core-Periphery Organizations," *Collective Intelligence*, (2012),

Cooren, François, Timothy Kuhn, Joep P Cornelissen and Timothy Clark, "Communication, organizing and organization: An overview and introduction to the special issue," *Organization Studies*, 32, 9, (2011), 1149-1170.

Crowston, Kevin and James Howison, "The social structure of free and open source software development," *First Monday*, 10, 2, (2005),

Dahlander, Linus and Lars Frederiksen, "The core and cosmopolitans: A relational view of innovation in user communities," *Organization Science*, 23, 4, (2012), 988-1007.

Donath, J., "Signals in social supernets," *Journal of Computer‐Mediated Communication*, 13, 1, (2007), 231-251.

Espinosa, J. Alberto, Jonathon N. Cummings and C. Pickering, "Time Separation, Coordination, and Performance in Technical Teams," *IEEE Transactions on Engineering Management*, 59, 1, (2012), 91-103.

Fairhurst, Gail, *Discursive leadership: In conversation with leadership psychology*, Sage Publications, Los Angeles, 2007.

Faraj, S., S.L. Jarvenpaa and A. Majchrzak, "Knowledge collaboration in online communities," *Organization Science*, 22, 5, (2011), 1224-1239.

Faraj, S. and S.L. Johnson, "Network exchange patterns in online communities," *Organization science*, 22, 6, (2011), 1464-1480

Faraj, Samer, Molly McLure Wasko and Steven L. Johnson, "The structure and processes of electronic knowledge networks," In *Advances in Management Information Systems, Knowledge Management: An Evolutionary View of the Field*, I. Becerra-Fernandez and D. Leidner (Ed.), M.E. Sharpe, Inc., Armonk, NY, 2008,

Ferrucci, David, "Build Watson: an overview of DeepQA for the Jeopardy! challenge," *Proceedings of the Proceedings of the 19th international conference on Parallel architectures and compilation techniques*, Vienna, Austria, 2010, 1-2.

Fleming, Lee and David M. Waguespack, "Brokerage, boundary spanning, and leadership in open innovation communities," *Organization Science*, 18, 2, (2007), 165-180.

Gerstner, Charlotte R and David V Day, "Meta-Analytic review of leader–member exchange theory: Correlates and construct issues," *Journal of applied psychology*, 82, 6, (1997), 827-844.

Goldsmith, John, "Unsupervised learning of the morphology of a natural language," *Comput. Linguist.*, 27, 2, (2001), 153-198.

Graen, George B and Mary Uhl-Bien, "Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain perspective," *The Leadership Quarterly*, 6, 2, (1995), 219-247.

Gunning, R., "The Fog Index after twenty years," *Journal of Business Communication*, 6, 2, (1969), 3-13.

Hackman, J Richard and Richard E Walton, "Leading groups in organizations," In *Designing effective work groups*, P. Goodman (Ed.), Jossey-Bass, San Francisco, 1986, 72-119.

Hart, Peter E, Richard O Duda and David G Stork, *Pattern classification*, John Wiley & Sons, Inc., New York, 2001.

Hoch, Julia E. and Steve W. J. Kozlowski, "Leading virtual teams: hierarchical leadership, structural supports, and shared team leadership," *Journal of applied psychology*, 99, 3, (2014), 390-403.

Hogg, Michael A., "A social identity theory of leadership," *Personality and Social Psychology Review*, 5, 3, (2001), 184-200.

Hollingshead, Andrea B., "Dynamics of leader emergence in online groups," In *Strategic uses of social technology: An interactive perspective of social psychology*, Z. Birchmeier, B. Dietz-Uhler and G. Stasser (Ed.), Cambridge University Press, Cambridge, 2011,

Hosmer, David W and Stanley Lemeshow, *Applied logistic regression*, John Wiley & Sons, Inc., Hoboken, NJ, 2005.

Huffaker, David, "Dimensions of Leadership and Social Influence in Online Communities," *Human Communication Research*, 36, 4, (2010), 593-617.

Johnson, Paul D and Marie T Dasborough, "Affective Events: Building Social Network Ties and Facilitating Informal Leader Emergence," In *Affect and Emotion: New Directions in Management: Theory and Research*, R. H. Humphrey (Ed.), Information Age Publishing, Inc., Charlotte, NC, 2008,

Jurafsky, D. and J.H. Martin, *Speech and language processing, 2nd edition*, Prentice Hall, 2008.

Jurafsky, Daniel and James H. Martin, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, Prentice Hall, Upper Saddle River, NJ, 2000.

Kankanhalli, Atreyi, Bernard C. Y. Tan and Wei Kwok-Kee, "Contributing Knowledge To Electronic Knowledge Repositories: An Empirical Investigation," *MIS Quarterly*, 29, 1, (2005), 113-143.

Kiesler, Sara, Robert Kraut, Paul Resnick and Aniket Kittur, "Regulating behavior in online communities," *Evidence-based social design: Mining the social sciences to build online communities*, (2012), 125-178.

Kietzmann, J.H., K. Hermkens, I.P. McCarthy and B.S. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media," *Business Horizons*, 54, 3, (2011),

Kincaid, J. Peter, Robert P. Fishburne, Jr., Richard L. Rogers and Brad S. Chissom, "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," Naval Technical Training Command Millington TN Research Branch, 1975.

Knoke, David and Song Yang, *Social network analysis*, Sage Publications, Inc, Thousand Oaks, California, 2008.

Kohavi, Ron, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Proceedings of the International joint Conference on artificial intelligence*, 1995, 1137-1145.

Kraut, Robert E and Paul Resnick, *Building Successful Online Communities: Evidence-Based Social Design*, The MIT Press, Cambridge, Massachusetts, 2011.

Lakhani, Karim and Eric von Hippel, "How Open Source software works: Free user-to-user assistance," *Research Policy*, 32, 6, (2003), 923-943.

Levina, N. and E. Vaast, "The emergence of boundary spanning competence in practice: implications for implementation and use of information systems," *MIS Quarterly*, 29, 2, (2005), 335-363.

Levina, Natalia and Manuel Arrigara, "Distinction and Status Production on User-Generated Content Platforms: Using Bourdieu's Theory of Cultural Production to Understand Social Dynamics in Online Fields," *Information Systems Research*, (Forthcoming),

Liu, C.H., "The effects of innovation alliance on network structure and density of cluster," *Expert Systems With Applications*, 38, 1, (2011), 299-305.

Long, J. Scott and Jeremy Freese, *Regression models for categorical dependent variables using STATA*, Stata Press, College Station, TX, 2006.

Mehra, Ajay, Martin Kilduff and Daniel J Brass, "The social networks of high and low self-monitors: Implications for workplace performance," *Administrative science quarterly*, 46, 1, (2001), 121-146.

Mitkov, R., *The Oxford handbook of computational linguistics*, Oxford University Press, Oxford, 2005.

Morgeson, Frederick P, D Scott DeRue and Elizabeth P Karam, "Leadership in teams: A functional approach to understanding leadership structures and processes," *Journal of Management*, 36, 1, (2010), 5-39.

Newman, M. E. J., "The structure and function of complex networks," *SIAM Review*, 45, 2, (2003), 167-256.

Nielsen, Finn Årup, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," arXiv:1103.2903v1 [cs.IR], (2011), 1-6.

O'Donnell, Mark, Gary Yukl and Thomas Taber, "Leader behavior and LMX: a constructive replication," *Journal of Managerial Psychology*, 27, 2, (2012), 143-154.

O'Mahony, S. and F. Ferraro, "The emergence of governance in an open source community," *The Academy of Management Journal*, 50, 5, (2007), 1079-1106.

Pang, B. and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, 2, 1-2, (2008), 1-135.

Pearce, Craig L and Henry P Sims, "Shared leadership: Toward a multi-level theory of leadership," In *Advances in interdisciplinary studies of work teams*, 7, Emerald Group Publishing Limited, 2000, 115-139.

Perry, Monica L, Craig L Pearce and Henry P Sims Jr, "Empowered selling teams: How shared leadership can contribute to selling team outcomes," *Journal of Personal Selling & Sales Management*, 19, 3, (1999), 35-51.

Pfeffer, Jeffrey and Robert B Cialdini, "Illusions of influence," In *Power and influence in organizations*, R. M. Kramer and M. A. Neale (Ed.), SAGE Publications, Inc., Thousand Oaks, California, 1998, 1-20.

Preece, Jenny, *Online Communities: Designing Usability, Supporting Sociability*, John Wiley & Sons, Chichester, 2000.

Reagans, Ray and Ezra W Zuckerman, "Networks, diversity, and productivity: The social capital of corporate R&D teams," *Organization science*, 12, 4, (2001), 502-517.

Reicher, Stephen, S. Alexander Haslam and Nick Hopkins, "Social identity and the dynamics of leadership: Leaders and followers as collaborative agents in the transformation of social reality," *The Leadership Quarterly*, 16, 4, (2005), 547-568.

Ren, Yuqing, F Maxwell Harper, Sara Drenner, Loren Terveen, Sara Kiesler, John Riedl and Robert E Kraut, "Building member attachment in online communities: Applying theories of group identity and interpersonal bonds," *MIS Quarterly*, 36, 3, (2012), 841-864.

Ren, Yuqing, Robert Kraut and Sara Kiesler, "Applying common identity and bond theory to design of online communities," *Organization Studies*, 28, 3, (2007), 377-408.

Robichaud, Daniel and François Cooren, *Organization and organizing: Materiality, agency and discourse*, Routledge, New York, 2013.

Royall, R. M., "The effect of sample size on the meaning of significance tests," *American Statistician*, 40, 4, (1986), 313-315.

Seidman, S.B., "Network structure and minimum degree," *Social Networks*, 5, 3, (1983), 269-287.

Shahaf, Dafna and Eyal Amir, "Towards a Theory of AI Completeness," *Proceedings of the AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2007, 150-155.

Sims Jr, Henry P, Samer Faraj and Seokhwa Yun, "When should a leader be directive or empowering? How to develop your own situational theory of leadership," *Business Horizons*, 52, 2, (2009), 149-158.

Sproull, L and M Arriaga, "Online communities," In *The Handbook of Computer Networks, Volume 3, Distributed Networks, Network Planning, Control, Management, and New Trends and Applications*, H. Bidgoli (Ed.), John Wiley & Sons, Inc., Hoboken, New Jersey, 2007, 898-914.

Sutanto, Juliana, Chuan-Hoo Tan, Boris Battistini and Chee Wei Phang, "Emergent Leadership in Virtual Collaboration Settings: A Social Network Analysis Approach," *Long Range Planning*, 44, 5-6, (2011), 421-439.

Sy, Thomas, Stéphane Côté and Richard Saavedra, "The contagious leader: impact of the leader's mood on the mood of group members, group affective tone, and group processes," *Journal of Applied Psychology*, 90, 2, (2005), 295.

Taylor, James R and Elizabeth J Van Every, *The emergent organization: Communication as its site and surface*, Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey, 1999.

Taylor, James R and Elizabeth J Van Every, *The situated organization: Case studies in the pragmatics of communication research*, Routledge, 2010.

von Krogh, Georg, Stefan Haefliger, Sebastian Spaeth and Martin W Wallin, "Carrots and rainbows: Motivation and social practice in open source software development," *MIS Quarterly*, 36, 2, (2012a), 649.

von Krogh, Georg, Ikujiro Nonaka and Lise Rechsteiner, "Leadership in organizational knowledge creation: A review and framework," *Journal of Management Studies*, 49, 1, (2012b), 240-277.

Walter, Frank, Michael S Cole, Gerben S van der Vegt, Robert S Rubin and William H Bommer, "Emotion recognition and emergent leadership: Unraveling mediating mechanisms and boundary conditions," *The Leadership Quarterly*, 23, 5, (2012), 977-991.

Wang, Danni, David A. Waldman and Zhen Zhang, "A meta-analysis of shared leadership and team effectiveness," *Journal of applied psychology*, 99, 2, (2014), 181-198.

Warmbrodt, John, Hong Sheng and Richard Hall, "Social Network Analysis of Video Bloggers' Community," *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, Waikoloa, HI, 2008, 1-9.

Wasko, M.M.L., R Teigland and S Faraj, "The provision of online public goods: Examining social structure in an electronic network of practice," *Decision Support Systems*, 47, 3, (2009), 254-265.

Wasko, Molly McLure and Samer Faraj, "Why Should I Share: Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice," *Management Information Systems Quarterly*, 29, 1, (2005), 35-47.

Weick, Karl E., *The social psychology of organizing*, Addison-Wesley, Reading, Mass., 1969.

Wickham, K.R. and J.B. Walther, "Perceived Behaviors of Emergent and Assigned Leaders in Virtual Groups," *International Journal of e-Collaboration (IJeC)*, 3, 1, (2007), 1-17.

Wolpert, David H and William G. Macready, "No free lunch theorems for optimization," *Evolutionary Computation, IEEE Transactions on*, 1, 1, (1997), 67-82.

Yoo, Youngjin and Maryam Alavi, "Emergent leadership in virtual teams: what do emergent leaders do?," *Information and Organization*, 14, 1, (2004), 27-58.

Yukl, Gary, "An evaluation of conceptual weaknesses in transformational and charismatic leadership theories," *The Leadership Quarterly*, 10, 2, (1999), 285-305.

Yukl, Gary, *Leadership in organizations*, Pearson Education, Inc., Upper Saddle River, NJ, 2010.

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

Zinsser, William, *On writing well: The classic guide to writing nonfiction*, Harper Perennial, 2006.

Appendix A: Summary Statistics and Validation

Table 7: Descriptive Statistics for Analysis Data Set

	Blender Artists		**Gearbox Software**		**Northern Sounds**		**All**	
Sample Size	n = 2,101		n = 331		n = 515		n = 2,947	
	Mean	**s.d.**	**Mean**	**s.d.**	**Mean**	**s.d.**	**Mean**	**s.d.**
Number of Posts	131.12	272.09	324.85	348.99	86.85	173.37	145.15	275.73
Centrality	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.01
Coreness	9.51	4.38	19.91	3.84	5.99	1.38	10.06	5.45
Boundary Spanning	0.53	0.19	0.52	0.16	0.69	0.15	0.55	0.19
Readability	6.26	2.36	5.47	2.32	6.28	1.87	6.18	2.29
Vocabulary Richness	17.86	7.60	10.06	4.96	26.24	11.30	18.45	9.22
Prototypicality	4.64	0.25	4.16	0.62	4.78	0.26	4.61	0.36
External Linking	0.22	0.20	0.13	0.12	0.22	0.26	0.21	0.21
Positive Sentiment	0.41	0.19	0.24	0.15	0.64	0.30	0.43	0.23

Table 8: Correlation Table for Analysis Data Set (n=2,947)

	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**
1. Online Community Leader										
2. Formal Role of Authority	0.33									
3. Group membership (1, 2, 3)	0.07	0.06								
4. Centrality	0.39	0.17	0.17							
5. Coreness	0.14	0.08	-0.08	0.19						
6. Boundary Spanning	0.00	0.06	0.30	0.10	0.15					
7. Readability	0.01	0.00	-0.02	-0.03	-0.15	-0.04				
8. Vocabulary Richness	-0.12	-0.06	0.25	-0.22	-0.57	0.08	0.34			
9. Prototypicality	-0.02	-0.01	0.03	-0.08	-0.38	0.04	0.20	0.42		
10. External Linking	0.03	0.03	-0.03	0.00	-0.12	-0.06	0.28	0.19	0.14	
11. Positive Sentiment	0.00	-0.01	0.28	-0.00	-0.20	0.21	-0.14	0.21	0.19	0.07

Correlations with absolute value 0.04 or greater are significant at $p < .05$

Table 9: Testing for Multicollinearity in Analysis Data Set Using Variance Inflation Factor

Measure	**VIF**	**1/VIF**
Formal Role of Authority	1.04	0.96
Centrality	1.10	0.91
Coreness	1.29	0.78
Boundary Spanning	1.90	0.53
Readability	1.67	0.60
Vocabulary Richness	1.29	0.78
External Linking	1.12	0.90
Prototypicality	1.15	0.87
Positive Sentiment	1.20	0.83
Mean VIF	**1.31**	

Table 10: 10-fold Cross Validation Using Analysis Data Set

Run	**RMSE**
1	0.10
2	0.15
3	0.14
4	0.16
5	0.11
6	0.14
7	0.15

Run	**RMSE**
8	0.11
9	0.10
10	0.11
Average RMSE	**0.13**

Table 11: Descriptive Statistics for Full Data Set

	Blender Artists		**Gearbox Software**		**Northern Sounds**		**All**	
	Mean	**s.d.**	**Mean**	**s.d.**	**Mean**	**s.d.**	**Mean**	**s.d.**
Sample Size	n = 10,264		n = 1,644		n = 2,488		n = 14,396	
Number of Posts	30.07	133.38	72.30	201.68	20.69	85.81	33.27	137.14
Centrality	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00
Coreness	3.48	3.87	6.91	7.72	2.63	2.08	3.73	4.43
Boundary Spanning	0.68	0.28	0.68	0.28	0.75	0.24	0.69	0.27
Readability	6.50	5.14	6.91	25.81	6.62	6.11	6.56	10.07
Vocabulary Richness	36.79	30.56	33.96	63.22	50.64	37.86	38.86	37.41
Prototypicality	4.66	0.64	4.22	2.81	4.81	0.49	4.63	1.12
External Linking	0.25	1.10	0.17	0.76	0.19	0.44	0.23	0.98
Positive Sentiment	0.43	0.45	0.29	0.45	0.57	0.45	0.44	0.46

Table 12: Correlation Table for Full Data Set (n = 14,396)

	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**	**11.**
1. Online Community Leader											
2. Formal Role of Authority	0.30										
3. Group membership (1, 2, 3)	0.03	0.03									
4. Number of Posts	0.38	0.14	0.00								
5. Centrality	0.40	0.17	0.08	0.75							
6. Coreness	0.17	0.10	-0.01	0.55	0.26						
7. Boundary Spanning	-0.03	0.00	0.10	-0.13	-0.02	-0.18					
8. Readability	0.00	0.00	0.01	-0.02	-0.01	-0.03	0.02				
9. Vocabulary Richness	-0.05	-0.03	0.12	-0.16	-0.08	-0.31	0.17	0.50			
10. Prototypicality	0.00	-0.01	0.01	-0.03	-0.01	-0.06	-0.02	0.04	0.16		
11. External Linking	0.00	0.00	-0.03	-0.01	0.00	-0.02	0.01	0.12	0.15	0.02	
12. Positive Sentiment	0.00	0.00	0.08	-0.03	0.00	-0.03	0.09	-0.04	0.05	-0.06	0.04

Correlations with absolute value 0.02 or greater are significant at $p < .05$

Table 13: Hierarchical Logistic Regression Model with Dependent Variable of Online Community Leader using Full Data Set

Measure	**Model A**	**Model B**	**Model C**
Group-level Coefficients			
Intragroup correlation	0.106	0.547	0.315
Participant-level Coefficients as Odds Ratios, Standard Errors in Parentheses			
Number of Posts	1.59*** (0.06)	1.12 (0.06)	1.11 (0.06)
Formal Role of Administrator or Moderator	72.32*** (35.33)	36.07*** (19.29)	36.85*** (20.29)
Centrality		1.16** (0.05)	1.14* (0.06)
Coreness		4.79*** (1.31)	3.07*** (0.85)
Boundary Spanning		0.41** (0.12)	0.41** (0.12)
Readability			1.35 (0.21)
Vocabulary Richness			0.05** (0.06)
External Linking			1.20 (0.14)
Prototypicality			6.04** (4.00)
Positive Sentiment			1.82* (0.53)
Goodness of Fit Indices			

Measure	**Model A**	**Model B**	**Model C**
Log likelihood	-243.1	-203.8	-193.7
Chi2	245.7	173.1	171.1
AIC	494.2	421.6	411.4
BIC	524.5	474.6	502.3
Comparison with previous model (Δ chi2)		78.57***	20.27**
n = 14,396; * p<0.05, ** p<0.01, *** p<0.001			

Table 14: Hierarchical Logistic Regression Model with Dependent Variable of Leaders using Analysis Data Set with Moderators and Administrators Removed

Measure	**Model A**	**Model B**
Group-level Coefficients		
Intragroup correlation	0.396	0.141
Participant-level Coefficients as Odds Ratios, Standard Errors in Parenthesis		
Centrality	1.60*** (0.12)	1.54*** (0.14)
Coreness	4.04*** (1.52)	2.66** (0.98)
Boundary Spanning	0.49** (0.11)	0.56** (0.12)
Readability		1.31 (0.19)
Vocabulary Richness		0.44* (0.15)
External Linking		1.78 (0.18)
Prototypicality		1.75* (0.41)
Positive Sentiment		1.43 (0.29)
Goodness of Fit Indices		
Log likelihood	-184.9	-176.5
Chi2	68.47	76.52
AIC	379.8	373.0
BIC	409.7	432.8
Comparison with previous model (Δ chi2)		16.80**
n = 2,925; Odds ratios; Standard errors in parentheses; * *p* < 0.05, ** *p* < 0.01, *** *p* < 0.001		

Table 15: Descriptive Statistics for all Online Community Leaders

	Blender Artists		**Gearbox Software**		**Northern Sounds**		**All**	
	n = 23		**n = 21**		**n = 15**		**n = 59**	
	Mean	**s.d.**	**Mean**	**s.d.**	**Mean**	**s.d.**	**Mean**	**s.d.**
Number of Posts	1088.17	1392.37	746.91	517.18	608.87	619.28	844.85	980.57
Centrality	0.02	0.04	0.02	0.02	0.03	0.04	0.02	0.03
Coreness	15.35	2.01	20.95	3.83	7.27	0.96	15.29	5.92
Boundary Spanning	0.54	0.19	0.47	0.15	0.72	0.13	0.56	0.19
Readability	6.12	1.67	6.25	2.23	6.49	1.87	6.26	1.91
Vocabulary Richness	9.34	4.30	9.98	6.41	13.93	6.37	10.73	5.88
Prototypicality	4.62	0.12	4.32	0.42	4.83	0.11	4.56	0.33
External Linking	0.23	0.19	0.17	0.10	0.38	0.34	0.24	0.23
Positive Sentiment	0.39	0.16	0.23	0.14	0.79	0.29	0.43	0.29

Table 16: Online Community Participants with Formal Roles of Authority

	Blender Artists		**Gearbox Software**		**Northern Sounds**		**All**	
	n = 4		**n = 5**		**n = 3**		**n = 12**	
	Mean	**s.d.**	**Mean**	**s.d.**	**Mean**	**s.d.**	**Mean**	**s.d.**
Number of Posts	967.25	682.11	668.80	322.02	450.33	619.73	713.67	526.92
Centrality	0.016	0.021	0.013	0.006	0.043	0.064	0.022	0.032
Coreness	16.00	0.00	22.20	1.79	7.00	1.00	16.33	6.39

	Blender Artists		Gearbox Software		Northern Sounds		All	
	n = 4		n = 5		n = 3		n = 12	
	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
Boundary Spanning	0.702	0.13	0.54	0.13	0.78	0.06	0.65	0.15
Readability	5.64	1.64	6.47	2.91	8.87	1.61	6.79	2.44
Vocabulary Richness	7.98	2.60	8.30	3.29	16.31	3.41	10.19	4.63
Prototypicality	4.62	0.17	4.24	0.48	4.86	0.028	4.52	0.40
External Linking	0.33	0.40	0.15	0.11	0.64	0.41	0.33	0.34
Positive Sentiment	0.31	0.24	0.22	0.20	0.74	0.34	0.38	0.31

Table 17: Comparison of Online Community Leaders with and without Formal Roles of Authority

Measure	**Online Community Leaders with No Formal Role n=47; Mean (s.d.)**	**Online Community Leaders also Moderator or Administrator n=12; Mean (s.d.)**	***t*-test for difference: diff., (*t* value)**
Number of Posts	878.3 (1067.9)	713.67 (526.92)	164.7 (319.2)
Centrality	0.02 (0.03)	0.02 (0.03)	-0.00 (0.01)
Coreness	15.0 (5.8)	16.33 (6.39)	-1.31 (1.92)
Boundary Spanning	0.54 (0.19)	0.65 (0.15)	-0.12 (0.06)
Readability	6.13 (1.75)	6.80 (2.45)	-0.67 (0.62)
Vocabulary Richness	10.9 (6.2)	10.20 (4.63)	0.67 (1.92)
External Linking	0.22 (0.18)	0.33 (0.31)	-0.12 (0.07)
Prototypicality	4.57 (0.32)	4.52 (0.40)	0.05 (0.11)
Positive Sentiment	0.45 (0.29)	0.38 (0.32)	0.06 (0.095)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; no significant differences found			

Table 18: Descriptive Statistics for Online Community Leaders Identified by One Participant

	Blender Artists		Gearbox Software		Northern Sounds		All	
	n = 21		n = 14		n = 11		n = 46	
	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
Number of Posts	885.91	931.92	727.36	607.08	352.27	343.87	710.043	751.36
Centrality	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.01
Coreness	15.29	2.10	20.50	4.49	7.00	1.00	14.89	5.75
Readability	6.24	1.63	5.92	1.70	6.33	2.05	6.16	1.72
Boundary Spanning	0.52	0.18	0.46	0.18	0.68	0.12	0.54	0.18
Vocabulary Richness	9.69	4.30	11.27	7.33	15.53	6.56	11.56	6.23
Prototypicality	4.63	0.11	4.43	0.37	4.83	0.09	4.61	0.26
External Linking	0.24	0.19	0.17	0.10	4.83	0.29	0.24	0.20
Positive Sentiment	0.38	0.15	0.23	0.14	0.72	0.29	0.41	0.26

Table 19: Descriptive Statistics for Online Community Leaders Identified by Two or More Participants

	Blender Artists		Gearbox Software		Northern Sounds		All	
	n = 2		n = 7		n = 4		n = 13	
	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
Number of Posts	3212.00	3924.44	786.00	300.66	1314.50	700.17	1321.85	1488.00
Centrality	0.09	0.12	0.02	0.02	0.09	0.04	0.05	0.06
Coreness	16.00	0.00	21.86	1.95	8.00	0.00	16.69	6.54
Boundary Spanning	0.76	0.10	0.48	0.11	0.81	0.08	0.63	0.19
Readability	4.78	2.02	6.90	3.09	6.93	1.37	6.58	2.49
Vocabulary Richness	5.68	2.75	7.40	2.98	9.51	3.24	7.78	3.09
Prototypicality	4.51	0.18	4.11	0.45	4.80	0.17	4.38	0.47
External Linking	0.13	0.07	0.18	0.11	0.47	0.47	0.26	0.29

Table 20: Comparison of Online Community Leaders Identified by One Participant to those Identified by Two or More Participants

Measure	**Identified by One Participant n = 46; Mean (s.d.)**	**Identified by Multiple Participants n = 13; Mean (s.d.)**	***t*-test for difference: diff., (*t* value)**
Number of Posts	710.0 (751.36)	1321.9 (1488.0)	-611.8* (-2.04)
Centrality	0.01 (0.01)	0.05 (0.06)	-0.04*** (-4.61)
Coreness	14.9 (5.75)	16.7 (6.54)	-1.8 (-0.97)
Boundary Spanning	0.54 (0.18)	0.63 (0.19)	-0.085 (-1.47)
Readability	6.17 (1.73)	6.58 (2.49)	-0.415 (-0.69)
Vocabulary Richness	11.57 (6.23)	7.78 (3.09)	3.78* (2.11)
External Linking	0.24 (0.21)	0.26 (0.29)	-0.02 (-0.29)
Prototypicality	4.62 (0.26)	4.38 (0.47)	0.235* (2.35)
Positive Sentiment	0.42 (0.26)	0.50 (0.39)	-0.08 (-0.87)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$			