Title	The influence of network structures of Wikipedia discussion pages on the efficiency of WikiProjects
Authors(s)	Qin, Xiangju, Cunningham, Pádraig, Salter-Townshend, Michael
Publication date	2015-10
Publication information	Qin, Xiangju, Pádraig Cunningham, and Michael Salter-Townshend. “The Influence of Network Structures of Wikipedia Discussion Pages on the Efficiency of WikiProjects” 43 (October, 2015).
Publisher	Elsevier
Item record/more information	http://hdl.handle.net/10197/9073
Publisher's statement	This is the author’s version of a work that was accepted for publication in Social Networks. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Social Networks (43 (2015)) https://doi.org/10.1016/j.socnet.2015.04.002
Publisher's version (DOI)	10.1016/j.socnet.2015.04.002

The Influence of Network Structures of Wikipedia Discussion Pages on the Efficiency of WikiProjects

Xiangju Qin[a], Pádraig Cunningham[a], Michael Salter-Townshend[b]

[a] *School of Computer Science & Informatics, University College Dublin, Dublin 4, Ireland*
[b] *Department of Statistics, Oxford University, 1 South Parks Road, Oxford OX1 3TG, UK*

Abstract

As a platform for discussion and communication, talk pages play an essential role in Wikipedia to facilitate coordination, sharing of information and knowledge resources among Wikipedians. In this work we explore the influence of network structures of these pages on the efficiency of WikiProjects. Project efficiency is measured as the amount of work done by project members in a quarter. The study uses the comments on WikiProject talk pages to construct communication networks. The structural properties of these networks are studied using ideas from social network theory. We develop three hypotheses about how network structures influence project effectiveness and examine the hypotheses using a longitudinal dataset of 362 WikiProjects. The evaluation suggests that an intermediate level of cohesion with a core of influential users dominating network flow improves effectiveness for a WikiProject, and that greater average membership tenure relates to project efficiency in a positive way. We discuss the implications of this analysis for the future management of WikiProjects.

Keywords:
Network Social Capital, Effectiveness, Wikipedia, Community governance, Longitudinal study, Leadership

1. Introduction

With the advent of Web 2.0, recent years have witnessed a growing popularity of a community-based peer production approach to software development and knowledge creation. Companies and non-profit organizations are increasingly relying on input from online communities to build knowledge and software artifacts. Well-known examples of peer production communities include Linux, Apache, Wikipedia, and OpenStreetMap. Different from traditional organizations which rely on markets or managerial hierarchy projects to organize production (Benkler, 2006), there exist no comparable hierarchy counterparts in online peer production systems (Ung and Dalle, 2010). In such systems, the primary purpose of project-like structure is to share resources (e.g., artifacts, wikis, mailing lists, norms) among participants. For instance, in Wikipedia, "WikiProjects" play an important role in sharing information and knowledge resources, coordinating collaborative activities for related topics. Wikipedia defines a WikiProject[1] as follows:

A WikiProject is a group of editors that collaborate on encyclopedic work at a collection of pages devoted to the management of a specific topic or family of topics within Wikipedia. A WikiProject is a group of people, not a set of pages, a subject area, or a category.

The success of online peer production systems have generated great interest among researchers to explore the mechanisms behind these systems, of which Wikipedia attracts the most attention. For instance, Adler and Alfaro (2007) proposed a trust quality metric to measure the reliability of Wikipedia content. Kittur and Kraut (2008) examined how Wikipedians improve the quality of articles through explicit and implicit coordination. Ung and Dalle (2010) explored the influence of the WikiProjects by examining project-based coordination activity, and found bursts of activity which appear to be related to individual leadership. Zhu et al. (2011) found strong evidence of shared leadership in Wikipedia, with editors in peripheral roles producing a large proportion of leadership behaviors. Nemoto et al. (2011) examined the influence of pre-existing social capital on the efficiency of collaboration among Wikipedia

[1] `http://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_Council/Guide&oldid=615488861`

editors, and found that higher social capital helps improve the efficiency of editors. While these studies shed light on important aspects that affect the success of peer production systems, a major limitation of these studies is that they fail to consider the broader social environment in which project participants coordinate and collaborate.

Despite the success as a means of knowledge sharing and collaboration, little is known about the mechanism behind community-based peer production from the perspective of social network analysis. Nevertheless, investigating social structure is a useful way to understand team practices since it allows researchers to explore questions with respect to coordination, control, socialization, continuity and learning - all topics of great interest for studies of collaborative groups (Crowston and Howison, 2005). In this study, we take a social network approach to investigate the impact of network structural properties on WikiProject effectiveness. We focus the study on the WikiProject unit as its primary purpose is to promote and facilitate coordination, collaboration, sharing of information and knowledge resources among participants interested in related topics to create and improve articles [2]. We measure the efficiency of a WikiProject as the amount of work done by project members in a quarter. We are interested in the following questions: Does the structure of communication networks related to a WikiProject affect its efficiency? If yes, what type of network structural properties will improve project efficiency?

To answer these questions, we investigate the relationship between network social capital and project efficiency in the context of Wikipedia. There is no universal and precise definition of network social capital (e.g., network closure (Coleman, 1988) versus structural holes (Burt, 1992) as social capital). Following Portes (1998), we define network social capital as the benefits network members secure from their membership in social networks or other social structures. We develop three hypotheses with respect to the influence of network structural properties on project efficiency and examine the hypotheses on a longitudinal dataset of 362 WikiProjects. The overall results suggest that an intermediate level of small-world structure with a core of influential users dominating network flow improves effectiveness for WikiProjects, and that greater average membership tenure relates to project efficiency in a positive way. This research provides insights into understanding the influence of network social capital on WikiProject efficiency and offers several practical implications for project management in Wikipedia.

The rest of this paper is organized as follows. The next section provides a discussion about the formation of communication networks and network resources, and a review of related work to develop hypotheses. In Section 3, we explain network measures and variables about project characteristics. Next, we discuss data collection and model specification. Section 5 presents the results, followed by an exploration of leadership behavior and language coordination in project communication. The last section presents discussion and conclusion.

2. Collaboration and Network Resources

2.1. Communication Network and Network Structures

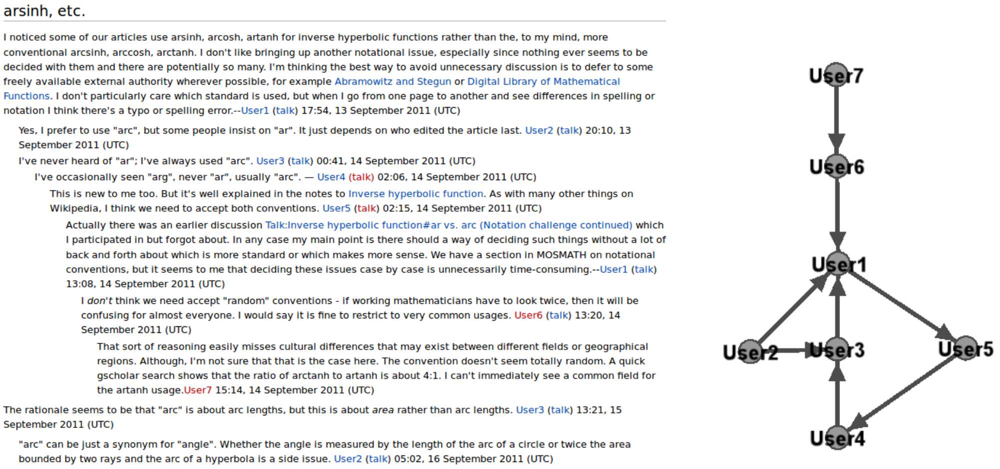
The primary purpose of WikiProject is to coordinate and organize the collaborative activities among project participants, as stated in Wikipedia[2]:

A WikiProject's pages are not used for writing encyclopedia articles directly, but as resources to help coordinate and organize the group's efforts at creating and improving articles. The discussion pages attached to a project page are a convenient forum for those involved in that project to talk about what they are doing, to ask questions, and to receive advice from other people interested in the group's work.

When faced with difficulties or in need of help, members generally turn to discussion pages for support from the community. New members can learn discipline, rules and regulations about Wikipedia, and how to make contributions by social learning and interaction with experienced pioneers. On many occasions, members have discussions on the discussion pages in order to reach consensus on controversial issues or make collective decisions regarding rules, regulations, and improvement of the system. The social relationships among participants facilitates coordination, the flow and sharing of information and knowledge resources across the whole community. In essence, social networking plays a foundational role in the functionality of WikiProjects in terms of facilitating information flow and the organization and coordination of the collaborative activities. The interactions among project participants in discussion pages form a communication network for the project.

[2] `http://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject&diff=614844785&oldid=614844781`

Specifically, the communication networks for WikiProjects can be constructed by parsing and coding the messages left on WikiProject talk pages recorded in the data dumps. For each message written by user A as a reply to user B's message, the two users were added as nodes into the network and a corresponding edge from A to B was created, with weight of the edge representing the number of messages A replied to B. By accumulating the communication on the talk pages of a WikiProject, we obtain its complete communication network from the inception of the WikiProject to the present time. We then get the communication network for a WikiProject in a specific quarter by extracting the nodes and edges from the overall network according to the timestamp associated with each edge. Figure 1 illustrates the strategy of how the networks are constructed using discussion topic "arsinh, etc." on the talk pages of WikiProject Mathematics[3].



arsinh, etc.

I noticed some of our articles use arsinh, arcosh, artanh for inverse hyperbolic functions rather than the, to my mind, more conventional arcsinh, arccosh, arctanh. I don't like bringing up another notational issue, especially since nothing ever seems to be decided with them and there are potentially so many. I'm thinking the best way to avoid unnecessary discussion is to defer to some freely available external authority wherever possible, for example Abramowitz and Stegun or Digital Library of Mathematical Functions. I don't particularly care which standard is used, but when I go from one page to another and see differences in spelling or notation I think there's a typo or spelling error.--User1 (talk) 17:54, 13 September 2011 (UTC)

Yes, I prefer to use "arc", but some people insist on "ar". It just depends on who edited the article last. User2 (talk) 20:10, 13 September 2011 (UTC)

I've never heard of "ar"; I've always used "arc". User3 (talk) 00:41, 14 September 2011 (UTC)

I've occasionally seen "arg", never "ar", usually "arc". — User4 (talk) 02:06, 14 September 2011 (UTC)

This is new to me too. But it's well explained in the notes to Inverse hyperbolic function. As with many other things on Wikipedia, I think we need to accept both conventions. User5 (talk) 02:15, 14 September 2011 (UTC)

Actually there was an earlier discussion Talk:Inverse hyperbolic function#ar vs. arc (Notation challenge continued) which I participated in but forgot about. In any case my main point is there should a way of deciding such things without a lot of back and forth about which is more standard or which makes more sense. We have a section in MOSMATH on notational conventions, but it seems to me that deciding these issues case by case is unnecessarily time-consuming.--User1 (talk) 13:08, 14 September 2011 (UTC)

I *don't* think we need accept "random" conventions - if working mathematicians have to look twice, then it will be confusing for almost everyone. I would say it is fine to restrict to very common usages. User6 (talk) 13:20, 14 September 2011 (UTC)

That sort of reasoning easily misses cultural differences that may exist between different fields or geographical regions. Although, I'm not sure that that is the case here. The convention doesn't seem totally random. A quick gscholar search shows that the ratio of arctanh to artanh is about 4:1. I can't immediately see a common field for the artanh usage.User7 15:14, 14 September 2011 (UTC)

The rationale seems to be that "arc" is about arc lengths, but this is about *area* rather than arc lengths. User3 (talk) 13:21, 15 September 2011 (UTC)

"arc" can be just a synonym for "angle". Whether the angle is measured by the length of the arc of a circle or twice the area bounded by two rays and the arc of a hyperbola is a side issue. User2 (talk) 05:02, 16 September 2011 (UTC)

(a) Discussion Thread

(b) Network

Figure 1: Schema of network construction

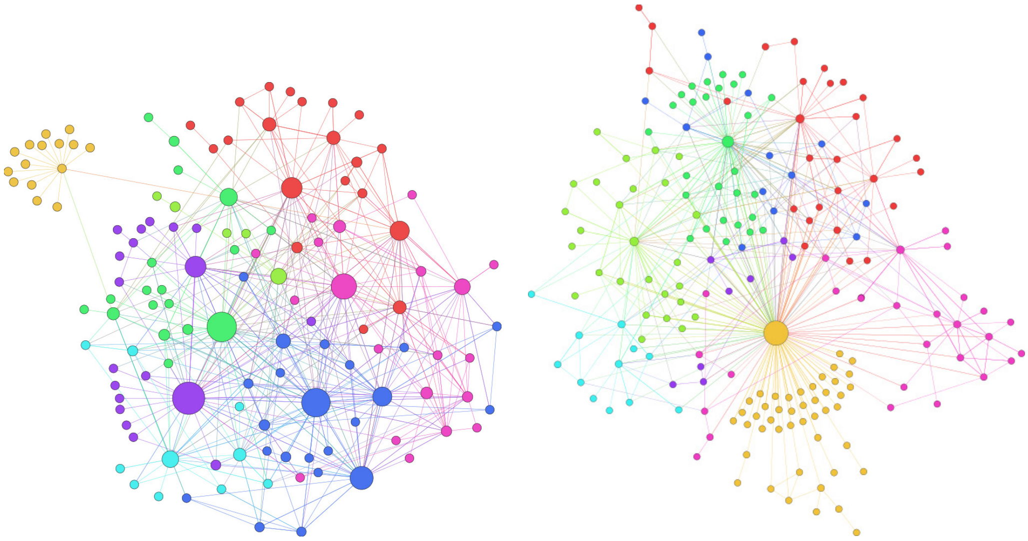
Figure 2 is a snapshot of communication networks for WikiProject Mathematics and Military History in the 18th Quarter (01/10/2006 – 31/12/2006). All the networks reported in this work are produced using Gephi software (Bastian et al., 2009). It is obvious from Figure 2 that a group of core members implicitly coordinate user activities and dominate network flow in both networks.

Because of their foundational role in facilitating the flow of information and knowledge resources among network members, network relationships have been described as network resources (Gulati, 1999; Ahuja, 2000). Similar to the technological and other collaboration networks, communication networks for WikiProjects are also associated with network benefits such as resource sharing and information diffusion. Resource sharing benefits enable network members to combine professional knowledge and accumulated skills, whereas information diffusion benefits can provide access to knowledge spillovers (Ahuja, 2000). Both benefits can potentially help improve members' work.

2.2. *Performance and Network Structures*

Theoretical and empirical work exploring the affects of organizational tenure on job performance has suggested that organizational tenure generally promotes performance (e.g., Reagans and Zuckerman (2001); Ng and Feldman (2010)). Reagans and Zuckerman (2001) performed a quantitative analysis on social networks, organizational tenure, and productivity of 224 corporate R&D teams, and found that average organizational tenure, network density, and network heterogeneity help improve team productivity. Their study suggests that R&D teams with heterogeneous networks and more senior members could enrich the research process and encourage coordination, thus promote greater productivity. Ng and Feldman (2010) investigated 350 empirical studies, and found that long-tenured workers generally have better in-role performance and citizenship performance, and that the tenure-performance relationship

[3] `http://en.wikipedia.org/wiki/Wikipedia_talk:WikiProject_Mathematics/Archive/2011/Sep#arsinh.2C_etc.`



(a) WP Mathematics, clustering coefficient: 0.499

(b) WP Military History, clustering coefficient: 0.544

Figure 2: Communication Networks for two WikiProjects in the 18th Quarter. The size of nodes indicates the influence of users measured in flow betweenness centrality, the color of nodes indicates community id.

becomes weak as organizational tenure increases. These studies suggest that organizational tenure can be used as a proxy for work experience or level of job-related knowledge, and that teams with more long-tenured members are more likely to enjoy an enhanced capacity for creative problem solving, better task distribution, and better performance. Based on these studies, we make the following hypothesis:

Ha *The higher the average membership tenure is, the higher the project efficiency is.*

Sociologic researchers generally agree that internal networks matter for group behavior and outcomes, and have studied the relationship between social network structures of teams and their performance in traditional organizations and online volunteer groups from different perspectives. For instance, Cummings and Cross (2003) studied the structural properties of 182 work groups in a global organization, and found that groups constrained by structural properties such as hierarchy and core-periphery perform worse in non-routine, complex tasks than groups with a more integrative structure. Their results suggested that, while certain group structures might be effective for diffusion, they might not be effective in an organizational setting where the tasks and required information are emergent. Crowston and Howison (2005) examined 120 project teams from SourceForge using the interactions among users in the bug tracking system, and studied social networks and communication structures of these projects. They found that teams vary widely in their communication centralization, from projects completely centered on one developer to projects that are highly decentralized and exhibit a distributed pattern of conversation between developers and active users. Kidane and Gloor (2007) studied the temporal communication patterns of developers and users in the Eclipse Java development project. Their preliminary results indicate that there is a correlation between communication structure (e.g., betweenness centrality) and productivity of open source developer teams. Zhu et al. (2012) investigated how distinct types of leadership behaviors and the legitimacy of the people who deliver them (people in formal leadership positions or not) influence the contributions that other participants make. Their results suggested that leadership behaviors exhibited by members at all levels significantly influence other members' motivation (in terms of making contributions to Wikipedia). These works motivate the following hypothesis regarding leadership behavior in group communication:

Hb *The existence of leadership behavior in project communication is positively related to project efficiency.*

Watts and Strogatz (1998) formalized small-world structure as cohesive clusters connected by a few bridging ties. In such networks, the dense clusters cultivate trust and close collaboration, whereas bridging ties bring fresh and nonredundant information to the clusters (Fleming et al., 2007). Network scholars generally agree that small-world networks facilitate information diffusion and the spread of epidemics. The benefits of small-world networks have attracted researchers to investigate how small-world structure influences economic and sociological outcomes of organizations and teams in a variety of contexts, including Broadway musicals (Uzzi and Spiro, 2005), U.S. regional patent coauthorship networks (Fleming et al., 2007), open source software projects (Singh, 2010). The most dominant argument states that small-world networks should enhance innovation and productivity (e.g., Uzzi and Spiro (2005); Singh (2010)). Uzzi and Spiro (2005) analyzed the small world network of the creative musicians who made Broadway musicals from 1945 to 1989, and found that the level of small worldliness of these artists affects their creativity in terms of the financial and artistic performance of the musicals they produced. Fleming et al. (2007) investigated the influence of regional small-world structures on innovation, and found that both shorter path lengths and larger connected components correlate with increased innovation. Singh (2010) studied the impact of the structure of community-level networks on the productivity of its member developers. By examining a longitudinal dataset of 4,279 projects from 15 different communities hosted at SourceForge, they found that the small-world properties of a community are positively and significantly related to the technical and commercial success of the software produced by its members. Motivated by these studies, we formulate the following hypothesis:

Hc *The small-world properties of communication network for a project are positively related to project efficiency.*

To summarize, previous studies suggest that network structural properties matter for group behavior and outcomes in traditional organizations and online volunteer groups. However, little is known about how network structures shape group collaboration and outcomes in online peer production communities. In this work, we study the mechanism behind efficient WikiProjects from the perspective of social interaction among contributors and its influence on project efficiency.

3. Measures

In this section, we introduce the dependent, independent and control variables we used for this study.

3.1. Dependent Variables: WikiProject Efficiency

Nemoto et al. (2011) measured the efficiency of collaboration among Wikipedia editors using the time taken between when an article got a previous promotion to when the article got promoted to a subsequent higher quality status. However, their efficiency measure is not always reliable in reflecting the actual efficiency of collaboration. In online collaborations like Wikipedia, due to the voluntary nature, it is very likely that there exists a big time gap between any two consecutive edits / contributions, which does not necessarily indicate inefficient collaboration.

In this study, the efficiency of a WikiProject is measured as the amount of work done by members to articles within project scope in a quarter in two aspects: (1) the total number of edits (i.e. edit count); (2) the edit longevity. The edit count is a very raw proxy of contributions by members to Wikipedia, as it generally neglects the quality of contributions. The edit longevity evaluates each contribution by combining its quality and quantity, and can be computed by using WikiTrust software (Adler and Alfaro, 2007).

It is well known in social science literature that network members can secure benefits from their membership in a network only after the network has been established. This is what Gulati and Gargiulo (1999) called a lag between network measures (i.e. network social capital) and the success / performance measure. Following this convention, we measure the dependent variable Efficiency_{jt} for project j in quarter t one quarter after the communication network was constructed in quarter $t-1$. For instance, if a communication network was constructed in 01/01/2005–31/03/2005, we measured the corresponding project efficiency in 01/04/2005–30/06/2005.

3.2. Independent Variables: Network Measures

Before proceeding, we need to introduce some terminology and notation. A graph is a mathematical representation of a network. More formally, an undirected graph G consists of a set of nodes (or vertices), $V = \{v_i | i \in [1, n]\}$, and a set of pairs of vertices called edges, $E = \{e_{ij} | i, j \in [1, n]; i \neq j\}$, termed as $G = (V; E)$. Where e_{ij} denotes a connection between vertex i and j. The neighborhood of a vertex i consists of its directly connected neighbors, denoted as $N_i = \{v_j | e_{ij} \in E\}$. We characterize the network structures of online teams using network measures, such as network density, network centralization, clustering coefficient and average path length. Let $k_i = |N_i|$ be the degree of vertex i, $\langle k \rangle$ being the average number of links per vertex in the network.

Network Density. This variable measures the ratio of existing connections in the network to the number of possible pairwise combinations of members, and takes values from zero to one, with larger values indicating increasing density. We include this measure as Yamaguchi (1994) suggested that the rate and extent to which information diffuses increases with the density of the network.

Network Centralization (Gini RW Betweenness). The centralization of a network can have an influence on its diffusion properties (Schilling and Phelps, 2007). The hub, star, or wheel networks are typical examples of highly centralized networks, in which members are tied to one or a few central nodes and communicate with each other through central node(s) (Freeman, 1979). We employ the Gini coefficient to measure how much the distribution of the centrality of the nodes is skewed. The Gini coefficient measure of centralization is very similar to Freeman's general formula for group degree and betweenness centrality (Freeman, 1979).

In this work, we employed random walk betweenness centrality (Newman, 2005) to measure the influence of members in network flow. Different from betweenness centrality and flow betweenness centrality which assume some kind of optimality in information transmission (shortest paths or maximum flow), random walk betweenness centrality quantifies how often a given vertex will fall on a random walk between another pair of vertices (Newman, 2005). The Gini coefficient of random walk betweenness centrality for communication network of project j in quarter t is computed as:

$$\mathrm{Gini}_{j,t} = \frac{2}{n^2 \times \mu} \sum_{x=1}^{n} \sum_{y=1}^{n} \left| C_B(n_x) - C_B(n_y) \right| \tag{1}$$

where μ is the average of the centrality for members, $C_B(n_x)$ and $C_B(n_y)$ are the random walk betweenness centrality of members x and y in project j for quarter t, n is the number of nodes in the network. The coefficient is 0 when the distribution is even (i.e., when all nodes in the network have the same number of links, indicating communication is evenly distributed among members), and is 1 in the case of a centralized network where one central member connects all other members. Large values of the measure implies several central members control and dominate the flow of the network, suggesting the presence of leadership behavior in the communication.

Clustering Coefficient ratio (CC ratio) and Average Path Length ratio (PL ratio). Clustering coefficient quantifies the extent to which two connected members in a network share a common third-party tie. Following Watts and Strogatz (1998), the clustering coefficient for node i, CC_i, is calculated as:

$$CC_i = \frac{2|e_{jk}|}{k_i(k_i - 1)} \tag{2}$$

where $v_j, v_k \in N_i, e_{jk} \in E$, k_i is the degree of vertex i. The clustering coefficient of the whole network is defined as the average of the clustering coefficients of all its vertices (Watts and Strogatz, 1998):

$$CC = \frac{1}{n} \sum CC_i \tag{3}$$

A higher clustering coefficient increases the information transmission of a network in terms of enabling information and knowledge to be exchanged and integrated. Although a densely connected network enables fast and reliable transmission of knowledge, the speed and integrity of knowledge diffusion across the network relies on its average path length (Schilling and Phelps, 2007). Average path length (PL) is a measure of the average of the shortest distance between any two members in the network. Shorter average path length would maximize the speed and minimize the decay in information transmission (Watts and Strogatz, 1998).

To quantify the degree of small worldliness for a network, Watts and Strogatz (1998) proposed a measure by combining the average path length and clustering coefficient of the actual network and that of the theoretical random network with the same number of nodes and links. Following Watts and Strogatz (1998), the clustering coefficient and average path length of a random network can be approximated as $CC\ random = \langle k \rangle / n$ and $PL\ random = ln(n)/ln(\langle k \rangle)$, respectively. For a network with small-world properties, its clustering coefficient ratio will exceed 1 (i.e. CC ratio = $CC/CC\ random \gg 1$) and average path length ratio will approximate 1 (i.e., PL ratio = $PL/PL\ random \approx 1$), which implies the network with a much higher clustering coefficient than the random network but a roughly equal short average path length. Alternatively, the larger the small world quotient (i.e., SWQ=CC ratio/PL ratio > 1), the more the network resembles a small world network (Watts and Strogatz, 1998; Uzzi and Spiro, 2005).

The communication networks used in this work are weighted and directed graphs. Network density was calculated directly for the networks, random walk betweenness centrality was calculated in the symmetrized networks. Moreover, following other studies (Uzzi and Spiro, 2005; Fleming et al., 2007), we calculated small-world properties in the giant component of the symmetrized networks.

3.3. Control Variables: Project Characteristics Measures

Total Number of Discussion Topics per quarter (Discussion topics). In online communities, members generally rely on discussion pages to communicate with others, ask for help and support, reach consensus on controversial issues, and make collective decisions regarding rules and regulations. A large amount of content posted on project-related talk pages is beneficial for the community in terms of enabling participants to arrive at a general understanding or get help for specific questions. We calculated the amount of group communication as the total number of discussion topics recorded in the edit history of project-related talk pages in the focal quarter.

Average Membership Tenure (Mean Tenure). In online communities like Wikipedia, it is often the case that members who have been active for a long time tend to be more experienced than new users. These active members play a fundamental role in the community in terms of spreading knowledge, information and experience across the whole community.

Since experience in Wikipedia transfers easily to projects, we measured membership tenure as the amount of time a member has been active in Wikipedia. Specifically, our measure of membership tenure consists of two parts: membership tenure before and after joining a WikiProject. The former is calculated as the number of days between the timestamp that the user made the 1st edit in Wikipedia and that this user joined a WikiProject. For any month in a quarter, if a user made at least one edit to any project related pages, we calculated the monthly project tenure of this user as the number of days in that month. We then accumulated the previous tenure and monthly project tenure up to the focal quarter to obtain each member's overall membership tenure in Wikipedia. The mean tenure of a WikiProject is calculated as the sum of the membership tenure of its members divided by the number of members in a quarter.

Turnover rate. Measured as the percentage of members who were active in previous quarter and not active in current quarter. We consider a member being active in a quarter if and only if the member made at least one edit to articles within project scope.

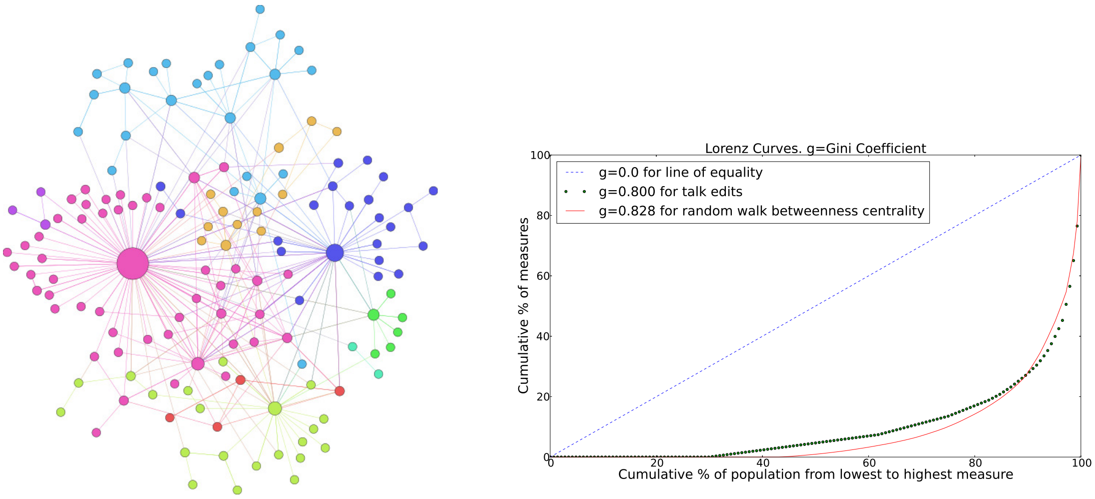
Gini Index of User Edits to Project Talk Pages (Gini Talk Edits). Efficient projects tend to have explicit or implicit leaders[4] to coordinate or organize group discussions and collaborative activities, to help reach consensus and build harmonious working environments. In WikiProjects, to some extent, leadership behavior can be captured by calculating the Gini coefficient for the distribution of the number of edits to project talk pages by project members.

3.4. Illustration of Gini Coefficient for Network Measures

In this study, we use the Gini coefficient to capture how much the distribution of some measures (i.e. betweenness centrality, talk edit statistics) among participants is skewed. In this section, to give an idea of what the Gini scores of the measures capture, we provide an illustration of these measures using the communication network of WikiProject Military History. This WikiProject has been very productive in editing articles and very active in its project talk pages over the time period considered in this study. Figure 3(a) is a snapshot of the communication network for the project in the 20th Quarter (01/04/2007 – 30/06/2007), Figure 3(b) graphically provides the distribution of the corresponding random walk betweenness centrality measures and talk edits statistics using Lorenz curve. In economics, the Lorenz curve (Lorenz, 1905) is often used to represent the inequality of wealth distribution, which shows the percentage y% of the assets assumed by the percentage x% of the population .

In Figure 3(a), the size of nodes indicates the influence of users measured in betweenness centrality and the color of nodes indicates community id. It is obvious from Figure 3(a) that a small group of core members implicitly coordinate activities and dominate network flow. By supporting the observation in Figure 3(a), we observe from Figure 3(b) that the distributions of centrality and talk edits are very skewed, and that about 20% of members account

[4] In this study, explicit leaders refer to project members who were nominated as project leaders or coordinators by the community, implicit leaders refer to those who earned their fame by being active and assumed leadership role without being nominated by the community.



(a) Communication Network

(b) Lorenz Curves

Figure 3: Graphical representation of mutual interactions and Gini coefficient for WP Military History in the 20th Quarter. Network statistics: #nodes:144, #links:357.

for more than 80% of communication and network flow, which suggests explicit or implicit leadership behavior in the communication. Table 1 provides detailed random walk betweenness centrality and talk edit statistics.

Table 1: Centrality and talk edit statistics for WP Military History in the 20th Quarter. Measures were sorted by centrality in descending order.

uid	Centrality	#Talk Edits	uid	Centrality	#Talk Edits	uid	Centrality	#Talk Edits	uid	Centrality	#Talk Edits
1	0.699	146	13	0.063	7	25	0.024	3	37	0.016	1
2	0.337	71	14	0.06	15	26	0.024	2	38	0.014	2
3	0.209	37	15	0.058	6	27	0.023	2	39	0.014	0
4	0.200	53	16	0.043	6	28	0.021	2	40	0.013	6
5	0.109	16	17	0.039	5	29	0.021	0	41	0.013	5
6	0.102	9	18	0.035	4	30	0.02	4	42	0.013	0
7	0.100	33	19	0.035	2	31	0.02	2	43	0.012	3
8	0.085	17	20	0.031	6	32	0.019	4	44	0.012	1
9	0.085	10	21	0.031	3	33	0.019	0	45	0.011	2
10	0.080	7	22	0.029	3	34	0.018	7	46	0.01	3
11	0.078	11	23	0.026	2	35	0.017	5	47–82	0.001–0.009	0–3
12	0.066	14	24	0.026	2	36	0.016	2	83–144	0.0	0–3
Gini coefficient for random walk betweenness centrality: 0.828, talk edits statistics: 0.800											

Consistent with the observations in Figure 3(b), it is obvious from Table 1 that the top 4 users dominate both measures, and that the measures of the top 2 users are much larger than those of others. The observations in this section suggest that it is reasonable to employ the Gini coefficient to capture the domination of communication by a few members.

4. Methods

4.1. Data Collection

As the English version of Wikipedia has the largest number of articles (about 4 million) and it is convenient to process the English edition compared to other language editions of Wikipedia, we rely on it to evaluate the hypotheses. Wikipedia regularly provides a complete copy of its data dump to the public, which attracts much attention from academia to explore its data. We downloaded enwiki data dumps from the Wikipedia website [5], wrote Python code to preprocess the data in order to construct communication networks for WikiProjects (by parsing the revision history of project talk pages). We parsed the historical edits of a project's member list to identify members and their joining time for each WikiProject. WikiProjects generally provide main pages or subpages to maintain the list of project members, any editors can join a project by adding one's username to the member list and then leave the project by removing the username from the list.

Project members generally claim an article in its scope by inserting project template into article talk page. We parsed the link to WikiProject in article talk pages and accumulated all articles tagged by a WikiProject up to a specific quarter to estimate its project scope. The edit longevity of each edit to articles was calculated using the WikiTrust software by Adler and Alfaro (2007). We then obtained the amount of work done by members to project scope articles in a quarter by accumulating all the edits or the edit longevity of edits contributed by members. To make the results more meaningful, we only included those WikiProjects that had at least 50 accumulated tagged articles, at least three members (the minimal size of group) and a weakly connected communication network. We considered a WikiProject active in a quarter if the size of its communication network in previous quarter met the basic requirement for small worldliness ($N >> \langle k \rangle$) (Watts and Strogatz, 1998). As a result, we obtained a longitudinal dataset of 362 WikiProjects with 4107 quarterly observations, each observation recording the characteristics, network measures and outcomes of a project in a quarter.

We used Python programming language (van Rossum and Drake, 2009) for data preprocessing, all the network measures were calculated using Networkx (Hagberg et al., 2008) packages for Python environment, while the regression analysis was performed using functions in **R** software (version 2.15.3) (R Core Team, 2013).

4.2. Model Specification

A preliminary analysis of the dataset revealed that some of the independent variables were positively skewed. To make the coefficients more comparable, as suggested by Gelman and Hill (2007), we performed a logarithmic transformation or z-score normalization on these independent variables. To explore the possible non-linear relationship between small-world properties and project efficiency, we included the squares of the small-world properties related variables in the model. Investigations also showed that the squares of these variables were highly correlated with their linear forms. We fitted a second order polynomial (using the poly function in R) and discussed the curvilinear relationship we found. We chose a second order polynomial (quadratic) over a smoothing spline to facilitate comparison with existing work (Uzzi and Spiro, 2005). To ease the interpretation, we represent the log-transformed and z-score normalization variables with "Log" and "Sc" prefix, respectively.

Although we include several project-related control variables to account for inherent differences among WikiProjects, there may still exist unobserved heterogeneity among projects. For example, some WikiProjects may regularly provide and update a list of open tasks that enumerate articles that need different types of improvement. These open tasks facilitate task-matching for interested participants and help focus their efforts to make contributions. Other unobserved characteristics such as motivation, the quality of involved editors and the atmosphere of WikiProject may potentially influence the effectiveness of WikiProject. To account for project-level unobserved differences, we include project-quarter random effect (i.e. $z_{j,t-1}$) in the model.

The nested structure of our data – quarterly observations nested within projects – suggests Hierarchical Linear Model (Bryk and Raudenbush, 1992) for the analysis. HLM is an advanced version of linear model which takes into account potential autocorrelation among time periods that are nested with the same project, while allowing us to examine the main effects of independent variables on dependent variables. The model specification is as follows:

$$\begin{aligned}\text{Efficiency}_{j,t} = \beta_0 + \beta_1 &\text{Discussion topics} + \beta_2\text{Mean Tenure} + \\ &\beta_3\text{Turnover rate} + \beta_4\text{Gini Talk Edits} + \beta_5\text{Density}_{j,t-1} + \\ &\beta_6\text{Gini RW Betweenness}_{j,t-1} + \beta_7\text{PL ratio}_{j,t-1} + \\ &\beta_8\text{CC ratio}_{j,t-1} + \beta_9\text{CC ratio}^2_{j,t-1} + \beta_{10}\text{SWQ}_{j,t-1} + \beta_{11}\text{SWQ}^2_{j,t-1} + z_{j,t-1}\end{aligned} \tag{4}$$

[5] http://dumps.wikimedia.org/enwiki/20130805/

Because the dependent variable measured as edit count is a count variable and takes only non-negative integer values, a Poisson regression approach is more appropriate (Hausman and Griliches, 1984). The presence of overdispersion (i.e. the variance exceeds the mean) on the edit count data suggested the need for the Negative Binomial specification (Hausman and Griliches, 1984). The Negative Binomial is a Poisson distribution whose rate parameter has been mixed with a Gamma distribution of the same mean. This creates a counts likelihood that has the same mean but a larger variance than the Poisson. The conditional mean of the Negative Binomial edit count function for project j in quarter t is described in Eq. (5):

$$\begin{aligned}\gamma_{j,t} = {} & E(\text{Edit Count}_{j,t}|\text{Discussion topics}, \text{Mean Tenure}, \text{Turnover rate}, \\ & \text{Gini Talk Edits}, \text{Density}_{j,t-1}, \text{Gini RW Betweenness}_{j,t-1}, \text{PL ratio}_{j,t-1}, \\ & \text{CC ratio}_{j,t-1}, \text{CC ratio}^2_{j,t-1}, \text{SWQ}_{j,t-1}, \text{SWQ}^2_{j,t-1}, z_{j,t-1}) \\ = {} & \exp(\beta_0 + \beta_1 \text{Discussion topics} + \beta_2 \text{Mean Tenure} + \\ & \beta_3 \text{Turnover rate} + \beta_4 \text{Gini Talk Edits} + \beta_5 \text{Density}_{j,t-1} + \\ & \beta_6 \text{Gini RW Betweenness}_{j,t-1} + \beta_7 \text{PL ratio}_{j,t-1} + \\ & \beta_8 \text{CC ratio}_{j,t-1} + \beta_9 \text{CC ratio}^2_{j,t-1} + \beta_{10} \text{SWQ}_{j,t-1} + \beta_{11} \text{SWQ}^2_{j,t-1} + z_{j,t-1})\end{aligned} \tag{5}$$

where variables are indexed across projects (j) and quarters (t); $z_{j,t-1}$ is the project-quarter random effect. We obtained our estimates using the lmer and glmer (for models with Edit Longevity and Edit Count as the dependent variable respectively) function in lme4 package (Bates et al., 2012) for R software. We check for violations of the assumption of the regression analysis and find no substantive violations.

Table 2: Descriptive Statistics (n=4107)

Variables	Mean	Std Dev	Min	Median	Max
Edit Longevity	37783.3	73371	1.5	15665.2	1014280.1
Edit Count	2183.3	4336.2	1	892	47360
Discussion topics	38.45	5.8	0	20	704
Mean Tenure	713.8	325.3	118	667.7	2305.7
Turnover rate	0.33	0.24	0	0.29	4
Gini Talk Edits	0.042	0.18	0	0.43	0.8
Density	0.14	0.11	0.008	0.11	1
Gini RW Betweenness	0.62	0.12	0	0.64	0.89
PL ratio	0.75	0.08	0.51	0.74	1.5
CC ratio	4.27	3.3	0	3.34	36.55
SWQ	5.68	4.3	0	4.52	66.2

5. Results

The descriptive statistics and the correlation matrix of the main variables are presented in Table 2 and Table 3, respectively. It is obvious from Table 2 that, the amount of work (i.e., Edit Longevity and Edit Count), Discussion topics, Mean Tenure are of reasonable variation and have a heavily right skewed distribution, we performed a log-transformation or z-score normalization on these variables. The dependent variable – Edit Count – is over dispersion (mean = 2183.3, Std Dev = 4336.20), suggesting that the Negative Binomial model is preferred. The mean of small world quotient is 5.68, which indicates the existence of small-world properties. The correlation between the two dependent variables (i.e., Log(Edit Longevity) and Log(Edit Count)) is positive and significant, suggesting that it is reasonable to use edit count as a measure of the amount of work done by project members. The correlation values are relatively low, except for the linear and squared terms of the CC ratio and SWQ variables.

Table 4 presents the results of HLM analysis in an incremental manner, with the upper and lower panel including the results when the Edit Longevity and Edit Count is used as the dependent variable, respectively. **Model 1** is the baseline model including all control variables, **Model 2** adds the network measures into **Model 1**, **Model 3** and

Table 3: Correlation matrix (n=4107)

	1	2	3	4	5	6	7	8	9	10	11	12	13
1.Log(Edit Longevity)	-												
2.Log(Edit Count)	0.91	-											
3.Sc Discussion topics	0.51	0.53	-										
4.Log(Mean Tenure)	-0.02	0.07	-0.08	-									
5.Sc Turnover rate	-0.26	-0.28	-0.12	0.00	-								
6.Sc Gini Talk Edits	0.46	0.49	0.53	-0.07	-0.16	-							
7.Sc Density	-0.34	-0.34	-0.33	-0.01	0.03	-0.33	-						
8.Sc Gini RW Betweenness	0.29	0.28	0.30	-0.12	-0.04	0.36	-0.50	-					
9.Sc PL ratio	0.15	0.17	0.19	0.08	-0.03	0.19	-0.27	-0.20	-				
10.CC ratio	0.47	0.48	0.76	0.00	-0.08	0.52	-0.37	0.34	0.20	-			
11.CC ratio^2	-0.10	-0.12	0.15	-0.03	0.03	-0.22	0.11	-0.11	-0.16	-0.00	-		
12.SWQ	0.45	0.46	0.73	-0.01	-0.08	0.50	-0.34	0.36	0.07	0.98	0.05	-	
13.SWQ^2	-0.12	-0.13	0.04	-0.03	0.03	-0.20	0.10	-0.12	-0.11	-0.08	0.90	-0.00	-

Model 4 add the two specifications of small worldliness into **Model 2**. A χ^2 goodness of fit analysis reveals that models with small worldliness fit the data better than the simpler models. Following Zuur et al. (2009), we checked for the existence of multicollinearity by computing the variance inflation factors (VIFs) using the vif function in the car package for **R** software, the vif values for **Model 4** are presented in Table 4. We observe that the vif values are quite small, suggesting multicollinearity is not a concern.

The coefficient of Sc Discussion Topics is positive and significant across the four models ($p<0.001$), implying that the amount of information and knowledge resources exchanging among participants has a positive impact on project effectiveness. The negative and significant coefficient for Sc Turnover rate ($p<0.001$) indicates that, controlling for other factors, projects with high levels of turnover in general complete less work in a quarter. The coefficient for LogMean Tenure is positive but insignificant when Log(Edit Longevity) is the dependent variable. The coefficient of the LogMean Tenure variable is positive and significant across the four models ($p<0.001$) when Edit Count is the dependent variable, suggesting that when holding other variables constant, greater average membership tenure is beneficial for project effectiveness. We observe that projects which are more senior in their membership achieve a higher level of efficiency, indicating the importance of experience for project efficiency. This provides support for **Ha**.

We characterize the existence of leadership behavior in WikiProjects using the possibility of having a few core members who dominate group communication (i.e. Gini Talk Edits) and network flow (i.e. Gini RW Betweenness). The coefficient for Sc Gini Talk Edits is positive and significant across the four models ($p<0.001$ when Log(Edit Longevity) is the dependent variable; $p<0.001$ in **Model 1**, **Model 2** and **Model 4**, $p<0.01$ in **Model 3** when Edit Count is the dependent variable), suggesting that the existence of explicit or implicit leadership behavior in group communication is positively related to project efficiency. The coefficient for Sc Gini RW Betweenness is positive and significant in **Model 2** and **Model 3** ($p<0.01$). These observations suggest that, controlling for other factors, the existence of explicit or implicit leadership behavior in group communication is positively related to project efficiency. This provides support for **Hb**. One possible implication for **Hb** is that having a core of experienced and dedicated editors to coordinate and organize group communication facilitates information diffusion and resource sharing in the network, which in turn improves project efficiency.

Hc predicted a positive relationship between small-world properties and project efficiency. To test the hypothesis, we firstly introduced the Sc PL ratio, CC ratio and CC ratio^2 terms into **Model 3**. The coefficient for PL ratio is positive and significant ($p<0.05$ and $p<0.01$ for model with Log(Edit Longevity) and Edit Count as the dependent variable, respectively), the coefficient for linear term CC ratio is positive and significant ($p<0.001$), the coefficient for its squared term is negative and significant ($p<0.05$ and $p<0.01$ for model with Log(Edit Longevity) and Edit Count as the dependent variable, respectively). We then introduced SWQ and SWQ^2 into **Model 4**. Again, the coefficient for linear term SWQ is positive and significant ($p<0.001$), and the coefficient for its squared term is negative and significant ($p<0.05$ and $p<0.01$ for model with Log(Edit Longevity) and Edit Count as the dependent variable, respectively). Both results provide partial support for **Hc**. The results suggest that, as the level of connectivity and internal cohesion of the communication network increases globally, the efficiency of WikiProject increases up to an optimal point, but decreases thereafter.

Table 4: Fixed-effect HLM results of Predicting Project Efficiency (ML estimates)

Variables	Model 1		Model 2		Model 3		Model 4		
	Coeff.	S. E.	Coeff.	S. E.	Coeff.	S. E.	Coeff.	S. E.	vif
Log(Edit longevity) as dependent variable, lmer model specification (#Obs: 4107, #WikiProjects: 362, #Quarters: 1–36)									
Intercept	8.751***	0.37	8.613***	0.367	8.730***	0.367	8.700***	0.366	
Sc Discussion topics	0.304***	0.024	0.287***	0.024	0.251***	0.025	0.260***	0.025	1.192
LogMean Tenure	0.049	0.06	0.076	0.059	0.061	0.059	0.064	0.059	1.024
Sc Turnover rate	-0.106***	0.012	-0.109***	0.012	-0.114***	0.012	-0.113***	0.012	1.024
Sc Gini Talk Edits	0.170***	0.017	0.166***	0.017	0.154***	0.017	0.157***	0.017	1.158
Sc Density†			-0.074***	0.018	-0.057**	0.019	-0.074***	0.018	1.238
Sc Gini Flow Between†			0.042**	0.014	0.045**	0.016	0.024	0.015	1.303
Sc PL ratio†					0.033*	0.014			
CC ratio†					7.048***	1.313			
CC ratio^{2†}					-1.929*	0.881			
SWQ†							6.341***	1.211	1.154
SWQ^{2†}							-2.034*	0.799	1.017
Fit									
AIC	9985		9946.1		9906.8		9918.1		
BIC	10041.9		10015.6		9995.3		10000		
logLik	-4983.5		-4962.1		-4939.4		-4946		
χ^2 / df vs. previous nested model			42.9***		45.3***		32.1***		
Variables	Model 1		Model 2		Model 3		Model 4		
	Coeff.	S. E.	Coeff.	S. E.	Coeff.	S. E.	Coeff.	S. E.	vif
Edit Count as dependent variable, glmer model specification (#Obs: 4107, #WikiProjects: 362, #Quarters: 1–36)									
Intercept	4.719***	0.294	4.610***	0.292	4.728***	0.291	4.691***	0.291	
Sc Discussion topics	0.324***	0.018	0.311***	0.018	0.279***	0.019	0.287***	0.018	1.179
LogMean Tenure	0.170***	0.033	0.185***	0.033	0.174***	0.033	0.178***	0.033	1.021
Sc Turnover rate	-0.103***	0.01	-0.106***	0.01	-0.111***	0.01	-0.110***	0.01	1.025
Sc Gini Talk Edits	0.178***	0.012	0.176***	0.012	0.165**	0.012	0.167***	0.012	1.148
Sc Density†			-0.055***	0.013	-0.040**	0.014	-0.055***	0.013	1.231
Sc Gini Flow Between†			0.034**	0.011	0.036**	0.012	0.018	0.011	1.297
Sc PL ratio†					0.030**	0.01			
CC ratio†					6.312***	0.97			
CC ratio^{2†}					-1.736**	0.648			
SWQ†							5.732***	0.892	1.148
SWQ^{2†}							-1.653**	0.586	1.018
Fit									
AIC	62812		62771		62711		62728		
BIC	62869		62840		62799		62811		
logLik	-31397		-31374		-31341		-31351		
χ^2 / df vs. previous nested model			45.4***		66.2***		46.4***		

Note: † one quarter lag for network measures.

Signif. codes: *** p<0.001, ** p<0.01, * p<0.05, ^{t} p<0.1.

Note that the coefficient for network density is negative and significant across the three models (p<0.001 in **Model 2** and **Model 4**; p<0.01 in **Model 3**), suggesting that when controlling for other factors, increasing the connectedness of the communication network for a project generally decreases its efficiency. One possible reason for this negative effect might be related to the nature of the communication networks and could be explained with Burt's theory (Burt, 1997): it takes time, effort and resources to establish and maintain social ties, and the cost of maintaining density communications in workgroups impacts performance.

Table 5: Bootstrapping analysis of highest density regions for main variables

Variables	Log(Edit Longevity) as dependent variable				Edit Count as dependent variable			
	75% hdr	95% hdr	75% hdr	95% hdr	75% hdr	95% hdr	75% hdr	95% hdr
LogMean Tenure	(0.03, 0.21)	(-0.04, 0.28)	(0.03, 0.23)	(-0.04, 0.31)	(0.20, 0.31)	(0.16, 0.35)	(0.20, 0.32)	(0.17, 0.35)
Sc Gini Talk Edits	(0.12, 0.18)	(0.10, 0.20)	(0.12, 0.18)	(0.10, 0.20)	(0.15, 0.18)	(0.13, 0.19)	(0.15, 0.18)	(0.14, 0.19)
Sc Gini RW Betweenness	(0.03, 0.07)	(0.02, 0.08)	(0.008, 0.04)	(-0.003, 0.05)	(0.03, 0.05)	(0.02, 0.05)	(0.008, 0.027)	(0.0003, 0.035)
Sc PL ratio	(0.02, 0.04)	(0.01, 0.06)			(0.02, 0.04)	(0.01, 0.05)		
CC ratio	(4.66, 8.69)	(3.21, 10.18)			(4.53, 8.01)	(3.31, 8.86)		
CC ratio^2	(-2.78, -0.75)	(-3.87, -0.02)			(-2.63, -0.66)	(-3.71, -0.25)		
SWQ			(4.15, 7.78)	(2.90, 8.78)			(4.05, 7.02)	(3.08, 7.86)
SWQ^2			(-2.89, -1.20)	(-3.87, -0.51)			(-2.67, -1.0)	(-3.41, -0.57)

5.1. *Robustness Checks*

We tested the robustness of the results regarding the coefficient estimates in Table 4. We employed a bootstrapping method to check the reliability of the regression coefficients upon which we formulated the hypotheses. Specifically, we used the sample function in the **base R** package to create 1000 replicates of the dataset with replacement for WikiProjects, performed regression analysis on these replicates using the same model specification as in Table 4 and recorded the corresponding coefficients in each round. Finally, we used the hdrcde package (Hyndman, 2008) for **R** software to help identify the highest density regions (hdr) of sampling distributions. The highest density regions of sampling distributions is a method proposed by Hyndman (1996) to find sophisticated confidence intervals for variables. Table 5 presents the 75% hdr and 95% hdr for main variables. We observe that: i) the coefficients for Sc Gini Talk Edits, Sc PL ratio, CC ratio, and SWQ are always positive and fall in the 95% hdr; ii) the coefficients for LogMean Tenure and Sc Gini RW Betweenness are always positive and fall in the 75% and 95% hdr for model with log(Edit Longevity) and Edit Count as dependent variable, respectively; iii) the coefficients for CC ratio^2 and SWQ^2 are always negative and fall in the 95% hdr. The observations suggest the robustness of the coefficient estimates.

5.2. *Effect of Small Worldliness*

Figure 4 graphically presents the bivariate relationships between small-world properties and project effectiveness. It is obvious from Figure 4 that the small worldliness of communication networks correlates positively with project efficiency, but the positive relationship continues only up to an optimal point after which the efficiency decreases markedly. The four plots suggest that a medium level of small worldliness produces the most beneficial small world effect on project efficiency. Either too little order (i.e. highly isolated network) or too much order (i.e. network with high interaction frequency) in the level of small worldliness has a negative impact on project effectiveness, which is consistent with findings by other researchers (e.g., the effect of small worldliness on musician performance by Uzzi and Spiro (2005)).

Essentially, the benefits of small-world structure are closely related to its working mechanism. Small-world networks are characterized by high levels of clustering among individuals and low average path lengths or social distance between the individuals (Watts and Strogatz, 1998). An intermediate level of small worldliness optimizes the trade-off between the benefits of information sharing and burdens of information overload for network members. On the other hand, as suggested by Lazer and Friedman (2007), information transfer and knowledge sharing is less likely to occur in a network with no or very low level of connectivity among members. A fully connected network would incur cost in that it takes time, effort and resources to establish and maintain social ties (Zhou et al., 2009). In the case of Wikiprojects, where communication networks with intermediate levels of connectivity have the most effective discussion and communication among members. This facilitates both access to and diversity of knowledge resources available to members and in turn improves project efficiency.

5.3. *Exploration of Leadership*

While the Gini measure of random walk betweenness centrality provides a brief overview of the temporal dynamics of a project and the possible leadership behavior in the project, it provides little insights into the peaks and troughs in the temporal evolution of centrality in user level in order to detect changes in group communication patterns. Specifically, the Gini measure provides little insights about the specific leadership exhibited in group communications. In

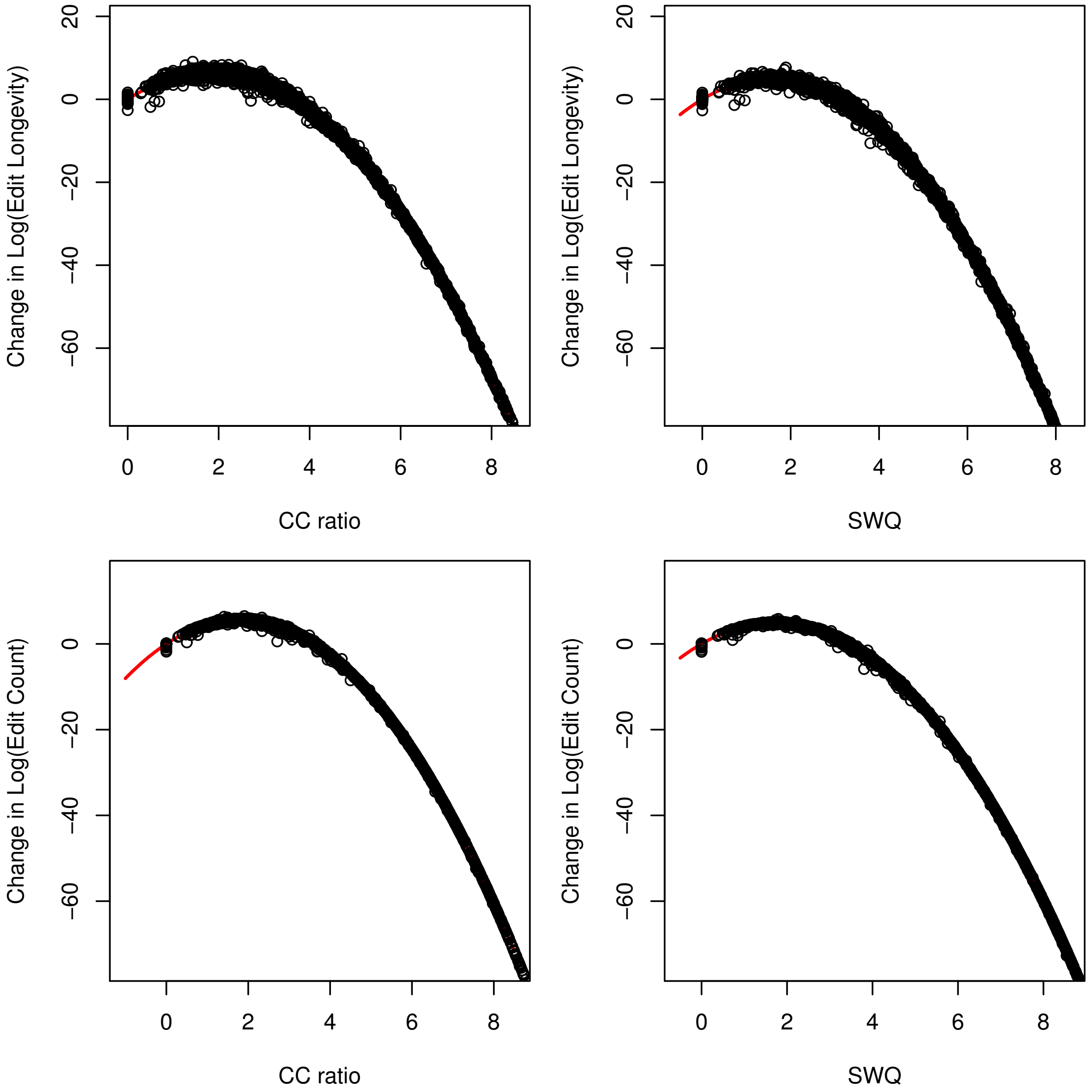


Figure 4: An Invert-U shaped relationship for small-world properties and project effectiveness. The red lines show the model fit for the relationship of CC ratio and SWQ to Log(Edit Longevity) on the upper panel and Log(Edit Count) on the lower panel. The y-values of the dots correspond to the observed Efficiency minus the Efficiency predicted by all other variables. Thus the scatter plots depicts the covariance of Efficiency with CC ratio and SWQ conditional on all other variables.

this section, we employ heatmap to visualize the change in centrality of project members over the lifecycle of a project and examine the inherent leadership behavior in the communication networks. Figure 5 illustrate team dynamics of selected healthy and less healthy WikiProjects. The two healthy projects were very active in project talk pages and very effective in improving project scope articles in the period considered in this study.

We make several observations from Figure 5: (1) the majority of members were inactive at any given quarter; (2) in healthy projects, there are a small group of highly central members, communicating with a large group of peripheral participants, indicating a core-periphery structure and leadership behavior in the projects; (3) in healthy projects, those highly central members tend to be active in project talk pages for a longer period of time; (4) by contrast, in

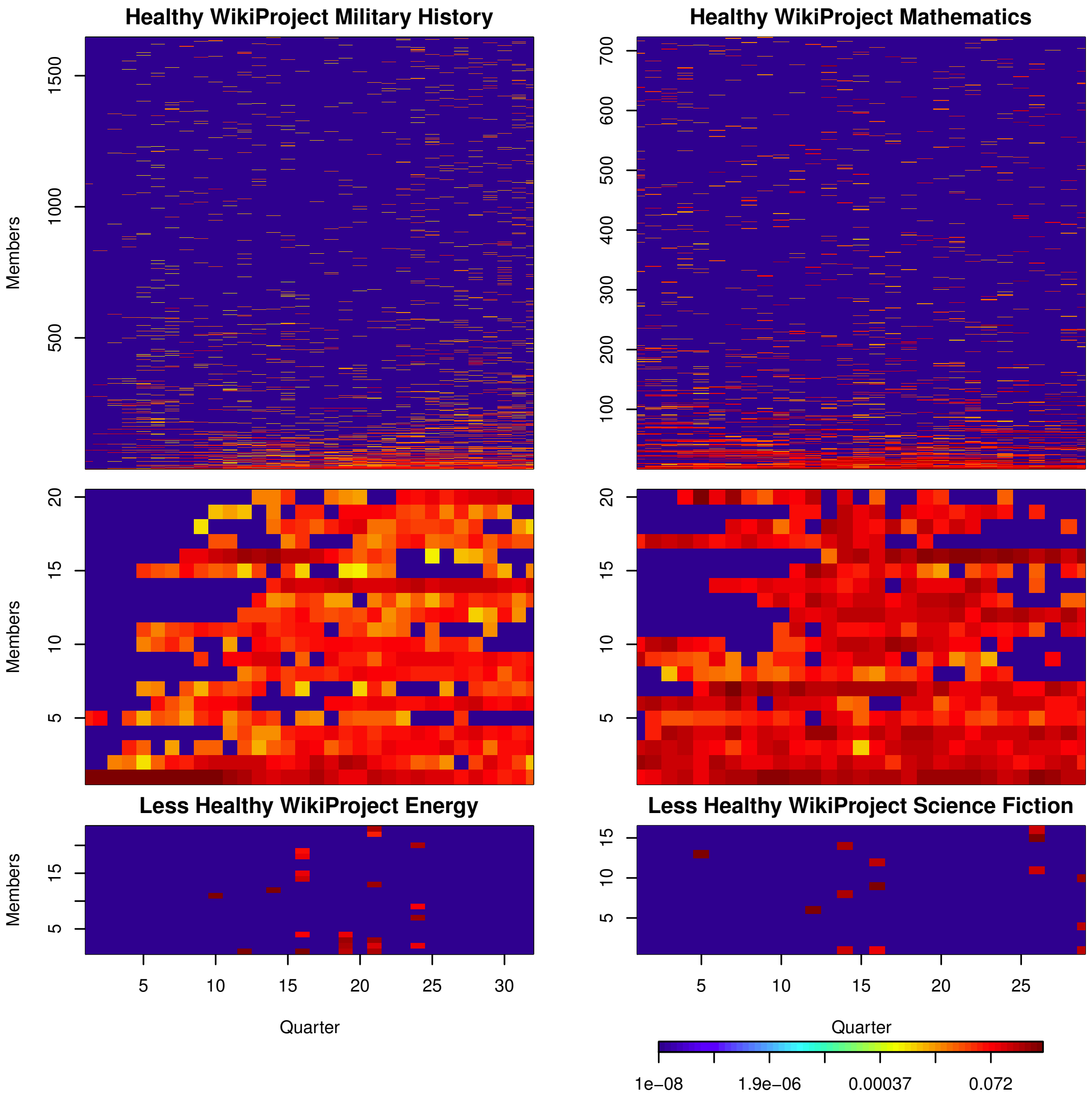


Figure 5: Heatmaps for the dynamics of random walk betweenness centrality in user level for four selected WikiProjects. The x-axis represents the index for Quarter, y-axis denotes the index for members. For the two healthy WikiProjects, a zoom-in visualization of the centrality for the top 20 active members is provided. Note that to smooth the visualization, the centrality measures are log-transformed.

less healthy projects, highly central members tend to locate randomly in the heatmap over time, suggesting there exists no long-term leadership in these projects to coordinate project communication. These observations suggest that healthy WikiProjects generally have a core of long tenure and central members to coordinate and organize group communication, which facilitates information diffusion and resource sharing in the network and in turn promotes project efficiency. These observations provide further support for **Hb**.

Table 6 provides further statistics for the 4 WikiProjects in Figure 5. We observe from Table 6 that on average, the Gini RW Betweenness and Gini Talk Edits of the two healthy WikiProjects are larger than those of the two less healthy ones; healthy WikiProjects generally have more quarterly observations than the less healthy ones; the member turnover rate of healthy WikiProjects is smaller than that of the less healthy ones, indicating that less healthy WikiProjects

Table 6: Statistics for the 4 selected WikiProjects in Figure 5

WikiProject	#Quarterly Obs	Gini RW Betweenness		Gini Talk edits		Turnover rate	
		Mean	std	Mean	std	Mean	std
WikiProject Military History	31	0.74	0.06	0.70	0.06	0.24	0.07
WikiProject Mathematics	28	0.58	0.04	0.59	0.04	0.28	0.08
WikiProject Energy	7	0.52	0.18	0.12	0.16	0.38	0.07
WikiProject Science Fiction	4	0.50	0.25	0.13	0.16	0.46	0.27

generally experience higher level of membership turnover than healthy ones, and that there might be some potential differences between team compositions of efficient and less efficient WikiProjects in terms of task distribution among members. These findings further confirm the observations in Figure 5.

5.4. Exploration of Language Coordination

Our second hypothesis stated that there is a positive relationship between project efficiency and the existence of leadership behavior in group communication. The idea being that those members in central / leadership positions play a fundamental role in facilitating information flow and the organization and coordination of the collaborative activities, which can potentially improve project efficiency. While the Gini measure of network centrality and user edits to project talk pages captures the potential leadership behavior in project communication, it provides very little insights into how these members earned and achieved their positions, what distinguished them from ordinary members. In this section, we explore the status differences among members by analysing their linguistic usages using the method of linguistic coordination suggested by (Danescu-Niculescu-Mizil et al., 2012).

Based on the exchange theory from sociology regarding power differences within social groups, Danescu-Niculescu-Mizil et al. (2012) provided a simple probabilistic measure which quantifies language coordination from a speaker to a target over a set of function words. Their results showed that the proposed linguistic coordination measure is successful in differentiating individuals with different power status. They defined the language coordination of a speaker a towards a target b over a function word category k as follows:

$$C^k(a \rightarrow b) = P(\varepsilon^k_{u_a \rightarrow u_b} | \varepsilon^k_{u_b}) - P(\varepsilon^k_{u_a \rightarrow u_b}) \tag{6}$$

where a is the speaker who coordinates towards the target b; $\varepsilon^k_{u_a \rightarrow u_b}$ is the event that the utterance of a exhibits k in its reply to b; $\varepsilon^k_{u_b}$ is the event that the utterance u_b (replied to a) exhibits k. The conversation set $S_{a,b}$ is defined over the exchanges $(a : u_a, b : u_b)$ which contain the words from a specific function word category k. The first probabilities can be factorized using Bayes' formula, and the the probabilities are estimated over $S_{a,b}$. For instance, in the conversational exchange between a speaker a and a target b, b used words from category k in all of its 15 replies to a; a used words from category k in 10 out of 12 replies to b. Then the probabilities can be estimated as: $P(\varepsilon^k_{u_a \rightarrow u_b}) = \frac{10.0}{12.0} = 0.83$, $P(\varepsilon^k_{u_b}) = \frac{15.0}{15.0} = 1.0$, $P(\varepsilon^k_{u_b} | \varepsilon^k_{u_a \rightarrow u_b}) = \frac{15.0}{10.0} = 1.5$. The coordination of a towards b over category k can be calculated as: $C^k(a \rightarrow b) = (\frac{1.5}{1.0} - 1.0) * 0.83 = 0.415$.

In the context of group conversations, the definition of linguistic coordination of a speaker a towards a group of targets B is given by:

$$C^k(a \rightarrow B) = P(\varepsilon^k_{u_a \rightarrow u_B} | \varepsilon^k_{u_B}) - P(\varepsilon^k_{u_a \rightarrow u_B}) \tag{7}$$

where the probabilities are estimated over the conversation set $S_{a,B}$. Similarly, the definition of linguistic coordination of a group of people A towards another group B is defined as the averaged coordination of speakers in A towards targets in B:

$$C^k(A \rightarrow B) = \langle C^k(a \rightarrow B) \rangle_{a \in A} \tag{8}$$

For this analysis, we selected 34 WikiProjects that have several hundreds to several thousands of members, and have a substantial amount of conversations in their project talk pages. Generally, researchers are interested in studying the language coordination between groups of people with power or status differences (e.g., admins vs non-admins, Justices vs. lawyers in (Danescu-Niculescu-Mizil et al., 2012)). Here we are interested in exploring whether members in peripheral positions (with low centrality measures) coordinate more towards members in central positions (with higher centrality measures) than towards members in non-central positions, and vice versa. We focus on the coordination between three groups of users: leader coordinators, non-leaders and ordinary members. We annotated those central and long-term members as leader coordinators, a random group of non-central members as non-leaders, the remaining members as a group of ordinary members. For consistency with prior work, we followed Danescu-Niculescu-Mizil et al. (2012) and used eight of the nine LIWC-derived function word categories (Pennebaker et al., 2007)[6] for the analysis. We calculated linguistic coordination measure for each WikiProject using the thread-based discussions in its talk pages.

Table 7: Statistics about aggregated language coordination

	Supporting H_{target}	Contradicting H_{target}	Supporting $H_{speaker}$	Contradicting $H_{speaker}$
#WikiProjects	24	10	0	34

We evaluate two hypotheses about linguistic coordination: (i) ordinary members coordinate more towards leader coordinators than towards non-leaders (H_{target}); (ii) leader coordinators coordinate towards ordinary members less than non-leaders coordinate towards ordinary members ($H_{speaker}$). Figure 6 shows the language coordination of the three groups for three WikiProjects. We observe from Figure 6 that (1) ordinary members coordinate more towards leader coordinators (central positions, high status) than towards non-leaders (peripheral positions, low status), supporting H_{target}; (2) leader coordinators coordinates more than non-leaders towards ordinary members, contradicting $H_{speaker}$. The observation is consistent with that of (Danescu-Niculescu-Mizil et al., 2012).

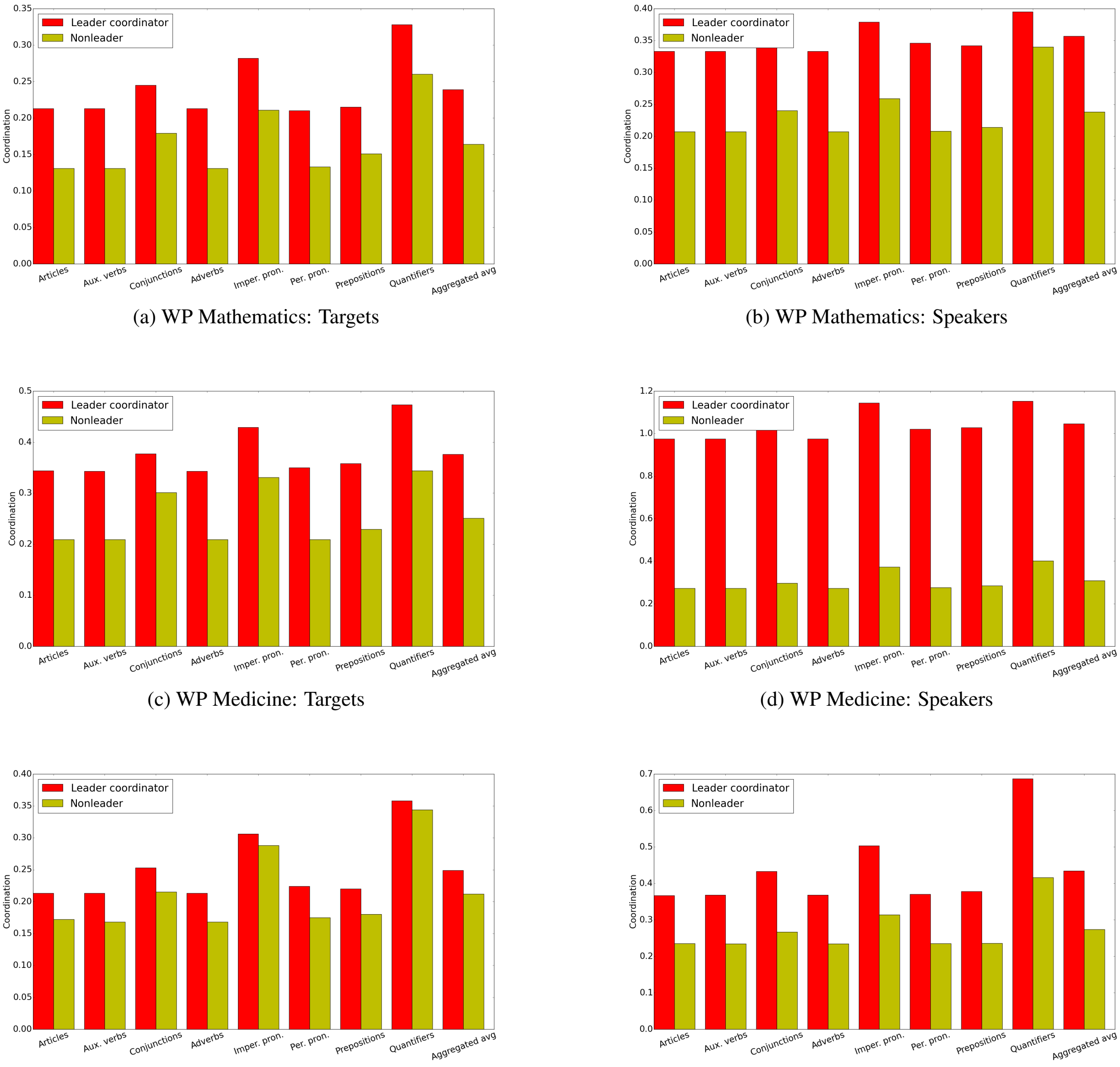
The overall statistics about the support of the two hypotheses for the 34 WikiProjects are given in Table 7. The coordination measure of the majority of the WikiProjects support H_{target}, indicating that it is reasonable to assume those central members are in leadership position. This observation also suggests the validity of our Gini measure for network centrality to quantify the presence of leadership behavior in the network. The linguistic coordination measure of all the 34 WikiProjects contradicts the hypothesis $H_{speaker}$. One possible explanation for the inconsistency of the observation with $H_{speaker}$ is that leader coordinators generally coordinate more than non-leaders towards members, and that these coordinators earn their central/leadership positions by actually coordinating and interacting with members. It is well-known that in online settings, new members can learn discipline, rules and regulations about the system, and how to make contributions by social learning and interaction with experienced pioneers. An intriguing but still unknown question arises here: whether the coordination of leader coordinators towards ordinary members will help members to learn more about the system and become experienced faster. Overall, the results suggest that, rather than being a place full of messy and contentious discussions, there is a group of leadership members who coordinate and organize the discussions in project talk pages.

6. Discussion and Conclusion

In this study, we investigated how the network structural properties of online teams engaged in knowledge creation relates to the efficiency. We formulated three hypotheses about how network social capital influences project efficiency, and evaluated the hypotheses on a longitudinal dataset of 362 WikiProjects. The results provided support for the first two hypotheses, and partially supported **Hc**. Specifically:

- In support of **Ha**, greater average membership tenure relates to project efficiency in a positive way.
- In support of **Hb**, the existence of leadership behavior in group communication is positively related to project efficiency.
- In partial support of **Hc**, the small-world properties of communication networks relate positively to project efficiency, but the positive relationship continues only up to an optimal point after which the efficiency decreases markedly.

[6] Available from the authors of the software upon email request.



(a) WP Mathematics: Targets

(b) WP Mathematics: Speakers

(c) WP Medicine: Targets

(d) WP Medicine: Speakers

(e) WP Military history: Targets

(f) WP Military history: Speakers

Figure 6: The aggregated language accommodation measurement using

6.1. *Implications*

This work has several implications for understanding knowledge creation in online communities, in particular Wikipedia. Firstly, the results suggest that, on average, online teams with more long-tenured members are more likely to enjoy an enhanced capacity for better performance. This finding suggests that WikiProject managers should employ automatic tools to identify members with specialized skills and valuable contributions, and give awards (e.g. barnstar-type recognition) to these dedicated members. The recognition of skills and achievements would encourage such members to stay longer with the project for the benefit of the community, and enable other new or inexperienced users to determine who best to approach if they have a query or problem. Secondly, the results indicate that having a core of influential users who assume explicit or implicit leadership roles in group communication promotes project efficiency. An implication of this for WikiProject management is that WikiProjects should nominate several socially active members as project coordinators or leaders. These members play a fundamental role in coordinating and organizing collaborative activities for the projects, explaining the community norms and policies, which facilitates information diffusion and resource sharing among participants and in turn improves project efficiency. Furthermore, these explicit or implicit leaders can help establish a harmonious working environment for project participants and promote social learning among members, which is essential to membership retention in online communities.

In addition, the results show that an intermediate level of small-worldliness is optimal for project effectiveness. This finding has a mixed message for Wikipedia editors. The first message is that communities (i.e. WikiProjects) with small-world structure are more interconnected by members who know each other well either through repeat collaboration or common third-parties. Since small-world structures are constructive to project effectiveness, WikiProjects should aim to achieve them by employing strategies such as improving membership retention. The second message of this finding is about Q&A in WikiProject talk pages. Recall that we construct the communication networks by parsing the mutual replying behavior in project talk pages. To obtain a connective and internally cohesive communication network, experienced members should be encouraged to respond to as many questions as possible and provide help or support for those in need. Having many questions replied indicates that project members get help and support from the community, and that information and knowledge resources are shared among members, which can potentially help improve members' work and in turn enhance project effectiveness.

6.2. Email Survey

To probe the validity of the results, we conducted a short email survey with a small group of experienced Wikipedians from several efficient WikiProjects. The survey consisted of three questions: (1) the functionality of WikiProjects, (2) the possible relationship between the efficiency of WikiProjects and communications in project talk pages, and (3) identification of well-organized and productive WikiProjects. We contacted about 30 experienced editors from selected active WikiProjects by leaving messages on their user talk pages, 10 of them responded and took the survey. The following is a summary of the survey:

- Generally, these members agree that WikiProjects, which they have interacted with, are well organized in terms of providing help and support for new members, coordinating collaborative and discussion activities, sharing necessary resources such as guidelines for article creation and promotion, reaching consensus and other aspects.

- Regarding question 2, many members mention that several key factors such as the involvement of a core group of enthusiastic and experienced members in the long term, having a system (i.e. WikiProject) to support project participants and single editors working on single articles, are essential to project efficiency.

- For question 3, the majority of them mention that the topic of a project can affect its efficiency, and that some WikiProjects tend to be less efficient due to the dynamic or contentious nature of the topic (e.g. articles on movie stars or politics). Due to this fact, these members generally identify WikiProjects related to scientific and static subjects (e.g., WikiProject Physics, Mathematics, Medicine, Military History, Aviation, Plants) to be well-organized and productive.

6.3. Limitations

This work has limitations. The first limitation of the work lies with the communication networks used. We only considered the communication networks encoded in project talk pages, while in Wikipedia, users can interact with others in pages such as user talk pages, article talk pages. Future work should consider communication networks that extend across multiple dimensions of user interaction. Secondly, the results suggested that efficient WikiProjects generally experience lower levels of member turnover than less efficient WikiProjects, it would be interesting to investigate the difference between team compositions of efficient and less efficient WikiProjects in terms of task distribution among members and its influence on group performance. Furthermore, the analysis of linguistic coordination reveals that those central and long-tenured members tend to coordinate more towards ordinary members, future work could explore whether such coordination help new and inexperienced members to learn more about the system and become experienced faster.

6.4. Conclusion

In this work, we empirically investigated the impact of network structural properties of communication networks for WikiProjects on project efficiency which was measured as the amount of work done by project members in a quarter. Despite the limitations, this work makes several contributions. First, it provides empirical evidence that moderate level of connectivity and internal cohesion of the communication network can positively affect project effectiveness. An intermediate level of small worldliness optimizes the trade-off between the benefits of information sharing and burdens of information overload for network members. Moreover, it empirically shows that online teams with leadership behavior and more long-tenured members are more likely to enjoy an enhanced capacity for better performance. Overall, the results show that an intermediate level of small-world structure with a core of long-tenured and influential members dominating network flow promotes project effectiveness. Thus, this work offers insights that extend our understanding of the influence of network structures on group performance in online communities.

References

Adler, B.T., Alfaro, L.D., 2007. A Content-Driven Reputation System for the Wikipedia, in: Proceedings of WWW2007, ACM, Banff, Alberta, Canada. pp. 261–270.

Ahuja, G., 2000. Collaboration Networks, Structural Holes, and Innovation: A Longitudinal Study. Administrative Science Quarterly 45, 425–455.

Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: An open source software for exploring and manipulating networks, in: Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM'09), AAAI, San Jose, California. pp. 361–362.

Bates, D., Maechler, M., Bolker, B., 2012. lme4: Linear mixed-effects models using S4 classes. URL: `http://CRAN.R-project.org/package=lme4`.

Benkler, Y., 2006. The Wealth of Networks: How Social Production Transforms Markets and Freedom. Yale University Press, New Haven,Connecticut.

Bryk, A.S., Raudenbush, S.W., 1992. Hierarchical Linear Models: Applications and Data Analysis Method. Sage Publications.

Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, MA.

Burt, R.S., 1997. A Note on Social Capital and Network Content. Social Networks 19, 355–373.

Coleman, J.S., 1988. Social capital in the creation of human capital. The American Journal of Sociology 94, S95–S120.

Crowston, K., Howison, J., 2005. The social structure of open source software development teams. First Monday .

Cummings, J.N., Cross, R., 2003. Structural properties of work groups and their consequences for performance. Social Networks 25, 197–210.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., Kleinberg, J., 2012. Echoes of Power: Language Effects and Power Differences in Social Interaction, in: Proceedings of WWW 2012, ACM, Lyon, France. pp. 699–708.

Fleming, L., King III, C., Juda, A.I., 2007. Small Worlds and Regional Innovation. Organization Science 18, 938–954.

Freeman, L., 1979. Centrality in Social Networks: Conceptual Clarification. Social Networks 1, 215–239.

Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, London.

Gulati, R., 1999. Network location and learning: The influence of network resources and firm capabilities on alliance formation. Strategic Management Journal 20, 397–420.

Gulati, R., Gargiulo, M., 1999. Where Do Inter-organizational Networks Come From? The American Journal of Sociology 104, 1439–1493.

Hagberg, A.A., Schult, D.A., Swart, P.J., 2008. Exploring network structure, dynamics, and function using NetworkX, in: Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA USA. pp. 11–15.

Hausman, J. andHall, B., Griliches, Z., 1984. Econometric models for count data with an application to the patents R&D relationship. Econometrica 52, 909–938.

Hyndman, R.J., 1996. Computing and Graphing Highest Density Regions. The American Statistician 50, 120–126.

Hyndman, R.J., 2008. hdrcde: Highest Density Regions and Conditional Density Estimation. R package version 2.09 URL: `http://CRAN.R-project.org/package=hdrcde`.

Kidane, Y.H., Gloor, P.A., 2007. Correlating temporal communication patterns of the Eclipse open source community with performance and creativity. Computational & Mathematical Organization Theory 13, 17–27.

Kittur, A., Kraut, R.E., 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination, in: Proceedings of CSCW-2008, ACM, San Diego, CA, USA. pp. 37–46.

Lazer, D., Friedman, A., 2007. The network structure of exploration and exploitation. Administrative Science Quarterly 52, 667–694.

Lorenz, M.O., 1905. Methods of measuring the concentration of wealth. Publications of the American Statistical Association 9, 209–219.

Nemoto, K., Gloor, P.A., Laubacher, R., 2011. Social capital increases efficiency of collaboration among Wikipedia editors, in: Proceedings of HT2011, ACM, Eindhoven, the Netherlands. pp. 231–240.

Newman, M., 2005. A measure of betweenness centrality based on random walks. Social Networks 27, 39–54.

Ng, T.W.H., Feldman, D.C., 2010. Organizational Tenure and Job Performance. Journal of Management 36, 1220–1250.

Pennebaker, J.W., Booth, R.J., Francis, M.E., 2007. Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program. URL: `http://www.liwc.net/`.

Portes, A., 1998. Social Capital: Its Origins and Applications in Modern Sociology. Annual Review of Sociology 24, 1–24.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`. ISBN 3-900051-07-0.

Reagans, R., Zuckerman, E.W., 2001. Networks, Diversity, and Productivity: The Social Capital of Corporate R&D Teams. Organization Science 12, 502–517.

Schilling, M.A., Phelps, C.C., 2007. Interfirm Collaboration Networks: The Impact of Large-Scale Network Structure on Firm Innovation. Management Science 53, 1113–1126.

Singh, P.V., 2010. The Small-World Effect: The Influence of Macro-Level Properties of Developer Collaboration Networks on Open-Source Project Success. ACM Transactions on Software Engineering and Methodology, 20, 6:1–6:27.

Ung, H., Dalle, J.M., 2010. Project management in the Wikipedia community, in: Proceedings of the 6th International Symposium on Wikis and Open Collaboration, ACM, Gdansk, Poland. pp. 13:1–13:4.

Uzzi, B., Spiro, J., 2005. Collaboration and Creativity: The Small World Problem. American Journal of Sociology 111, 447–504.

van Rossum, G., Drake, F.L., 2009. *The Python Language Reference*. URL: `http://docs.python.org/2.7/reference/index.html`.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393, 440–442.

Yamaguchi, K., 1994. The flow of information through social networks: Diagonal-free measures of inefficiency and the structural determinants of inefficiency. Social Networks 16, 57–86.

Zhou, J., Shin, S.J., Brass, D.J., Choi, J., Zhang, Z.X., 2009. Social networks, personal values, and creativity: Evidence for curvilinear and interaction effects. Journal of Applied Psychology 94, 1544–1552.

Zhu, H., Kraut, R., Kittur, A., 2012. Effectiveness of shared leadership in online communities, in: Proceedings of CSCW-2012, ACM, Bellevue, Washington. pp. 407–416.

Zhu, H., Kraut, R.E., Wang, Y.C., Kittur, A., 2011. Identifying shared leadership in Wikipedia, in: Proceedings of CHI 2011, ACM, Vancouver, BC, Canada. pp. 3431–3434.

Zuur, A., Ieno, E., Walker, N., Saveliev, A.A., Smith, G., 2009. Mixed Effects Models and Extensions in Ecology with R. Springer Publishing, New York City.