Social Capital Increases Efficiency of Collaboration among Wikipedia Editors

Keiichi Nemoto
Fuji Xerox Co., Ltd.
Tokyo, Japan
keiichi.nemoto@fujixerox.co.jp

Peter A. Gloor
MIT Center for Collective Intelligence
Cambridge MA, USA
pgloor@mit.edu

Robert Laubacher
MIT Center for Collective Intelligence
Cambridge MA, USA
rjl@mit.edu

ABSTRACT
In this study we measure the impact of pre-existing social capital on the efficiency of collaboration among Wikipedia editors. To construct a social network among Wikipedians we look to mutual interaction on the user talk pages of Wikipedia editors. As our data set, we analyze the communication networks associated with 3085 featured articles – the articles of highest quality in the English Wikipedia, comparing it to the networks of 80154 articles of lower quality. As the metric to assess the quality of collaboration, we measure the time of quality promotion from when an article is started until it is promoted to featured article.

The study finds that the higher pre-existing social capital of editors working on an article is, the faster the articles they work on reach higher quality status, such as featured articles. The more cohesive and more centralized the collaboration network, and the more network members were already collaborating before starting to work together on an article, the faster the article they work on will be promoted or featured.

Categories and Subject Descriptors
H.5.3 Group and Organization Interfaces: Computer-supported cooperative work, H.5.1 Multimedia Information Systems: Evaluation/methodology

General Terms
Management, Human Factors

Keywords
Social capital, Social network analysis, Social networks, Social media, Collaboration, Community governance, Wikipedia, Open source projects, Time-to-market.

1. INTRODUCTION
It has been frequently argued that social capital, in analogy to other forms of capital, offers benefits to those who have it [23], [7]. However, it has been notoriously difficult to measure the amount of social capital an individual or group has accumulated. Existing methods include survey- and assessment-based approaches [31], [27]. In this paper we investigate this question in a well-documented and measurable environment, exploring how pre-existing social capital among Wikipedians influences the efficiency of their work. In particular, we investigate how a pre-existing network of mutual ties of support among collaborating authors influences "time-to-market" of Wikipedia articles from the beginning of an article to the highest level of quality.

Prior studies inside firms have shown the role of informal ties in shaping organizational performance [17]. Until recently, studies examining such informal networks were undertaken using surveys asking respondents to identify the nature of their connections with colleagues [12]. More recently, collaboration networks have been constructed based on e-mail exchanged. Motivated by Aral et. al [5], who showed that social network position in the e-mail network and performance of executive recruiters were highly correlated, we expect that social network structure of Wikipedians will also predict their performance in getting new articles to the highest level of quality.

2. RELATED WORK
Wikipedia is an example of a new form of Internet enabled group production. It has variously been termed open source production [4], peer production [6], crowdsourcing [14], collaborative innovation [13], and Internet-enabled collective intelligence [20].

Three streams of literature inform our work: studies that examine how interaction patterns affect group performance in traditional organizations; studies of open source software development; and studies of Wikipedia.

Many studies have examined the linkage between social network structure and performance in business organizations [3], [12], [21], [26], [5]. These studies have shown that different network structures are correlated with high performance of different work tasks. Of note is Uzzi and Spiro's finding [26] that for the creative work of developing Broadway shows, a mix of some network embeddedness, in the form of collaborators who had worked together previously, and some network diversity, in the form of newcomers to the team, was associated with critical and box office success in this creative field.

Open source software (OSS) development was the first prominent example of Internet enabled, voluntary collaboration. Many social and organizational scientists have examined OSS. The primary focus of this research has been on how OSS development teams get work done and what motivates individuals to participate [18], [11]. Some have noted that quality is often higher in open source projects than in traditional software development [28].

Studies examining what makes for more effective OSS development have been produced by practitioners (for a review see [1]). This research has focused on organizational practices, software architecture, project management, and development processes. One finding from this work is that an effective structure is the onion model, which features a small number of

core developers, along with a peripheral group of contributing developers, bug reporters, and users.

A study of OSS that takes a social/organization science perspective shows that a stable, centralized network structure is more effective for teams whose primary task is fixing bugs, while a structure that features fluctuations in centrality is more effective for teams whose primary task is generating new code [15].

Anthony et al. [4] is one of the first studies to employ an independent metric to assess quality, though the metric used in this study measures the quality of individual editor contributions, rather than the quality of articles as a whole. This study posited the share of an editor's contributions that remained in the current version of Wikipedia as a measure of the quality of that editor's contributions. Editors were categorized along two dimensions—registered vs. non-registered; and according to the number of contributions they had made. The study found the highest quality edits were made by two groups: registered editors who make many edits and anonymous editors who make few. An implicit story emerged from this research—that the best articles included a core of experienced editors along with contributions from people who could provide tidbits of specialized expertise.

Subsequent studies have used a similar metric. Adler and Alfaro [2] extended this approach by measuring the amount of time editors' contributions stood. Priedhorsky et al. [22] then extended it further by also taking into account the number of readers who viewed editors' contributions.

With maturation of the Wikipedia community's article evaluation project [29], researchers had access to a tool for evaluating the quality of Wikipedia's articles. 2,714,054 articles have been rated on a nine-level scale as of May 2010. The quality of articles, as assessed by independent reviewers, is strongly positively correlated with the ratings provided by Wikipedia community [16].

Wilkinson and Huberman [30] was one of the first studies to exploit the article evaluation project in a systematic way. After controlling for article age and size, this study found that featured articles had more edits and more editors than a random sample of other Wikipedia articles.

Kittur and Kraut [16] examined how explicit and implicit coordination were associated with changes in the quality of Wikipedia articles over time. This study found that articles where editorial work was more concentrated, and thus which relied more on implicit coordination, improved more than the norm. Explicit coordination, in the form of activity on article talk pages, also improved quality, but only when the number of editors was manageable. This study suggested that in early stages of development, having a small number of editors to set an article's "direction, structure, and scope" was important. Once those tasks were completed, it was possible for a larger number of editors to make effective contributions.

Liu and Ram [19] used clustering analysis to identify six primary roles that Wikipedia editors played, based on the constellation of tasks they typically performed. This study then undertook another clustering analysis to identify five primary types of Wikipedia articles, based on the mix and volume of tasks undertaken by each type of editor. This analysis showed that the highest quality articles by far were those where "all-around editors," who were adept at every task, assumed the greatest role.

The picture that emerges from this research is that effective open source production occurs when the efforts of a core group of experienced contributors are augmented by occasional additions from low volume contributors. These findings also suggest that a pre-existing collaboration network with a centralized core/periphery structure might be more efficient in getting articles to the highest level of quality.

3. METHODS

In our project we study the collaboration networks among the editors of the English Wikipedia. While many Wikipedia readers are only aware of the main text pages where the articles reside, and perhaps the talk pages associated with the articles, there are also a very large number of Wikipedia user pages—effectively personal home pages for each registered Wikipedia editor. Wikipedia user pages work much like Facebook. The main user page displays whatever personal information the user wishes to share, plus all the awards the user may have received from fellow Wikipedians. The main user page is complemented by a user talk page, where users, just like on the Facebook Wall, discuss the articles they are working on, debate topics of general Wikipedia interest, and exchange social messages.

The number of edits on article and article talk pages on the English Wikipedia peaked in 2007, however the number of edits on user talk pages has still been growing [8]. Thus, while the level of activity on article and article talk pages is in decline, interpersonal activity on user talk pages is expanding. This means that direct interaction between registered Wikipedians is becoming more and more important. Analysis of user talk pages effectively allows a look at social capital among Wikipedians to see its impact on how Wikipedians get their work done.

3.1 Measuring Article Quality

Quality assessments of Wikipedia articles are mainly performed by members of WikiProjects (see 3.3) (http://en.wikipedia.org/wiki/Wikipedia:WikiProject). The 7-point quality score ranges from "Stub"-Class (lowest) to "Start" to "C" to "B" to "GA"(Good article) to "A" to "FA"(Featured article) (highest). Once an article reaches A-Class, it is considered "complete", although edits will continue to be made. GA and FA assessments are made by external panels, rather than by WikiProjects[1]. Before getting higher-level assessments, an article typically progresses through several levels. For instance, the article "Atom" was assessed as Stub quality on Oct. 8, 2001, Start on Sep. 20, 2002, C-Class on Sep.18, 2004, B-Class on Aug. 19, 2006, GA on Feb. 10, 2008, and FA on Feb. 12, 2008[2].

We collected the featured articles list from Wikipedia on Nov. 17, 2010, the good articles list on Nov. 20, 2010, and the B-Class articles on Nov. 29, 2010 from the English Wikipedia. The

[1] http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

[2] http://outreach.wikimedia.org/wiki/ Life_of_an_Article

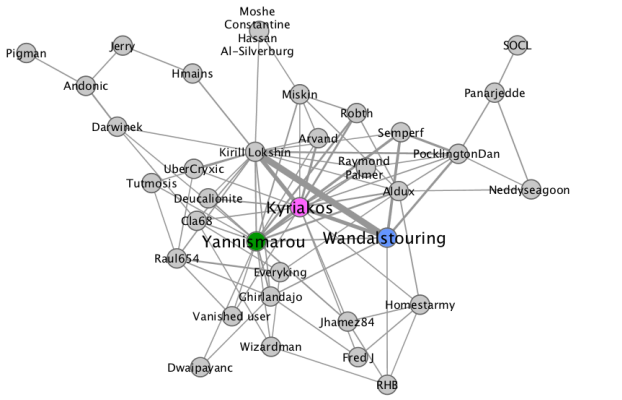
whole list included 3085 "FA-Class (featured)" articles[3], 10058 "GA-Class (good)" articles[4], and 70096 "B-Class" articles.

To track the article evolution history, we also collected the time when an article got a promotion or demotion. Since some articles did not list the exact promotion date, we were only able to collect 3080 FA-Class articles, 10051 GA-Class articles, and 69627 B-Class articles.

In our analysis we decided to focus on the transition of articles from Start to B, from B to GA, and from GA to FA level, as these were the dominant status changes that most articles went through. For example, 47 percent of all FA articles were promoted from GA level, while only 6 percent of FA articles directly came from A-Class. Similarly, 41 percent of all GA articles were directly promoted from B-Class level.

3.2 Collaboration Network Construction

To construct the collaboration network we utilized the user talk pages, employing an approach similar to Crandall et. al. [10], constructing a link between editors A and B if A and B worked on the same article, and editor A left a comment on the talk page of B (or vice versa). We only looked at registered users because they have their own User and User talk page on Wikipedia. Anonymous IP users are eliminated as well as "bot" users, which are robots written by Wikipedians to do repetitive cleanup tasks. Overall, anonymous IP users, bot users, and registered users make 23%, 3%, and 74% of all article edits, respectively.



Roman-Spartan War [edit]

Hi Yanni, I withdrew the Roman-SPartan War from FAC and I put it on Peer Review at WPMILHIST. So if you have the time and don't mind could you please go and leave some suggestions becuase they always seem to be of good value and worth while. Thanks in advance and also I forgotto congratulate you on El Greco becoming a FA. It's great that you took the article of the greatest Greek painter to FA level. Kyriakos 05:49, 4 January 2007 (UTC)

Figure 1. Collaboration network constructed by comments on user talk pages and friendly comment about the article "War against Nabis" on the user talk page of Yannismarou[5]

We assumed a link existed between two authors if they exchanged at least one message on each other's user talk pages. Figure 1 illustrates the collaboration network of the featured article "War against Nabis" from November 18, 2006 to February 17, 2007. This article became a featured article only three months after it was created, an unusually rapid ascent. The two most active editors were Kyriakos (pink) with 41% of all edits (266 out of 651) and Wandalstouring (blue) with 35% of all edits (228 out of 651). In the graph a node represents a user working on the article, and an edge between two users on the social graph is drawn if user A and user B have exchanged at least one comment on their respective talk pages.

The bottom of figure 1 shows a comment that Kyriakos (pink) wrote on the talk page of Yannismarou (green), who is the fourth most active editor with 3 % of all edits (19 out of 651). This comment illustrates the way that editors discuss work on user talk pages in general and also the friendly relationship that exists between these two editors.

To measure the structure of the collaboration network, we calculated Group Degree Centrality (GDC) and Clustering Coefficient (CC) for each collaboration network. GDC measures the degree of centralization of the network in terms of the distribution of the number of the direct connections among actors. GDC reaches its maximum of 1 when one actor connects all other actors, and the other actors connect only to this one (star graph), while the index reaches its minimum of 0 when all degrees are equal. High GDC of the collaboration network would suggest that there are a few influential actors who have substantially more ties than the rest. We also calculated other centrality metrics such as betweenness centrality, but found best results using GDC.

The Clustering Coefficient [32] measures the degree of cohesiveness. This index reaches its maximum of 1 when any of two actors sharing one neighbor are connected, while it reaches its minimum value of 0 when any of two actors sharing one neighbor are disconnected. A high clustering coefficient suggests that the group of actors form a cohesive clique.

3.3 Are Articles Part of a Wiki Project?

As control variables we measure two factors that – in addition to the network structure – may also influence the "time-to-market" of articles. They make use of WikiProjects project management pages on Wikipedia that have emerged to organize editorial activity on Wikipedia[6]. This allows us to investigate whether structured project organization improves the success of an article.

The two factors we include are the importance of the article as rated by WikiProjects' members and the number of WikiProjects that cover a particular article. Each project rates the importance of an article on a 4-point scale from "Low" to "Top". Several WikiProjects may identify the same article as a topic of interest and focus of their efforts. To assess the effect of WikiProjects,

[3] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles

[4] http://en.wikipedia.org/wiki/Wikipedia:Good_articles

[5] http://en.wikipedia.org/wiki/User_talk:Yannismarou/Archive_4 #Roman-Spartan_War_2

[6] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Council/Guide

we collected the number of WikiProjects that had identified an article we analyzed as being of importance as well as the importance ratings they had assigned to this article. If there are differences between the importance ratings assigned to a particular article by different WikiProjects, we use the highest rating.

3.4 Dependent Variable: Time to Article Promotion

The metric used to assess the performance is *time to article promotion*, that is, the time between when an article got a previous promotion to when the article got promoted to one-step higher quality class. This metric measures how fast a group of editors can lift an article by one quality level. We measure two main promotion times, promotion from B-Class to GA and promotion from GA-Class to FA.

3.5 Survival Analysis

To assess the potential impact of pre-existing collaboration networks on the performance of teams working on articles, we test whether the pre-existing collaboration network pattern is correlated with the performance of Wikipedia editors for article improvement work in terms of the completion speed (or promotion rate). To assess the potential impact of pre-existing collaboration networks, we collected the featured (FA-Class) articles and good (GA-Class) articles for which we were able to identify the date of promotion to at GA-Class and B-Class. At first, we collected the editors working on an article from previous promotion to next promotion (period B in Figure 2). Then we constructed the pre-existing collaboration network among these editors of the article from 1 year before the previous promotion (GA-Class or B-Class) to that previous promotion date (period A in Figure 2.). This collaboration network represents the pre-existing social capital of the editors trying to get the article promoted to higher quality (FA-Class or GA-Class) during period B in Figure 2.

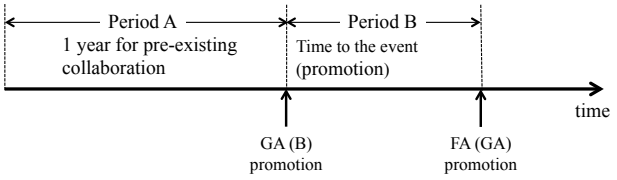


Figure 2. Explanation of pre-existing network construction

For our analysis we employ an event-history (or survival) model. We define the event time as promotion time from the previous level. The event time of a FA-Class article is the time from GA-Class article to FA-Class article, and the event time of a GA-Class article is the time from B-Class article to GA-Class article. This means that we cannot observe the survival time of the articles that are not yet promoted to FA-Class or GA-Class articles. We account for these articles by including them in the censored sample because the event (FA or GA-Class article promotion) did not (yet) happen before the termination of this study.

We use a hazard rate model of the likelihood of a promotion event at time t, conditional on it not having been completed earlier. To test the effect of pre-existing collaboration network variables to the promotion rate, we employ the Cox Proportional Hazard Model [9], which is written as

$$h(t|X) = h_0(t)exp\ (\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots)$$

where $h(t|X)$ represents promotion rate at t, and $h_0(t)$ is the baseline completion rate when all independent variables are set to zero, and the x's are the covariates.

We defined the pre-existing collaboration ratio, R_{pc}, as

$$R_{pc} = N_p \ / \ N_{all}$$

where N_p is the number of editors who contributed to the article in period B and were already collaborating in period A, and N_{all} is the total number of contributors to the article in period B. Editors who do not have any ties with other editors in period A are eliminated. Therefore, $R_{pc} = 1.0$ means that all editors working on the article in period B are connected in a pre-existing collaboration network in period A.

As independent variables, we measured the pre-existing collaboration network ratio R_{pc}, group degree centrality (GDC), and clustering coefficient (CC) as characteristics of a pre-existing collaboration network. We do not take into account the weight and directionality of ties for this analysis.

In exploratory observation of the data, we detected collaboration networks that carried on after the effective completion of an article. Figure 3 illustrates the editing activity of the article "War against Nabis," which was used in Figure 1 to illustrate our method of network construction. The number of edits on the article page fluctuated significantly, and then fell to a low level after mid-January 2007. But the Wikipedians who collaborated on this article continued to leave roughly the same number of messages on each other's user talk pages, even after activity on the "War against Nabis" article had fallen almost to zero. This suggests that the relationship between the editors continued even after work on this article was mostly completed. Some of this activity may be been connected work undertaken by editors after the article was complete but before completion of the article rating process which led to its promotion to FA.

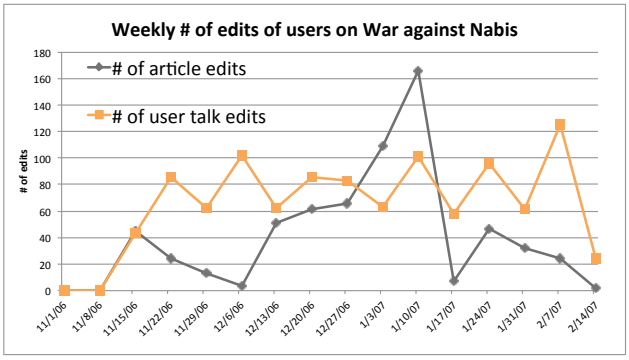


Figure 3. Editing activity of article "War against Nabis" from creation until it was a featured article

In a separate analysis described in section 5.2 we assess the potential impact of prior collaboration networks on the performance of teams working on new articles. We measured the strength of the prior collaboration network a group brought to its editorial work on a new article. This was done by segregating the group of Wikipedians who contributed to a FA-Class article into two sub-groups (figure 4). We collected editors who worked on the article between when the article was created

and when the article got FA-Class promotion. Prior user talk activities are collected from the beginning of the Wikipedia to the creation of the article.

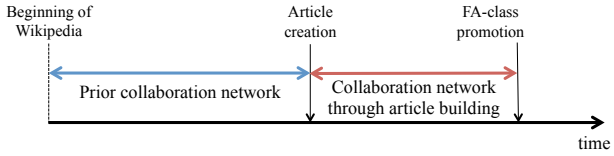


Figure 4. Explanation of team membership analysis. Priro ties through user-talk pages are taken as indication for prior collaboration between editors of an article.

4. HYPOTHESES

In earlier research, we have observed the emergence of collaborative innovation networks or COINs [13], where a small, tightly connected team of creators drives innovation. In larger networks, COINs are typically identifiable as cliques, fully connected subgraphs visible as clusters within the full network. We speculate that more closely connected collaborators on a Wikipedia article will perform better than less connected groups, and thus achieve featured article status more quickly. Motivated by ad hoc observations of COIN-like grouping in early, exploratory analysis of the data, we formulated the following hypotheses.

H1: The more connected the editors of an article are, as indicated by higher clustering coefficient of their collaboration network, the better they are capable of executing complex tasks, such as quality promotion from B to GA or GA to FA.

To test the role of pre-existing social capital, measured as prior collaboration on new articles, we formulate the second hypothesis.

H2: Teams having a high degree of pre-existing social capital — measured by a high proportion of team members being part of a pre-existing collaboration network and high cohesiveness of their pre-existing collaboration network — will get the articles they are working on faster to higher quality level.

5. RESULTS

Table 1 displays mean and (standard deviation) values of three categories of article quality promotions: Start to B, B to GA, GA to FA level. The elapsed time is the time between when an article was promoted to the previous quality level and when the article was promoted to the next higher quality level. In other words, elapsed time is the time it took the editors to lift an article by one quality level. The collaboration networks were constructed during each elapsed time.

We eliminated articles, which have less than one day for their promotion in order to ignore wrong assessments by mistake or vandalisms. We also eliminated articles, which have only one or two editors in their collaboration network, because calculation of the relevant network metrics calculation requires at least three network members

Using analysis of variance (or ANOVA) we found that all the variables in table 1 (elapsed time, number of editors, Group degree centrality (GDC), Cluster coefficient (CC)) of collaboration network during the article editing were significantly different between the three article promotion groups (Start to B, B to GA, GA to FA). GDC and CC of Start to B are lowest among those three categories of article quality promotions. One interesting observation is that GDC of GA-FA is smaller than that of B-GA, while CC of GA-FA is bigger than that of B-GA. Using Scheffe's method to account for multiple comparisons, we confirmed that any two GDCs out of three categories are significantly different. CC of GA-FA is significantly higher than that of Start-B and B-GA; however, CCs between Start-B and B-GA have insignificantly difference. This means that each of the three article categories has a different type of collaboration network, with the articles obtaining the highest level of perfection (GA-FA) having the highest CC.

According to the WikiProject article quality-grading scheme, the higher the class of the article, the more difficult and complex the article-improving task is.

This observation suggests that the collaboration network patterns among editors are associated with the difficulty and complexity of tasks they are working on. The most complex tasks, such as GA-FA promotion, have best embeddedness but less centralized structure than less complex tasks, such as B-GA promotion. Therefore, this result supports hypothesis 1.

Table 2 compares mean and (standard deviation) of promoted articles (B-GA or GA-FA) against not (yet) promoted articles (B-B or GA-GA). We again focus on promotion from B to GA, and from GA to FA level. In the case of non-promoted articles, the elapsed time measures the time from promotion to B, or GA respectively, to the end of the observation period (2010-12-01 00:00:00).

Looking at promotion to FA (GA-FA), we found that the collaboration networks of promoted sets of articles – the end product of successful collaboration – have a significantly more centralized and cohesive network than the not promoted articles. This result suggests that high-performance collaboration is associated with centralized and cohesive network structure, thus again supporting our hypothesis 1.

Looking at promotion from B-Class to GA-Class (B-GA), and comparing the non-promoted (B-B) samples, we found that the collaboration networks for the promoted articles had more centralized and cohesive structure (the cohesiveness finding was not statistically significant). This observation is consistent with the table 1 result, where the clustering coefficient of B-GA is not significantly larger than that of Start-B. This result suggests that B-GA quality improvement work is associated only with the centralized collaboration pattern; however, GA-FA quality improvement, the more complex task, requires more cohesive cliques as well.

Table 1. Mean and (standard deviation) of collaboration networks for articles moving up in quality levels. Elapsed time is the time between two promotions of an article, e.g. from Start to B, or from B to GA.

Article quality status	Start-B	B-GA	GA-FA	p (anova)
N	7175	2786	1379	
Elapsed time for promotion (days)	514.4676 (327.2266)	248.6111 (292.7814)	189.9306 (205.4879)	***
Number of editors	42.72 (60.908)	56.56 (119.1018)	41.76 (94.99766)	***
Group Degree Centrality (GDC)	0.3931 (0.2672)	0.4928 (0.3221)	0.4466 (0.2247)	***
Clustering Coefficient (CC)	0.2200 (0.2329)	0.2235 (0.2528)	0.3254 (0.2172)	***

*** $p < 0.001$

Table 2. Collaboration network metrics of promoted and not promoted articles

Article quality status	B-B	B-GA	p (anova)	GA-GA	GA-FA	p (anova)
N	39883	2786		7600	1379	
Elapsed time (days)	978.3565 (401.6305)	248.6111 (292.7814)	***	766.2037 (367.5333)	189.9306 (205.4879)	***
Number of editors	65.29 (102.867)	56.56 (119.1018)	***	51.43 (106.2683)	41.76 (94.99766)	**
Group Degree Centrality (GDC)	0.3755 (0.2310)	0.4928 (0.3221)	***	0.4043 (0.2250)	0.4466 (0.2247)	***
Clustering Coefficient (CC)	0.2269 (0.1957)	0.2235 (0.2528)		0.2817 (0.2159)	0.3254 (0.2172)	***

** $p < 0.01$, *** $p < 0.001$

Table 3. Descriptive statistics and correlation between variables for event-history model for FA-Class promotion from GA-Class (N=7900)

Variables	Mean	S.D.	Min.	Max	1	2	3	4
1. Survival time	665.509	414.399	3.2303	1672.5				
2. Pre-existing collaboration ratio R_{pc}	0.4253	0.2082	0.0417	1	-.57***			
3. Group Degree Centrality	0.3976	0.2713	0	1	-.18*	.13***		
4. Clustering coefficient	0.2400	0.2600	0	1	-.17***	.30***	-.17***	
5. Number of WikiProjects	1.164	1.2635	0	15	-.28***	.17***	.03*	.06***

* $p < 0.05$, *** $p < 0.001$

Table 4. Descriptive statistics and correlation between variables for event-history model for GA-Class promotion from B-Class (N=33033)

Variables	Mean	S.D.	Min.	Max	1	2	3	4
1. Survival time	918.75	447.06	1.0838	1680.6				
2. Pre-existing collaboration ratio, R_{pc}	0.2834	0.1688	0.0224	1	-.54***			
3. Group Degree Centrality	0.3741	0.2895	0	1	-.15***	.15***		
4. Clustering coefficient	0.1702	0.2304	0	1	-.09***	.23***	-.14***	
5. Number of WikiProjects	0.5121	1.1201	0	23	-.43***	.20***	.03***	.06***

*** $p < 0.001$

Tables 3 and 4 display the means, standard deviations, minimum, and maximum of the variables used in the survival analysis and correlations among these variables for the articles that had GA-Class promotion dates for FA-Class articles and B-Class promotion dates for GA-Class promotion, respectively. Survival time is the time between when an article got B-GA-Class promotion to when the article got GA-FA-Class promotion. For the articles, which were not promoted by the end of the observation period (censoring), survival time shows the time between B- respectively GA-Class promotion to the end of the observation. Pre-existing collaboration ratio, GDC, and CC represent network metrics of the pre-existing collaboration network, constructed from 1 year before the previous promotion to the GA respectively B promotion date, that is when a group of editors started GA-FA or B-GA work. The correlations between variables are weak except between survival time and pre-

existing collaboration ratio. This is not surprising because the more time it took for an article to get promotion, the more opportunities editors got to participate in editing that article. As a consequence, prior collaboration ratio decreases.

Table 5. Cox proportional hazard model

Dependent Variable	FA-Class promotion rate			GA-Class promotion rate		
N	7900			33033		
Number of events	1369			2503		
Pre-existing period	1 year			1 year		
	exp(coef.)	se(coef.)	p	exp(coef.)	se(coef.)	p
WikiProject Importance	1.273	0.0376	***	1.196	0.0292	***
Number of WikiProjects	0.683	0.0449	***	0.973	0.0295	
Pre-existing Collaboration network variables						
Pre-existing collaboration ratio, R_{pc}	2.532	0.0269	***	2.278	0.0138	***
Group Degree Centrality (GDC)	1.074	0.0314	*	1.165	0.0201	***
Clustering Coefficient (CC)	1.345	0.0256	***	1.113	0.0173	***

$^{*}p < 0.05$, $^{***}p < 0\ 001$

5.1 Survival Analysis Results

Table 6 lists the number of articles that we included in our survival analysis in each category of importance as determined by WikiProject members. For FA-Class promotion samples, the importance value was taken at the time the articles were of GA-Class level. For GA-Class promotion samples, the importance value was taken when the articles got B-Class level. This variable was incorporated into the Cox proportional hazard model shown in Table 5.

Table 6. Article importance determined by WikiProject.

WikiProject importance	Promoted and not promoted GA articles (N=7900)	Promoted and not promoted B articles (N=33033)
Top (4)	528	1690
High (3)	1093	2639
Mid (2)	1990	2964
Low (1)	1475	1810
N/A (0)	2814	23930

The Cox proportional hazard model illustrates the effect of the pre-existing collaboration network on work efficiency getting the article to the next-higher level. To compare coefficients among variables, we standardized independent variables with transforming mean to 0 and standard deviation as 1. Therefore, the coefficients (β) in Table 5 should be interpreted as the effect, where |β-1| represents the percentage increase or decrease in promotion rate associated with a one-standard-deviation increase in the independent variable.

The model predicts change in article quality (e.g. FA-Class promotion) based on pre-existing collaboration network variables, using WikiProject importance and number of Wikiprojects per article as control variables. 1369 articles out of 7900 GA-Class articles got promoted to FA-Class, and 2503 articles out of 33033 B-Class articles got promoted to GA-Class by the end of this study.

As expected, WikiProject importance increases the promotion rate of both FA-Class (27% increase) and GA-Class (20% increase) promotion, which suggests Wikipedians focus on the higher importance articles rather than lower importance articles in terms of the "time-to-market". On the other hand, number of WikiProjects reduces FA-Class promotion significantly (32% decrease). The number of WikiProjects may be associated with the range of topics the article covers, which means the article is dependent on several topics or disciplines. Therefore, the more different WikiProjects identified the article as of importance, the longer it took the editors to lift the quality to the highest level.

We found that pre-existing collaboration ratio R_{pc} significantly increases the promotion rate of both FA-Class and GA-Class promotions, by a factor of 2.5 (150% increase) for FA promotion, and by a factor of 2.3 (130% increase) for GA promotion with a one-standard-deviation increase. This result suggests that the more the editors collaborate before they started working together, the more likely the article will get promoted. This means that hypothesis 2 is true also: pre-existing social capital dramatically increases the productivity of Wikipedians in producing highest-quality work.

In addition, group degree centrality (GDC) and clustering coefficient (CC) of the pre-existing collaboration network also significantly increase the rate of both FA-Class and GA-Class promotions. For the FA-Class promotion rate, the coefficient of GDC (7.4% increase) is smaller than that of CC (35% increase), which means that the cohesiveness of a clique has more impact on the quality of work than the centralization of the network structure. This result supports hypothesis 2. On the other hand, for the GA-Class promotion rate, the coefficient of GDC (17% increase) is larger than that of CC (11% increase), which means that in this realm, the centralized network structure has a bigger impact than the cohesiveness of the clique. This result is consistent with the results in Table 1 and 2. Both in pre-existing (Period A in Figure 2) and article editing (Period B in Figure 2) collaboration network, the embeddedness of editors in collaboration network is correlated with high productivity of editors. On the other hand, both in the pre-existing and the article editing collaboration networks, a centralized network structure has more impact on the productivity of B-GA

promotion work than GA-FA promotion work. These results suggest that the form of collaboration patterns is associated with the complexity and difficulty of tasks editors work on, again confirming hypothesis 1.

5.2 Comparison with Prior Collaboration Networks

We did an additional analysis, comparing for each featured article the collaboration network from the beginning of Wikipedia with the collaboration network during the article creation process. This means that we looked at prior collaboration among article editors before they started work on a featured article. We compared this prior collaboration network with the collaboration network from article creation to FA promotion.

We distinguished between featured articles written by teams of editors who were collaborators before starting work on the article vs. articles written exclusively by editors who had not previously been communicating on each other's talk pages. In this analysis, we grouped the featured articles by the year they were created (2002, 2003, 2004, 2005, 2006, 2007), since we wanted to compare articles during time periods when the impact of collaboration networks was likely to be the same. We surmised that impact of collaboration networks in the early years of Wikipedia (e.g. 2002 and 2003) would likely be different from their impact in later years (e.g. 2006 and 2007), after Wikipedia had become a well-established institution.

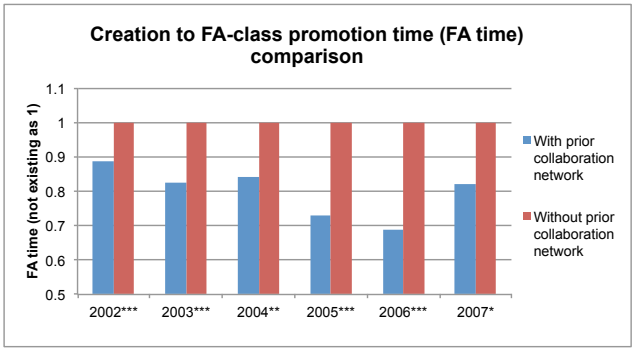


Figure 5. FA time (time between when an article was created to when the article got promoted) goes down if there is a prior user talk collaboration network.

(*p<0.1 ** p<0.05 ***p<0.01)

Figure 5 visually summarizes the findings from this analysis, which showed that the mean time to featured article (FA time) was significantly lower for articles where a prior collaboration network was in place. The prior collaboration ratio—the percentage of editors in the pre-existing collaboration networks as a share of the overall editorial team is strongly negatively correlated with performance. The more a group of editors have prior collaboration ties, the faster the article they are working on reaches featured article status. This again confirms hypothesis 2.

6. DISCUSSION

We also undertook cluster analysis to identify groups who collaborated on multiple articles, using the bicomponent cluster algorithm in JUNG based on [25]. A separate network was created for this clustering analysis, with a tie between two authors assumed if they had collaborated on at least five articles. This cluster analysis allowed us to identify teams that had collaborated on multiple articles.

We employed this cluster-detection algorithm to find groups among the editors. Our clustering algorithm identified three clusters. Figure 6 displays the network of the 136 editors who collaborated on at least 5 articles.

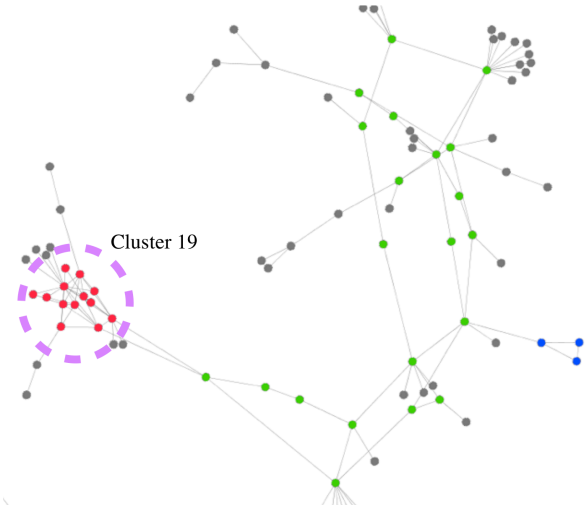


Figure 6 pre-existing collaboration networks among editors, editors subsequently collaborating on at least 5 articles

Highlighted in Figure 6 is cluster 19, which includes 14 editors. When we looked at the activities of this group, we found that it had contributed to 31 articles about hurricanes. This illustrates that groups of editors build lasting networks to collaborate on multiple articles. Seven members of cluster 19 were in a prior collaboration network even before the formation of this cluster.

To test the quality of the collaboration by these 14 editors, we prepared the equivalent set of featured articles which have similar creation date, number of editors, and number of edits. We examined the mean FA time between 2 sets using Welch's t-test. We found that the mean time of these 2 sets of articles is significantly different (N=31, $p < 0.05$). On average, 31 articles done by cluster 19 took 822 days to be featured status while the other set of articles took 1077 days. The group of editors working on several articles together who is also well connected in their collaboration network makes an article reach featured level faster.

Our analysis shows that pre-existing social capital indeed reduces "time-to-market" of articles. We examined the working of informal collaboration networks in Wikipedia, inferring links if editors of the same article comment on the user talk pages of other editors who are working on that article. Including the analysis of prior collaboration networks allows for new insights into what Kittur & Kraut [16] call "implicit coordination". Our results suggest that the higher the connectivity among editors, measured as density of the collaboration network, and the more centralized the team working on the article, the faster the article will reach featured status. We found that implicit coordination works best if the core group of editors of an article is well

connected, and is embedded in a dense network of more peripheral collaborators.

Furthermore, Wikipedia articles that are most rapidly promoted appear to be created by groups of editors who have previously worked together on other articles. The social capital they have built-up by collaborating before seems especially important in the early phases of article definition and team organization. Once the general direction of the article is set, the team then appears able to absorb new contributors effectively.

This structure is similar to the "onion model" that has emerged in open source software, where a group of core developers are responsible for setting overall direction and goals, and undertaking the lion's share of new development, while at the same time benefitting from smaller contributions that other members of the community provide—some development tasks, bug fixes, and user input to indentify bugs and inform future development.

7. LIMITATIONS, FURTHER WORK AND CONCLUSIONS

As a next step we intend to further examine collaboration network evolution over time and to focus on collaboration networks that extend across multiple articles.

An open question is how broadly our results apply outside of Wikipedia. Our study has shown that Wikipedia has an implicit social organization of its own, comprised of networks of collaborators who work together closely. It might be that the most active Wikipedians operate under an implicit set of rules that have evolved within the Wikipedia community and that their practices cannot be generalized to other open source platforms or to traditional organizations.

Nevertheless, we believe that our results give first indications of the role of social capital for teams in organizations where members are collaborating virtually without much face-to-face contact. In the same way that social network surveys made visible the importance of the informal organization within large corporations, so might analysis of collaboration networks on Wikipedia provide first steps towards making the role of social capital in organizations measurable: social capital indeed seems to increase organizational efficiency.

8. ACKNOWLEDGEMENTS

We thank Stephanie Woerner for her excellent suggestions for improvement based on an earlier draft of this paper.

9. REFERENCES

[1] Aberdour, M. "Achieving Quality in Open Source Software," IEEE Software, January-February 2007.

[2] Adler, B. T., and Alfaro L. de. "A content-driven reputation system for the Wikipedia." Proceedings of the 16th International Conference on the World Wide Web, 2007, pp. 261–270.

[3] Ahuja, M. K. Carley, K.M. "Network Structure in Virtual Organizations," Organization Science. 1999 (10:6). (pp. 741-757)

[4] Anthony, D., Smith, S. W., and Williamson, T. "Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of Wikipedia," Working Paper, Department of Computer Science, Dartmouth College, November 2005.

[5] Aral, S., Brynjolfsson, E. Van Alstyne, M. W., Information, Technology and Information Worker Productivity: Task Level Evidence (June 2007). NBER Working Paper Series, Vol. w13172, pp. -, 2007. Available at SSRN: http://ssrn.com/abstract=993403

[6] Benkler, Yochai. The Wealth of Networks: How Social Production Transforms Markets and Freedom, Yale University Press, New Haven and London, 2007.

[7] Bourdieu, P. The forms of capital. In J. Richardson (Ed.) Handbook of Theory and Research for the Sociology of Education (New York, Greenwood), 241-258. 1986.

[8] Chi, E.H., Kittur, N., Pendleton, B., and Suh, B. "Long Tail of user participation in Wikipedia," http://asc-parc.blogspot.com/2007/05/long-tail-and-power-law-graphs-of-user.html, 2007.

[9] Cox, D. R. "Regression Models and Life Tables". Journal of the Royal Statistical Society Series B 34 (2): 187–220. 1972

[10] Crandall, D. Cosley, D. Huttenlocher, D. Kleinberg, J. Suri, S. Feedback Effects between Similarity and Social Influence in Online Communities. ACM KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

[11] Crowston, K., and Scozzi, B. "Coordination practices for bug fixing within FLOSS development teams." Proceedings of the First International Workshop on Computer Supported Activity Coordination (CSAC), 2004, Portugal.

[12] Cummings, J. N. Cross, R. "Structural properties of work groups and their consequences for performance." Social Networks, (25:3), 2003, pp.197-210.

[13] Gloor, P. Swarm Creativity: Competitive Advantage Through Collaborative Innovation Networks. Oxford University Press, New York, 2005.

[14] Howe, J. "The Rise of Crowdsourcing," Wired (14:06), June 2006.

[15] Kidane, Y., and Gloor, P. "Correlating temporal communication patterns of the Eclipse open source community with performance and creativity." Computational & Mathematical Organization Theory. (13:1) 17 - 27, 2007.

[16] Kittur, A., and Kraut, R. E. "Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination." In Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work—CSCW '08. New York: ACM Press, 2008.

[17] Krackhardt, D. Hanson, J.R., "Informal Networks: The Company Behind the Chart," Harvard Business Review, July-August 1993.

[18] Lakhani, K. R., and von Hippel, E. “How open source software works: "Free" user-to-user assistance.” Research Policy, (32:6), 2003, pp. 923-943.

[19] Liu, J. Ram, S. “Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Data Quality” (December 8, 2009). 19th Workshop on Information Technologies and Systems, pp. 175-180, December 2009. Available at SSRN http://ssrn.com/abstract=1565682. Published in Proceedings of 19th Workshop on Information Technologies and Systems, Phoenix, December 2009, pp. 175-180.

[20] Malone, T. Laubacher, R. Dellarocas, C. “The Collective Intelligence Genome.” Sloan Management Review (51:3) Spring, 2010.

[21] Porter, K.A. Bunker Whittington, K.C. Powell, W.W. “The institutional embeddedness of high-tech regions: Relational foundations of the Boston biotechnology community”. S. Breschi & F. Malerba (Eds.), Clusters, Networks, and Innovation. Oxford, UK: Oxford University Press, 2005.(pp. 261-296)

[22] Priedhorsky, R. Chen, J. K. Lam, S. K. Panciera, K. Terveen, L. Riedl, J. Creating, Destroying, and Restoring Value in Wikipedia. ACM GROUP’07, November 4–7, 2007, Sanibel Island, Florida, USA. 2007.

[23] Putnam, R.D. Bowling Alone: America's Declining Social Capital, Journal of Democracy - Volume 6, Number 1, January 1995, pp. 65-78.

[24] on Visual Analytics Science and Technology Sacramento: IEEE, 2007. (pp. 163-170)

[25] Tarjan, E.R. “Depth first search and linear graph algorithms,” SIAM J. Computing (1:2) 1972 (pp.146-160)

[26] Uzzi, B. Spiro, J. “Collaboration and Creativity: The Small World Problem.” American Journal of Sociology. (111:2), September 2005, pp. 447–504.

[27] Van Der Gaag, M. Tom A.B. Snijders, T.B.A. The Resource Generator: social capital quantification with concrete items. Social Networks 27, 2005. 1–29.

[28] von Hippel, E. “Innovation by User Communities: Learning From Open-Source Software,” Sloan Management Review, July 2001.

[29] WikiProject Council/Guide/WikiProject: Assessment, 2010. http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Council/Guide/WikiProject#Assessments_in_practice.

[30] Wilkinson, D, and Huberman, B. “Cooperation and Quality in Wikipedia,” WikiSym’07, October 21–23, 2007, Montreal, Canada

[31] Worldbank. Measuring Social Capital. http://go.worldbank.org/A77F30UIX0 (retrieved Jan 16, 2011)

[32] Watts, D.J., and S.H. Strogatz, 1998: Collective dynamics of ‘small-world’ networks. Nature, 393, 440-4