6-2011

Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Article Quality

Jun Liu

Sudha Ram

Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Article Quality

JUN LIU and SUDHA RAM, University of Arizona

The quality of Wikipedia articles is debatable. On the one hand, existing research indicates that not only are people willing to contribute articles but the quality of these articles is close to that found in conventional encyclopedias. On the other hand, the public has never stopped criticizing the quality of Wikipedia articles, and critics never have trouble finding low-quality Wikipedia articles. Why do Wikipedia articles vary widely in quality? We investigate the relationship between collaboration and Wikipedia article quality. We show that the quality of Wikipedia articles is not only dependent on the different types of contributors but also on how they collaborate. Based on an empirical study, we classify contributors based on their roles in editing individual Wikipedia articles. We identify various patterns of collaboration based on the provenance or, more specifically, who does what to Wikipedia articles. Our research helps identify collaboration patterns that are preferable or detrimental for article quality, thus providing insights for designing tools and mechanisms to improve the quality of Wikipedia articles.

Categories and Subject Descriptors: H.m [**Information Systems**]: Miscellaneous

General Terms: Design, Management

Additional Key Words and Phrases: Wikipedia, collaboration pattern, article quality

ACM Reference Format:
Liu, J. and Ram, S. 2011. Who does what: Collaboration patterns in the Wikipedia and their impact on article quality. ACM Trans. Manag. Inform. Syst. 2, 2, Article 11 (June 2011), 23 pages.
DOI = 10.1145/1985347.1985352 http://doi.acm.org/10.1145/1985347.1985352

1. INTRODUCTION

Wikipedia is one of the most heralded success stories of peer collaboration. Free distribution, constant updates, and broad and diverse coverage have made it one of the ten most visited Web sites on the Internet [Dondio and Barrett 2007]. Wikipedia has been cited increasingly more often in the press as a source on historical facts and figures [Lih 2004]. It has even been cited in court cases [Cohen 2007]. The ubiquity of Wikipedia make the quality of Wikipedia a critical issue since Wikipedia can "act as a megaphone, amplifying the (sometimes incorrect) conventional wisdom" [Rosenzweig 2006]. False and incorrect information can easily be propagated to millions of potential readers world-wide.

Is Wikipedia indeed a reliable source of information? Although the everyone-can-edit idea of Wikipedia seems "bizarre" [Stvilia et al. 2005] and sounds like "a recipe for chaos" [Louridas 2006], researchers have found that the quality of Wikipedia articles

This work was supported in part by research grants from the National Science Foundation (IIS-0455993), the National Science Foundation Cyberinfrastructure Program (#EF-0735191), and the Science Foundation of Arizona.

Authors' addresses: J. Liu (corresponding author), S. Ram, MIS Department, University of Arizona, Tucson, AZ; email: jliu@email.arizona.edu.

DOI 10.1145/1985347.1985352 http://doi.acm.org/10.1145/1985347.1985352

to be surprisingly good. A much discussed article from *Nature* [Giles 2005] compares Wikipedia with the Britannica Encyclopedia and argues that, despite its anarchical function, the former comes close to the latter in terms of the accuracy of its science entries. The surprisingly high quality of Wikipedia has spurred supporters to hail it as a victory of the "wisdom of the crowds" [Kittur and Kraut 2008]. Critics of Wikipedia, on the other hand, have never stopped attacking it since "no one officially stands behind the authenticity and accuracy of any information in Wikipedia" [Denning et al. 2005]. Indeed, there is sufficient evidence that makes it entirely reasonable to question the quality of Wikipedia. For instance, Wikipedia is assigning quality grades including Featured Articles, A-class, Good Articles, B-class, C-class, etc., to its articles. As of March 2010, only 2,777 out of a total of 2,994,903 articles on the English Wikipedia are slated to be featured articles, articles that are "professional, outstanding, and thorough" [Wikipedia 2010]. It is therefore unreasonable to simply assume that Wikipedia is a completely reliable or unreliable. Wikipedia articles vary widely in quality. Many prior studies such as Lih [2004], McGuinness et al. [2006], and Wilkinson and Huberman [2007] have proposed approaches to distinguishing high-quality articles from the unreliable ones. Labeling articles with a quality grade undoubtedly helps make their readers aware of low- or high-quality content. Nevertheless, more and more people are using Wikipedia as an information source despite being aware of its unreliability [Luyt et al. 2008]. Hence, the focus of research on Wikipedia quality should be shifted to understanding why Wikipedia articles are different in quality and working toward a more sophisticated solution to enhance Wikipedia article quality, since it is increasingly becoming a major intellectual influence on many of its users.

Our research is therefore motivated by two questions: (1) Why are some Wikipedia articles of high quality while others are not; and (2) how we can improve the quality of Wikipedia articles? Several studies have attempted to find answers to the *why* question. Wikipedia relies on the open-source model [Hendry et al. 2006]. It is thus tempting to draw parallels between Wikipedia and Open-Source Software (OSS) projects. Researchers have attributed the success of many OOS projects to a balance between "centralization" and "decentralization" [Gacek and Arief 2004]. Similarly, studies including Kittur and Kraut [2008] and Ortega et al. [2008] proved that high-quality Wikipedia articles rely on centralization, that is, the edits are concentrated within a small group of core contributors, while researchers such as Lih [2004] and Wilkinson and Huberman [2007] have proposed decentralization measures including "rigor" (total number of edits made for the article) and "diversity" (total number of unique editors for the article) as Wikipedia article quality indicators since "given enough eyeballs all bugs are shallow" [Lih 2004]. A serious problem with these prior studies is the narrow focus on "easily accessible aggregate data" [Kane and Fichman 2009] such as number of edits. The fact that Wikipedia is easy to edit does not mean that all contributors edit the same way, or with the same intensity. In a single edit, a contributor can insert a number of sentences or just change a single word. Aware or not, most previous studies treated the development of content on Wikipedia as a collaboration by a group of people making homogeneous contributions. Almost none of them has delved deep into the unique and often implicit collaborative processes behind the development of Wikipedia articles. A lot of critical knowledge about contributors and their collaboration is consequently lost in this simplification of collaboration in Wikipedia. As a result, while these previous studies have begun shedding light on the *why* question, that is, why Wikipedia articles vary widely in quality, they are not informative about quality development of Wikipedia articles, in particular, the effects of collaboration on Wikipedia article quality. Without a deeper understanding of collaboration in Wikipedia, they have left the question of *how* we can improve the quality of Wikipedia articles largely unanswered.

Our research investigates what different roles a contributor can play for a given Wikipedia article, how contributors assuming different roles collaborate, and what the relationship is between the collaboration and Wikipedia article quality. Here, collaboration is defined as "the process of shared creation" [Schrage 1990]. Collaboration behind the development of Wikipedia consists of various actions performed by different contributors. There are existing studies that identify user roles in Wikipedia. For instance, Anthony et al. [2009] showed that high-quality content in Wikipedia comes from users playing two different roles, that is, *zealots*, registered users with a strong interest in reputation and high level of participation and *good Samaritans*, unregistered, anonymous, and occasional contributors. The authors classified the contributors based on the assumption that the age or "survival ratio" [Adler and Alfaro 2007] of a contribution indicates the quality of the contribution. Luyt et al. [2008], however, questioned the assumption by confirming the first-mover effect whereby material added by early edits tends to stay longer. Our research is similar to Anthony et al. [2009] in that we believe that the quality of a Wikipedia article depends to a large extent on who the contributors are. However, we do not attempt to distinguish trustworthy contributors from unreliable ones since existing studies such as Luyt et al. [2008] suggested that it is impossible to automatically do so based on the contributors' previous contributions. Instead, we identify the different roles a contributor can play for a specific Wikipedia article based on her actions. Traditional research on collaborative writing [Fitzgerald 1987; Ede and Lunsford 2001; Bracewell and Witte 2003] has proved that different types of authors often display various patterns of revision. For instance, expert professional writers often first make more meaning-related revisions and then change the style or other surface features, while less competent writers focused primarily on surface changes [Fitzgerald 1987], implying that the pattern of revision and the quality of writing are inherently related [Jones 2008]. Our research shows that although the unique affordances of Wikipedia (e.g., virtually no barriers to entry) allow a large number of volunteer contributors to contribute in a seemingly anarchic way, the contributors still play various roles in collaboration, displaying different patterns of actions. Our study also extends the prior studies that measured the concentration of edits within the distribution of contributors. Given an article, we investigate whether a type of action (e.g., sentence insertions) was concentrated within a group of contributors assuming a specific role. Based on this, we uncover implicit collaboration patterns, each of which represents a unique way of collaboration in Wikipedia. Using statistical tests, we demonstrate that the various collaboration patterns have different impacts on article quality. Identifying collaboration patterns that are preferable or detrimental for Wikipedia article quality helps provide insights into how we can devise tools and mechanisms to improve the quality of Wikipedia articles.

The article proceeds as follows. In Section 2, we provide an overview of our research. We describe the provenance of Wikipedia articles, in particular the various actions that can be performed by contributors on each article, in Section 3. Using the provenance, we then present the major findings of our research, including the roles, collaboration patterns, and the relationship between collaboration patterns and article quality in Sections 4, 5, and 6. In Section 7, we further validate our findings. This is followed in Section 8 by a discussion on the implications of this study to enhance the quality of Wikipedia articles. We conclude our article in Section 9 with a discussion of future research.

2. RESEARCH OVERVIEW AND DATA COLLECTION

Figure 1 shows the theoretical framework that forms the foundation for this study. The framework we propose is based upon the input-process-output model [McGrath 1984] that has been used widely in research on collaboration in both traditional groups

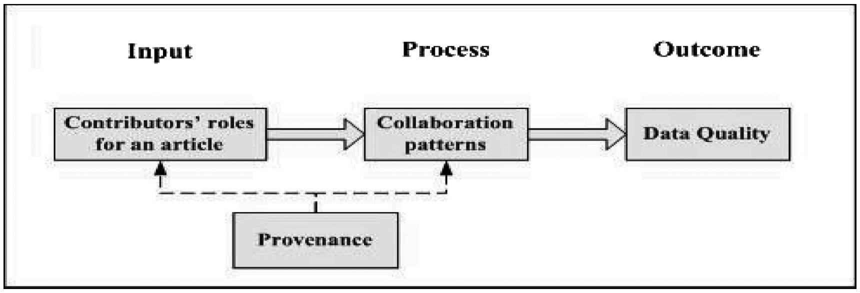


Fig. 1. Overview of our research.

[Littlepage et al. 1995] and virtual teams [Pinsonneault and Caya 2005]. The fundamental logic underlying this framework is that team members playing different roles (the input) produce different collaboration patterns in the process of developing a Wikipedia article (the process), which in turn has an impact on the outcome of the group collaboration, that is, the quality of the article in our case (the output). In our research, we first identify various roles played by contributors for a given article. The role of a contributor may vary from one article to another since a contributor has different levels of expertise on different topics [Anthony et al. 2009], which makes our research different from existing research such as Anthony et al. [2009] and Stvilia et al. [2008] that investigated the type of individual contributors in the whole Wikipedia community. The arrow from "contributor's role for an article" to "collaboration patterns" shown in Figure 1 does not represent a causal relationship. It indicates that we used the contributors' roles as inputs to uncover a number of collaboration patterns. We identify the roles of contributors and collaboration patterns based on the provenance (defined in the next section) of each Wikipedia article. We then examine the quality of the articles to determine the impact of collaboration patterns on quality of the Wikipedia articles. In the rest of this article, we discuss each of the components of our research in more detail. Before that, we describe the sample dataset of articles we used to derive the roles and collaboration patterns.

We took advantage of Wikipedia's article assessment project, which has organized the evaluation over 900,000 articles into various grades of quality. These quality ratings range from lowest to highest and are termed Stub, Start, C-class, B-class, Good Articles (GA), A-class, and Featured Articles (FA). Wikipedia provides formal guidelines for assessing the quality of Wikipedia articles. Consistent with Wang and Strong's multidimensional definition of data quality [Wang and Strong 1996], the Wikipedia community views the quality of its articles as multidimensional. For instance, the Featured Article quality assessment criteria include: (1) well-written; (2) comprehensive; (3) well-researched and verifiable by including references; (4) neutral; (5) stable, not changing often; (6) compliance with Wikipedia style guidelines; (7) having appropriate images with acceptable copyright status; and (8) having appropriate length and focusing on the main topic. Obviously, measurement of these dimensions such as "well-written," "comprehensive," or "well-researched" relies on the subjective perception of individuals. As a result, it is difficult or even impossible to develop article quality measures that are perfectly objective and neutral. To increase the objectivity and neutrality of the quality assessment, Wikipedia has developed different mechanisms that center around two major themes: (1) relying on the consensus of a large number of reviewers, and (2) constraining the influence of the significant contributors to the article. Two levels, GA and FA, are assessments made "externally" by those who are not the significant contributors to the article. Once an article has been nominated and posted on the FA candidate page, all editors can review the article, choose to support or oppose a nomination according to the FA quality criteria, and provide their arguments. A contributor to the article is allowed to give her opinions, but she must indicate if she has been a significant contributor to the article. For a nomination to be promoted to FA status, consensus among reviewers and nominators must be reached that it meets the FA criteria. The FA director "Raul654" or one of his three delegates determine whether there is consensus and have the final say. Similarly, a GA candidate can be promoted to the GA status after it has been reviewed and approved by reviewers based on the GA quality criteria. While any registered editor can nominate an article as a GA candidate, a GA candidate cannot be reviewed by a user who is the nominator or who has made significant contributions to the article. The quality assessments of articles for the other levels (e.g., B-class or C-class articles) are performed by members of WikiProjects. A WikiProject manages a specific topic or family of topics within Wikipedia. It is composed of a collection of articles and a number of editors who collaborate on these articles. To assign a quality grade to an article, a project needs to reach a consensus. Contributors who contribute a lot of content to an article are normally excluded from the assessment of the article. An existing study [Kittur and Kraut 2008] has tested the validity of the Wikipedia quality ratings by requesting external raters to rate the quality of a number of articles. The article ratings from external raters and the Wikipedia quality ratings are significantly correlated (Spearman' rho = .54, $p < .001$). Hence, we believe that the Wikipedia quality ratings, though not guaranteed to be completely objective and neutral, are critical indicators of Wikipedia article quality. With the intention to study the relationship between collaboration and the quality of Wikipedia articles, we selected an equal number of articles of different Wikipedia designated quality levels.

As of March 2010, of the over 3 million Wikipedia articles, 2,777 articles are categorized as featured articles (FA) by Wikipedia's quality assessment teams. An additional 8,247 are listed as good articles (GA). There are also 73,226 B-class articles and 55,021 C-class articles. We collected a sample of 1600 articles including 400 featured articles, 400 good articles, 400 B-class, articles, and 400 C-class articles from the English Wikipedia in March 2010. We did not select articles with the "stub" or "start" status since it is not that meaningful to study the collaboration in the articles that are stubs or were just created. The Wikipedia quality grades also include A-class representing a transitional status between good articles and featured articles. However, only 647 A-class articles appear on Wikipedia, and they are concentrated in a few domains such as military history. Hence, we did not include A-class articles in our sample. We recognize that specific topic areas or domains may impact articles' quality. For example, empirical studies that compared Wikipedia with other encyclopedias found that Wikipedia is quite reliable when it comes to science topics [Giles 2005] while history articles in Wikipedia were found to be less accurate than those in other encyclopedias [Rector 2008]. To control for the effects of topic areas, we collected articles from various Wikipedia topics using a stratified sampling approach. We started our data sampling with featured articles. The English Wikipedia has classified a total of 2,777 featured articles into 31 mutually exclusive categories including biology, law, politics and government, etc. The number of featured articles we randomly selected from each category is in rough proportion to the number of featured articles in the category. As an example, we randomly selected 6 featured articles from the domain of law since 39 of the 2,777 feature articles are law articles. We then randomly selected the same number of good articles (GA), B-class, and C-class articles from the same domain. If a featured article we sampled belonged to a WikiProject, we randomly selected a GA, a B-class, and a C-class article from the same project. If it was not part of a specific project, we identified a WikiProject in the domain and then randomly selected a GA, a B-class, and a C-class article from the project. In this way, we ensured that the quality assessments of the

Table I. Definition of Actions that Affect a Wikipedia Article

Type of actions	Explanation
Sentence insertion	Insertion of a sentence
Sentence modification	Modification or rewording of an existing sentence
Sentence deletion	Deletion of a sentence
Link insertion	Linking of a word within an existing sentence to an article (a link to another Wikipedia article or to external Internet articles)
Link modification	Modification of an existing link (can be a change of the URL or the name of the link)
Link deletion	Deletion of an existing link
Reference insertion	Adding a reference or creation of an inline citation
Reference modification	Modification of an existing reference
Reference deletion	Deletion of a reference
Revert	Reverting an article to a former version

sampled articles represent the consensus of different WikiProjects. For the purposes of tracking specific actions a contributor performed on an article, we extracted the various versions of each article. Since we attempted to study the impact of collaboration on article quality, we collected only the versions of an article from its creation to the time point at which the article was assigned its current quality grade. We also noticed that when editing an article, contributors often saved intermediate results, thus performing multiple consecutive edits. Hence, before processing the versions, we filtered them, keeping only the last of consecutive versions by the same contributor. We also did not include the version from the edits that were later reverted.

3. DATA PROVENANCE OF WIKIPEDIA ARTICLES

Data provenance refers to the source and processing history of data. We tracked and used the provenance of Wikipedia articles. In our previous research [Ram and Liu 2007], we defined the concept of provenance using the W7 model. We employ a subset of this model by tracking *who* does *what*, that is, every action performed by each contributor that affects the life of a Wikipedia article from its creation to the present time. We built on Pfeil et al.'s categorization [Pfeil et al. 2006] and developed our classification of actions. Pfeil et al. [2006] defined the following categories: add information, clarify information, delete information, add link, fix link, delete link, format, grammar, mark-up language, style/typography, spelling, reversion, and vandalism. While this detailed categorization is useful for understanding the Wikipedia editing process, it was impossible to automatically identify some of the actions. For example, it is difficult to determine that a specific action entails "clarify information." Hence, we use a high-level category called "sentence modification" to represent the clarification, grammar, and spelling change that affects a sentence. We added three categories including reference creation, modification, and deletion. These reference-related actions can be automatically identified and are crucial for Wikipedia article quality. Table I summarizes the various actions that can affect the life of a Wikipedia article from its creation to the present time. Our categorization of actions is similar to the one described in Arazy et al. [2010]. "Sentence creation" in our categorization corresponds to "add" in Arazy et al. [2010], "sentence modification" to "proofread," and "sentence deletion" to "delete." A major difference between our categorization and Arazy et al.'s is that we added "reference insertion/modification/deletion" as separate categories in addition to actions related to links because having a sufficient number of references is an important quality indicator. One of Wikipedia quality criteria states that an article needs to be "supported by inline citations where appropriate." Moreover, adding references and adding links appear to have different knowledge requirements. As will later be discussed, we found that references or inline citations are usually added by the contributors who inserted the sentence while other contributors often did not bother or were unable to add references. Links are different. Contributors can add links to a sentence even when they are unaware of the source of the sentence. To track and harvest the various actions performed by each contributor on article, we compared different versions of the article based an algorithm proposed in Adler and Alfaro [2007]. Our research goes one step further than Pfeil et al. [2006] and Ehmann et al. [2008] and takes the size of a contribution into consideration by determining the actions at the granularity of sentences. Inserting a large number of sentences or just one sentence will obviously have different impacts on a Wikipedia article, though both of them would be categorized as “add information” according to Pfeil et al. [2006]. In our approach, each contribution or edit made by a contributor may include not only different types of actions but a number of actions of the same type (e.g., a contributor can perform five sentence insertions and three sentence modifications in a single edit).

4. IDENTIFICATION OF CONTRIBUTOR ROLES IN WIKIPEDIA

Identifying the roles played by each contributor helps us understand the sources of quality variance in Wikipedia. We classify Wikipedia contributors based on clustering their actions performed on specific article. To do this, we employ the K-means clustering technique.

(1) Inputs to clustering:

If we use $P = \{p_1, p_2, \ldots, p_n\}$ to represent a set of Wikipedia articles and $E_i = \{e_{i1}, e_{i2}, \ldots, e_{im}\}$, to represent a set of contributors who have contributed to an article $p_i \in P$, then we model the actions performed by a contributor e_{ij} on the article p_i as a vector

$$\overrightarrow{act_{e_{ij},p_i}} = \langle a^1_{e_{ij},p_i} / a^T_{e_{ij},p_i}, a^2_{e_{ij},p_i} / a^T_{e_{ij},p_i}, \ldots, a^{10}_{e_{ij},p_i} / a^T_{e_{ij},p_i} \rangle, \tag{1}$$

where $a^1_{e_{ij},p_i}, a^2_{e_{ij},p_i}, \ldots, a^{10}_{e_{ij},p_i}$ represent the number of each of the 10 types of actions (see Table I) performed by the contributor e_{ij} to a given article p_i, and $a^T_{e_{ij},p_i}$ represents the total number of actions performed by the contributor to the article. We counted the numbers of different types of actions performed by a contributor on a specific page and then normalized them by dividing them by the total number of actions performed by the contributor on the page. The inputs to clustering thus include $m \times n$ such vectors, each of which includes the normalized number of different types of actions performed by a contributor on a specific page. Although we do not attempt to determine the quality of each action since it seems impossible to automatically do so according to Luyt et al. [2008], we do take the survival of the actions into consideration. We do not count actions that were reverted or deleted within the next five edits.

The 1600 sample articles contain a total of 824,738 such vectors. We also noticed that 82.74% of contributors had less than 4 actions for a given article. We categorized these contributors as *casual contributors* for the article and did not include them in clustering. As a result, the data used for clustering include 142,329 vectors.

(2) Repeated K-means algorithm:

Since we intended to identify mutually exclusive roles played by contributors for a specific article, we used the widely used K-means algorithm as the base method to partition the input vectors. A well-known disadvantage of K-means is that it requires the number of clusters, *k*, to be specified a priori. To address this problem, we applied the K-means method repeatedly using *k* values ranging from 2 to 10. Here, we set 10 as the maximum k value to avoid a trivial classification of roles. For each *k* value, we first evaluated the quality of the clustering results using evaluation functions proposed in He et al. [2004], that is, cluster compactness (*Cmp*), cluster separation (*Sep*), and combined measure of overall cluster quality (*Ocq*), to evaluate both the intracluster

Table II. Summary of Contributors

Cluster Number	Size	Description of actions by contributors	Role Label
1	31,236 (21.9%)	Engaging in many types of actions including sentence creations, modifications, and deletions and link and reference creations, modifications and deletions.	All-round Contributors
2	6,574 (4.6%)	Focusing on reverts.	Watchdogs
3	13,062 (9.2%)	Focusing on sentence creations and seldom engaging in other actions.	Starters
4	40,717 (28.6%)	Focusing on three types of actions: sentence creations, link creations and reference creations.	Content Justifiers
5	36,703 (25.8%)	Focusing on sentence modifications	Copy Editors
6	14,037 (9.9%)	Focusing on removing sentences, references and links	Cleaners

homogeneity and intercluster separation of the clustering result. The definitions of these functions are given next.

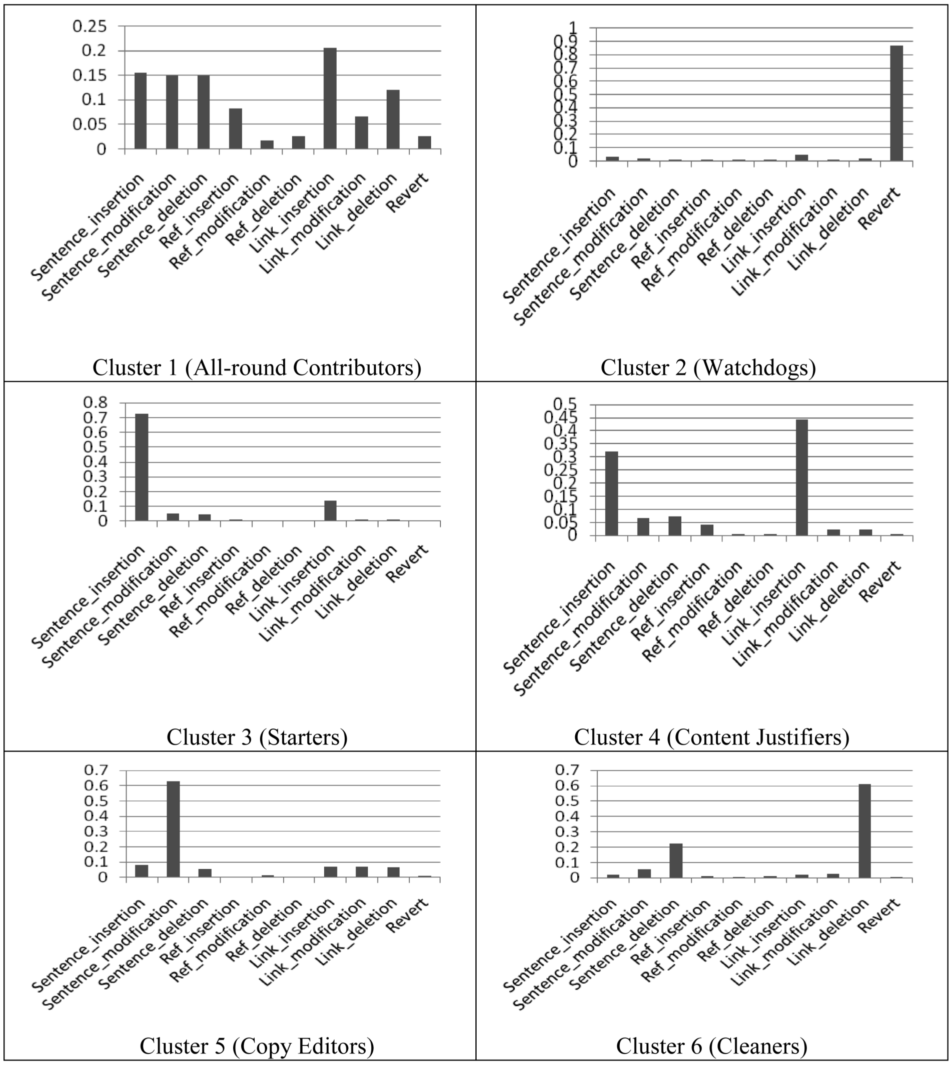
$Cmp = \frac{1}{c}\sum_{i}^{c}\frac{v(c_i)}{v(X)}$, where C is the number of clusters generated on the dataset X, $v(c_i)$ is the deviation of the cluster c_i and $v(X)$ is the deviation of the dataset X. $v(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d^2(x_i, x)}$, where $d()$ is a distance measure between two vectors, N is the number of members in X, and $\bar{x}$ is the mean of X.

$Sep = \frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c} \exp\left(-\frac{d^2(x_{c_i}, x_{c_j})}{2\sigma^2}\right)$, where C is the number of clusters, σ is a Gaussian constant, x_{c_i} is the centroid of the cluster c_i, and $d(x_{c_i}, x_{c_j})$ is the distance between the centroid of c_i and the centroid of c_j.

$Ocq = 0.5 \times Cmp + 0.5 \times Sep$. The lower the Ocq value, the better quality of the overall output clusters.

In our study, 6 was the optimal number of clusters generated from the dataset because Ocq had the lowest value at k = 6. Table II shows a summary of the six clusters that were generated. Each cluster includes a number of vectors, each of which represents a contributor's actions on a specific article. Figure 2 represents the centroid of each of the six clusters, which conspicuously indicates that contributors belonging to different clusters play different roles for a given article by focusing on certain types of actions. For instance, on average, 86.60% of actions performed by contributors that belong to cluster 2 are reverts. 72.19% of actions performed by those belong to cluster 3 are sentence insertions. There are also contributors, such as those belonging to cluster 1, who are more all-round contributors and performed various actions.

We assigned a role label to each of these clusters to designate the role played by the contributors for an article. Given a specific article, we categorized the contributors whose action vectors belong to cluster 1 as *all-round contributors* since they were engaged in almost all types of actions. Contributors with vectors in cluster 2 were labeled as *watchdogs* since most of the actions they performed were reverts. Cluster 3 includes the vectors of contributors who created sentences while seldom engaging in other actions, and these contributors were hence called *starters*. Contributors whose vectors belong to cluster 4, on the other hand, not only created sentences, but justified them with links and references. They were therefore classified as *content justifiers*. Cluster 4 includes the vectors of *copyeditors* who contributed primarily through modifying existing sentences. Finally, those who primarily focused on removing incorrect sentences, references, and links were termed *cleaners*. Thus, a contributor for a given Wikipedia article could assume one of these six roles or could be a *casual contributor*. A contributor can play different roles for different articles. As shown in Table II, a large



Cluster 1 (All-round Contributors)

Cluster 2 (Watchdogs)

Cluster 3 (Starters)

Cluster 4 (Content Justifiers)

Cluster 5 (Copy Editors)

Cluster 6 (Cleaners)

Fig. 2. Centroids of clusters.

percentage (more than 75%) of the contributors for a given article are either all-round editors, content justifiers, or copyeditors while there are fewer starters, cleaners, and watchdogs.

5. IDENTIFICATION OF COLLABORATION PATTERNS

As the next step, we wanted to investigate how contributors assuming different roles implicitly collaborate with each other on each Wikipedia article. For instance, there may be Wikipedia articles where starters create a large chunk of text and then copyeditors are relied upon to modify it; or articles where all-round contributors form a core group that insert much of the content and then continuously modify their own and other people's insertions. We attempted to identify collaboration patterns among the contributors with different roles. We used clustering to group Wikipedia articles based on roles and actions performed by contributors on these articles.

Table III. Description of Clusters and Their Corresponding Collaboration Patterns

Cluster Number	Number of Articles	Collaboration pattern description
1	330	Content justifiers dominated in sentence insertions (account for 72% of sentence insertions), reference insertions (68%), and link insertions (81%). They also made Casual contributors played an important role in sentence, link and reference modifications (30%, 24%, and 23% respectively).
2	361	Three types of contributors made a large percentage of sentence insertions. All-round contributors conducted 45% of sentence insertions. Starters performed 32% of sentence insertions. Content justifiers also made 21% of sentence insertions. All-round contributors played an important role in other actions. They made 49% of sentence modifications, 66% of sentence deletions and 68% of reference insertions, 47% of reference modifications, 69% of reference deletions, 35% of link insertions, 35% of link modifications, and 42% of link deletions.
3	296	Compared with other clusters, casual contributors played a more important role. Casual contributors contributed 42% of sentence insertions and 55% of sentence modifications. They also conducted many reference insertions and modifications (56% and 51% respectively). Cleaners carried out 58% of sentence deletions and 52% of link deletions. Content justifiers made 25% of sentence insertions and 28% of reference insertions. Starters also made 21% of sentence insertions.
4	351	All-round contributors dominated. They made 75% of sentence insertions, 59% of sentence modifications, 79% of sentence deletions, 88% of reference insertions, 56% reference modifications, 83% reference deletions, 68% of link insertions, 52% link modifications and 63% link deletions.
5	262	Starters dominate sentence insertions (58%). Causal contributors played important roles in sentence modifications (37%). They are also responsible for 17% of reference insertions and 24% of link insertions. All-round contributors made 36% of reference insertions.

Note: It is to be noted that watchdogs performed most of the reverts (at least 78%) for pages in all of the clusters. Also, copyeditors made 20%–22% percent of sentence modifications for articles in all of these clusters.

We used the 1600 sample articles described previously in our dataset. We identified the collaboration pattern of an article by examining what actions were performed by whom. If we use $P = \{p_1, p_2, \ldots, p_n\}$ to represent a set of Wikipedia articles, the collaboration among contributors with different roles for the article $p_i \in$ P is represented as a vector

$$\overrightarrow{col}_{p_i} = \langle a^k_{p_{i,j}} / a^T_{p_{i,j}} \rangle, j = 1..10, k = 1..7, \tag{2}$$

where $a^T_{p_{i,j}}$, $j = 1..10$, represents the total number of one type of action (e.g., sentence insertion) that affected an article p_i, and $a^k_{p_{i,j}}$, $k = 1..7$, the total number of one type of action (e.g., sentence creation) performed by one type of contributor (e.g., all-round contributors) to the article. In essence, for each type of action that affected an article, we tracked what percentage of this action was made by each type of contributor (e.g., starters, all-round contributors, etc.). We constructed the vectors for all of the selected 1600 articles and used them as input to the repeated K-means clustering algorithm described in Section 4. Once again, we set k, the number of clusters, to vary from 2 to 10. The repeated K-means algorithm resulted in 5 as the optimal number of clusters. Table III shows the characteristics of the 5 clusters, and each cluster represents a collaboration pattern. Each article is thus assigned a collaboration pattern.

6. RELATIONSHIP BETWEEN COLLABORATION PATTERNS AND ARTICLE QUALITY

Next, we examined the quality of articles in each cluster described in Table III. We examined the correlation between the Wikipedia designated quality grade and the

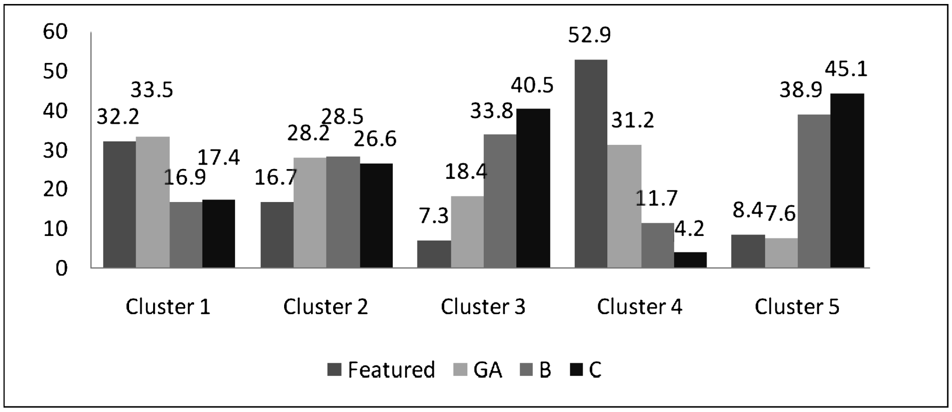


Fig. 3. Quality of articles belonging to different clusters.

collaboration pattern for each article in each cluster. The quality of articles that belong to different clusters is shown in Figure 3. The collaboration patterns and article quality are significantly correlated (*Kendall's Tau-c* = .43, $p < .001$, *Spearman's rho* = .49, $p < .001$). For instance, articles that belong to cluster 4 (where all-round editors dominated) are often of high quality with 52.9% of them being designated as featured articles and 31.2% as good articles by Wikipedia. Articles that belong to cluster 1 are also of relatively high quality. 32.2% of them are featured articles and 33.5% are good articles. Articles that belong to cluster 2 where all-round contributors, content justifiers, and starters all made a large percentage of sentence insertions are diverse in quality. The quality of articles in cluster 3 (where casual contributors played a dominant role) and in cluster 5 (where starters dominated sentence creations), on the other hand, is often questionable.

Determining the impact of collaboration patterns on article quality requires us to control the effects of exogenous factors. We controlled for five factors: (1) number of edits; (2) number of unique editors; (3) article age; (4) article length; and (5) number of unique administrators. The number of unique contributors who have contributed to the article and the number of edits the article went through have long been considered to be determinants of Wikipedia article quality in existing research. We also control for the age of articles. The quality of a Wikipedia article improves gradually. The age of an article thus reflects the article's maturity and can be associated with quality. We operationalized article age by counting the number of days from its creation to the date it was assigned the current quality grade. The article length may also be an indicator of Wikipedia article quality. It has been found that article length is a good predictor of whether an article will be a featured article [Blumenstock 2008]. We operationalized article length by counting the number of words in each article. To address the concern that administrators may have an influence over which articles get selected as Featured Articles or receive high ratings, we controlled for the number of unique administrators who have contributed to the articles. As shown in Table IV, we compare articles displaying different collaboration patterns along the different control variables. As expected, the articles that belong to cluster 4 where all-round contributors dominated have a larger number of unique administrators than articles in other clusters since a large percentage (30.8%) of all-round contributors are administrators in the Wikipedia community. They also have a slightly larger average number of edits and unique editors than other articles. These articles, however, on average have shorter article age. That is probably because we operationalized article age as the number of days from the

Table IV. Comparison of Articles in Different Clusters along Control Variables

Collaboration Pattern	Number of edits		Number of unique editors		Article Age		Article length		Number of unique administrators	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Cluster 1	474.7	440.4	324.2	294.0	2196.3	630.3	2293.7	1260.1	37.8	35.4
Cluster 2	477.8	311.8	321.1	208.1	2398.5	542.4	1713.5	954.7	34.0	26.1
Cluster 3	436.3	304.4	313.9	212.7	2317.9	570.6	1546.3	1288.4	31.5	26.8
Cluster 4	479.3	286.4	331.2	286.4	2065.3	564.9	2194.9	1027.7	45.7	24.9
Cluster 5	474.7	440.4	324.2	293.9	2256.1	459.5	1284.1	844.3	29.9	21.9
Overall	469.3	352.4	323.1	257.8	2245.5	552.9	1813.4	1059.0	36.1	26.7

Table V. Correlation between Variables

Variable	CP	NUE	NE	AA	AL	NUA	AQ
Collaboration Pattern (CP)	1						
Number of unique editors (NUE)	0.089*	1					
Number of edits (NE)	0.124*	0.986*	1				
Article age (AA)	–0.128*	–0.055	–0.076	1			
Article length (AL)	0.364*	0.222*	0.250*	–0.186*	1		
Number of unique administrators (NUA)	0.241*	0.830*	0.834*	–0.115*	0.301*	1	
Article Quality (AQ)	0.495*	0.135*	0.226*	–0.078	0.427*	0.226*	1

*Correlation is significant at the 0.01 level.

creation of an article to the date the article was assigned its current quality grade. Many articles that belong to cluster 4 are FAs or GAs, and these articles obtained their FA or GA status often earlier in their lifetime, which suggests that article age may not be a good indicator of article quality.

To assess the impact of collaboration patterns on article quality when controlling for the confounding variables, we used Multinomial Logistic Regression (MLR), a statistical model used to determine the dependence of a nominal variable on one or more predictor variables. The dependent variable, *article quality*, was recorded as a 4-category variable: C-class (recorded as “1”), B-class (“2”), GA (“3”), and FA (“4”). The predictor or independent variables include collaboration pattern and the control variables. We used a scale from 1 to 5 to represent the various collaboration patterns that are associated with articles of different quality with 5 representing the collaboration pattern of articles in cluster 4 that on average have the highest quality and 1 the collaboration pattern of articles in cluster 5 that have the lowest quality.

Table V describes the correlations between different variables (including article quality, collaboration pattern, and the control variables) we used in the MLR model. Since some of these variables including collaboration pattern and article quality are categorical, we used Spearman’s rank correlation coefficient (Spearman’ rho) to indicate the correlations between these variables. As shown in Table V, only weak correlations exist between collaboration pattern and the control variables including number of unique editors and number of edits (0.089 and 0.124, respectively), but we still included them in the MLR model since the correlation is statistically significant at $p < 0.01$. The two control variables that are moderately correlated with collaboration pattern include number of unique administrators (0.241) and article length (0.364). Article length is more correlated with article quality than number of unique administrators (0.427 and 0.226, respectively). The only control variable we did not include in the MLR model is article age since the correlation between article age and article quality is insignificant (–0.078), and there exists only a weak negative correlation between article age and collaboration pattern (–0.128). The correlation between number of unique editors and number of edits is extremely high (0.98). Due to multicollinearity concerns, we were not able to include both controls in the same model. Hence, we tested the model twice,

Table VI. Model Fitting Information for the MLR Models (the link function is logit)

Model	The chi-square test for goodness-of-fit				
	–2 log likelihood intercept only	–2 log likelihood final model	Chi-square	Degree of freedom	Significance
Model 1	2778	2220	558.0	21	.00
Model 2	2778	2293	486.5	21	.00
	Pseudo R-square				
	Cox and Snell R^2		Nagelkerke R^2		
Model 1	.427		.455		
Model 2	.385		.410		

Table VII. Likelihood Ratio Tests for Individual Effect in the MLR Models

Model	Variable	–2 Log Likelihood of reduced Model	Likelihood Ratio Tests		
			Chi-Square[a]	Degree of freedom	Sig.
Model 1	Intercept	2292	.000	0	
	Article length	2335	43.52	3	.000
	Number of unique editors	2327	35.40	3	.000
	Number of unique administrators	2374	82.29	3	.000
	Collaboration Pattern	2470	178.47	12	.000
Model 2	Intercept	2220	.000	0	
	Article length	2265	45.02	3	.000
	Number of unique edits	2327	96.91	3	.000
	Number of unique administrators	2398	106.26	3	.000
	Collaboration Pattern	2452	166.00	12	.000

Note: [a]The chi-square statistic is the difference in –2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

alternating between number of unique editors and number of edits. We named the model with number of unique editors as one of the independent variables Model 1, and the model that includes number of edits as Model 2.

The model fitting information for the MLR models is shown in Table VI. In assessing the overall model fit, the goodness-of-fit measure compared the predicted probabilities to the observed probabilities. Smaller values of the –2 log likelihood measure indicated better model fit [Hair et al. 1998]. Table VI presents a good model fit of Model 1 on the basis of variables including collaboration patterns, number of editors, article length, and number of unique administrators ($\chi^2 = 558.0$, $p < 0.01$), which indicates that the MLR model adequately describes the quality difference of the articles. The resulting model accounted for a significant amount of variance (Cox and Snell $\text{R}^2 = 0.427$ and Nagelkerke $\text{R}^2 = 0.455$). The results also indicate a goodness of fit for Model 2.

Likelihood ratio tests on each of the independent variables shown in Table VII revealed that collaboration pattern contributed significantly to both models when confounding factors including number of edits, number of unique editors, number of unique administrators, and article length are controlled. As expected, number of unique administrators is a quality indicator since administrators are the editors that have a good track record of contributions. The likelihood ratio tests show that collaboration pattern has a significant impact on article quality when we controlled the variable number of unique administrators. Article length has been proved to be an indicator of article quality. Number of unique editors and number of edits also have a significant effect on Wikipedia articles. The idea that more edits made by more editors lead to higher article quality is in general true. However, the path to quality improvement may differ from one article to another. The change of the chi-square statistics shown in Table VII indicates that collaboration pattern is a distinctive factor that makes a more significant impact on the quality of Wikipedia articles.

Table VIII. Summary of the Domain Samples

Domain	FA	GA	B-class	C-class	Avg. number of edits		Avg. number of editors
Geosciences	53	66	289	122	275		194
Computing	13	37	593	131	484		277
Politics	217	423	328	302	352		228
	All-round Contributors	Watchdogs	Starters	Content Justifiers	Copy Editor	Cleaners	Rand Statistic[a]
Geosciences	25.7%	4.7%	11.7%	23.5%	22.6%	11.8%	0.901
Computing	17.9%	5.0%	12.4%	24.5%	31.5%	8.7%	0.854
Politics	25.6%	6.7%	11.3%	20.5%	28.6%	7.3%	0.890
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Spearman's rho[b]	Sig.
Geosciences	56	77	222	46	129	0.30	0.000
	(10.6%)	(14.5%)	(41.9%)	(8.7%)	(24.3%)		
Computing	71	86	343	12	262	0.22	0.000
	(9.2%)	(11.1%)	(44.3%)	(1.6%)	(33.9%)		
Politics	211	191	372	196	300	0.27	0.000
	(16.6%)	(15.0%)	(29.3%)	(15.4%)	(23.6%)		

Note: [a]The Rand statistic represents the level of agreement between two classifications of contributors. [b]The Spearman's rho measures the correlation between the collaboration pattern and the Wikipedia designated quality grade.

7. FURTHER VALIDATIONS

A distinguishing feature of our study, compared with prior research such as Ehmann et al. [2008] that has examined the collaborative process which builds articles, is the significantly larger sample size of articles. Using the stratified sampling method to select 1,600 articles from different domains, we believe the findings of this study to be generalizable to Wikipedia as a whole. Here, we further validated the robustness of our findings in two different ways.

7.1. A Multiple Sample Validation

We first validated our findings using different samples from three domains including geosciences, computing, and politics. Each of these samples includes all featured articles (FA), good articles, (GA), B-class articles and C-class articles, except those with less than 50 unique editors or less than 100 edits. Articles in these three domains are different in quality, average number of edits, and unique editor (See Table VIII). Although often deemed controversial and less reliable [Korfiatis et al. 2006], politics articles on average are of surprisingly high quality. In comparison, computing articles have a larger number of edits and larger number of unique editors but lower quality.

To validate the roles we identified, we employed K-means clustering to cluster the contributors working on a given article that belong to each of the three domains into six clusters. For the purpose of determining if the six roles we identified in Section 4 are consistent across different domains, we adopted the cluster validation method proposed in Tibshirani and Walther [2005] to determine if the same set of roles can be assigned to the clusters. Our method consists of three steps. First, based on the cluster results using articles in one domain (e.g., computing), we obtained a classification C that comprises six clusters of contributors who worked on a given article in the domain. Second, the centroids (i.e., the centers of clusters) resulting from the repeated K-means performed on the 1,600 sample articles were used to obtain a classification D that also comprises six clusters of contributors who performed actions on articles in one domain (e.g., computing). We assigned a contributor who worked on an article in the domain to a cluster if the vector representing the contributor's actions on the article has the shortest Euclidean distance to the centroid of the clusters. Third, in order to determine the consistency of the roles across domains, we assessed the concordance between the two classifications, C and D, using the Rand statistic [Rand 1971], a metric that reflects the proportion of classification agreement between two classifications of the same objects, examining all pair-wise comparisons of objects. This metric ranges from 0 to 1 with 1 indicating perfect agreement. The Rand statistic values for articles sampled from the domain of geosciences, computing, and politics are, respectively, 0.901, 0.854, and 0.890, which leads us to claim that the roles identified in our study are consistent across these domains. We assigned the role labels described in Section 4 each cluster in C. As shown in Table VIII, computing articles have a lower percentage of all-round contributors (17.9%) than articles belonging to the other two domains, which may help explain why computing articles are in general of lower quality.

Next, we attempted to validate if the collaboration patterns that we identified in Section 5 were correlated with quality for articles that belonging to these three domains. Again, we used the centroids resulting from the clustering described in Section 5 to assign a collaboration pattern to each of the articles in the three domains. As shown in Table VIII, the collaboration patterns and quality of the article in the three domains are significantly correlated. For instance, 15.4% of politics articles belong to cluster 4 where all-round contributors dominated, while only 1.6% of computing articles and 8.7% of geosciences articles belong to cluster 4. It is hence not surprising that politics articles have a higher percentage of featured or good articles than those belonging to the other two domains. We also conducted the Multinomial Logistic Regression (MLR) with article quality as dependent variable and collaboration pattern and number of unique editors (or number of edits) as independent variables. The MLR analysis indicated that collaboration patterns have significant impact on the quality of articles belonging to different domains. Hence, our results are not dependent on the domain; rather, they generalize across domains in Wikipedia.

7.2. Validation Using Articles with External Quality Measures

To address the concern that editors may influence the rating of their own articles [Arazy and Nov 2009], we validated our model on an alternative dataset that employed external measures of article quality. We used the dataset that was obtained from an unpublished survey [Press 2006] advertised at the AISWorld mailing list (operated by the Association for Information Systems and serving information systems researchers) and was used in Arazy and Nov [2009]. It is to be noted that in this dataset, the external ratings were obtained from a small number of reviewers. Therefore they may not be as reliable as the Wikipedia ratings that are based on a large number of reviewers.

In the AISWorld survey, fifty scholars answering an advertisement rated the accuracy and completeness of 50 Wikipedia articles of their choice (the topic of these articles was science and technology) on a 5-point Likert scale. For each article, we took the sum of the accuracy and completeness measures. We noticed that the distribution of the sum of the accuracy and complete measures of the articles is highly skewed. For instance, only seven articles have a sum that is less than 5. We hence recorded the quality of these articles as a 4-category variable: “1” if the sum of the accuracy and completeness of an article is smaller than 5, “2” if the sum is 5 or 6, “3” if the sum is 7 or 8, and “4” if the sum is 9 or 10. We assigned a role to each contributor working on an article in the dataset using the centroids resulting from the repeated K-means performed on the 1,600 sample articles. We then used the centroids resulting from the clustering described in Section 5 to assign a collaboration pattern to each of the 50 articles. Again, we used a scale from 1 to 5 to represent the various collaboration patterns that are associated with articles of different quality with 5 representing the collaboration pattern of articles in cluster 4 and 1 the collaboration pattern of articles in cluster 5. We

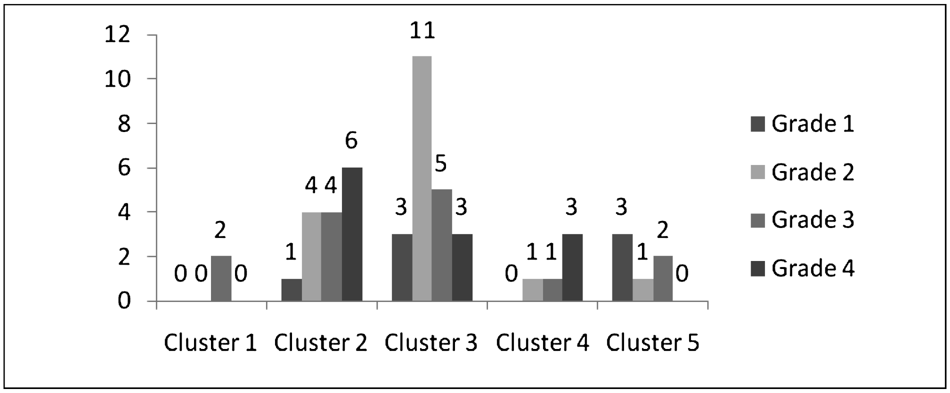


Fig. 4. Quality of the sampled articles belonging to different clusters.

found that the collaboration patterns and quality of the articles are significantly correlated (Spearman's = 0.428, $p < 0.001$). Figure 4 shows the relationship between the collaboration patterns and the quality grades of the 50 articles. As shown in Figure 4, a majority of articles that belong to cluster 3 where casual contributors played an important role are of low quality, while articles that belong to cluster 4, cluster 1, and cluster 2 are of better quality. It is noteworthy that there are more articles in cluster 3 and cluster 2 than those in other clusters. One possible reason may be that when the survey was conducted in 2006, many of the sampled articles were still in the early stage of their development, and articles in the early stage may tend to display certain collaboration patterns, which suggests that the collaboration pattern of an article may evolve over time, a possibility that merits future research.

8. IMPLICATIONS

Our research helps answer the question *why* Wikipedia articles vary in quality. It points to a new direction toward understanding the factors driving Wikipedia article quality: Article quality depends on different types of contributors, that is, the roles they play, and the way they collaborate. Currently, the Wikipedia community is manually assigning quality grades including FA, GA, B-class, C-class, etc., to Wikipedia articles. The sheer size of Wikipedia (over three million entries in the English Wikipedia currently) makes assigning and maintaining the quality grades a strenuous task. The statistical tests described in Section 6 have demonstrated that the collaboration patterns identified in our research are effective in determining Wikipedia article quality. More than 84% of articles that belong to cluster 4 where all-round contributors dominated are featured or good articles. More than 83% of articles in cluster 5 where starters dominated sentence insertions and more than 73% of articles that belonged to cluster 3 where casual contributors played an important role are B-class or C-class articles. Based on collaboration patterns and other metrics such as number of edits and unique editors, Wikipedia members can assign a preliminary quality rating to each article. They can then focus more attention on the ones (e.g., articles that belong to cluster 2) whose collaboration pattern does not clearly indicate their quality.

Our research is also intended to provide insights into *how* we can improve the quality of Wikipedia articles. Providing answers to the how question requires us first to investigate why different collaboration patterns impact article quality differently. It is thus worth further studying the characteristics of different patterns and their impact on article quality. Here, we compare articles that belong to the clusters in terms of two variables: *number of references* and *modification ratio*. According to Wikipedia,

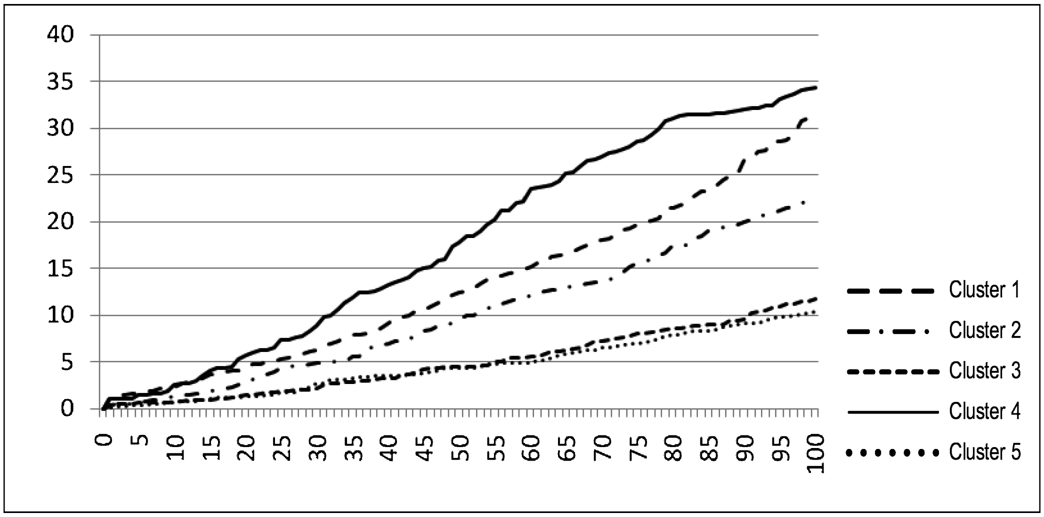


Fig. 5. Average number of references of articles that belong to different clusters.

featured and good Wikipedia articles must be verifiable against reliable sources and be supported by references to all sources of information. A sufficient number of references are thus a crucial prerequisite for high-quality articles. We counted the number of references of each article by computing the difference between number of reference insertions and deletions. Modifications are also critical for Wikipedia article quality. A modification often leads to an increment in article quality. Here, we consider the variable modification ratio, a normalization of number of sentence modifications by dividing by number of sentence insertions.

Figure 5 represents the average number of references for the articles that belong to different clusters at different points during their lifetime. The x-axis represents the relative positions of all the edits in the life of the articles, whereas the y-axis represents the average number of references for articles that belong to the different clusters. Articles vary in number of edits. In order to average the number of references of multiple articles at different points during their lifetime for each article, we used points from 0 to 100 to represent the relative positions of all edits that occurred to the article during its lifetime. For example, if the number of references of an article is 3 until its 6th edits and there are 200 total edits for that article, we first compute the relative position for all the edits. The relative position of the 6th edit is 3 (i.e., 6/(200/100)), and we record 3 as its number of references at the position 3. As shown in Figure 5, articles that belong to cluster 4 (where all-round contributors dominated) and cluster 1 (where content justifiers dominated sentence insertions) have a larger number of references than those belonging to other clusters during their lifetime. A conspicuous problem with certain collaboration patterns such as cluster 5 (where starters dominated sentence insertions) and cluster 3 (where casual contributors played a dominant role) is the lack of references in the articles. We noticed that 76.7% of all references for the 1,600 articles were added by the contributors who inserted the sentences to which the references are linked. It is thus reasonable to believe that the difference in number of references of articles belonging to different clusters may depend on who inserted the sentences. All-round contributors and content justifiers who dominated the sentence insertions for articles in cluster 4 and cluster 1, respectively, not only inserted sentences but justified them with references. On the other hand, the starters who dominated sentence insertions for articles in cluster 5 and casual contributors who inserted about half of

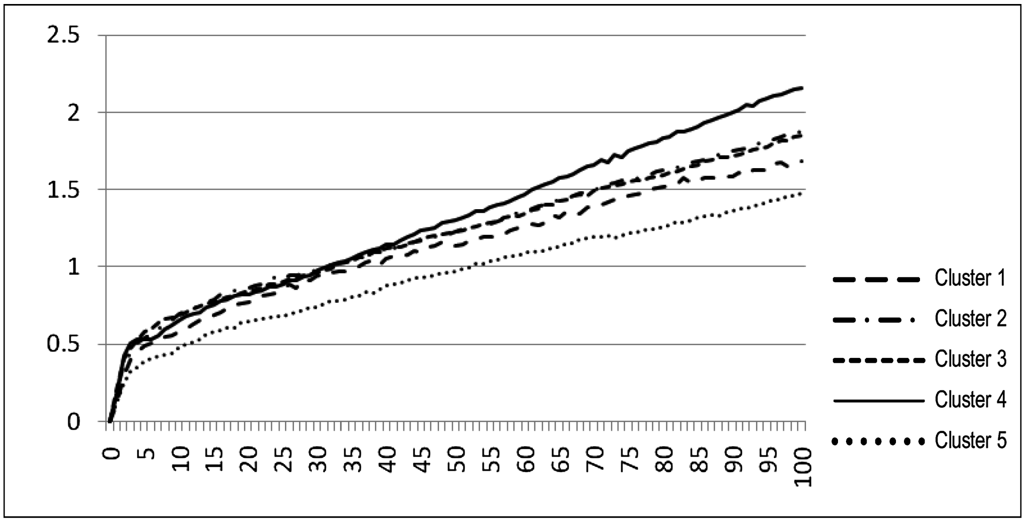


Fig. 6. Average modification ratio for articles belonging to different clusters.

the sentences for articles in cluster 3 tended to create sentences without citing sources, while other people often did not bother (or were unable) to identify the sources of these sentences. For articles belonging to cluster 2, three types of contributors, including all-round contributors, content justifiers, and starters, all made a large percentage of sentence insertions. As a result, the average number of references for articles in cluster 2 is smaller than that of articles belonging to cluster 4 and cluster 1 but larger than that of articles in cluster 3 and cluster 5.

Next, we compared the articles that belong to different clusters with respect to modification ratio (ratio between the number of sentence modifications and sentence insertions). As shown in Figure 6, articles in cluster 5 (where starters dominated sentence insertion) have a lower modification ratio than those developed in other patterns after the first few edits because starters inserted sentences but seldom modified sentences. Articles that belong to other clusters have a similar modification ratio during the early stage of their life. Then the line representing the modification ratio of articles in cluster 4 (where all-round contributors dominated) diverged from the rest after about one-third of their lifetime, indicating the these articles on average have a higher modification ratio than those in other clusters. This is probably because all-round editors not only created sentences and justified them with links and references, but also corrected and expanded the sentences, and reviewed and improved them after their sentences were modified by other people. This kind of "self-policing" accounts for 30.1% of modifications made by all-round editors. As described previously, articles that belong to cluster 1 have relatively high quality. 32.2% of them are featured articles and 33.5% are good articles. However, these articles often do not have a large modification ratio. As shown in Figure 6, the average modification ratio of the articles in cluster 1 is only higher than that of the articles in cluster 5, but lower then articles belonging to the other clusters during their lifetime. A possible reason is that the content justifiers who made most of the sentence insertions for these articles inserted sentences, added references and links, but seldom modified sentences inserted by themselves and other people. These articles may still have the potential for improvement if the content justifiers could conduct "self-policing" on the sentences they inserted.

Based on the preceding analysis, understanding which type of contributors dominated sentence insertions is critical information for determining Wikipedia article quality. The general characteristic of editing in many Wikipedia articles is the incremental development of individual elements [Jones 2008]. Contributors add new information by inserting sentences. These contributors and other contributors then work on the sentences by modifying them, adding links and references, or building on them. Wikipedia relies on volunteer contributors to correct errors of the inserted sentences, which is not always effective. An existing study [Luyt et al. 2008] proved that errors made in earlier edits were often found to be retained in the latest version of Wikipedia articles. Our study also indicates that contributors often did not bother (or were unable) to add references for sentences inserted by other people. "Self-justification" and "self-modification" are therefore critical for article quality. Contributors assuming different roles behaved differently when it comes to "self-justification" and "self-policing". Sentences inserted by all-round contributors are often of high quality since these contributors not only inserted sentences, but justified them with links and references and modified their own and other people's sentences. As a result, the articles belonging to cluster 4 where all-round contributors dominated are often of high quality. Articles that belong to cluster 5, on the other hand, are often low quality since the starters who inserted most of the sentences for these articles seldom conducted self-justification and self-policing.

Our current study extends existing research that focused on finding the determinants of Wikipedia article quality. Studies such as Adler and Alfaro [2007] and Anthony et al. [2009] attempted to distinguish reliable contributors from unreliable ones based on their previous contributions. The researchers assumed that a contribution is of high quality if it has been retained in the current version. Luyt et al. [2008] questioned the assumption by proving that earlier contributions tend to survive longer anyway. We also found the assumption questionable since the quality of Wikipedia articles is a multidimensional concept, and a long-lived contribution may be content-wise reliable but not high quality. For example, the article "Ethology" was assessed to be a C-class article since it needed "additional citations for verification." A contributor named "Outspan" inserted over 90 sentences to the article but provided only one reference in August 2007. Many of these sentences are retained in the current version, and this contributor thus would probably be deemed reliable according to the existing research. In fact, the contributor is at least partially responsible for the article's quality problem (namely, the lack of references). Similar to the existing research, we also intended to find out what types of contributors tend to provide high-quality content. However, we did not assess the reliability of a contributor based her previous contributions since it is impossible to automatically do so according to Luyt et al [2008]. We also did not separate experts from less competent contributors. Even though Wikipedia recognizes "administrators" who have a higher status in the Wikipedia community, that status cannot be easily connected to domain expertise because there is no system in Wikipedia to confirm the expertise of any contributor, and the online environment enables users to invent personas [Jones 2008]. Rather, we classified the contributors based on their explicit actions on a given article. Our findings are consistent with existing theories in the field of collaborative writing that different types of authors often display different patterns of contribution and the pattern of contribution and the quality of writing are inherently related [Fitzgerald 1987; Ede and Lunsford 2001; Bracewell and Witte 2003]. This action-based approach to finding user roles can potentially be used to reveal the self-organized social structure often found in other Web 2.0 applications. For instance, existing research on tagging systems has classified different tagging behaviors, categorized the users based on their behaviors, and explicated factors leading to the different tagging behaviors [Sen et al. 2006; Thom-Santelli et al. 2008].

Our findings also extend existing research that proves that high-quality Wikipedia articles rely on edit centralization [Kittur and Kraut 2008; Ortega et al. 2008] or contribution inequality [Arazy and Nov 2010]. Consistent with the existing research, our study shows that high-quality articles need to have a group of “leaders” or “core contributors” who made a majority of the contributions. As discussed previously, articles in cluster 3 where casual contributors played an important role are often of low quality. Our research moves one step further by indicating that there are different types of contributors, and who the “core contributors” are has a significant impact on the quality of the article. For instance, Wikipedia articles where all-round contributors dominated are often of high quality, while articles where starters inserted most of the sentences are often of low quality.

Our findings have important implications for improving Wikipedia article quality. The current approach adopted by Wikipedia to bolster the quality of its articles is to add layers of control. For example, nowadays, changes made to entries about living people will become live only when they’ve been vetted by a Wikipedia administrator. However, by adding layers of control, Wikipedia has developed a kind of bureaucracy that may possibly dissuade some people from participating. Consequently, it has been suggested that this has led to a slowing down of Wikipedia’s growth in recent years [Manjoo 2009]. Our research points to a new direction toward improving Wikipedia article quality. Our observations show that self-justifications and self-policing are important since it takes extra effort to add references and correct errors in sentences created by other people. Hence, instead of adding layers of control that could eventually hurt Wikipedia, we believe that it is crucial to develop software tools that are targeted towards gently nudging contributors to assume different roles and support self-justification and self-policing. For instance, starters and casual contributors often inserted sentences without providing references. This observation calls for a software tool (e.g., a pop-up window) that alerts contributors to justify their inserted sentences by adding links and references after they insert a number of sentences. A problem with starters, casual contributors, and even content justifiers is the lack of self-policing. It is therefore necessary to develop mechanisms that motivate the contributors to revisit the article, review their inserted sentences, and respond to other contributors’ modifications. As an example, we can send these contributors messages requesting them to verify the sentences they inserted whenever these sentences are modified by other people.

Our contribution extends beyond the specific Wikipedia context and has implications for virtual organizations in other contexts. Our study indicates the existence of de facto roles and the importance of an emergent leadership even in such an open environment as Wikipedia that virtually has no barrier to entry. Since activities in virtual organizations are often voluntary, identifying and developing effective “core contributors” are likely to be a critical success factor for the performance of any virtual organization. Our research points out two features of effective contributors: (1) They tend to perform diversified actions; and (2) they conduct self-policing, monitoring, and respond to changes made to their contributions. The first feature is related to the concept of functional diversity. It is well-known that functional diversity often leads to positive outcomes such as faster product development times [Eisenhardt and Tabrizi 1995], greater innovation [Ruef 2002], and greater team performance [Peters and Karren 2009]. The functional diversity described in the extant research often refers to the distribution of team members across a range of functional assignments. However, virtual organizations normally rely on voluntary participants performing different actions and often lack explicit functional assignments. Functional diversity in virtual organizations thus involves individuals performing diverse types of actions. Our research indicates that when the contributors, especially those “core contributors” perform a diversity of actions, team performance is often improved. There is potential to extend this finding to virtual communities in other contexts. Our findings regarding policing and monitoring can also inform research on virtual organizations. The effect of monitoring has been considered an important aspect of social capital that is critical for the success of a community [Bowles and Gintis 2002]. Virtual communities, often without principals that can directly supervise other community members, rely on mutual and self-monitoring for community governance. Our research hence suggests the need to develop social mechanisms to induce the community members to conduct self-policing and monitoring.

9. CONCLUSION

This study makes three major contributions. First, our study contributes to the research on wiki-based collaboration. Recent studies such as Bryant et al. [2005] and Majchrzak [2009] suggested that Wikipedia represents a emerging genre, not only as an information resource, but of collaboration, calling for developing theories regarding wiki-based collaboration. However, the tendency to use aggregate measures such as number of edits or unique editors has hindered researchers from gaining profound understanding about collaboration on Wikipedia articles. Our research is one of the first that delved deep into the implicit collaborative processes and classified contributors based on the actions they performed on Wikipedia articles. We further identify a number of collaboration patterns, each of which represents a distinct way in which contributors assuming different roles collaborate. Our research thus lays a foundation for developing new theories regarding wiki-based collaboration. Second, our research proves that contributors' collaboration pattern is a critical factor driving the quality of Wikipedia articles. We identify patterns that are preferable or detrimental for quality: Articles developed using patterns where all-round editors played a dominant role are often of high quality, while patterns where starters and casual contributors dominate are often associated with low quality. Third, our research provides insights about how to improve the quality of Wikipedia articles. Our study indicates that self-justification and self-policing are critical for Wikipedia article quality, suggesting the need to develop mechanisms that alert contributors to add references and encourage them to revisit the article and improve their inserted sentences. We believe our research paves the way for developing new software tools for collaboration for Wikipedia to encourage specific role setting and collaboration patterns to improve the quality of articles.

In future research, we will focus on a variety of other social-network-based collaboration issues in Wikipedia. Wikipedia articles are developed by voluntary contributors through collaboration. As contributors work on the same article or across several articles, they weave a network of relationships that act as channels which facilitate the flow of information and technical know-how. Our understanding of how social networks affect performance in the context of Wikipedia, however, remains unclear because the specific elements of network structure that influence team performance have yet to be identified. Investigating this issue is important because it potentially has important implications for optimal team composition and high-quality articles.

We are also extending our research to enterprise wikis. The success of Wikipedia makes wikis an increasingly popular knowledge management solution in organizations. A study [Economist Intelligence Unit 2007] shows that over 30% of the surveyed organizations make use of wiki technology or plan to do so in the future. While enterprise wikis are mostly free from vandalism or malicious edits, the approach of collectively created content also presents certain shortcomings, and critical voices exist that question the quality of the created information in wikis [Lykourentzoua et al. 2010]. In fact, after an initial period of promise and trial, due to quality concerns, many companies are not so satisfied with their adoption of wiki technology [Bughin 2007]. We believe that a mechanism that stimulates active participation, encourages or creates teams of people with different explicit roles, and encourages self-justification and self-policing is panacea to the quality woes plaguing deployment of enterprise wikis.

REFERENCES

Adler, B. T. and Alfaro, L. D. 2007. A content-driven reputation system for the wikipedia. In *Proceedings of the 16th International Conference on World Wide Web.* ACM Press, New York, 261–270.

Anthony, D., Smith, S., and Williamson, T. 2009. Reputation and reliability in collective goods. *Rational. Soc. 21*, 3, 283–306.

Arazy, O. and Nov, O. 2010. Determinants of wikipedia quality: The roles of global and local contribution inequality. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW).* 233–236.

Arazy, O., Stroulia, E., Ruecker, S., Arias, C., Fiorentino, C., Ganev, V. and Yau, T. 2010. Recognizing contributions in wikis: Authorship, categories, algorithms, and visualizations. *J. Amer. Soc. Inf. Sci. Technol. 61*, 6, 1166–1179.

Blumenstock, J. 2008. Size Matters: Word count as a measure of quality on wikipedia. In *Proceedings of the 17th International Conference On World Wide Web.* 1095–1096.

Bowles, S. and Gintis, H. 2002. Social capital and community governance. *Econ. J. 112*, 483, 419–436.

Bracewell, R. J. and Witte, S. P. 2003. Tasks, ensembles, and activity: Linkages between text production and situation of use in the workplace. *Writt. Comm. 20*, 4, 511–559.

Bryant, S. L., Forte, A., and Bruckman, A. 2005. Becoming wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work.* 1–10.

Bughin, J. R. 2007. How companies can make the most of user-generated content. *McKinsey Quart.*, 1–4.

Cohen, N. 2007. Courts turn to wikipedia, but selectively. *New York Times* (1/29/07).

Denning, P., Horning, J., Parnas, D., and Weinstein, L. 2005. Wikipedia risks. *Comm. ACM 48,* 12, 152–152.

Dondio, P. and Barrett, S. 2007. Computational trust in web content quality: A comparative evaluation on the wikipedia project. *Informatica 31*, 151–160.

Economist Intelligence Unit. 2007. Serious business: Web 2.0 goes corporate. http://replyweb20.files.wordpress.com/2008/01/web_20_goes_corporate.pdf

Ede, L. and Lunsford, A. 2001. Collaboration and concepts of authorship. *J. Mod. Lang. Assoc. Amer. 116*, 2, 354–369.

Ehmann, K., Large, A., and Beheshti, J. 2008. Collaboration in context: Comparing article evolution among subject disciplines in wikipedia. *First Monday 13*, 10.

Eisenhardt, K. and Tabrizi, B. 1995. Accelerating adaptive processes: Product innovation in the global computer industry. *Admin. Sci. Quart. 40*, 84–110.

Fitzgerald, J. 1987. Research on revision in writing. *Rev. Educ. Res. 57*, 4, 481–506.

Gacek, C. and Arief, B. 2004. The many meanings of open source. *IEEE Softw. 21*, 1, 34–40.

Giles, J. 2005. Internet encyclopedias go head to head. *Nature 438,* 7070, 900–901.

Hair, J., Anderson, R., and Tatham, R. 1998. *Multivariate Data Analysis.* Prentice-Hall, Upper Saddle River, NJ.

He, J., Lan, M., Tan, C. L., Sung, S. Y., and Low, H. B. 2004. Initialization of clusters refinement algorithms: A review and comparative study. In *Proceedings of the International Joint Conference on Neural Networks*. 297–302.

Hendry, D., Jenkins, J., and Mccarthy, J. 2006. Collaborative bibliography. *Inf. Process. Manag. 42*, 3, 805–825.

Jones, J. 2008. Patterns of revision in online writing: A study of wikipedia's featured articles. *Writt. Comm. 25*, 262–289.

Kane, G. C. and Fichman, R. G. 2009. The shoemaker's children: Using wikis to improve is research, reaching, and publication. *MIS Quart. 33,* 1, 1–22.

Kittur, A. and Kraut, R. 2008. Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. ACM Press, New York, 37–46.

Korfiatis, N., Poulos, M., and Bokos, G. 2006. Evaluating authoritative sources using social networks: An insight from wikipedia. *Online Inf. Rev. 30*, 3, 252–262.

LIH, A. 2004. Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism,* 16–17.

LITTLEPAGE, G., SCHMIDT, G., WHISLER, E., AND FROST, A. 1995. An input-process-output analysis of influence and performance in problem-solving groups. *J. Personality Social Psychol. 69*, 5, 877–889.

LOURIDAS, P. 2006. Using wikis in software development. *IEEE Softw. 23,* 2, 88–91.

LUYT, B., TAY, C. H., LIM, H. T., AND CHENG, K. H. 2008. Improving wikipedia's accuracy: Is edit age a solution. *J. Amer. Soc. Inf. Sci. Technol. 59,* 2, 318–330.

LYKOURENTZOU, I., PAPADAKI, K., VERGADOS, D., POLEMI, D., AND LOUMOS, V. 2010. CorpWiki: A self-regulating wiki to promote corporate collective intelligence through expert peer matching. *Inf. Sci. 180*, 1, 18–38.

MAJCHRZAK, A. 2009. Comment: Where is the theory in wikis? *MIS Quart. 33*, 1, 18–20.

MANJOO, F. 2009. Is wikipedia a victim of its own success? *Time Mag. 174.*

MCGRATH, J. 1984. *Groups: Interaction and Performance*. Prentice-Hall, Englewood Cliffs, NJ.

MCGUINNESS, D., ZENG, H., DA SILVA, P., DING, L., NARAYANAN, D., AND BHAOWAL, M. 2006. Investigations into trust for collaborative information repositories: A wikipedia case study. In *Proceedings of the Workshop on Models of Trust for the Web*.

ORTEGA, F., GONZALEZ-BARAHONA, J., AND ROBLES, G. 2008. On the inequality of contributions to wikipedia. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, IEEE Computer Society, 304–304.

PETERS, L. AND KARREN, R. 2009. An examination of the roles of trust and functional diversity on virtual team performance ratings. *Group Organiz. Manag. 34,* 4, 479–504.

PFEIL, U., ZAPHIRIS, P., AND ANG, C. S. 2006. Cultural differences in collaborative authoring of wikipedia. *J. Comput.-Mediat. Comm. 12,* 1, 88–113.

PINSONNEAULT, A. AND CAYA, O. 2005. Virtual teams: What we know, what we don't know. *Int. J. e-Collab. 1*, 3, 1–16.

PRESS, L. 2006. Unpublished wikipedia web survey results. http://bpastudio.csudh.edu/fac/lpress/wikieval/

RAM, S. AND LIU, J. 2007. Understanding the semantics of data provenance to support active conceptual modeling. In Lecture Notes in Computer Science, vol. 4512. Springer, 17–29.

RECTOR, L. 2008. Comparison of wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Ref. Serv. Rev. 36*, 1, 7–22.

ROSENZWEIG, R. 2006. Can history be open source? Wikipedia and the future of the past. *J. Amer. Hist.*, 117–146.

RUEF, M. 2002. Strong ties, weak ties and islands: Structural and cultural predictors of organizational innovation. *Industr. Corp. Change 11*, 3, 427–449.

SCHRAGE, M. 1990. *Shared Minds: The New Technologies of Collaboration*. Random House, New York.

SEN, S., LAM, S., RASHID, A., COSLEY, D., FRANKOWSKI, D., OSTERHOUSE, J., HARPER, F., AND RIEDL, J. 2006. Tagging, communities, vocabulary, evolution. In *Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work.* 181–190.

SUROWIECKI, K. 2004. *The Wisdom of Crowds*. Doubleday, New York.

STVILIA, B., TWIDALE, M., AND SMITH., L. 2005. Information quality discussions in wikipedia. Tech. rep. ISRN UIUCLIS-2005/2+CSCW, University of Illinois.

STVILIA, B., TWIDALE, M., SMITH., L. AND GASSER, L. 2008. Information quality work organization in wikipedia. *J. Amer. Soc. Inf. Sci. Technol. 59,* 6, 983–1001.

THOM-SANTELLI, J., MULLER, M. AND MILLEN, D. 2008. Social Tagging Roles: Publishers, Evangelists, Leaders. In *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems*. 1041–1044.

WIKIPEDIA. 2010. Wikipedia: Version 1.0 editorial team/assessment. http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

WILKINSON, D. M. AND HUBERMAN, B. A. 2007. Assessing the value of cooperation in wikipedia. *First Monday 12*, 4.

Received November 2010; revised March 2011; accepted April 2011