Data & Knowledge Engineering 115 (2018) 80–93

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Using big data and network analysis to understand Wikipedia article quality

Jun Liu [a], Sudha Ram [b,*]

[a] Dakota State University, USA
[b] University of Arizona, USA

ABSTRACT

The research reported in this paper focuses on the question of why Wikipedia articles are different in quality. Since these articles are developed in an open and social environment, our work investigates if the social capital of contributors plays a role in determining the quality of the articles. We focus on three major types of social capital with respect to teams of contributors working on Wikipedia articles: internal bonding, external bridging and functional diversity. Through a social network analysis of these articles based on a dataset extracted from its edit history, our research finds that all three types of social capital have a significant impact on their quality. In addition, we found that internal bonding interacts positively with external bridging resulting in a multiplier effect on article quality. The findings of our research have implications for developing automated techniques for quality assessment of Wikipedia and also provide insights into improving quality of these articles.

1. Introduction

Data explosion is an inevitable trend in the age of Web 2.0 as people are connected more than ever. Data are generated faster than ever, and to date about 2.5 quintillion bytes of data are created daily, which gives birth to the widely circulated concept Big Data. One example of socially generated Big Data is Wikipedia. As of September 2017, the English Wikipedia alone includes more than 5,472,000 articles. Wikipedia represents a new way of "big collaboration" that has never been seen before. It deviates substantially from the traditional paradigm of collaborative development which can be summarized with the saying "too many cooks spoil the broth" [1]. This traditional paradigm assumes that only a small and selected circle of experts should be allowed to develop high-quality information, while the vast majority of practitioners are expected to provide little improvement. Wikipedia follows and then goes beyond the open-source paradigm that facilitates mass collaboration. It fosters an anarchic environment where anyone can edit any information in Wikipedia, and no one officially stands behind the authenticity and accuracy of the information, which makes it much more "open" than open-source software projects such as Linux and Apache that often maintain a hierarchical social structure that includes owners, core developers, peripheral contributors, etc.

The goal of this research is to investigation the impact of collaboration and social relationships between Wikipedia contributors on the quality of Wikipedia article. Despite its seemingly chaotic everyone-can-edit mechanism, the quality of Wikipedia articles is surprisingly good. A much discussed article from Nature [76] 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. On the other hand, critics of Wikipedia never have trouble finding low quality Wikipedia articles. As of September 2017, only 5140 featured articles out of 5,481,729 (<0.1%) articles on the English Wikipedia are slated to be featured articles, articles that are professional, outstanding and thorough.

Why does the mass collaboration in Wikipedia result in articles of different quality? Research such as [2] has proved that it is

* Corresponding author.
E-mail addresses: jun.liu@dsu.edu (J. Liu), sram@email.arizona.edu (S. Ram).

https://doi.org/10.1016/j.datak.2018.02.004

Available online 16 February 2018
0169-023X/© 2018 Elsevier B.V. All rights reserved.

insufficient to attribute the quality of Wikipedia to aggregate measures such as number of edits or number of editors. To decipher the quality issue of Wikipedia, Arazy et al. [3] stress the need to develop Wikipedia-centric collaborative theory. Hence, the objective of the paper is to apply the well-grounded theory of social capital to the context of Wikipedia and investigate the impact of social capital on the quality of its articles. Organizational researchers have proved that social capital is a powerful factor explaining actors' success in a number of arenas. For instance, social capital has been proved to facilitate inter-unit resource exchange and product innovation [4], the creation of intellectual capital [5], and cross-functional team effectiveness [6]. Since content creation on Wikipedia is a social process, we can view a group of contributors collaborating on an article as a virtual team. Then the team's social capital, i.e., the composition of the team and the configuration of team members' social relationships within a team and in the social structure of the whole Wikipedia community, may provide a critical focus for differentiating teams that are more effective than others. This in turn may help us distinguish high quality articles from low quality ones. We differentiate between three different types of social capital including Bonding and Bridging capital [7] and functional diversity [8] and develop Wikipedia-centric theories that explicate the impact of each on article quality.

2. Literature review

Investigation of antecedents to the quality of Wikipedia articles started with the analysis of article level metrics (e.g., article length and number of contributors). Stvilia et al. [9] proposed seven different metrics for article quality on the basis of author roles and article profiles and found that these metrics were in fact instrumental in determining quality. Hu, Lim, Sun, Lauw, & Vuong [10] proposed three quality measurement models that make use of the interaction between articles and their contributors. They found that while the authority of the contributors did matter, article length was an important antecedent to quality. Ref. [11] showed that a simple count of the number of words in the article is a reliable metric for quality. Wilkinson & Huberman [12] proved the impact of successful collaboration on article quality by showing that it has a strong correlation with the number of edits. Ransbotham et al. [13] used the number of contributors and the centrality of authors in the content-collaboration network to assess their effect on the value of the Wikipedia content. They found that the number of contributors is related to value of the content in a curvilinear fashion – increasing up to a point and decreasing thereafter-while centrality - both global and local - contributed significantly to the article quality. They also found that article age negatively moderated the effect of number of contributors and centrality.

It has been argued that some characteristics of the group who contribute to a Wikipedia article would affect the quality of the article content. Arazy & Nov [14] used the concepts of local inequality and global inequality to differentiate between different levels of author contribution - at the level of the article as well Wikipedia as a whole - to a particular article. They found that global inequality exerts significant positive impact on article quality while local inequality has an indirect effect and is mediated by coordination measured in terms of the number of words on discussion pages." Liu & Ram (2011) classified contributors on the basis of their roles in editing articles and found that articles in which all-round editors played a dominant role are often of high quality. Arazy et al. [3] later found that Wikipedia participants' role taking behavior is turbulent across roles. At the organizational level, however, there are a highly stable set of emergent roles, despite the absence of traditional stabilizing mechanisms such as pre-defined work procedures or role expectations. They conceptualized this dualism as "Turbulent Stability". Ransbotham & Kane [15] assessed the effect of membership turnover both on the probability of an article being promoted to featured status and being demoted thereafter and found that that it is optimal to have a mix of both new and long-term contributors to attain and maintain featured status and moderate level of turnover is in fact beneficial for article quality. They also found that the balance between male and female contributors seems to be another contributing factor; it is negatively related to the comprehensiveness of Wikipedia content. Qin & Cunningham [16] quantified the contributions of authors on the basis of edit longevity and the centrality of the author in talk and co-author networks and found that these metrics were in fact instrumental in assessing quality of articles.

Some studies have examined the impact of user coordination and communications on Wikipedia article quality articles. Viegas et al. [17] manually annotated the posts in discussion pages of 25 articles into 11 different categories and found that requests for co-ordination form the majority of these postings thus confirming that these pages were used for strategic planning of edits. Kittur & Kraut [18] distinguished between explicit co-ordination in which editors plan to build an article and implicit co-ordination in which a subset of editors set the direction by doing majority of work. They found that implicit coordination was more useful when there were more contributors to the article and both types of coordination improved quality when the article was in the formative stage. Schneider, Passant, & Breslin [19] proposed an additional four categories for annotation and reinforced the findings that talk pages were primarily used for coordination. Stvilia et al. [20] manually annotated the discussion pages of 30 featured articles and 30 random articles to identify the differences in the types of information quality problems that are discussed. Kane [21] conducted an in-depth study of the collaboration processes of a Wikipedia article on 2007 Virginia Tech Massacre. They annotated the discussion posts of the article and surveyed the editors about their experiences in editing the article. The author found that metrics including type of contributor activity, number of anonymous contributors and top contributor experience had a significant effect on article quality, upon testing these metrics on a set of 188 featured articles. Arazy et al. [22] found when Wikipedia group members' disagreements – originally task-related – escalate into personal attacks or hinge on procedure, these disagreements impede group performance

Research that is most similar to ours is the ones that investigate Wikipedia content production and user collaboration through social network analysis. By analyzing an affiliation network incorporating both authors and articles, Kane & Ransbotham [23] found that the position of an article in the affiliation network is associated with the quality of the article. De La Robertie et al. [24] constructed a co-edit network for Wikipedia article and found that the higher the quality of an article is, the more frequent the interactions between co-editors are. No reported research so far has tried to assess the impact of social capital on the quality of Wikipedia articles. Since content creation on Wikipedia is a social process, we argue that relationships formed between different contributors play an important role in

determining the quality of articles. By modeling the relationships between different article editors in terms of their joint contribution to the focal article and to other articles on Wikipedia, we provide a way of quantifying its social capital. We also take into account the positional value of a contributor in the overall collaborator network and assess its impact on article quality. Note that the concept of "teams" in the context of Wikipedia refers to the implicit collaborations formed as a result of working on one or more articles.

3. Theoretical background and hypotheses development

Social capital is a contextual complement to human capital that provides an explanation for why certain individuals or groups are more successful than others [25]. Putnam [7] defines social capital as "features" of a social organization that facilitate coordination and cooperation. It is reflected in the configuration of team members' social relationships within the social structure of the team itself, as well as, in the broader social structure of the organization to which the team belongs [26]. The primary proposition of the social capital theory is that given two individuals who have the same skill set and ability, the one that is connected well performs better than the other. Following Putnam [7] that made a distinction between bonding vs. bridging social capital, we focus primarily on two "features" of teams including "internal bonding" and "external bridging". Bonding social capital is about the *internal* ties that cement members within groups. Bridging social capital, on the other hand, is about the *external* ties linking the team with other teams.

Prior literature has showed that these two forms of social capital have an independent effect on the performance of teams. Baker & Iyer [27] found that cohesive networks of direct connections between producers improve communication that stabilizes prices in securities exchange. Coleman [28] found that networks with close internal connections facilitate sanctions that make it less risky for people to trust each other in case of rotating-credit associations. M. Granovetter [29] also argues that the chance of deceit is considerably reduced in networks with close connections. While all these studies emphasize the importance of internal bonding, there is a significant body of literature on the effect of ties outside a focal team on its performance. Rosenthal [30] studied the performance of teams in Midwest manufacturing teams and found that successful teams are those whose members have less constrained networks beyond the team. Baum, Calabrese, & Silverman [31] found that biotechnology companies that have multiple kinds of alliance partners have more patents and faster revenue growth. In case of small groups, lab experiments have shown that resources of a team accumulate in the form of brokers, people who connect otherwise disconnected partners [32–34]. There is also evidence of the interaction effect of these two forms. Reagans & Zuckerman [35] found that R&D teams in which members span structural holes and have a dense communication pattern within the team have higher performance. Dyer & Nobeoka [36] showed that Toyota promotes co-ordination among diverse suppliers by investing in infrastructure to facilitate knowledge exchange.

Besides internal bonding and external bridging, we investigate the impact of "functional diversity" – measured in terms of the roles played by various contributors to the focal article - which has been considered an important aspect of social capital in existing research such as [8].

3.1. Effects of internal bonding

Internal bonding" is closely related to the concept of network closure that was defined by Burt as the extent to which members of a team are connected to each other. Networks that have strong internal bonding affect the team performance in three ways. Firstly, information flows better within a team than between teams [37] and hence strong internal bonding reduces the risk of information asymmetry between team members. Secondly, well connected teams trust each other and can easily arrive at a consensus on behavioral norms [28,29,38]. Previous research on open source networks has showed that members of a team interact more within the team than outside [39,40]. Moreover, networks are known to institutionalize social relationships [41]. Thirdly, a network with more direct connections help its members develop a shared understanding of problems they are working on, and hence reduces the possibility of conflicts [27,42]. Moreover as social interaction becomes continuous in time, it reinforces the need for cooperation [43].

In the Wikipedia context, a team with strong internal bonding may have three major advantages. First, internal bonding results in more bounded solidarity, stronger reciprocity norms, greater trust between members, and a shared identity [26]. Contributors are thus more willing and motivated to invest time, energy, and effort in contributing to the article and sharing information with others. They also feel more obligated to contribute to the article because of their attachment to peers [44]. Second, in Wikipedia, tasks such as, planning the structure of an article require significant coordination and interaction among contributors [17,18]. A cohesive team provides a comprehensive point of view about how an article should be developed and helps smooth coordination and cooperation. Third, internal bonding is effective for preserving and maintaining resources that can affect article quality. Ransbotham & Kane [15] have showed that the average experience of contributors is an important antecedent to preserving the article quality. Social capital in cohesive teams sanctions self-serving behaviors, diminishes the probability of opportunism [13], and reduces the cost of monitoring [26]. A cohesive team of contributors working on a Wikipedia article may be effective in preserving and maintaining the quality of the article.

Internal bonding social capital accumulates when its members have strong ties. In Wikipedia, these ties often occur and are enhanced through repeated collaborations. We make a distinction between internal repeated collaborations, i.e., repeated collaborations of team members on the targeted article, and external repeated collaborations, i.e., collaborations of team members on other articles. Homophily has been proposed as an integral part of internal bonding [45]. The theory of homophily predicts that people are more likely to have strong ties with individual similar to themselves. Here we focus on the homophily of team members in their social relationships. The interpersonal ties or connections between the members of a team are reinforced when that team is composed of those who are similar in social relationships (i.e., who have worked with many of the same collaborators besides the focal team members) [8]. Wikipedia contributors who are homogeneous in social relationships can be expected to have similar interests and beliefs and are more likely to

develop shared understanding of a topic [15,46], which promotes smooth coordination and increases the effectiveness in conflict resolution. We hence propose the following hypotheses.

Hypothesis 1 (H1). *The level of internal repeated collaborations of a team of contributors working on a Wikipedia article will have a positive impact on the quality of Wikipedia articles.*

Hypothesis 2 (H2). *The level of external repeated collaborations of a team of contributors working on a Wikipedia article will have a positive impact on the quality of Wikipedia articles.*

Hypothesis 3 (H3). *Given a team of contributors working on a Wikipedia article, the team members' homophily in social relationships will have a positive impact on the quality of the article.*

3.2. Effects of external bridging

Although bonding social capital is beneficial for teams, it alone may not be sufficient to elicit their optimal performance. Overly strong ties among team members may make them rigid and cognitively locked-in Refs. [13,15,47], which can potentially prevent the entry of new members with fresh ideas, subsequently resulting in incomplete and one-sided content. Bridging social capital is complementary to bonding social capital in that it facilities the diffusion of information and allows combining ideas or knowledge from different sources [48,49]. The bridging social capital of a team can increase when its members participate in other teams and establish "external ties" with contributors outside the team, thus bridging the focal team with other teams. Burt [25] argues that the value of cooperation in a team is amplified by the brokerage role played by the team members who have contacts with other teams. In the context of Wikipedia, some contributors may play a brokerage role by linking teams that would otherwise not be connected thus bringing in knowledge from diverse sources. In this context we investigate the effect of Structural holes that can have a significant effect on the article quality.

Structural holes define the number of non-redundant ties that an actor has and hence defines the boundary spanning behavior of individuals. Individuals who span more boundaries are expected to bring in multifarious knowledge because of the variegated information sources to which they are connected to. Because of their position at the interface of different groups, they learn first about the activities of these groups and can bring that knowledge soon to the table [50]. The strategic value created by the actors spanning structural holes is clearly evident in the success of loan officers in commercial banks in bringing a deal to a closure [51], higher probability of early stage investments surviving to IPO in venture-capital firms with joint-investment networks [52] and in higher patent output in organizations that hold broker positions in the joint venture network [31].

In the context of Wikipedia, the structural holes spanned by an article's contributor in the collaboration network can affect its quality in three different ways. Firstly, the quality requirements for articles on Wikipedia mandate that information provided on articles should always be verifiable. Contributors spanning structural have access to a variety of sources in the most economical way and can be the first point of contact for any team member looking for reliable knowledge to be incorporated into the article. Secondly, a contributor, who spans structural holes, having worked in different kinds of teams, is also likely to have experienced different points of view, and hence would in a better position to reconcile those opinions and provide a neutral point of view to the article. Finally, given that comprehensiveness of information is one of the important antecedents to quality of articles, contributors spanning structural holes can bring in knowledge from diverse sources and can incorporate new insights helping to improve the completeness of an article. They can reach other parts of collaborative network easily and can reduce the latency involved in searching for new knowledge. While the content of all articles in Wikipedia is freely available, bounded rationality of the contributors can make them search for external knowledge through the team members first [53]. Moreover the private benefits that these contributors can accrue – like being promoted to an administrator in this case – can motivate them to engage in collective action with other contributors working on the article [54]. Therefore we hypothesize that

Hypothesis 4 (H4). *The number of structural holes spanned by the team of contributors working on a Wikipedia article will have a positive impact on the quality of the article.*

3.3. Effects of interaction between internal bonding and external bridging

The previous two sections have discussed the independent effects of bonding and bridging capital on the article quality. Although each of them can contribute to quality significantly, it is possible that strongly cemented internal collaboration on an article, where the contributors have strong external connections, can have a multiplier effect on the quality of the article. While bridging social capital can provide a pipe through which external knowledge can be incorporated into the article, strong internal bonding may facilitate smooth incorporation of that knowledge because of the trust among the team of contributors. Burt [55] argues that network closures and structural holes complement each other. Reagans & Zuckerman [35] in their study of R&D divisions in eight different industries have shown that high performance teams are those in which members span multiple structural holes and strong relations within the team provide a platform for communication and coordination. Similar results were reported in the case of film and television writers [56]. In short, the intellectual capital of a member who spans structural holes may amplify the effect of social capital of his team members on the article quality. Thus we hypothesize that

Hypothesis 5 (H5). *The interaction between internal repeated collaborations and the structural holes spanned by the contributors will have a positive impact on quality of Wikipedia articles.*

Hypothesis 6 (H6). *The interaction between external repeated collaborations and the structural holes spanned by the contributors will have a positive impact on Wikipedia articles.*

Hypothesis 7 (H7). *The interaction between team members' homophily in social relationships and the structural holes spanned by the contributors will have a positive impact on Wikipedia articles.*

3.4. Effects of functional diversity

Apart from investigating internal bonding and external bridging, we also consider other social features of teams. In this respect, perhaps the simplest predictor of team performance is team size, which has long been proved to be correlated with article quality since "given enough eyeballs all bugs are shallow" [57]. Here we focus on functional diversity. Functional diversity normally refers to the distribution of team members across a range of functional roles (often defined by the enterprise as manufacturing, engineering, marketing, finance, human resources, etc.). Functionally diverse teams take independent responsibility for different parts of the project and provide a comprehensive shape to it in the end [58,59]. Keck & Tushman [60] have shown that executive teams in an organization that are heterogeneous are both stable and are willing to accept change easily. In the context of internet, Peters & Karren [59] have shown that functional diversity has a direct impact on the performance of virtual teams.

In Wikipedia, virtual teams working on articles include voluntary participants performing different actions and lack explicit functional assignments. Functional diversity in the Wikipedia context thus involves individuals performing diverse types of actions. Functionally diverse teams can enhance article quality in two different ways. Firstly, the quality of an article is judged not only on the basis of its content but also on the writing style, presentation, references to reliable sources and organization. This requires independent responsibility on the part of contributors. While the same contributor can be involved in functionally diverse roles, we can still count them as belonging to diverse teams and independently contributing to various dimensions of quality. In Liu & Ram (2011), the authors proved that articles where all-round contributors (i.e., contributors who performed different actions) dominated are often of high quality. Kane [21] has proved that the diversity of contributor activity - measured in terms of the ratio of minor edits to the total edits - is positively related to the article quality. Secondly, previous research has shown that knowledge sharing among structurally diverse groups enhances the team performance [61]. In Wikipedia, all round contributors who perform diverse edits have their contributions clearly palpable on the article. Impressions of higher quality of these contributions can motivate other editors to either move in the same direction or collaborate with those contributors. In either case it results in quality enhancement. Therefore we hypothesize the following.

Hypothesis 8 (H8). *The functional diversity of a team of contributors working on a Wikipedia article will have a positive impact on the quality of the article.*

4. Methodology and data

To test our hypotheses, we sampled Wikipedia articles of different quality, and we describe our approach to data collection.

4.1. Data collection

We took advantage of Wikipedia's article assessment which has organized the evaluation of more than 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. The Wikipedia community has developed comprehensive criteria for determining its article quality. 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. To increase the objectivity and neutrality of the quality assessment, Wikipedia has developed different mechanisms to increase the objectivity and neutrality of its quality ratings. These mechanisms 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. 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. GA candidates must undergo a similar review process to be promoted to the GA status. The quality assessments of B or C-class articles were performed by 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 excluded from the assessment of the article. An existing study [18] has proved that the article ratings from external raters and the Wikipedia quality ratings are significantly correlated (Spearman' rho = .54, $p < .001$). Hence, we believe that the 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.

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 August 2014. Previous research shows that topic 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 [76] while history articles in Wikipedia were found to lack accuracy, breadth and depth [77]. 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 its featured articles into 31 mutually exclusive categories including biology, politics, sports, 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. Based on the set of featured articles we selected, we then randomly selected the same number of good articles (GA), B-class, and C-class articles from the same domain.

In order to investigate how people collaborate in the Wikipedia context, we drew upon the article parsing method proposed in Ref. [62] and developed a novel algorithm for parsing through the edit history. Our algorithm takes the edit history of a Wikipedia as an input and then compares every two versions of the article before and after the contributors made the edit. It generates information including who inserted or edited which sentence and in addition to sentence insertions and modifications, it indicates if a contributor has performed other actions such inserting a link or a reference. This information is critical for measuring internal bonding capital as well as function diversity. More specifically, contributors performed actions on an articles that result in a sequence of versions v_0, v_1, ..., v_n. Our algorithm computes, for each version v_i, $i = 0...n$, the *live text* of the version. The *live text* consists of all sentences in v_i, and each word in each sentence is labeled with the identifier (ID) of the edit in which the word was added. For each sentence in the live text, we also track the links and references associated with it and label them with an edit ID. A unique feature of Wikipedia is that it allows users to revert the article to one of its previous version. To identify the reverts, we keep the *dead text*, which includes sentences that were previously deleted. Similarly we associated an edit ID with each word in each sentence in the dead text. To capture the actions that result in version v_k, we compare v_k with its previous version v_j. We first split v_k into a number of sentences, and we also track the links and references associated with each sentence. Then we match the sentences in v_k with those in the live text of v_j. We adopt the text matching algorithm described in Ref. [62] that evaluates the similarity of every two sentences that belong to v_k and v_j respectively to find the retained, newly created, modified and deleted sentences, thus determining the number of sentence insertions, modifications or deletions. Correspondingly, we identify the newly created, modified or deleted links or references associated with each sentence. The inserted or modified sentences may also result from restoring previously deleted sentences. Hence, we compare the inserted sentences with each sentence in the dead text to determine if the action is a revert and if so, to which version the article has been reverted. Finally, we save the live text including the sentences, links and references associated with v_k and update the dead text if necessary. When we applied this algorithm in this study, we note the following details. Since we attempted to study the impact of social capital 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 versions that were reverted later.

The algorithm described above ushers us into a new level of understanding of the collaborative process of Wikipedia article development. Different from previous research (e.g., [13] that relies on the information made available by Wikipedia and focus on analyzing a social network constructed based on whether two contributors have worked on the same articles, our research is one of the first that investigates how Wikipedia contributors collaborated within an individual article, on a sentence level, thus generating internal bonding social capital at a very fine grained level.

4.2. Construction of social networks

We constructed social networks for Wikipedia articles for developing measures for assessing social capital. The algorithm described in the previous subsection enabled to extract not only the actions a contributor performed on an article but also the actions carried out on each individual sentence in the article. Based on this sentence-level knowledge, we constructed a *local collaboration network* for each article in our data sample. Again each node represents an author. An edge exists between two contributors who worked on at least one common sentence. A local collaboration network is a weighted network. The weight assigned to an edge linking two nodes represents the number of common sentences on which the two authors worked. The nodes in each network are the individual contributors.

Following existing research such as [13], we also constructed an overall social network on the basis of whether two contributors have ever worked on the same article. This overall social network is called the *overall collaboration network* in our research, and is a weighted network. A Wikipedia contributor was linked to another contributor if they worked on the same article. The weight represents the total number of articles on which these two contributors collaborated.

4.3. Dependent variable

Article Quality (AQ): We operationalize the article quality as the grade assigned to an article. We use four quality grades Featured Article (FA), Good Article (GA), B and C for our analysis. The validity of using these ratings as a proxy for quality has been firmly established in the previous studies [18,21,63]. While this metric cannot be guaranteed as a perfect indicator of quality, the fact that all the previous studies have found it to be at least 85% accurate alleviates any concerns about its validity.

4.4. Independent variables

***Internal repeated collaborations (IC)*:** We determined the level of internal repeated collaborations of the team g based on the *local*

collaboration network developed for the article *p*. The level of internal repeated collaboration (IC) was computed as $IC = \frac{\sum_{i=1}^{n_e} \frac{w_i}{n_s}}{n_p}$, where n_e represents the number of edges of the local network, and w_i is a weight of each edge and represents the number of common sentences on which a pair of team members worked, n_s represents the total number of sentences in the article, and n_p the total number of pairs that exist in the team *g* that worked on an article *p*.

***External repeated collaborations (EC)*:** To compute the level of external repeated collaborations of the team *g* that worked on an article *p*, we used the *overall collaboration network*. Given a pair of members m_i, m_j within *g*, if there is an edge that links the vertices representing m_i and m_j in the overall network, a weight ω_i of the edge represent the total number of common articles on which the pair worked, including the article *p*. We hence computed the level of external repeated collaboration (EC) as $EC = \frac{\sum_{i=1}^{n_p} \omega_i - 1}{n_p}$, where n_p represents the total number of pairs that exist in the team.

***Homophily in social relationships (HS)*:** We followed [8] and used the Jaccard similarity coefficient to represent the similarity in the relations of two members within the team g. The Jaccard similarity coefficient J_i of each pair of team members c_j, c_k is given as $J_i = \frac{M11}{M10+M01+M11}$, where $M11$ represents the total number of contributors both c_j and c_k have worked with on at least one article, $M10$ represents the total number of contributors only c_j has worked with, and $M01$ the total number of contributors only c_k has worked with. We then computed a measure of the homophily in social relationships for the team *g* as $HS = \frac{\sum_{i=1}^{n_p} J_i}{n_p}$, where n_p represents the total number of pairs that exist in the team. Again we used the *overall collaboration network* for computing this measure.

***Structural Holes (SH)*:** Based on the *overall collaboration network*, we use the constraint measure of structural holes since it can help us identify the extent to which a particular actor is constrained in terms of access to external knowledge [64]. Following [64], we define this measure as $C_i = \sum_j \left(p_{ij} + \sum_q p_{iq} + p_{qj} \right)$, where *i* is the ego for which we are calculating the metric, *j* is the set of all the alters of *i*, *q* is the set of all nodes to which both *i* and *j* are connected, and p_{ij} is the proportional strength of *i* and *j* given by $\frac{z_{ij}}{\sum_q z_{iq}}$, where *z* is the corresponding entry in the adjacency matrix. We used the average of this metric across all actors to compute the article level measure.

Functional diversity (FD): We computed the functional diversity of team *g* by adopting the functional diversity measure proposed in Ref. [65]. Existing research such as [2,66] proposes slightly different categorizations of actions that can be performed on a Wikipedia article. In this research, we consider five types of actions including 1) adding sentences, 2) modifying sentences, 3) deleting sentences, 4) adding links, and 5) adding references. The algorithm defined in Section 4.1 enables us to extract the different types of actions performed by contributors on each Wikipedia article. We computed the percentage of members who performed each type of action, p_i, i = 1...5. The functional diversity of the team was then computed as $FD = \prod_{i=1}^{5} p_i$. A high value of this measure indicates high functional diversity.

4.5. Control variables

We control for extraneous variables that can possible affect the quality of the article. These control variables were identified on the basis of prior established research on Wikipedia Articles [11,13,21].

Number of Edits (NE): Previous research has shown that high quality articles tend to have a large number of edits. Kane [21] showed that high quality articles tend to have a lot of minor edits since the editors are concerned with presenting the article in a proper format. We control for this factor by including the number of edits.

***Number of Unique Contributors (NC)*:** Articles which have more contributors tend to be of high quality [12,67]. Greater numbers of contributors bring multifarious knowledge to the articles and prevent errors of omission. They can also help in correcting wrong content; identify missing information and reverting vandalism [21]. In addition additional contributors bring in alternate points of view that are debated and incorporated into the article. On the flip side, recent research has shown that the marginal value of adding contributors decreases with the addition of each contributor and also affects the quality negatively beyond a point [13]. In either case, the number of contributors does have a direct effect on the quality of article.

Article Age (AA): Ransbotham et al. [13] have showed that the impact of collaborative activities on Wikipedia may be negatively moderated by the age of the article. This is because the knowledge and experience gained through previous collaboration of the contributors might create a knowledge lock-in that prevents the entry of new contributors and hence new knowledge [68]. Ransbotham & Kane [15] have showed that the average experience of the contributors, beyond a certain value, in fact negatively affects the chance of the focal article being promoted to featured article status. We operationalize the article age as the number of days that elapsed between the day on which it was assigned a particular quality grade to the day on which it first appeared on Wikipedia.

Article Length (AL): While Wikipedia has strict guidelines on the length of articles, researchers have found in the past that lengthy articles tend to be of higher quality because they comprehensively cover the breadth of the topic [11,13]. We operationalized the article length as the number of words in the article at the time it was assigned a grade.

Anonymity of Contributors (ANON): In Wikipedia, contributors can either register before editing any article or can anonymously edit the article. In case the editor is not registered, automatic bots assign their IP address as the identity. Previous research has shown that the number of anonymous edits correlated negatively with the number of controversial revisions of an article signifying that these contributors might be valuable to the article page when they are adding or refining content [69]. However, Kane [21] has showed that the number of anonymous contributors is in fact negatively related to the quality of article. In either case anonymous contributors do

have an effect on the article quality. We operationalize this variable by taking the percentage of anonymous contributors calculated as the ratio of anonymous contributors to the total number of contributors.

Information Presentation (IP): Content that is well-written and presented lucidly tends to have higher value and hence perceptions of higher quality. In fact one of the featured article criteria of Wikipedia says that the article should follow the style guidelines. Articles in Wikipedia can have up to six levels of nested sections. Also articles that have more images give impressions of high quality. We operationalize this factor as the ***number of images (NI)*** and the ***maximum section level (MS)*** reached in the article.

Reading Complexity (RC): Articles written in a sophisticated style tend to give perceptions of higher quality to the readers. Kane [21] surveyed the top 1% of editors for a featured article and found that these contributors were more concerned about presenting the content clearly, compellingly and professionally rather than making it more appropriate and accurate. We use the automated readability index - calculated as $(4.71 \times \textit{letters/words}) + (0.5 \times \textit{words/sentences})$ - 21.43 - to control for this factor [70]. While there are other metrics to measure reading complexity, this index is well suited for automated processing of large data [13].

4.6. Hierarchical multinomial logistic regression

The dependent variable in our case is categorical, and hence we use multinomial logistic regression to test our hypotheses. Before feeding the independent variables and the control variables into our model, we provide the descriptive statistics of the variables. The first part of Table 1 shows the means and standard deviations of all the variables used in our model. The second part of Table 1 indicates the correlations between these variables. As shown in Table 1, only the correlation between "number of unique contributors" and "number of edits" and the correlation between "number of unique contributors" and "anonymity of contributors" are greater than 0.7. Due to the multi-collinearity concerns, we did not include the variable "number of unique contributors" in our logistic regression model.

To evaluate the impact of each type of social capital on the quality of Wikipedia articles, we conducted hierarchical multinomial logistic regression analysis. Model 0 includes only the control variables and was used as our baseline. Model 1 adds the three internal bonding variables including "Internal repeated collaborations (IC)", "External repeated collaborations (EC)"and "Homophily in social relationships (HS)". We added the "Structural Holes (SH)" measure to Model 2 and the variable "Functional Diversity (FD)" to Model 3. Finally, we investigated the impact of interactions between internal bonding and external bridging by incorporating the three interaction terms into Model 4.

Table 2 summarizes the model fit information and provides a statistical analysis of the effect of each independent or control variable. The results of Model 0 show that, among all the control variables, "Maximum Section Level" does not have a significant effect on the quality of Wikipedia articles.

Table 3 shows the estimated multinomial logistic regression coefficients for the models. An important feature of the multinomial logistic model is that it estimates *k-1* coefficients, where *k* is the number of levels of the outcome variable. In our case, we set "featured articles" to be the referent group and estimated the impact of the independent and control variables on C-class articles vs. featured articles (FAs), B-class articles vs. FAs, and GAs vs. FAs. As shown in Table 3, in Model 0, both "number of edits" and "article age" have a significant positive coefficient for C-class articles compared with FAs. This indicates that C-class articles are more likely to have a larger number of edits and longer article age than FAs. As mentioned previously, we computed the measures up to the point when an article was assigned its current quality grade. Many featured articles obtained their status earlier in their lifetime while there are articles that remain at a low quality level even after they undergo a large number of edits, which indicates that aggregate and simplified measures such as number of edits and article ages are often not reliable indicators of Wikipedia article quality.

Model 1 incorporates the internal bonding measures including "Internal repeated collaborations (IC)", "External repeated collaborations (EC)" and "Homophily in social relationships (HS)". It shows a significant improvement over Model 0 with respect to the model fitting statistics including the AIC and −2 Log likelihood. As shown in Table 2, all these three measures have significant impact on the dependent variable, i.e., article quality. Table 3 further shows that the coefficients associated with these three variables are all significant and negative. Since we used the quality grade "featured article (FA)" as the reference, a negative coefficient means that an increase in the measure reduces the probability of the article being of low quality (e.g. C-class) vs. being an FA. These results support Hypothesis 1, 2 and 3. The level of internal repeated collaborations, the level of external repeated collaborations, and team members' homophily in social relationships all have a significant impact on the quality of Wikipedia articles. Among these three variables, the level of internal repeated collaborations has a more significant impact on articles than the other two internal bonding measures, as evidenced by its larger Wald Chi-square value.

Table 2 shows that adding the structural hole measure to Model 2 further enhances the overall model fit, and Table 3 shows a negative coefficient of the variable, which in fact means that featured articles are more likely to have larger structure hole measures than articles belonging to other classes. The results of Model 2 hence support our fourth hypothesis that the bridging social capital of a team, measured by the average structural holes spanned by contributors, has a significant impact on article quality. As shown in Table 2, the internal bonding capital appears to have more impact on article quality than bridging social capital. This result is a bit surprising as related research in team performance often finds contradictory evidence. Even in the context of Wikipedia, Okoli & Oh [71] raised concern about "relational inertia" and "cognitive lock-in", which may prevent the participation of new members with fresh ideas and result in incomplete and one-sided content. We believe that the anyone-can-edit philosophy of Wikipedia may explain why bonding social capital is more critical for article quality than bridging social capital. This philosophy enables casual contributors to introduce new ideas and knowledge even when the team members are strongly bonded. Benefits from team members being cohesive in social space still accrue due to its effect on shared understanding and coordination among members.

The results of Model 3 indicate that functional diversity has a positive effect on the quality of Wikipedia articles, thus supporting Hypothesis 8. This effect of functional diversity is not as large as internal bonding and external bridging but is significant with $p < 0.05$.

Table 1
Descriptive statistics.

	Independent variables					Control Variables							
	Internal repeated collaborations (IC)	External repeated collaborations (EC)	Homophily in social relation-ships (HS)	Structural Holes (SH)	Functional diversity (FD)	Number of Edits (NE)	Number of Unique Contributors (NC)	Article Age (AA)	Article Length (AL)	Anonymity of Contributors (ANON)	Number of Images (NI)	Maximum Section Level (MS)	Reading Complexity (RC)
Mean	0.0182	6.23	0.0209	0.0046	$4.94*10^{-5}$	490.13	129.76	1240.70	19,204.26	30.39	40.23	5.85	17.96
Std	0.0064	7.11	0.0229	0.0070	$2.82*10^{-5}$	582.87	158.26	880.56	14,429.43	19.60	60.26	3.03	3.45
IC	1.0000												
EC	−0.1003	1.0000											
HS	0.2220	−0.2032	1.0000										
SH	0.1109	−0.1265	0.2396	1.0000									
FD	0.0209	0.0123	0.0695	0.0147	1.0000								
NE	0.0056	−0.0597	0.1808	−0.0705	−0.0173	1.0000							
NC	0.0455	−0.0734	0.2897	−0.0899	−0.0187	*0.7963*	1.0000						
AA	−0.0127	0.0800	0.2138	−0.0928	−0.0359	0.4390	0.5006	1.0000					
AL	0.2073	−0.1387	0.3259	0.0972	0.0567	0.3004	0.3015	0.2350	1.0000				
ANON	−0.0167	0.0708	0.2073	−0.0821	−0.0500	0.5995	*0.7105*	0.4126	0.1959	1.0000			
NI	0.1914	−0.1439	0.3234	0.1409	0.0365	0.0875	0.1338	0.0813	0.5323	0.0353	1.0000		
MS	0.0384	0.0096	0.1275	0.0144	−0.0111	0.1265	0.1225	0.1034	0.2458	0.1132	0.1254	1.0000	
RC	−0.1279	0.0483	−0.0962	−0.0576	0.0246	−0.1255	−0.1351	−0.1001	−0.2507	−0.0890	−0.1821	−0.1144	1.0000

*The article age was measured in number of days.
**The article length was measured in number of words.

Table 2
Model fit and analysis of effects.

	Model 0	Model 1	Model 2	Model 3	Model 4
Model Fit Statistics::					
AIC	3635.86	3053.685	2982.17	2978.92	2963.88
−2 Log L	3587.86	2987.685	2910.17	2900.92	2867.88
Analysis of Effects:	***Wald Chi-Square***				
Number of Edits (NE)	35.8119**	31.3229**	28.6067**	28.4474**	27.9664**
Article Age (AA)	74.6819**	95.1583**	90.2088**	88.6415**	86.815**
Article Length (AL)	163.4151**	136.854**	139.6467**	135.5432**	139.6058**
Anonymity of Contributors (ANON)	20.635**	21.0231**	15.7738**	15.3443**	20.9083**
Number of Images (NI)	110.3277**	64.9584**	61.8631**	61.4883**	60.9641**
Maximum Section Level (MS)	1.2148	7.027	6.0227	5.4689	7.0961
Reading Complexity (RC)	34.2586**	29.2757**	28.276**	28.369**	28.4925**
Internal repeated collaborations (IC)		183.4824**	182.6738**	182.0613**	105.7958**
External repeated collaborations (EC)		60.3069**	63.4922**	59.8566**	46.4981**
Homophily in social relationships (HS)		101.7879**	52.3521**	51.5038**	54.2395**
Structural Holes (SH)			45.0877**	45.4124**	47.8963**
Functional diversity (FD)				9.1845*	8.9482*
Interaction SH and IC					1.0384
Interaction SH and EC					6.1922
Interaction SH and HS					25.6204**

*$p < 0.05$; **$p < 0.01$.

The bridging social capital of a team accumulates when team members contribute to articles on different related or unrelated topics. In essence, bridging social capital indicates the "diversity" of contributors in terms of interest and ideas and is related to generation of alternative ideas and multiple interpretations of information [55], which leads to enhanced article quality. Functional diversity represents a different kind of "diversity", which has more significant impact on Wikipedia article quality.

Burt [72] argues that actors who span structural holes in strong internally bonded teams amplify the value of collaboration. We hence investigate the interaction between external bridging and internal bonding. Since we used three measures to denote internal bonding capital, we hence developed Model 4 and investigated the interaction of external bridging, as measured by the structural hole, with each of three internal bonding measures, including "Internal Repeated Collaborations", "External Repeated Collaborations" and "Homophily in Social Relationship". The results of Model 4 indicate one of the interaction variables, i.e., interaction between "Homophily in Social Relationships" and external bridging has significant impact on the quality of Wikipedia articles, thus supporting Hypothesis 7. Although the interaction between "External Repeated Collaborations" and external bridging is statistically insignificant in distinguishing the four quality classes, the results shown in Table 3 indicates that the estimated coefficient of the interaction for C-class articles relative to featured articles is significant, indicating that the higher interaction between "External Repeated Collaborations" and external bridging is more likely to result in featured articles vs. C-class articles and thus partially supporting Hypothesis 6. The results of Model 4 also indicate that the interaction between "Internal Repeated Collaborations" and external bridging does not have significant impact on article quality, which means that in the context of Wikipedia, contributors having larger external bridging capital may not collaborate more on individual articles.

Burt [72] also argues that a team composed of multiple corporate functions that spans more structural holes has faster access to information sources. Functional diversity in the corporate setting normally refers to the distribution of team members across a range of functional roles (often defined by the enterprise as manufacturing, engineering, marketing, finance, human resources, etc.). However, in Wikipedia, contributors does not assume pre-specified functional roles, and given a contributor, we compute the functional diversity by investigating what type or types of action the contributor performed most. Based on our data analysis, we did not find evidence that the interaction between functional diversity and internal bonding or external bridging has a significant impact on Wikipedia article quality.

5. Discussion and contributions

We have analyzed the effect of social capital on the quality of 1600 randomly selected articles on Wikipedia belonging to four different quality grades. We hypothesized about the impact of three different forms of social capital on article quality. Consistent with our hypotheses, we find that all the three forms have a significant positive impact on the quality of the article. In short, our results support the core idea that social capital can have a profound effect on the quality of user generated content in a collaborative environment. These results point to new opportunities and avenues for researchers and managers.

5.1. Theoretical contributions

This paper has several implications for theory. Firstly, our work builds on [13] by providing a more comprehensive and granular view of social capital in terms of conceptualization of different forms of social capital and interactions between these forms. Previous studies have demonstrated that Wikipedia promotes a new way of collaboration that calls for new theoretical models [73]. We develop and test a theoretical framework that helps determine the quality of Wikipedia articles based on the well-grounded social capital theory. This can serve as a basis for researchers investigating the success of open source collaborative environments.

Table 3
Coefficient estimations in multinomial logistic regression.

	Quality	Model 0	Model 1	Model 2	Model 3	Model 4
Number of Edits (NE)	C-Class	0.000783**	0.000827**	0.000666*	0.000671*	0.000619*
	B-Class	0.00072**	0.00082**	0.000711**	0.000716**	0.000679*
	GA	−0.00046*	−0.00033	−0.00043	−0.00042	−0.00046
Article Age (AA)	C-Class	0.000824**	0.0013**	0.00125**	0.00123**	0.00125**
	B-Class	0.000146	0.000523**	0.000457**	0.000443**	0.000461**
	GA	0.000074	0.000361**	0.000287*	0.000272*	0.000275*
Article Length (AL)	C-Class	−0.00014**	−0.00014**	−0.00014**	−0.00014**	−0.00015**
	B-Class	−0.00005**	−0.00006**	−0.00006**	−0.00006**	−0.00006**
	GA	−0.00005**	−0.00005**	−0.00005**	−0.00005**	−0.00005**
Anonymity of Contributors (ANON)	C-Class	0.00238	0.0112	0.00409	0.00319	−0.00157
	B-Class	0.0149**	0.0217**	0.0156*	0.0152*	0.0163*
	GA	0.019**	0.0245**	0.0194**	0.0186**	0.0181**
Number of Images (NI)	C-Class	−0.0108**	−0.00756**	−0.00698**	−0.00702**	−0.00746**
	B-Class	−0.0337**	−0.0268**	−0.026**	−0.0259**	−0.0258**
	GA	−0.00397**	−0.00225	−0.00172	−0.00172	−0.00184
Maximum Section Level (MS)	C-Class	0.0117	0.1159*	0.1005	0.094	0.1117*
	B-Class	0.0264	0.1295**	0.1167*	0.1119*	0.1325**
	GA	0.00502	0.1025*	0.093*	0.0871*	0.1107*
Reading Complexity (RC)	C-Class	0.102**	0.1047**	0.0991*	0.1012*	0.1018*
	B-Class	0.1031**	0.1075**	0.1044**	0.104**	0.0997*
	GA	−0.0305	−0.0288	−0.0338	−0.0337	−0.037
Internal repeated collaborations (IC)	C-Class		−255**	−262.3**	−261.9**	−245.2**
	B-Class		−210.8**	−217.2**	−216.7**	−217.1**
	GA		−79.0794**	−82.9054**	−82.5101**	−81.785**
External repeated collaborations (EC)	C-Class		−0.0823**	−0.0864**	−0.0845**	−0.0773**
	B-Class		−0.0683**	−0.0732**	−0.0721**	−0.0887**
	GA		−0.0957**	−0.101**	−0.0981**	−0.1202**
Homophily in social relation-ships (HS)	C-Class		−167.5**	−123.8**	−122.3**	−166.3**
	B-Class		−141.8**	−117.2**	−117.1**	−101.2**
	GA		−99.3182**	−76.1286**	−75.2756**	−66.0605**
Structural Holes (SH)	C-Class			−464**	−468**	−987.2**
	B-Class			−280.2**	−278.5**	−271.3**
	GA			−166.3**	−166.8**	−271.4**
Functional diversity (FD)	C-Class				−0.0949*	−0.0939*
	B-Class				−0.036	−0.0346
	GA				−0.0704*	−0.0697*
Interaction SH and IC	C-Class					13,493.3
	B-Class					1100.8
	GA					2187.4
Interaction SH and EC	C-Class					−17.7127*
	B-Class					−12.6671
	GA					3.5161
Interaction SH and HS	C-Class					−44470.9**
	B-Class					103.4
	GA					−7014.7*

*p < 0.05; **p < 0.01.

Secondly, we provide an extensive list of metrics for quantifying different forms of social capital at a fine grained level. We provide a novel way of extracting these metrics using the edit history of Wikipedia articles. These metrics can be used in further research either separately or as a whole. Moreover they can be customized as per the needs of the situation. Kane [21] distinguished between minor and major edits and proved that both of them are instrumental to article quality. Our work delves deeper by identifying the different types of edits and quantifying their joint effect on quality. Our work is also the first to investigate the effect of structural holes on the success of Wikipedia articles.

Finally, the fact that the interaction between different forms of social capital has a profound effect on article quality should be of interest to researchers trying to develop theories on the success of open source systems. Putnam [7] defined social capital as the features of the organization and distinguished between bonding and bridging capital. Burt [72] argued that the interaction effect of these two forms of social capital can have a multiplier effect on the outcome. We combined these two forms with functional diversity and proved that it has an effect on article quality. In short our work provides a comprehensive framework for the examination of antecedents to the quality of user generated content in peer production environments.

5.2. Methodological contributions

This paper also makes three significant contributions on the methodological front. Firstly, we contribute to the extant literature on network analysis techniques by providing a method that can be used to compare the effects of two different types of networks in the same environment. Previous studies on network effects in open source environments [74,75] have used only one type of network. Our

conceptualization of local and overall collaboration networks can be extended to study the success of other open source environments such as blogs, tweets, and software. Moreover we also provide a way of examining the interaction effects of these two types of networks that can also be extended to other environments.

Secondly, we demonstrate how measures of social capital can be extracted from Wikipedia activity big data. We examined the implicit collaboration at the sentence level on each article and across articles. We also provide a novel way of operationalizing these metrics from theory which can be replicated by researchers working in the area of social networks.

Finally, our study reasserts the importance of single-mode networks in the study of open source environments. Recent studies in this area [13,74], have proved the usefulness of two-mode networks. While we acknowledge the importance of this method, our study proves that single mode networks can be studied more intensely by defining customizable metrics at a very fine grained level. This should provide a new method for researchers studying networks.

5.3. Managerial contributions

The findings of this paper should be of particular interest to managers in firms experimenting with Wikis to create knowledge repositories. We suggest two broad implications for them. As mentioned previously, more than 60% of the Fortune 1000 firms fail to create knowledge repositories with Wikis. The results of our study - which suggest that internal bonding has a significant positive effect on article quality - should provide clues for managers that collaboration within a team is an important antecedent to the quality of an article. Therefore managers controlling teams that are responsible for creating a certain knowledge repository should look for ways to motivate the contributors to collaborate effectively. Also the fact that external bridging is also important for article quality should make managers encourage their team members to establish ties with other teams that are also using wikis for knowledge creation. In addition, our results also emphasize the need to establish contacts with multifarious teams such that redundant information doesn't flow from the external contacts. Finally, given the fact that functional diversity has an important role to play in article quality, managers can either explicitly define the roles of team members in constructing an article or can let the team decide who will take the responsibility for each kind of edits such that a team contributing to an article will always have people assigned to various aspects of quality improvement. Also encouraging each of the contributors to play multiple roles in editing an article can prove to be instrumental for the success of Wikis.

Finally, the results of our study should also prompt firms to build tools that can be used for monitoring the social capital parameters in real time. Such a tool can help identify possible reasons for failure to build certain knowledge repositories such that corrective action can be taken. Moreover, such a tool can serve the purpose of automating quality of Wikipedia articles. Also such a tool can help firms compare the success of traditional and open source knowledge repositories.

6. Conclusion and future work

The interesting results of our study point to several areas of future research. Firstly, our study only investigates the effects of networks - among the contributors of an article and Wikipedia as a whole-on the article quality. However we do not know how the article evolves as it accumulates more social capital. Also it could be possible that social capital of the article co-evolves with the content and quality. An investigation into this process using a dynamic panel data set can lead to more searing insights into the evolution of social capital. This should prompt researchers to conduct a field study on the evolution of social capital in enterprise wikis. Such a study would also assess the impact of networks on various stages of article formation.

Secondly, while we have argued that grade assigned to an article is a reliable metric for the value of user generated content, it could be possible that we can use other metrics as a substitute for market value. Ransbotham et al. [13] used the monthly views as a proxy for value. Future research can use both these metrics in a multivariate regression to see which of the two is more reliable. Such a study would also un-earth the dynamics of interplay between quality and number of views.

Thirdly, interactions also happen on the talk pages of Wikipedia. Exploratory studies on the use of these talk pages [17,19,20] show that these pages are used for strategic planning of edits. It could be possible that social capital accumulates through these discussions. A study on how these discussions affect article quality can possibly unleash another dimension of social capital on Wikipedia. Such a study would also have implications for theory.

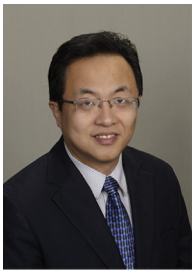
In conclusion, we provided a theoretical basis for investigating the quality of Wikipedia articles from the social capital perspective. We also added a new dimension to social capital on the basis of prior research on functional diversity. Our results show that each of the three forms of social capital has a profound effect on the quality of the article. Also it was found that internal bonding interacts positively with external bridging resulting in a multiplier effect on article quality.

References

[1] A. Neus, P. Scherf, Opening minds: cultural change with the introduction of open-source collaboration methods, IBM Syst. J. 44 (2) (2005) 215–225.
[2] J. Liu, S. Ram, Who does what: collaboration patterns in the wikipedia and their impact on article quality, ACM Trans. Manag. Inform. Syst. (TMIS) 2 (2) (2011) 11.
[3] O. Arazy, J. Daxenberger, H. Lifshitz-Assaf, O. Nov, I. Gurevych, Turbulent stability of emergent roles: the dualistic nature of self-organizing knowledge coproduction, Inf. Syst. Res. 27 (4) (2016) 792–812.
[4] F.X. Molina-Morales, M.T. Martínez-Fernández, Social networks: effects of social capital on firm innovation, J. Small Bus. Manag. 48 (2) (2010) 258–279.
[5] E. Bueno, M.P. Salmador, Ó. Rodríguez, The role of social capital in today's economy: empirical evidence and proposal of a new model of intellectual capital, J. Intellect. Cap. 5 (4) (2004) 556–574.
[6] M. Akdere, P.B. Roberts, Economics of social capital: implications for organizational performance, Adv. Develop. Hum. Resour. 10 (6) (2008) 802–816.
[7] R.D. Putnam, Bowling Alone: America's Declining Social Capital, The City Reader, 1995, pp. 120–128.

[8] P.V. Singh, Y. Tan, V. Mookerjee, Network Effects: the Influence of Structural Social Capital on Open Source Project Success, SSRN eLibrary, 2008. Retrieved from, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1111868.
[9] B. Stvilia, M.B. Twidale, L.C. Smith, L. Gasser, Assessing information quality of a community-based encyclopedia, in: Proceedings of the International Conference on Information Quality, vol. 11, 2005. Retrieved from, http://mailer.fsu.edu/~bstvilia/papers/quantWiki.pdf.
[10] M. Hu, E.P. Lim, A. Sun, H.W. Lauw, B.Q. Vuong, Measuring article quality in wikipedia: models and evaluation, in: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, 2007, pp. 243–252. Retrieved from, http://dl.acm.org/citation.cfm?id=1321476.
[11] J.E. Blumenstock, Size matters: word count as a measure of quality on wikipedia, in: Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 1095–1096. Retrieved from, http://dl.acm.org/citation.cfm?id=1367673.
[12] D.M. Wilkinson, B.A. Huberman, Assessing the Value of Coooperation in Wikipedia, 2007. Arxiv preprint cs/0702140. Retrieved from, http://arxiv.org/abs/cs/0702140.
[13] S. Ransbotham, G.C. Kane, N. Lurie, Network characteristics and the value of collaborative user-generated content, Market. Sci. 31 (3) (2012) 387–405.
[14] O. Arazy, O. Nov, Determinants of Wikipedia quality: the roles of global and local contribution inequality, in: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, 2010, pp. 233–236. Retrieved from, http://dl.acm.org/citation.cfm?id=1718963.
[15] S. Ransbotham, G.C. Kane, Membership turnover and collaboration success in online communities: explaining rises and falls from grace in Wikipedia, MIS Q. 35 (3) (2011) 613–627.
[16] X. Qin, P. Cunningham, Assessing the Quality of Wikipedia Pages Using Edit Longevity and Contributor Centrality, 2012. Arxiv preprint arXiv:1206.2517. Retrieved from, http://arxiv.org/abs/1206.2517.
[17] F.B. Viegas, M. Wattenberg, J. Kriss, F. Van Ham, Talk before you type: coordination in wikipedia. System sciences, 2007, in: HICSS 2007. 40th Annual Hawaii International Conference on, 2007, 78–78.
[18] A. Kittur, R.E. Kraut, Harnessing the wisdom of crowds in wikipedia: quality through coordination, in: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, 2008, pp. 37–46.
[19] J. Schneider, A. Passant, J.G. Breslin, A Content Analysis: How Wikipedia Talk Pages Are Used, 2010.
[20] B. Stvilia, M.B. Twidale, L.C. Smith, L. Gasser, Information quality work organization in Wikipedia, J. Am. Soc. Inf. Sci. Technol. 59 (6) (2008) 983–1001.
[21] G.C. Kane, A multimethod study of information quality in wiki collaboration, ACM Trans. Manag. Inform. Syst. (TMIS) 2 (1) (2011) 4.
[22] O. Arazy, L. Yeo, O. Nov, Stay on the Wikipedia task: when task-related disagreements slip into personal and procedural conflicts, J. Assoc. Inform. Sci. Technol. 64 (8) (2013) 1634–1648.
[23] G.C. Kane, S. Ransbotham, Research Note—content and collaboration: an affiliation network approach to information quality in online peer production communities, Inf. Syst. Res. 27 (2) (2016) 424–439.
[24] B. de La Robertie, Y. Pitarch, O. Teste, Measuring article quality in wikipedia using the collaboration network, in: Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on, IEEE, 2015, pp. 464–471.
[25] R.S. Burt, The social capital of structural holes, New Econ. Sociol. Develop. Emerg. Field (2002) 148–190.
[26] H. Oh, M.H. Chung, G. Labianca, Group social capital and group effectiveness: the role of informal socializing ties, Acad. Manag. J. (2004) 860–875.
[27] W.E. Baker, A.V. Iyer, Information networks and market behavior, J. Math. Sociol. 16 (4) (1992) 305–332.
[28] J.S. Coleman, Social capital in the creation of human capital, Am. J. Sociol. (1988) 95–120.
[29] M. Granovetter, Economic action and social structure: the problem of embeddedness, Read. Econ. Sociol. (1985) 63–68.
[30] E. Rosenthal, Social networks and team performance, Team Perform. Manag. 3 (4) (1997) 288–294.
[31] J.A.C. Baum, T. Calabrese, B.S. Silverman, Don't go it alone: alliance network composition and startups' performance in Canadian biotechnology, Strat. Manag. J. 21 (3) (2000) 267–294.
[32] K.S. Cook, R.M. Emerson, M.R. Gillmore, T. Yamagishi, The distribution of power in exchange networks: theory and experimental results, Am. J. Sociol. (1983) 275–305.
[33] R. Emerson, K. Cook, Power, equity and commitment in exchange networks, Am. Sociol. Rev. 43 (5) (1978) 721–739.
[34] B. Markovsky, D. Willer, T. Patton, Power relations in exchange networks, Am. Sociol. Rev. (1988) 220–236.
[35] R. Reagans, E.W. Zuckerman, Networks, diversity, and productivity: the social capital of corporate R&D teams, Organ. Sci. (2001) 502–517.
[36] J.H. Dyer, K. Nobeoka, Creating and managing a high-performance knowledge-sharing network: the Toyota case, Strat. Manag. J. 21 (3) (2000) 345–367.
[37] L. Festinger, Social Pressures in Informal Groups: a Study of Human Factors in Housing, Stanford University Press, 1963.
[38] J.S. Coleman, Foundations of Social Theory, Belknap Press, 1994.
[39] P.V. Singh, The small-world effect: the influence of macro-level properties of developer collaboration networks on open-source project success, ACM Trans. Softw. Eng. Meth. 20 (2) (2010) 6.
[40] J. Xu, Y. Gao, S. Christley, G. Madey, A topological analysis of the open souce software development community, in: System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on, 2005 (p. 198a–198a). Retrieved from, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1385642.
[41] M. Mann, The Sources of Social Power: a History of Power from the Beginning to AD 1760, vol. 1, Cambridge University Press, 1986.
[42] W.E. Baker, The social structure of a national securities market, Am. J. Sociol. (1984) 775–811.
[43] M. Fraser, S. Dutta, Throwing Sheep in the Boardroom: How Online Social Networking Will Transform Your Life, Work and World, John Wiley & Sons Inc, 2008.
[44] P. Bourdieu, Sociology in Question, vol. 18, Sage Publications Ltd, 1993.
[45] Y.C. Yuan, G. Gay, Homophily of network ties and bonding and bridging social capital in computer-mediated distributed teams, J. Comp. Med. Commun. 11 (4) (2006) 1062–1084.
[46] J.S. Brown, P. Duguid, Organizational learning and communities-of-practice: toward a unified view of working, learning, and innovation, Organ. Sci. (1991) 40–57.
[47] M. Gargiulo, M. Benassi, Trapped in your own net? Network cohesion, structural holes, and the adaptation of social capital, Organ. Sci. (2000) 183–196.
[48] S.P. Borgatti, Centrality and network flow, Soc. Network. 27 (1) (2005) 55–71.
[49] L. Leydesdorff, Betweenness centrality as an indicator of the interdisciplinarity of scientific journals, J. Am. Soc. Inf. Sci. Technol. 58 (9) (2007) 1303–1319.
[50] R.S. Burt, The social capital of opinion leaders, Ann. Am. Acad. Polit. Soc. Sci. 566 (1) (1999) 37–54.
[51] M.S. Mizruchi, L.B. Stearns, Getting deals done: the use of social networks in bank decision-making, Am. Sociol. Rev. (2001) 647–671.
[52] J.M. Podolny, Networks as the pipes and prisms of the Market1, Am. J. Sociol. 107 (1) (2001) 33–60.
[53] A. Nerkar, S. Paruchuri, Evolution of R&D capabilities: the role of knowledge networks within a firm, Manag. Sci. (2005) 771–785.
[54] M. Olson, The Logic of Collective Action: Public Goods and the Theory of Groups, Harvard University Press, Cambridge, MA, 1965.
[55] R.S. Burt, The network structure of social capital, Res. Organ. Behav. 22 (2000) 345–423.
[56] W.T. Bielby, D.D. Bielby, Organizational mediation of project-based labor markets: talent agencies and the careers of screenwriters, Am. Sociol. Rev. (1999) 64–85.
[57] A. Lih, 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, 2004.
[58] J.S. Bunderson, K.M. Sutcliffe, Comparing alternative conceptualizations of functional diversity in management teams: process and performance effects, Acad. Manag. J. (2002) 875–893.
[59] L. Peters, R.J. Karren, An examination of the roles of trust and functional diversity on virtual team performance ratings, Group Organ. Manag. 34 (4) (2009) 479–504.
[60] S.L. Keck, M.L. Tushman, Environmental and organizational context and executive team structure, Acad. Manag. J. (1993) 1314–1344.
[61] J.N. Cummings, Work groups, structural diversity, and knowledge sharing in a global organization, Manag. Sci. (2004) 352–364.

[62] T. Adler, L. de Alfaro, A Content-driven Reputation System for the Wikipedia, Technical Report ucsc-crl-06-18, School of Engineering, University of California, Santa Cruz, 2006. Available at: http://works.bepress.com/luca_de_alfaro/3.
[63] G.C. Kane, S. Ransbotham, Collaborative development in wikipedia, 2012. Arxiv preprint arXiv:1204.3352. Retrieved from, http://arxiv.org/abs/1204.3352.
[64] B. Ronald, Structural Holes: the Social Structure of Competition, Harvard, Cambridge, 1992.
[65] R. Reagans, E. Zuckerman, B. McEvily, How to make the team: social networks vs. demography as criteria for designing effective teams, Adm. Sci. Q. 49 (1) (2004) 101–133.
[66] O. Arazy, E. Stroulia, S. Ruecker, C. Arias, C. Fiorentino, V. Ganev, T. Yau, Recognizing contributions in wikis: authorship categories, algorithms, and visualizations, J. Am. Soc. Inf. Sci. Technol. 61 (6) (2010) 1166–1179.
[67] P. Ball, The more, the wikier, Nature 27 (2007).
[68] D.A. Levinthal, J.G. March, The myopia of learning, Strat. Manag. J. 14 (S2) (1993) 95–112.
[69] A. Kittur, B. Suh, B.A. Pendleton, E.H. Chi, He says, she says: conflict and coordination in Wikipedia, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2007, pp. 453–462.
[70] E.A. Smith, R.J. Senter, Automated Readability Index. CINCINNATI UNIV OHIO, 1967. Retrieved from, http://www.stormingmedia.us/37/3727/0372766.html.
[71] C. Okoli, W. Oh, Investigating recognition-based performance in an open content community: a social capital perspective, Inf. Manag. 44 (3) (2007) 240–252.
[72] R.S. Burt, Structural holes versus network closure as social capital, Social Capital. Theory Res. (2001) 31–56.
[73] S. Oreg, O. Nov, Exploring motivations for contributing to open source initiatives: the roles of contribution context and personal values, Comput. Hum. Behav. 24 (5) (2008) 2055–2073.
[74] R. Grewal, G.L. Lilien, G. Mallapragada, Location, location, location: how network embeddedness affects project success in open source systems, Manag. Sci. (2006) 1043–1056.
[75] W. Oh, S. Jeon, Membership herding and network stability in the open source community: the Ising perspective, Manag. Sci. 53 (7) (2007) 1086–1101.
[76] J. Giles, Internet Encyclopaedias go head to head, Nature (2005) 900–901.
[77] L. Rector, Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles, Ref. Serv. Rev. 36 (1) (2008) 7–22.



Jun Liu is Assistant Professor of Information Systems in the College of Business & Information Systems Management at Dakota State University. He received a Ph.D. in Management (MIS) from the University of Arizona. His research interest include is data management, Big Data analytics, social network analysis, and healthcare data analytics. He has published research papers in journals such as Journal of Database Management, ACM Transactions on MIS, Information Systems Frontiers, etc.

Sudha Ram is Anheuser-Busch Endowed Professor of MIS, and Entrepreneurship & Innovation in the Eller College of Management at the University of Arizona. She has joint faculty appointment as Professor of Computer Science. She is the director of the Advanced Database Research Group (ADRG) and co-director of INSITE: Center for Business Intelligence and Analytics (www.insiteua.org) at the University of Arizona. Dr. Ram received a Ph.D. from the University of Illinois at Urbana-Champaign in 1985. Her research is in the areas of Enterprise Data Management, Business Intelligence, Large Scale Networks and Data Analytics. Her work uses different methods such as machine learning, statistical approaches, ontologies and conceptual modeling. Dr. Ram has published articles in such journals as Communications of the ACM, IEEE Intelligent Systems, IEEE Transactions on Knowledge and Data Engineering, Information Systems, Information Systems Research, Management Science, and MIS Quarterly. Her research has been highlighted in several media outlets including NPR news, She was a speaker for a TED talk in December 2013 on “Creating a Smarter World with Big Data”.