The impact of initial group characteristics on quality in online communities of creation

AMIRA REZGUI, Syracuse University
NICOLAS JULLIEN, LEGO-Marsouin, IMT Atlantique
KEVIN CROWSTON, Syracuse University

Open projects aimed at creating new knowledge (also known as online communities of creation) are increasingly central in the production of new and innovative knowledge. In our research study, we are interested in the impact of initial group characteristics on the quality of the output. We studied in particular Wikipedia in different three languages: Arabic, Romanian and Thai. Our results confirm the importance of the initial project group. We found a positive impact of a large initial group formed by members with an intermediate level of diversity in the focus of their editing but having an equal and lower level of longevity.

CCS Concepts: • **Human-centered computing → Computer supported cooperative work**; **Empirical studies in collaborative and social computing**;

Additional Key Words and Phrases: Online communities of creation, group diversity, group longevity, contribution, reputation.

ACM Reference format:
Amira Rezgui, Nicolas Jullien, and Kevin Crowston. 2019. The impact of initial group characteristics on quality in online communities of creation. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 111 (November 2019), 21 pages.
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Understanding how group outcomes are affected by group members' diversity (i.e., the extent to which group members share or have different attributes) is a question of growing importance for organizations. In particular, research in traditional organizations has linked group characteristics and member diversity to group performance [23]. The impact of member diversity on performance is equally important for groups engaged in virtual cooperation for knowledge production [22, 48, 56]. However, in this setting, where interaction is mediated by collaborative technologies, the negative effects of difference in some individual factors may be curtailed because the technologies can provide visual anonymity and equal opportunity to participate. Similarly, their positive effects may be enhanced because of additive capabilities such as coordination support and information technologies.

A distinctive feature of many online groups is their self-organizing nature. Unlike work groups or even virtual groups in organizations whose memberships are determined by organizational design and managerial oversight, group composition in self-organized online groups is driven by members' voluntary participation. Therefore, group composition will depend on how members attract other,

2573-0142/2019/11-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

which can in turn affect how effective their collaborative efforts will be. As a result, the diversity of the members of the group in the early stages of the collaboration may be particularly important to attract new members and thus start a self-reinforcing cycle of growth.

In this paper we study the impact of diversity of an initial group in the setting of Wikipedia, chosen as a prominent example of an online epistemic community or 'community of creation' [52]. We consider as a group the editors who have contributed to a particular article and as the outcome, the article itself. Studies of Wikipedia have shown a link from group characteristics to article quality. We extend this work by considering the impact of the initial group. Based on findings about group composition from prior research, we develop a set of theoretical propositions about the impact of composition and diversity of the initial group of editors. As initial group characteristics are not visible in the history of contribution to a particular article, a novel feature of our theorizing is that we examine the contribution of the contributors in their careers working on other articles before the creation of the article in question.

2 LITERATURE REVIEW

2.1 Group Formation

Group formation is a pervasive social phenomenon and the mechanisms through which groups form have been a subject of interest in diverse fields. Social psychology and sociology have examined mechanisms determining the formation of naturally-occurring groups within organizations [51]. groups and other organizational structures fall along a continuum between institutional formality and informal emergence [1]. Strategically, a project group in organizations is typically formed by a manager who assigns individuals to the group based on their characteristics. In contrast, in case of online communities of creation like Wikipedia, groups that produce knowledge (i.e., an article) self-assemble. Self-assembly is a process in which a disordered system with pre-existing components forms an organized structure with interaction among the components, without external direction.

In this context, [44] indicated that group formation is a result of a strategic decision of individuals who seek to find and join a group that will satisfy their needs. For example, editors in Wikipedia are motivated to choose articles and collaborators that provide the possibility to reach the intended benefits of participation. However, it is difficult to predict the success of the future project because of the uncertainty in judging the work style of group members and the true objective quality of the project. To deal with this uncertainty, initial members may rely on others' experience (long or short) to have an idea about the likely collaboration success. For instance, previous experience provides some hints about the capabilities and skills possessed by others and how much new members can bring new and fresh ideas. Consequently, editors can assess others' work quality and the risks involved in working with existent or new editors.

2.2 Article Lifecycle and Initial Members

Articles in Wikipedia do not evolve linearly. Kittur and Kraut [31] found that articles go through different phases, and they explicitly demonstrated that in the initial phases, in the infancy of the article, "data suggest that it is important to have a small number of contributors setting the direction, structure, and scope of the article" (p. 44). As with open source software [14, 58], it can be said that the initiators of such projects are its 'natural' core members. Therefore, studying the structure from the initial stage can give an idea about core members. Kane et al. [29] went a step further, describing a three-phase life cycle for an article, the first one being the grouping period, where members start contributing in a non-coordinated way, but from which some of these initial members will emerge and become core in their volume of contribution.

2.3 Core vs. Peripheral Group Members

A complication in defining an initial group for a Wikipedia article is the great diversity in the volume of contributions from different editors. There is an extensive body of literature that describes the core-periphery structure in online communities (specifically, peer-production communities) [4]. This structure is characterized by a dense, cohesive core and a sparse, unconnected periphery [6]. The majority of contributors situated at the community's periphery are not very active and involved in only a few tasks, while a small portion of contributors (the core members) are more active and take additional responsibilities [5, 6, 14, 15, 18, 41, 52]. The conceptualization of this core-periphery structure is concentrated on participants' power within the community constituting the core in term of activities and contributions [41].

Crowston and Shamshurin [14] mentioned three approaches to identify the core group: self-report, e.g., on a project website, comparatively high level of contribution, and core-and-periphery social network analysis. The first approach is simplest, where the core is defined as those individuals who officially designate themselves as core, and the periphery are all other contributors. The second approach is defined from the amount of contributions. Online projects typically demonstrate a very skewed distribution in terms of levels of contribution: a few members contribute a lot (core members) while the others contribute only a little (the periphery). By analogy with [7], Crowston and Shamshurin [14] defined the core group as the members who contribute the most such that the sum of their contributions constitute 1/3 or more of the total number of contribution. In the third approach, the core is defined from the social interactions. According to [6], a core-and-periphery network "entails a dense, cohesive core and a sparse, unconnected periphery". group members can be partitioned into the core defined as a very interconnected group and the periphery as the disconnected group. Crowston and Shamshurin [14] found that these three approaches resulted in very similar cores and suggested that future research can use the simplest approach.

In Wikipedia too, there is a core-periphery structure. Prior research in Wikipedia has described core members' duties [10, 45], the quantity and types of activities they perform [34, 38], factors affecting a participant's promotion to a leadership role [9, 36], as well as on the relationship between editor's position within the social network and promotion decisions [13]. Some research has expanded to the two levels of core and periphery by describing different roles that editors can fill in producing an article (as distinct from the formal roles defined by the Wikipedia project) [3, 5, 17]. These roles are characterized for example by different levels and types of contributions [3]. Note that vandals (a negative but not uncommon role) typically have low edit counts, meaning that they will be members of the periphery. And other work suggests that all these roles are needed for an article to be of quality [29].

Taken together, this research suggests that the core members of a group will be most determinative of the quality of the article, with the peripheral members playing an important but secondary role. Further, editors may fill different kinds of roles in an article, each making an important contribution.

2.4 Wikipedia Article Quality

In our study, we consider how characteristics of a group affect the group's outcome. We assess the quality of the outcome according to the criteria of the group itself (i.e., does the production meets the standards expected by the producers). Specifically, we rely on Wikipedia's own article rating system, which has rules to define different quality levels[1]. While there are variations between different languages, one can distinguish three main classes of articles: basically completed articles, labeled "good" or "featured" articles in the English Wikipedia, articles with content but which need improvement (labeled "B" or "C" articles) and articles that are not yet satisfactory (labeled

[1]https://en.wikipedia.org/wiki/Wikipedia:Content_assessment.

"starting" or "stub" articles). There is evidence that the internally-applied quality class matches independent quality assessments made by outside reviewers [31] (i.e., that the quality scores are valid as a measure of quality).

2.5 Predictors of Article Quality

We next consider what characteristics of a group predict a good group outcome. Considering groups more generally, [57] provided a recent review of the most prevalent effectiveness models using an international perspective. This research has detailed in particular the characteristics of individual members that are predictive of effectiveness, such as competence, experience, orientation to the task or personality.

Considering Wikipedia more specifically, prior research has identified a variety of factors that lead to a high-quality article. A simple approach simply considers the volume of inputs, e.g., "featured" articles have been found to have more editors and more edits than non-featured [60]. Of course, it is well known that not all edits are a contribution to the article: some are deliberate vandalism and others are simply not helpful.

Other research considers differences among contributing editors, such as editor experience. For instance, studies have found a positive impact on quality of having contributions from editors who have experience from contributing to articles in different categories and also with formal administrative positions in Wikipedia [4]. By exploring German Wikipedia, [53] showed that high-quality articles are not necessarily written by a large number of people, but the most important ones are written by contributors with a reputation for high-quality contributions. Most of the time, as showed by [25], this reputation, or what these authors called the 'authority' of the contributors, comes from the fact that they have contributed to high-quality articles in the past. It is worth noting that Wikipedians who have contributed to featured / good articles often note these contribution on their personal pages [35]), and many will keep contributing only to good articles [40].

Another important factor is the group's experience working together. For instance, Carillo and Okoli [11] found a positive impact of editors' prior interactions, as well as their number (though interaction is not defined in the article). Liu and Ram [39] found positive impacts of the editors' social capital as derived from internal bonding (i.e., having worked together before) as well as the previously mentioned external bridging (i.e., working on diverse other articles) and group size.

2.6 Summary

The research on group formation suggests that the characteristics of the current membership of a group can impact who subsequently chooses to join the group. In the case of online groups though, the characteristics of the core members (those who make the bulk of the contributions) are expected to have more of an impact. For Wikipedia, where there is not a defined list of group members, the initial group can be defined as those active in the formative stages of an article. Finally, research on Wikipedia in particular has shown a connection between the characteristics of the editors and article quality, as well as providing tools to measure quality. This paper will extend this work by examining specifically the impact of initial core group characteristics and quality.

3 HYPOTHESIS DEVELOPMENT

In this section we develop the specific hypotheses that we will test in the paper. We start by defining more specifically who we consider as the initial core group members. We then develop a set of hypotheses linking the characteristics of those members and the resulting article quality.

3.1 Defining Initial Core Members

As noted above, the initial members of a Wikipedia article can be defined as those who contribute in the formative stages of an article [29, 31]. Specifically, we propose to describe the an article' life-cycle as shown in Figure 1. The first contribution is the time an editor proposes the article as a new and original knowledge product. The period after the first contribution can be divided into 2 sub-phases. The initial "teaming" phase is the period after the first contribution during which an initial set of group members self-select themselves and start the project. We call those members who contribute actively in this period the initial core members. The phase after the grouping period we call the "active" period, during which the project starts to be more and more active and productive. The rates of contribution, contributors and efficient creation increase and the article quality improves, up to the best quality (FA or GA) for the best articles.

We hypothesis that the shift from the grouping period to the active period can be determined empirically by looking for a transition in the pattern of contributions. Inspired by the theory of innovation diffusion by [50], by research on open-source software production organization [28, 33], and by the articles already mentioned about Wikipedia, we defined the transition between phase 1 and phase 2 as the moment when the volume of production accelerates. Mathematically speaking, that means when the the second derivative turns positive (the first inflection point).

Finally, we name the period before the creation of an article the "learning" period. It is the period during which the core members of the grouping period may have learned to edit Wikipedia articles through their engagement with other articles. We collect information about the core members' activities during this period that might predict the long-term success of the article.

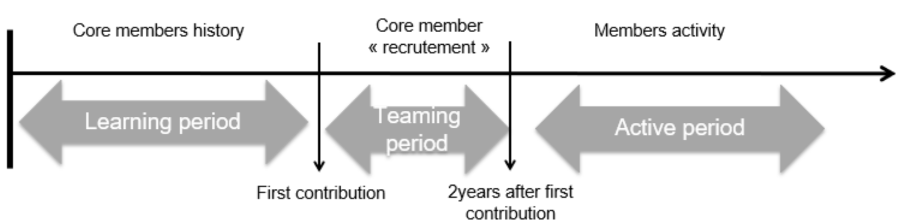


Fig. 1. Article life cycle periods

3.2 Effects of Initial Core Member Characteristics on the Future Quality of the Article

In this section, we develop a series of hypotheses about the effects of initial core member group characteristics on the quality of the Wikipedia articles they edit. As noted in the literature review above, prior work has identified a range of characteristics that are believed to be relevant to group effectiveness. However, as we are following the idea that editors are self-assembling groups based on their observation of initial members, we focus on characteristics that will be observable by the participants as they start to work together on the article (i.e., activities from the learning period in Figure 3.1). (We assess quality at the point of data collection, i.e., at the current end of the active period.)

Given this focus, we chose from Verhoeven et al. [57] variables that can be identified in the past activity of the core members to build our hypotheses, specifically longevity or tenure as a Wikipedia editor (i.e., how long the editors have been active) [48], past effort (i.e., how many edits the editors have made) and the proportion of contributions made specifically to good or featured articles, which we label as "reputation". Tenure and past effort are interesting both as a proxy for experience and as an indication that the editor is likely to know how the rules are built and negotiated. Reputation indicates both experience in building an effective project as well as being an indication of position within the Wikipedia community more generally [40]. For each of these factors, we consider the effect of both the overall average of the group members and the diversity of the level across the initial core group.

3.2.1 H1: Longevity. The first question to be addressed is whether it is better to bring together a group of old-timers or newcomers to build a good article. On one hand, evidence from virtual organizations has shown that although old-timers are more experienced and skilled than newcomers, their effort is generally lower [59]. On the other hand, newcomers may discover a new article by chance, but more often, they edit because the topic is in the center of their interest [27], so can contribute their knowledge of the field. However, they have little experience with the community, meaning that they may not know how a Wikipedia article has to be written and how to coordinate the work. Overall, we expect that the advantages more experienced editors will win out. We therefore hypothesize that:

H1.1: There is a positive relationship between average longevity of the initial core group of editors during the learning period and article quality.

We next consider the impact of having editors with different levels of editing experience and familiarity with the Wikipedia environment, its policies and its rules [20]. Differences are likely to result in varied attitudes and opinions about how and what to contribute to a particular article. Old-timers and newcomers may have different views, for example, on article scope and interpretations of Wikipedia policies. Overall, we expect diversity in longevity to reduce communication and social integration and to increase conflicts [61]. Longevity diversity may also have a negative impact on day-to-day collaboration [47, 47, 49]. If an article receives contributions with diverse styles and approaches from editors with diverse experiences, intensive coordination would be needed bridge these differences [47]. We therefore hypothesize that:

H1.2: There is a negative relationship between the diversity of longevity of the initial core group during the learning period and article quality.

3.2.2 H2: Past Efforts. In Wikipedia, a classic proxy for modeling the effort of the contributors is the number of edits they have made (see the discussion by [19]). Similarly, one of the best proxies for article quality is the number of edits [42], as it is for editor's impact [46]. Previous research on online collaboration has suggested that having a core group of active participants who outline the task structures can benefit task coordination and motivate contribution from peripheral members [16]. Simple counts are clearly over-simplification of effort, as edits are not all equal: there might be hours of work behind an edit to collect information and structure a new contribution or just a few seconds, e.g., to correct a typo. Indeed, an edit can be of poor quality, even an act of vandalism, and consequently be a poor guarantee of the benefit the editor will bring to the project. Unfortunately, it is difficult to know exactly how much work an edit represents (we discuss vandalism below). As a consequence, we retained the number of edits as the best trade-off between information provision and simplicity of understanding to measure the involvement and activity of the editors. We therefore hypothesize that:

H2.1: There is a positive relationship between the average numbers of edits by the initial core group during the learning period and article quality.

Effort diversity indicates differences in previous contributions. We take these differences as indications that there are editors with different patterns of contribution. As noted in the review above, these patterns can be interpreted as different roles in the creation of articles, which are important for article quality. We therefore hypothesize that:

H2.2: There is a positive relationship between diversity of numbers of edits by the initial core group during the learning period and article quality.

3.2.3 H3: Reputation. We noted above the importance of the experience of editors in contributing to high-quality articles, which we view as a proxy for the reputation of the editor. We therefore hypothesize that:

H3.1: There is a positive relationship between average reputation of the initial core group during the learning period and article quality.

On the other hand, Anthony et al. [2] found that the highest-quality contributions come from the vast numbers of editors who contribute infrequently. So having some users with a lower reputation level may be a good thing. Consequently, we hypothesize that:

H3.2: There is a non-monotonic relationship between diversity of reputation of the initial core group during the learning period and article quality: articles with an intermediate reputation diversity will have the highest quality.

The hypotheses of our study are summarized in Table 1.

Table 1. Hypothesized relationship between core members characteristics and article quality in Wikipedia.

	Variable	Hypotheses
H1: Longevity	average	H1.1: positively related to article quality
	diversity	H1.2: negatively related to article quality.
H2: Previous effort	average	H2.1: positively related to article quality
	diversity	H2.2: positively related to article quality
H3: Reputation	average	H3.1: positively related to article quality
	diversity	H3.2: non-monotonic relationship (intermediate diversity is better)

4 METHODOLOGY

4.1 Data Gathering

In this paper, we examine articles in three different Wikipedias from 2004 to 2016. The unit of analysis for the study is the article. In particular, we analyzed a sample of articles from 3 different language Wikipedias: Arabic, Romanian and Thai. The three languages were chosen to increase the culturally diversity of the sample (3 regions of the world, 3 alphabets) to avoid the cultural specificity that studying just one language might create (e.g., just English). By examining articles from particular languages separately and comparing across them, our methodology could be used to examine cultural specificity, though that is a topic for future research (we have included in an Appendix the differences between languages).

To create a balanced data set for analysis, we preferentially sampled the quality articles ("Featured Articles" and "Good Articles", approximately 1% of the total number) and we randomly sampled the same number of non-quality (B and C) articles. We used the MediaWiki API to retrieve article quality. The data our study is balanced between the three different language: 448 Arabic, 500 Romanian and 500 Thai.

For each language version processed, we retrieved the "*pages-meta-history.xml.7z*" XML dump file from the Wikimedia download website. It contains for every wiki page in that language version, the complete information about every single revision. We processed the XML records using WikiDat. WikiDAT [2] is a tool for Wikipedia data analytics, based on Python and R [43]. It is aimed at creating an extensible toolkit for Wikipedia and to automate the extraction of Wikipedia data. We stored the data into 5 different tables of a MySQL database (page, people, revision, revision hash, logging).

[2]https://github.com/glimmerphoenix/WikiDAT

4.2 Determining Timing of Article Phases

In section 3.1 we presented a three-phase model of the life-cycle of a Wikipedia article. A first step in the research was to empirically determine the timing for the initial phase of the article, the teaming period. In principle, the timing could be determined for each article separately, but we instead computed it overall based on the average across projects.

We defined the point at which articles moved from one phase to another by looking for the point at which there was a change in the growth rate of project characteristics, specifically number of contributions, contribution size and quality (see Figure 2). Those three variables give us information about editors' activities. The intuition is that the initial period for the article is characterized by a certain rate of growth of these variables, and that the period ends when the growth rate changes.

A breakpoint occurs in a continuous function when the concavity of the graph changes at that point. So, we need to identify where a curve goes from concave upward to concave downward (or vice versa). Mathematically, the breakpoints showing an acceleration of an effect are determined when the sign of the second derivative of the function changes from negative to positive. To control for noise in the data (local perturbation such as the variation of contributions during summer), we took the average by month and we smoothed our data using the Loess function in R statistical language (a technique used to detect trends in noisy data). The detected inflection point in the smoothed data is shown in the curve given by R (the blue circle as can be seen in Figure 2; we added the red line as a separator to make the breakpoints more visible). According to graphs seen in Figure 2, the first breakpoints appear after around 24 months. Therefore, we fixed the teaming period (i.e., the period used to pick initial members) as the 2 years after the first contribution.

In choosing the timing of the teaming period, we opted for an easily described round number rather than an overly precise number. The goal of identifying the teaming period is simply to identify the initial core group members. Changing the duration by a small amount (e.g., adding another month) would likely result in the entire initial group gaining a few members (those who contributed only in the added period). However, the core group, those who contribute the majority of the edits, will likely stay nearly the same because it is unlikely that many (or even any) editors will have contributed enough during the added period to move in or out of the core.

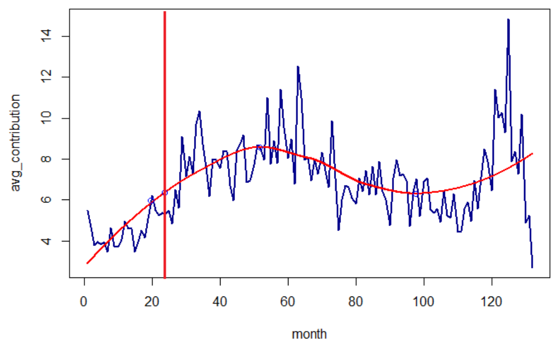
4.3 Defining the Core Members

Having defined the initial phase of the article life cycle, we next identify the initial core members, the "small number of contributors" whose participation matters in the initial phase [31]. To identify core members, we adopted the second of the the three approaches to identify a core group reviewed above [14]. Specifically, we defined the core members as the editors who contributed the most edits and whose sum total edits amounts to 1/3 or more of the total number of edits made during the initial phase. As contributions are unequal, the core is much smaller than 1/3 of the total number of editors. The first approach proposed by Crowston and Shamshurin [14] was not feasible since Wikipedia editors do not officially designate themselves as core members and the third was not feasible because of the lack of a defined communications network for the editors. Further, the chosen approach is coherent with Kittur and Kraut [31]'s findings. We note that because vandals tend to have low edit counts [30, 62], they are quite unlikely to be included in the core.

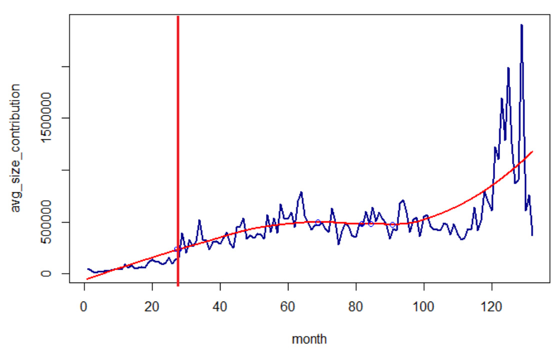
4.4 Operationalization of the Variables

4.4.1 Dependent Variable. The dependent variable in our study is article quality assessed at the end of the study period. To measure the quality of Wikipedia articles, we used each article's class as assessed by the project's own community through a peer-review process. This measure is Wikipedia's internal quality categorization schema and has been used as a measure of quality

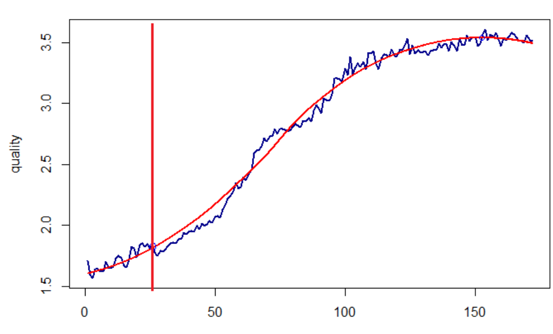
Fig. 2. The detection of the first breakpoints in the evolution of the averages in different characteristics related to both FA and GA articles (months are ordered from the first one till the last one)



(a) Number of contributions per month



(b) Contribution size per month



(c) Quality per month

in many Wikipedia studies [4, 26, 31, 32, 37, 38, 54, 60]. As said before, there are six classes, or quality levels, to which articles can be assigned. In descending order of quality, the article classes are "Featured Article" (FA), "Good Article" (GA), B, C, Start and Stub. Articles in the classes Start and Stub contain very little information and have very few editors. Because we were interested in studying editors' behaviors, we dropped articles in these categories from the analysis and kept only FA, GA, B and C articles. After this filtering, we grouped articles into two categories, FA and GA ($quality = 1$) versus B and C ($quality = 0$).

4.4.2 Independent Variables. The independent variables of our study are group size and diversity of longevity (time experience), previous effort and reputation, of the 1/3 most active editors of the studied article during the learning period. For each core group member, we computed individual-level metrics using the history of their revisions on other articles before the creation date of the article in question. As a result, if an editor is a core member for two articles in our sample, her data will not be the same unless these two articles were created at the same moment. Finally, we computed the article-level metrics (averages and diversity) for each article from the individual-level measures. Table 2 explains how these variables were constructed.

Table 2. Operationalization of variables

Variables	**Individual-level metric**	**Project-level metric**
Longevity	The number of days elapsed from a member's first edit in Wikipedia to the most recent revision [12].	Average and variation coefficient of the distribution of longevity of editors [21].
Previous effort	The length of edits made by each editors before current article	Average and variation coefficient of the distribution of the length of edits of editors [21].
Reputation	We compute the reputation of authors based on their contribution to excellent pages (# excellent participation). We then compute the rating of a page based on the reputation of the contributing authors [53]. $rep_{pb} = \frac{\#excellentparticipation}{\#participation}$	Average and variation coefficient of the distribution of editors' reputation [21].

4.4.3 Control Variable. We included one control variable in our study, namely group size. Prior research has shown that group size impacts quality in two ways. Large groups ensure a large set of opinions and knowledge and a faster time to correct errors and discover incomplete information [55]. In addition, small groups often lack the resources that large groups can extend, which makes it difficult to give additional resources to produce an article within Wikipedia as a collective action [63]. Group size is defined as the total number core editors who have contributed to the article during the teaming period.

5 RESULTS

5.1 Descriptive Statistics and Data Exploration

We start by presenting some basic statistics and exploratory plots of the data. Table 3 presents the descriptive statistics using the median, the mean and the standard deviation of all the variables. Fig. 4 reports the correlation between these variables. Note that the length of contributions and reputation have low correlations with other variables but group size is negatively correlated with average experience, which in turn is negatively correlated with the various measures of variation. One interpretation of these results is that when average experience is low, the group seems to be less homogeneous on other measures. We will return to the question of colinearity when we analyze the data.

Table 3. Descriptive statistics of the variables used in our study to describe the core member groups of our articles.

Statistic	Median	Mean	Standard deviation
Average previous effort	1.167e+07	5.513e+0	1.443637e+08
Average reputation	66.11	93.50	94.5
Average longevity	2542	2711	1104
Previous effort diversity	1.616	1.534	0.802
Reputation diversity	1.3488	1.3215	0.755
Longevity diversity	0.8188	0.8193	
Group size	11	12.55	10.0
Article quality	Count		
Quality article (1)	748		
Non-quality article (0)	700		

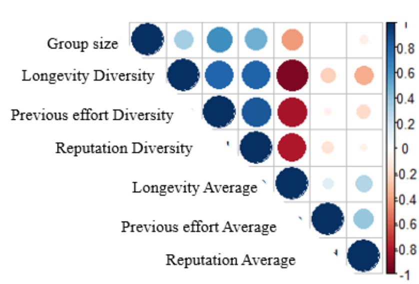


Fig. 4. Correlation among independent variables

Before starting the analysis, we wanted to assess whether the variables were related to quality in a monotonic way. For instance, considering the group size variable, is it the bigger the group, the better, or is a larger group less effective in terms of producing an article of good quality (perhaps coordination problems appear beyond a certain size)? We therefore carried out a visual inspection of the variables. Figures 5, 6 and 7 show plots of the variables with the quality of the article as the y-axis (articles of quality 1 are at the top of the graph, and articles of quality 0 at the bottom). The blue crosses represents the average quality level at different steps of the variables.

Figure 5 shows that most of the articles have a of group size under 40 (as shown by the distribution of black dots). The plot shows a near monotonic relationship between article quality and (core) group size that plateaus by about 40 editors (though there aren't many articles with more than that many editors). Looking at Figure 6, longevity diversity seems to present a nearly monotonic relationship with article quality, but a negative one. However reputation diversity is visually non-monotonic with article quality. Finally, Figure 7 shows that the averages of the variables do not seem to have a clear relationship with quality.

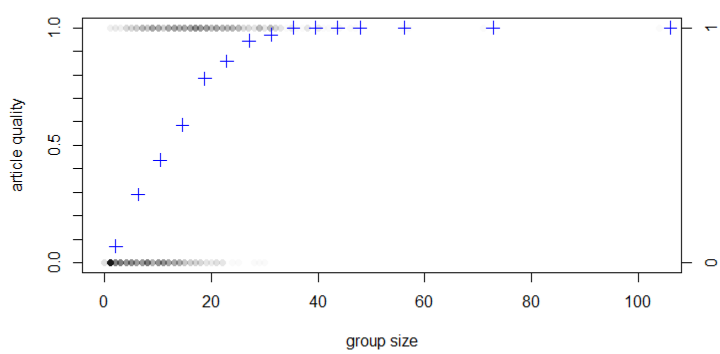


Fig. 5. Plot of quality vs. group size

5.2 Data Analysis

We carried out several analyses to test the hypotheses and explore the relation between the initial core group attributes and performance. First, we performed a regression. The dependent variable in our case study is binary, and hence we used logistic regression to test our hypotheses about the impact of the different variables [24]. We also used Random Forest as another test of the relative importance of the proposed variables [8]. Finally we carried out a Decision Tree analysis to extract rules to predict quality article from the variables. We present each of these findings in turn.

5.2.1 Logistic regression. First, the proposed variables were analyzed using logistic regression. Table 7 presents regression results with all the variable and selected squared forms to test for a non-linear relationship. Figure 4 suggests that some of the variables are highly correlated, which could cause problems for the analysis. To assess the impact of co-linearity, we ran analysis in which we introduced the variables one by one. As there were no changes in the signs of the coefficients as the variables were introduced, we can interpret the results of the full regression, even with co-linear variables.

We next discuss each hypothesis in turn.

- Our results contradict H1.1: average longevity is a negative predictor of article quality. It appears to be advantageous for article quality to have newer editors.
- Our results support H1.2: longevity diversity appears with significant and negative effects ($\beta = -7.447 * **$) in the model. This finding implies the desirability of similar longevity levels

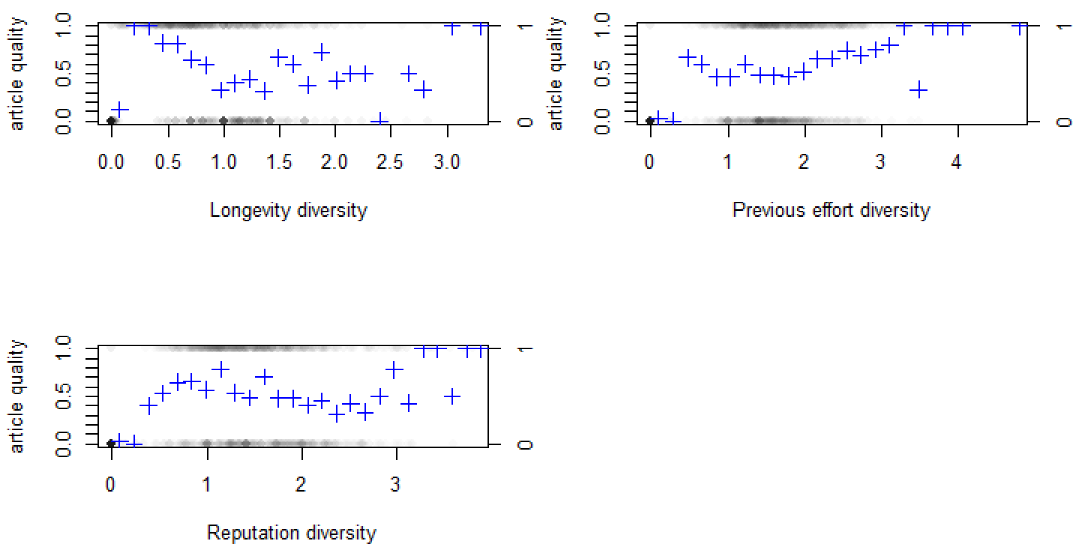


Fig. 6. Plots of quality vs. diversity of different variable

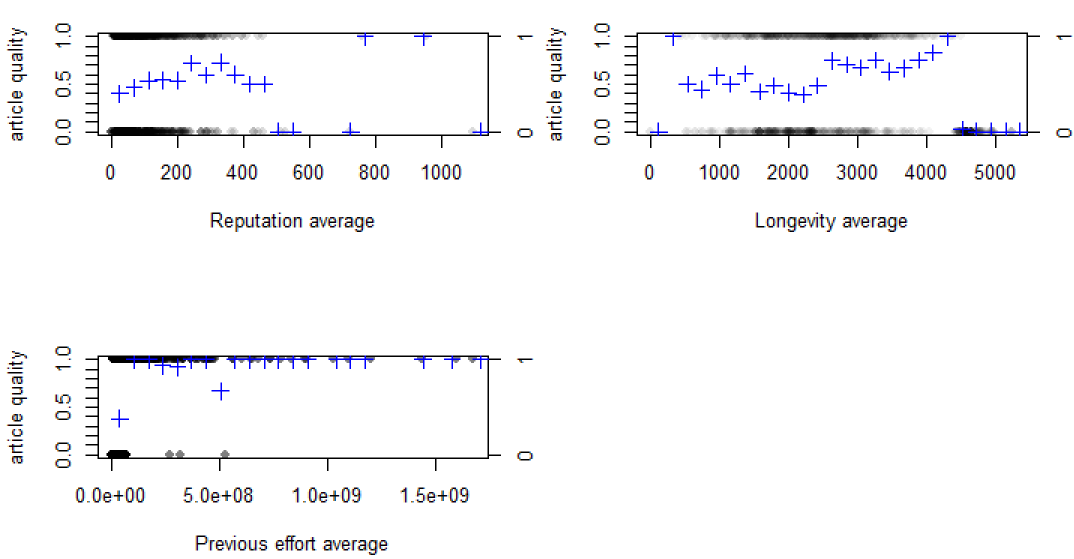


Fig. 7. Plots of quality vs. averages of different variable

Table 4. Logistic regression model of the average and diversity variables for the three languages combined plus squared diversity variables to test non linear effects.

Variables	Estimate	Std. Error	z value	Pr(> $|z|$)
Average longevity	-1.335e-03	3.330e-04	-4.010	6.07e-05 ***
Longevity diversity	-7.447e+00	1.306e+00	-5.701	1.19e-08 ***
Squared longevity diversity	1.545e+00	3.408e-01	4.535	5.76e-06 ***
Average previous effort	4.469e-08	6.699e-09	6.671	2.54e-11 ***
Previous effort diversity	1.957e+00	1.000e+00	1.956	0.050417 .
Squared previous effort diversity	-1.624e-01	2.508e-01	-0.648	0.517294
Average reputation	-6.951e-03	1.960e-03	-3.546	0.000391 ***
Reputation diversity	2.323e+00	1.064e+00	2.183	0.029013 *
Squared reputation diversity	-7.236e-01	2.586e-01	-2.798	0.005144 **
Group size	2.226e-01	2.069e-02	10.759	< 2e-16 ***

during grouping period. At the beginning of the core member recruitment process, recruiting editors with heterogeneous time experience (Longevity) is not desirable.

- Our results support H2.1: average previous effort does predict quality. As we expected, having editors with more experience as part of the initial group does predict a quality result.
- Our results do not support (though they do not contradict) H2.2: the coefficient for diversity of effort (previous amount of contribution) is positive as predicted, but the p value for this coefficient is just above 0.05, meaning that the hypothesis is not supported.
- Our results to do not support H3.1. The average reputation (i.e., the proportion of edits given to high-quality articles) is negatively related to quality. Apparently it is better for quality to have editors who have spent their time editing lower quality articles.
- Our results support H3.2. The diversity of reputation-based page appeared as an important variable with a positive effect (2.323e+00 *) and negative effect for its squared form (-7.236e-01 **) (see Table 7). These results show that the increase in "reputation" diversity increases the quality until an optimum after which the quality decrease. The maximum impact is reached at a value of about 1.585.

Finally, our results confirm that group size has a positive and significant impact on article quality (0.2226**). This finding can be understood as the fact that the group which attracts contributors during the initial period has a better chance of success, which is related to the fact that a well-known indicator of article quality is the number of contributors who have participated in its creation.

5.2.2 Random Forest Analysis. To select the most important variables that predict the outcome's quality, we carried out a Random Forest analysis. This method quantifies the importance of an attribute to the quality of the prediction. The variable importance plot is the main output of the random forest algorithm. The technique works by creating a number of decision trees to predict an outcome from the input variables (in our case, quality from the initial group characteristics). Algorithms for constructing decision trees usually work top-down, by choosing an attribute at each step that best splits the set of observations. The quality of the split is measured using the decrease of node homogeneity, e.g. the difference between the class homogeneity in the parent node and the child nodes. The class homogeneity is measured using entropy or Gini impurity measure. A large decrease indicates that the attribute is relevant. The average decrease of node homogeneity is taken over all splitting nodes and over all trees used to construct an ensemble classifier. Generally, the higher the value of the important measures, the stronger relationship with the predicted attribute (article quality).

We used the prepared data as presented in the previous section, split into training (70 % observations) and testing (30 % observations). Fig. 8 displays the importance of each variable as predicted by the random forest model. These results shows that the order of importance between the proposed variables is group size (by far), diversity of reputation, average longevity and diversity of previous effort. The f-measures for the Random Forest was 0.809, indicating good performance in predicting the quality level.

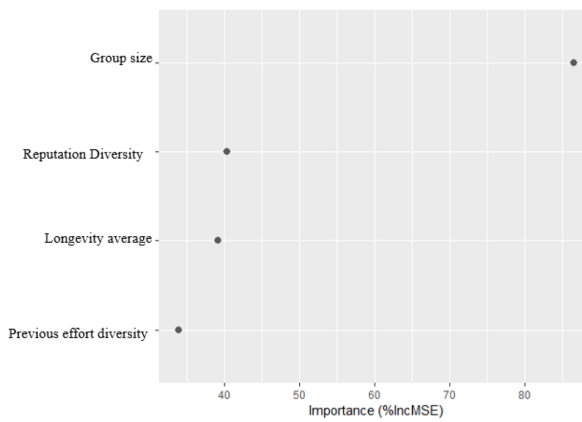


Fig. 8. Contribution of the variables to the increase of the explicable capacity of the model (for instance, Group size variable increases the efficiency of the model by nearly 100%)

5.2.3 Decision Tree. Finally, to develop a more interpretable sense of how the group characteristics predict article quality, we carried out a decision tree analysis. Fig. 9 displays the different rules we needed to predict the quality of an article. The f-measures for the Decision Tree was 0.786, again indicating good performance in predicting the quality level.

To summarize, an article is likely to be of high quality if any of the following hold:

- core group size > 14.5 (though there are few such articles) or
- 14.5 >= core group size > 5.5 and
  - reputation diversity < 1.174 and longevity average >= 4410 or
  - reputation diversity >= 1.174 and previous effort diversity >= 1.258

In other words, articles with large core groups (15 or more) are usually of high quality and articles with small core groups (5 or fewer) are usually of low quality. For intermediate sizes, groups with low reputation diversity and high average longevity or higher reputation and previous effort diversity are likely to be of high quality.

6 DISCUSSION

Our work makes a contribution to the online collaboration literature. First, we have proposed a method to assess the relationship between characteristics of an initial core group and the quality of the article they go on to create. To do so, we measure their activities before they started working together, using only accessible attributes in the context of virtual teaming, specifically experience (or longevity), previous effort and reputation. This study thus develops a novel method to define the initial group for a Wikipedia article and to test the relation between that groups' characteristics

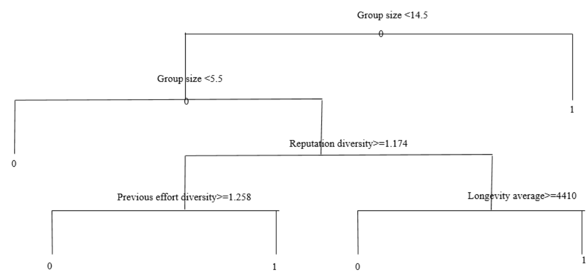


Fig. 9. Decision Tree applied to the proposed variables

and article quality. The results show that the distinction of initial members from all members is useful to understand antecedents of success.

The results shown in section 5 indicate that in terms of average longevity, it seems better to have newcomers to address the challenge of creating a new piece of knowledge, rather than old-timers, contradicting H1.1. Further, the data show that the quality is better if the group is homogeneous in terms of diversity of longevity, as suggested by H1.2. We note that this result stands in contrast to the findings of [48], which posit that high longevity disparity leads to high productivity, suggesting that conflicts between newcomers and old-timers strongly impacts the prospects of an article. Our findings are consistent with an interpretation that members in online volunteer groups are treated differently according to their experience. Experience in Wikipedia is sometimes viewed as conferring social status. Experienced editors and newcomers may refer disparagingly to each other, which may cause conflict in Wiki editing that reduces performance. Such conflicts at the beginning of article creation, i.e., during the initial growth period, may be particularly damaging to online volunteer groups. When members get frustrated, they are more likely to leave or stop contributing to the group effort. High longevity diversity of core member increasing conflict is consistent with prior research on offline groups. So in the first step of teaming period, having more equal levels of experience distributed among members seems to predict a higher quality article.

Our results suggest also that high quality articles are not necessarily written by reputed contributors, those with a higher experience in contributing to quality articles, in contradiction with H3.1. However, in agreement with H3.2, good quality articles seem to have contributors with different level of reputation cooperate during the initial period, even if the squared diversity reputation indicates that too much diversity is not good either.

Finally, for an article to have chance of success, it has to be taken care of by committed contributors, who have done a lot of edits in the past (as hypothesized in H2.1). However, the effect of diversity, while in the hypothesized direction, was not significant.

In summary, having from the beginning of the work diverse profiles of contributors not in terms of tenure on the project but rather in the past focus of their editing, seems to be a good thing, providing that the average of the previous edits is high. This result is consistent with an interpretation that having the different types, or roles of contributors from the beginning is a good indicator of success for and article project.

The decision tree modeling helps to understand, or refine, the characteristics of the efficient teams. It seems from the decision tree that there are two type of efficient groups: groups of old-timers even in terms of ratio of contribution to good articles, and more diverse people in terms of longevity and of efforts. A proposed explanation is that a group of old-timers who are used to developing articles (who may even know each others and agree on starting a new project), from one hand, and an more ad-hoc group of people who aggregate on a project they individually are interested in. This may explain that groups bigger than 14.5 seem to always reach quality 1: they are project with a lot of contributors per se, which is known to be good for quality.

7 LIMITATIONS AND FUTURE DIRECTIONS

In addition to the refinement in the calculation of the characteristics of the groups, several extensions can be proposed to this work. First, we would like to test our hypotheses on a larger number of articles and compare it with other articles languages. We could also test the relationship between different patterns of editing and article quality over time.

Second, our results could be biased by articles' topics and the method for estimating quality, namely relying on the self-classification. Further research is required in order to validate the findings across alternative samples on different topics, and test the model when using more robust estimates of article quality (preferably a measure more nuanced than a simple binary score).

Third, regarding the characteristics of the article group, we based our only on the proposed new group attributes. Adding other variables from prior research (e.g., checking whether people have worked together before the starting the target article) would refine our understanding of how the core teams are built. Still, even with the small number of variables, the high level of detection shows that these simple measures seem quite solid to identify a good group.

Fourth, in the current paper we gathered data from diverse language Wikipedias that we analyzed collectively. However, separate analyses show that there are differences between languages in the importance of variables, suggesting the need for theorizing cultural or work practice differences in initial group formation.

Finally, knowledge shared among online community differs from one community to another. Trace data of Open Source Software differ from one of Wikipedia. So in future work, we will try to generalize our findings to other kinds of online communities of creation.

8 CONCLUSION

In this paper, we show the importance of identifying the initial group composition for online open collaboration. Although most existing research on online collaboration has focused on motivation, governance, and social structure, our results suggest that the attributes of initial group members are another important factor that influences the success of these groups.

REFERENCES

[1] Ahuja, M. K. and Carley, K. M. (1999). Network structure in virtual organizations. *Organization Science*, 10(6):741–757.

[2] Anthony, D., Smith, S. W., and Williamson, T. (2009). Reputation and reliability in collective goods: The case of the online encyclopedia wikipedia. *Rationality and Society*, 21(3):283–306.

[3] Arazy, O., Nov, O., and Ortega, F. (2014). The [wikipedia] world is not flat: on the organizational structure of online production communities. In *Proceedings of the ECIS*. AIS.

[4] Arazy, O., Nov, O., Patterson, R., and Yeo, L. (2011). Information quality in wikipedia: The effects of group composition and task conflict. *Journal of Management Information Systems*, 27(4):71–98.

[5] Arazy, O., Ortega, F., Nov, O., Yeo, L., and Balila, A. (2015). Functional roles and career paths in wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1092–1105. ACM.

[6] Borgatti, S. P. and Everett, M. G. (2000). Models of core/periphery structures. *Social networks*, 21(4):375–395.

[7] Bradford, S. (1948). Documentation. londres: Crosby lockwood. reimpresión en (1950).

[8] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[9] Burke, M. and Kraut, R. (2008). Mopping up: modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 27–36. ACM.

[10] Butler, B., Sproull, L., Kiesler, S., and Kraut, R. (2007). Community effort in online groups: Who does the work and why? In Weisband, S., editor, *Leadership at a distance: Research in Technologically Supported Work*. Lawrence Erlbaum,

Mahwah, NJ.
[11] Carillo, K. and Okoli, C. (2011). Generating quality open content: A functional group perspective based on the time, interaction, and performance theory. *Information & Management*, 48(6):208–219.
[12] Chen, J., Ren, Y., and Riedl, J. (2010). The effects of diversity on group productivity and member withdrawal in online volunteer groups. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 821–830. ACM.
[13] Collier, B. and Kraut, R. (2012). Leading the collective: Social capital and the development of leaders in core-periphery organizations. *arXiv preprint arXiv:1204.3682*.
[14] Crowston, K. and Shamshurin, I. (2017). Core-periphery communication and the success of free/libre open source software projects. *Journal of Internet Services and Applications*, 8(1):10.
[15] Crowston, K., Wei, K., Li, Q., and Howison, J. (2006). Core and periphery in free/libre and open source software team communications. In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, volume 6, pages 118a–118a. IEEE.
[16] Daniel, S., Agarwal, R., and Stewart, K. (2009). The effects of diversity in global, distributed collectives: A study of user participation in open source projects. *Forthcoming at Information Systems Research*.
[17] Forte, A., Larco, V., and Bruckman, A. (2009). Decentralization in wikipedia governance. *Journal of Management Information Systems*, 26(1):49–72.
[18] Gasson, S. and Purcell, M. (2018). Legitimate peripheral participation in hybrid foss community innovation. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.
[19] Geiger, R. S. and Halfaker, A. (2013). Using edit sessions to measure participation in wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 861–870, New York, NY, USA. ACM.
[20] Halfaker, A., Kittur, A., and Riedl, J. (2011). Don't bite the newbies: how reverts affect the quantity and quality of wikipedia work. In *Proceedings of the 7th international symposium on wikis and open collaboration*, pages 163–172. ACM.
[21] Harrison, D. A. and Klein, K. J. (2007). What's the difference? diversity constructs as separation, variety, or disparity in organizations. *Academy of management review*, 32(4):1199–1228.
[22] Hess, C. and Ostrom, E. (2007). Understanding knowledge as a commons. *From theory to Practice*.
[23] Horwitz, S. K. and Horwitz, I. B. (2007). The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of management*, 33(6):987–1015.
[24] Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.
[25] Hu, M., Lim, E.-P., Lim, E.-P., Sun, A., Lauw, H. W., and Vuong, B.-Q. (2007). Measuring article quality in wikipedia: Models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pages 243–252, New York, NY, USA. ACM.
[26] Jones, J. (2008). Patterns of revision in online writing: A study of wikipedia's featured articles. *Written Communication*, 25(2):262–289.
[27] Jullien, N. (2012). What we know about wikipedia: a review of the literature analyzing the project (s)(2012). *Marsouin working papers. Available at SSRN 2053597*.
[28] Jullien, N., Stol, K.-J., and Herbsleb, J. D. (2019). A Preliminary Theory for Open Source Ecosystem Micro-economics. In Fitzgerald, B., Mockus, A., and Zhou, M., editors, *Towards Engineering Free/Libre Open Source Software (FLOSS) Ecosystems for Impact and Sustainability*, Communications of NII Shonan Meetings. Springer.
[29] Kane, G. C., Johnson, J., and Majchrzak, A. (2014). Emergent life cycle: The tension between knowledge change and knowledge retention in open online coproduction communities. *Management Science*, 60(12):3026–3048.
[30] Kiesel, J., Potthast, M., Hagen, M., and Stein, B. (2017). Spatio-temporal analysis of reverted wikipedia edits. In *Eleventh International AAAI Conference on Web and Social Media*.
[31] Kittur, A. and Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46. ACM.
[32] Kittur, A. and Kraut, R. E. (2010). Beyond wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 215–224. ACM.
[33] Koch, S. (2008). Effort modeling and programmer participation in open source software projects. *Information Economics and Policy*, 20(4):345–355.
[34] Kriplean, T., Beschastnikh, I., and McDonald, D. W. (2008). Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 47–56. ACM.
[35] Kuznetsov, S. (2006). Motivations of contributors to wikipedia. *SIGCAS Comput. Soc.*, 36(2).
[36] Leskovec, J., Huttenlocher, D. P., and Kleinberg, J. M. (2010). Governance in social media: A case study of the wikipedia promotion process. In *ICWSM*, pages 98–105.
[37] Lipka, N. and Stein, B. (2010). Identifying featured articles in wikipedia: writing style matters. In *Proceedings of the 19th international conference on World wide web*, pages 1147–1148. ACM.

[38] Liu, J. and Ram, S. (2011). Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems (TMIS)*, 2(2):11.

[39] Liu, J. and Ram, S. (2018). Using big data and network analysis to understand Wikipedia article quality. *Data & Knowledge Engineering*, 115:80–93.

[40] McGuinness, D. L., Zeng, H., Da Silva, P. P., Ding, L., Narayanan, D., Bhaowal, M., et al. (2006). Investigations into trust for collaborative information repositories: A wikipedia case study. In *Proceedings of the Workshop on Models of Trust for the Web*.

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

[42] Ortega, F. (2009a). *Wikipedia: A Quantitative Analysis*. Doctoral dissertation.

[43] Ortega, F. (2009b). *Wikipedia: A quantitative analysis*. PhD thesis, PhD thesis. Universidad Rey Juan Carlos, Madrid,.(Cit. on p.).

[44] Owens, D. A., Mannix, E. A., and Neale, M. A. (1998). Strategic formation of groups: Issues in task performance and team member selection. *Research on managing groups and teams*, 1:149–165.

[45] Panciera, K., Halfaker, A., and Terveen, L. (2009). Wikipedians are born, not made: a study of power editors on wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 51–60. ACM.

[46] Priedhorsky, R., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L., and Riedl, J. (2007). Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 259–268. ACM.

[47] Ren, R. and Yan, B. (2017). Crowd diversity and performance in wikipedia: The mediating effects of task conflict and communication. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6342–6351. ACM.

[48] Ren, Y., Chen, J., and Riedl, J. (2015). The impact and evolution of group diversity in online open collaboration. *Management Science*.

[49] Robert, L. and Romero, D. M. (2015). Crowd size, diversity and performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1379–1382. ACM.

[50] Rogers, E. M. (1976). New product adoption and diffusion. *Journal of consumer Research*, 2(4):290–301.

[51] Ruef, M., Aldrich, H. E., and Carter, N. M. (2003). The structure of founding teams: Homophily, strong ties, and isolation among us entrepreneurs. *American sociological review*, pages 195–222.

[52] Rullani, F. and Haefliger, S. (2013). The periphery on stage: The intra-organizational dynamics in online communities of creation. *Research Policy*, 42(4):941–953.

[53] Stein, K. and Hess, C. (2007). Does it matter who contributes: a study on featured articles in the german wikipedia. In *Proceedings of the eighteenth conference on Hypertext and hypermedia*, pages 171–174. ACM.

[54] Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. (2008). Information Quality Work Organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001.

[55] Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

[56] Tan, Y., Mookerjee, V., and Singh, P. (2007). Social capital, structural holes and team composition: Collaborative networks of the open source software community. *ICIS 2007 Proceedings*, page 155.

[57] Verhoeven, D., Cooper, T., Flynn, M., and Shuffler, M. L. (2017). Transnational team effectiveness. *The Wiley Blackwell Handbook of the Psychology of Team Working and Collaborative Processes*, pages 73–101.

[58] Von Krogh, G., Spaeth, S., and Lakhani, K. R. (2003). Community, joining, and specialization in open source software innovation: a case study. *Research Policy*, 32(7):1217–1241.

[59] Wang, L. S., Chen, J., Ren, Y., and Riedl, J. (2012). Searching for the goldilocks zone: trade-offs in managing online volunteer groups. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 989–998. ACM.

[60] Wilkinson, D. M. and Huberman, B. A. (2007). Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on Wikis*, pages 157–164. ACM.

[61] Williams, K. and O'Reilly III, C. (1998). Demography and diversity in organisations: A review of 40 years of research in bm staw and ll cummings (eds) research in organisational behaviour vol. 20. *Jai Pres, Connecticut*.

[62] Yang, D., Halfaker, A., Kraut, R., and Hovy, E. (2017). Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010.

[63] Yang, W., Liu, W., Viña, A., Tuanmu, M.-N., He, G., Dietz, T., and Liu, J. (2013). Nonlinear effects of group size on collective action and resource outcomes. *Proceedings of the National Academy of Sciences*, 110(27):10916–10921.

APPENDIX

Presentation of the results for each language

Table 5. Logistic regression model of the average, diversity and squared diversity variables for the Arabic language.

Variables	Estimate	Std. Error	z value	Pr(> $\|z\|$)
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.966e+00	2.704e+00	-0.727	0.46725
group size	2.940e-01	5.820e-02	5.052	4.38e-07 ***
previous effort average	8.385e-08	3.392e-08	2.472	0.01345 *
reputation average	2.632e-02	2.615e-02	1.006	0.31427
Longevity average	-1.719e-03	7.099e-04	-2.422	0.01545 *
previous effort diversity	-3.379e+00	2.497e+00	-1.353	0.17600
reputation diversity	1.154e+01	4.458e+00	2.588	0.00965 **
Longevity diversity	-7.617e+00	3.700e+00	-2.059	0.03951 *
squared previous effort diversity	2.845e-01	4.696e-01	0.606	0.54471
squared reputation diversity	-1.933e+00	1.104e+00	-1.752	0.07979 .
squared Longevity experience	1.319e+00	9.858e-01	1.338	0.18087

Table 6. Logistic regression model of the average, diversity and squared diversity variables for the Thai language.

Variables	Estimate	Std. Error	z value	Pr(> $\|z\|$)
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	8.680e+00	2.107e+01	0.412	0.6804
group size	1.867e-01	7.792e-02	2.396	0.0166 *
previous effort average	2.377e-07	9.706e-08	2.449	0.0143 *
reputation average	8.252e-02	5.054e-02	1.633	0.1025
Longevity average	-7.478e-03	4.325e-03	-1.729	0.0838 .
previous effort diversity	1.193e+00	9.112e+00	0.131	0.8958
reputation diversity	1.411e+01	9.869e+00	1.430	0.1527
Longevity diversity	-1.732e+01	1.336e+01	-1.297	0.1946
squared previous effort diversity	9.413e-01	2.415e+00	0.390	0.6967
squared reputation diversity	-3.727e+00	2.781e+00	-1.340	0.1802
squared Longevity diversity	6.849e-01	2.947e+00	0.232	0.8162

Table 7. Logistic regression model of the average, diversity and squared diversity variables for the Romania language.

Variables	Estimate	Std. Error	z value	Pr(> $|z|$)
(Intercept)	2.213e+00	3.040e+00	0.728	0.466678
group size	2.567e-01	2.970e-02	8.644	< 2e-16 ***
previous effort average	2.284e-08	5.602e-09	4.077	4.57e-05 ***
reputation average	-3.889e-03	1.887e-03	-2.061	0.039341 *
Longevity average	-1.494e-03	6.244e-04	-2.392	0.016751 *
previous effort diversity	1.233e+00	1.495e+00	0.824	0.409742
reputation diversity	2.300e+00	1.485e+00	1.550	0.121251
Longevity diversity	-8.329e+007	2.295e+00	-3.629	0.000285 ***
squared previous effort diversity	1.924e-01	4.082e-01	0.471	0.637451
squared reputation diversity	-8.491e-01	3.808e-01	-2.230	0.025749 *
squared Longevity diversity	1.397e+00	8.156e-01	1.712	0.086829 .