2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Diverse Teams Tend to Do Good Work in Wikipedia (but Jacks of All Trades Don't)

Jürgen Lerner
Dept. Computer and Information Science
University of Konstanz
Konstanz, Germany
juergen.lerner@uni-konstanz.de

Alessandro Lomi
Institute of Computational Science
University of Lugano
Lugano, Switzerland
alessandro.lomi@gess.ethz.ch

***Abstract*—We define network-based indicators of diversity for Wikipedia teams and users. A team of Wikipedia contributors is diverse to the extent that its members edit different articles. An individual contributor is a "jack of all trades" to the extent that she edits articles that are rarely co-edited by the same other users. For both indicators of team and individual diversity we propose a model-based normalization in which we compare observed values to expected values in a random graph model that preserves expected degrees of users and articles. Using data on all articles in the English-language edition of Wikipedia, we show that diverse teams tend to write high-quality articles, but articles written by teams containing jack of all trades contributors tend to be of lower quality. These findings are robust to several alternative explanations for article quality. We also show that the proposed model-based normalization of network indicators outperforms an ad-hoc normalization via cosine similarity.**

***Index Terms*—online peer-production, collaboration networks, team diversity, team performance, Wikipedia**

I. Introduction

Past decades have witnessed the emergence of open peer-production systems in which self-organizing teams of volunteers contribute to the production of public goods [1]–[3]. Successful examples include open-source software communities, and the free online encyclopedia Wikipedia – the system that we examine in this paper. The high productivity of open production communities has sometimes been explained by the "wisdom of the crowds" argument [4], [5]. Self-organizing teams can draw from large pools of contributors with potentially diverse and complementary background knowledge, experience, and capabilities. We would thus expect that diversity of teams relates positively to their productivity [6], [7].

Such expectations on the virtue of diversity are somewhat at odds with findings from organizational sociology demonstrating that category spanning, or a broad niche width of producers, is typically associated with lower evaluation of products by relevant audiences [8]–[11]. According to this literature "jacks of all trades" are "masters of none" and hence typically associated with outcomes of lower quality [10]. Consistent with this view, studies have found, for example, that interdisciplinary research proposals typically get discounted by funding agencies [12]. In this paper we claim that it is crucial to distinguish between the diversity of a *team*, and the diversity of interests of its *individual* members.

How does diversity affect quality of the output produced by peer-production systems? We address this question by examining how diversity of teams and users in Wikipedia affects the qualiry of Wikipedia articles. We derive indicators of diversity from the structure of the 2-mode network connecting users to the articles they edit. We define two users as having high distance if they contribute mostly to different articles. A group of users jointly contributing to an article, in turn, is said to have high team diversity if it is composed of users with high average pairwise distance. Thus, teams with high diversity represent atypical combinations [13] of users who normally contribute to different articles, see Fig. 1 for illustration.

A complementary aspect of diversity that we develop in this paper concerns the extent to which individual users have disparate interests – i. e, the extent to which users resemble "jacks of all trades." Two Wikipedia articles have high distance if they are written mostly by different users. An individual user, in turn, has disparate interests to the extent that she contributes to articles with high average pairwise distance. Thus, users with disparate interests contribute to atypical combinations of articles that are not normally co-edited by the same users. In this paper, we sometimes write *individual diversity* to refer to the extent to which the interests of individual users are disparate.

We normalize diversity indicators via random graph models that control for the observed degrees of users and articles (i. e, their activity and popularity), but otherwise have no clustering into latent topics or knowledge disciplines. We show that these model-based indicators consistently outperform their respective counterparts obtained via an ad-hoc normalization.

Using data on all articles of the English-language edition of Wikipedia, the main empirical result of our study is that team-level diversity increases article quality, but individual-level diversity of team members decreases output quality. We show that these findings are robust with respect to several control variables, membership of articles in main topic areas, indicators of team composition and role diversity, and to varying criteria for article quality.

This work has been supported by Swiss National Science Foundation (FNS Project Nr. 100018_150126) and Deutsche Forschungsgemeinschaft (DFG Grant Nr. LE 2237/2-1).

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

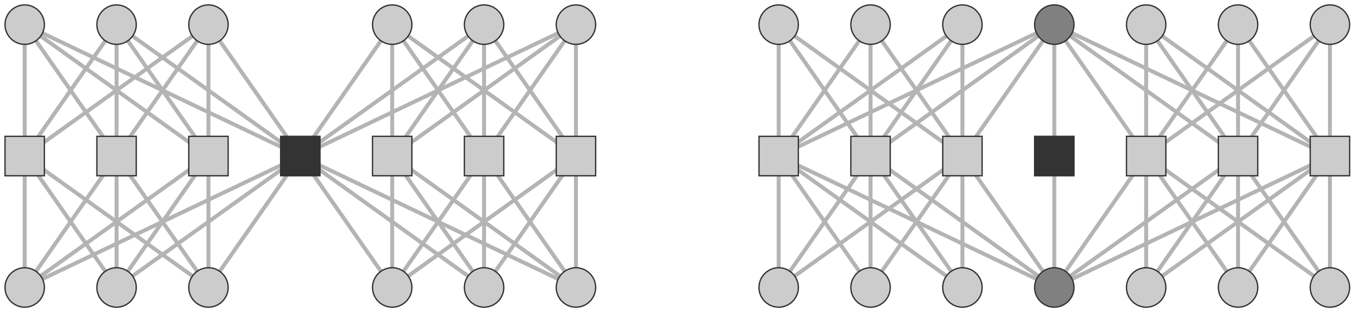

Fig. 1. Two fictitious bipartite collaboration networks. Wikipedia users (circles) contribute to articles (squares). Both networks have two distinct dense clusters that may represent latent topics or knowledge areas. *Left:* the black-colored article is written by members from different clusters and therefore has high team-level diversity; each of its contributors edits text mostly within one cluster and therefore has a relatively low individual-level diversity. *Right:* the two dark-gray contributors in the middle edit articles from both clusters and therefore have high individual diversity (i. e., they embody the notion of "'jack of all trades"'); the black-colored article is written by two users that contribute to exactly the same set of articles and therefore has low team diversity. Empirical results from this paper suggest that the black-colored article on the left-hand side has a higher probability to be of high quality than the black-colored article on the right-hand side.

II. Related Work and Hypotheses

The relation between work-team diversity and performance has attracted considerable attention [14], [15], at least since the path-breaking work of [16] on organizational demography. Extant research demonstrates that the relation between team diversity and performance depends on the aspects of diversity selected for consideration, and on a number of contextual or mediating factors [15], [17], [18]. Indeed, diversity itself is a diverse concept, as it may refer to heterogeneity in social or demographic characteristics, attitudes, view points, interests, backgrounds, experiences, tenure, function, position, status, or role. In this paper we analyze (mostly) the effect of an activity-based definition of diversity, where we say that Wikipedia teams are diverse if they are composed of users that typically contribute to a variety of different articles.

Effects of diversity in Wikipedia have been studied from a number of perspectives [19]. For instance, it has been shown that diversity moderates the effect of team size on performance in the WikiProject "Film" [20] and that cognitive diversity moderates the effect of task conflict on article quality [21]. Liu and Ram [22] analyzed how role diversity of Wikipedia teams affects output quality and Shi et al. [23] demonstrated that Wikipedia teams composed of politically diverse users produce articles of higher quality in the domains of politics, social issues, and science. Ren et al. [24] showed, among others, that interest diversity of the members of a WikiProject positively affects its productivity (i. e., the total number of edits) and that tenure disparity affects productivity in a curvilinear fashion: positive up to a point and negative afterwards; [25] also found that groups with intermediate tenure diversity make good decisions.

Previous work that analyzed the effect of diversity of individual users on article quality includes [26]–[28]. *Editors' diversity* or *versatility* has been defined in [26]–[28] via the entropy of editors' contributions to top-level categories and diversity of a *team of editors* has been measured via the variance in the number of contributions and tenure. It has been shown by an analysis of the Polish-language and German-language editions of Wikipedia, that both, versatility of editors and team diversity have a positive effect on article quality. It is noteworthy that we find diversity of individual users to be negative for article quality. In comparing our results with those from [26]–[28] we have to take into account several differences in the operationalization of the tests, where the most fundamental difference seems to be in the definition of diversity of users and teams (see next paragraph). Further differences can be found in the use of control variables and in the Wikipedia language edition from which the data has been gathered (English vs. German and Polish).

In our work we define diversity as a function of the 2-mode network connecting contributing Wikipedia users to the articles they write. The team of users editing a particular article is said to have high diversity if its members contribute to different sets of articles and it has low diversity if its members contribute mostly to the same articles. Thus, a diverse team is formed by an "atypical combination" [13] of users, while a team with low diversity consists mostly of users that typically work together. Likewise, an individual user is said to have diverse interests if she contributes to articles that are rarely co-edited by the same users, that is, if she contributes to an atypical combinations of articles.

Consistently with previous work on teams, we hypothesize that team diversity is positive for article quality, since diverse teams can draw from a larger pool of complementary background knowledge or experience, leading to the hypothesis

H_1 the higher the diversity of the *team* of users writing a Wikipedia article, the *higher* the probability that this article has high quality.

On the other hand, drawing on the jack-of-all-trades (and master-of-none) argument [10], we hypothesize that the presence of team members with disparate interests will have a negative effect on article quality:

H_2 the higher the average *individual* diversity in the team of users writing a Wikipedia article, the *lower* the probability that this article has high quality.

III. Data and Methods

Wikipedia makes its whole database publicly available at https://dumps.wikimedia.org/. For this paper we use data from the dump of the English-language edition of Wikipedia from January 1st, 2018.

A. Outcome Variable: Article Quality

The outcome variable that we consider in this paper is the *quality* of an article: an article is of high quality if it is a *featured article* (FA)[1]. There are 5,225 FA out of 5.5 million articles. FA is the highest in Wikipedia's quality classes: featured articles (FA), A-class, good articles (GA), B-class, C-class, start-class, stub-class. Wikipedia's article evaluations are often used in academic research [22], [29]–[32] and have been found to be consistent with external evaluations [29]. As a robustness check, we estimate models for article quality applying the weaker criterion in which an article is defined to be of high quality if it is featured or a good article (GA), which applies to 32,194 articles.

The units in our analysis are all Wikipedia articles and we test hypotheses by logistic regression where the binary outcome variable on articles is either the FA indicator or the FA-or-GA indicator. Explanatory variables on articles, including team diversity and individual diversity, are introduced in the remainder of this section. Since the notion of diversity of a one-person team is undefined, we only analyze articles to which at least two users contribute in a non-negligible way (see Sect. III-B below), which are 4,245,902 articles.

B. The Wikipedia Collaboration Network

The data structure that we use to compute team diversity and individual diversity is the weighted 2-mode network connecting Wikipedia users to the articles they edit. For a user u and an article a we define the weight $w(u,a)$ to be equal to the amount of text (measured in the number of bytes [28]) that u has added to a. (If u has never contributed to a, we set $w(u,a)=0$.) We consider only registered users (thus, we discard "anonymous" users identified by IP-addresses) and we also discard "bots" (software scripts that perform routine tasks [33]). When computing the amount of text added by a user to an article, we further discard contributions that are reverted in the very next revision, ensuring that users who make large but inappropriate contributions are not considered as main contributors. More sophisticated ways to weight individual contributions based on how long they survive (e. g., [34]–[39]) are not considered in this paper.

The amount of contributions of users to articles is very skewed in the sense that few users contribute a lot and most users contribute very little. To ensure that team diversity is mostly a function of the main contributors, we apply the following filtering. For each article a we order its contributors $u_1,\dots,u_k$ such that

$$w(u_1,a) \geq w(u_2,a) \geq \cdots \geq w(u_k,a) \ ,$$

breaking ties arbitrarily, and define $w(a)=\sum_{j=1}^{k} w(u_j,a)$ to be the total amount of contributions made to article a. We then compute the index i_0 such that

$$i_0 = \min\{i\colon \sum_{j=1}^{i} w(u_j,a) \geq 0.95 \cdot w(a)\} \ ,$$

which is the minimum index i such that users $u_1,\dots,u_i$ contribute at least 95% of the total contributions to article a. We then keep only $u_1,\dots,u_{i_0}$ as contributors of a, that is, we set the weights $w(u_i,a)$ for $i=i_0+1,\dots,k$ to zero. In doing so, we discard users making marginal contributions, while keeping 95% of the total contributions to every article. We write $U(a)$ for the set of users making nonzero contributions to article a (after the above-mentioned filtering), U for the set of all users, and A for the set of all articles.

C. Team Diversity

Assume that for any two users u and v we are given distances $dist(u,v)$ that encode differences in their interests (We will define two such distance functions below). The *team diversity* of an article a with $|U(a)| \geq 2$ is defined as the weighted average pairwise distance of its users, where we set $w_{uv;a} = w(u,a) + w(v,a)$ as the weight of the pair (u,v) for article a. In formulas, team diversity is defined by

$$team.diversity(a) = \frac{\sum_{u \neq v \in U(a)} dist(u,v) \cdot w_{uv;a}}{\sum_{u \neq v \in U(a)} w_{uv;a}} \ . \quad (1)$$

Thus, in computing the average, we give more weight to the distance of pairs of users that strongly contribute to the given article.

Distances between users are defined via a measure of similarity – or overlap - of interests.

$$c_{uv} = \sum_{a \in A} w(u,a) \cdot w(v,a) \ . \quad (2)$$

The raw measure c_{uv} captures to what extent u and v co-edit the same articles, but it is not normalized and by chance alone we would expect that pairs of more active users have higher overlap. A first way to normalize c_{uv} is the so-called "cosine similarity"

$$sim.cosine(u,v) = \frac{c_{uv}}{\|u\| \cdot \|v\|} \ ,$$

where the 2-norm $\|u'\|$ of a user u' is defined by

$$\|u'\| = \sqrt{\sum_{a \in A} w(u',a) \cdot w(u',a)} \ .$$

The respective distance of two users u and v who collaborate on at least one article is then defined by

$$dist.cosine = -\log(sim.cosine(u,v)) \ ,$$

and by substituting $dist.cosine$ for $dist$ in Eq. (1) we obtain the first indicator of team diversity, denoted by $team.diversity.cosine$.

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

D. Model-based Normalization of User Distance

Even though the cosine similarity is an established measure used to assess the similarity of vectors, it might have its drawbacks in a network setting. As a matter of fact the activity of users, as well as the popularity of articles, is very skewed and expected overlap in interests is likely to be influenced by these characteristics. For example, suppose two users contribute to all articles. Then these could not have no overlap in interests. Likewise, if there was an article to which every user contributed, then any two users would necessarily overlap in this article. Clearly, degrees of users and articles can influence similarity and distance.

Rather than applying an arbitrary (although established) normalization, as in the cosine similarity, we assess whether the observed overlap c_{uv} is higher or lower than *expected*, given the degrees of users and articles. Such an approach has been applied, e. g., by [13] who normalized the overlap c in journal citations of scientific papers to the z-score measure

$$zscore(c) = \frac{observed(c) - expected(c)}{standard\ deviation(c)}. \qquad (3)$$

To estimate expectation and variance [13] randomized networks with an edge-switching algorithm that preserves in-degrees and out-degrees. Since this procedure seems not to generalize in a straightforward manner to weighted networks, we apply instead a variant of the so-called *fitness model* [40], [41], as introduced in the following.

Let $d_a = \sum_{u \in U} w(u,a)$ denote the weighted degree of an article $a \in A$, let $d_u = \sum_{a \in A} w(u,a)$ denote the weighted degree of a user $u \in U$, and let $D = \sum_{a \in A} d_a = \sum_{u \in U} d_u$ denote the sum of degrees on either side of the 2-mode network. Adapting ideas of [40], [41], we define that random weights W_{ua} for $(u,a) \in U \times A$ are drawn independently from a distribution with expectation

$$\mathbb{E}(W_{ua}) = \frac{d_u \cdot d_a}{D} \ . \qquad (4)$$

This random graph has the property that the expected degrees of users and articles (i. e., activity and popularity) are equal to their observed degrees [40], [41]. Besides this property, the random graph has no clustering and, thus, no latent topics or knowledge disciplines. We use this null model to assess whether the interests of users overlap more or less than expected, given their degrees and given the degrees of articles.

To reproduce skewed distributions of edge weights, we draw random weights W_{ua} for $(u,a) \in U \times A$ from a Pareto distribution

$$Prob(W_{ua} \leq w) = 1 - \left(\frac{w_{ua}^{(\min)}}{w} \right)^{\alpha} \quad ,$$

with shape parameter $\alpha = 3$ and scale parameter $w_{ua}^{(\min)} = \frac{2 \cdot d_u \cdot d_a}{3 \cdot D}$, which defines a distribution with the required expectation, given in Eq. (4). Since random edge weights are independent by assumption, we can compute the expectation and variance of c_{uv} from Eq. (2) analytically and obtain for the respective z-score, defined in Eq. (3)

$$z(c_{uv}) = \frac{c_{uv} - \frac{d_u \cdot d_v}{D^2} \cdot \sum_{a \in A} d_a^2}{\frac{d_u \cdot d_v}{D^2} \cdot \sqrt{\sum_{a \in A} d_a^4}} \ .$$

The z-score $z(c_{uv})$ is a measure of similarity that is positive if the interests of u and v overlap more than expected and negative if they overlap less than expected. The respective measure of user-user distance is defined by

$$dist.zscore(u,v) = -\text{sign}(z(c_{uv})) \cdot \log(1 + |z(c_{uv})|)$$

and by substituting $dist.zscore$ for $dist$ in Eq. (1) we obtain the second indicator of team diversity denoted by $team.diversity.zscore$ as an alternative to $team.diversity.cosine$.

E. Individual Diversity

The other main variables used in this paper are indicators for whether the individual contributors of an article have diverse interests. We say that a user has high *individual diversity* if she contributes to articles that are usually not co-edited by the same person. Users with high individual diversity represent "Jacks of all trades" [10], "polymaths", "Renaissance men" [26], or interdisciplinary users with diverse interests.

Diversity of individual interests is computed using the same formulas as team diversity of articles after transposing "users" and "articles" in the 2-mode network. Thus, two articles a and b have high distance if they are mostly edited by different users (again we apply two normalization variants based on cosine similarity and based on z-scores computed with the fitness model). A user, in turn, has high individual diversity if she contributes to pairs of articles with high distance. Articles are then assigned the weighted average individual diversity of their contributors.

F. Control Variables

When assessing the influence of diversity on article quality we must take into account that articles vary largely in basic characteristics that have strong effects on the probability to be featured, or good and the we use as control variables in our models. The variables that have the highest predictive power for the quality of articles are indicators of article size [42] and the amount of work invested in writing the article. We use the *length* (number of bytes) of articles, their *age* (time since the first edit), *number of edits*, *team size* (number of unique contributors), and *number of reverts*. Links to information sources can be indicative of quality; we use the *number of intra-wiki links*, *number of external references*, and *number of inter-language links*. Characteristics of the text and potential appeal are captured by the *number of sections at level one and two*, *number of images*, *number of templates*, *average number of characters per word*, and *average number of words per sentence* (the last two variables capture the so-called reading complexity [31]). Embedding into Wikipedia's category structure is measured by the *number of categories* of the article, the *average size of its categories*, and the average *granularity* of categories (that is, their distance from the root category [32]). Further, more specific control variables are introduced in the following two sections.

G. Checking Against Simpson's Paradox: Main Topic Areas

Wikipedia articles are about very different topics and it might be that articles in different areas have different mean levels of diversity but also different probabilities to be featured. This could lead to spurious findings where a global relation among two variables is reversed in any sub-group of data (referred to as *Simpson's paradox* [43]–[45]). To test the robustness of our findings against this conjecture we assign articles to one or several of 21 top-level categories (TLC) [46], as described in [32]. We then extend our models by 21 binary indicator variables that are one if the article is in the respective TLC. We also estimate a more complex model in which we interact our variables of interest (team diversity and average individual diversity) with all 21 TLC indicators.

H. Interest Diversity vs. Role Diversity

A last potentially confounding effect against which we test our findings is related to the conjecture that team diversity (as we define it in this paper) might just be a byproduct of varying *composition* of the team. Different users have very different levels of activity and tend to perform different *tasks* or play different *roles* [22]. These differences could influence both, team diversity and article quality.

We first define three numerical variables, assigned to users, that capture activity levels in three different tasks. For a user u we define

- $provide.content(u)$ as the total amount of text added by u to any Wikipedia article (this variable is the weighted degree d_u in the 2-mode network, defined above);
- $edit.content(u)$ as the total number of edits done by u to any article;
- $coordinate(u)$ as the total number of edits done by u to any non-article page in Wikipedia (that is, to talk pages, user pages, project pages, templates, categories, etc).

For each of these three variables, we characterize an article by the average of this variable taken over the team of contributors and by the coefficient of variation of this variable, within the team. More precisely, let x denote any of the three variables defined above and let a be an article. We define

$$avg.x(a) = \sum_{u \in U(a)} x(u)/|U(a)|$$

$$var.x(a) = \frac{\sqrt{\frac{\sum_{u \in U(a)}[x(u)-avg.x(u)]^2}{|U(a)|-1}}}{avg.x(a)} \quad ,$$

where we get a missing value if $avg.x(a) = 0$.

I. Nomalization of variables

Before estimating models we transform variables that have skewed distributions (whose names are prefixed by "log1p" in the following parameter tables) by the mapping $x \mapsto \log(1 + x)$. For each variable we subtract its mean and divide by its standard deviation. This normalization makes parameter sizes better comparable.

IV. Results and Discussion

Diverse teams vs. jacks of all trades: Table I reports estimated parameters of logit models for the probability that articles are featured (FA). All seven models include the control variables introduced in Sect. III-F. The first three models (prefixed with *Team*) include *team.diversity.cosine*, *team.diversity.zscore*, or both. We find that both of these indicators, when added separatelly to the null model, have a positive effect on the FA-probability, consistent with Hypothesis H_1 stating that diverse teams tend to do good work. We observe that the parameter of *team.diversity.zscore* is larger. We also find that the model including the z-score based measure is better with respect to the model fit indicators AIC and BIC (recall that lower values indicate a better model fit), than the model with the cosine-normalized team diversity. Including both variables in the same model (*Team.C+Z*) reverses the effect of *team.diversity.cosine* to the negative but leaves *team.diversity.zscore* positive. Thus, the model-based normalization yields an indicator of team diversity that shows a stronger effect and leads to a better model fit and a more robust finding.

The next three models (prefixed with *Ind*) include indicators for the average *individual.diversity.cosine*, the average *individual.diversity.zscore*, or both. We find that both of these indicators have a negative effect on the FA-probability, consistent with Hypothesis H_2 stating that Jacks of all trades tend to do poor work. We observe that the parameter of *individual.diversity.zscore* is considerably larger in absolute value. Again we find that the model including the z-score based measure is better with respect to the model fit indicators AIC and BIC, than the model with the cosine-normalized individual diversity. Including both variables in the same model (*Ind.C+Z*) reverses the effect of *individual.diversity.cosine* to the positive but leaves *individual.diversity.zscore* negative. Thus, we find again that the model-based normalization yields an indicator of individual diversity that shows a stronger effect and leads to a better model fit and a more robust finding.

Last but not least, the right-most model (*Team+Ind.Z*) includes the z-score based measures for team diversity and individual diversity. Both effects remain qualitatively the same: a high team diversity is positive for article quality and a high individual diversity is negative.

Defining FA and GA as high-quality articles: Table II reports findings for models that have exactly the same explanatory variables as those from Table I but whose binary outcome variable is the indicator whether articles are featured (FA) or good (GA). All findings remain qualitatively the same: a high team diversity is positive for article quality, a high individual diversity is negative for article quality, and the z-scored based measures lead to stronger and more robust effects and to a better model fit. Thus, our findings are robust to a weaker, more inclusive, definition of article quality.

TABLE I

LOGISTIC REGRESSION FOR FA-PROBABILITY. ESTIMATED PARAMETERS AND STANDARD ERRORS (IN BRACKETS). ALL PARAMETERS ARE SIGNIFICANTLY DIFFERENT FROM ZERO AT THE 1% LEVEL. EFFECTS RELATED TO OUR HYPOTHESES ARE IN **BOLD FONT**.

	Team.C	Team.Z	Team.C+Z	Ind.C	Ind.Z	Ind.C+Z	Team+Ind.Z
(Intercept)	−11.23 (0.06)	−11.28 (0.07)	−11.31 (0.07)	−11.20 (0.06)	−11.16 (0.06)	−11.24 (0.06)	−11.23 (0.06)
log1p.length	2.70 (0.04)	2.72 (0.04)	2.67 (0.04)	2.50 (0.04)	2.31 (0.04)	2.28 (0.04)	2.47 (0.04)
age	1.06 (0.03)	1.01 (0.03)	0.99 (0.03)	1.04 (0.03)	0.99 (0.03)	0.99 (0.03)	0.96 (0.03)
log1p.#edits	2.18 (0.06)	2.26 (0.06)	2.25 (0.06)	2.05 (0.06)	2.04 (0.06)	2.00 (0.06)	2.22 (0.06)
log1p.#reverts	1.10 (0.03)	1.10 (0.03)	1.09 (0.03)	1.10 (0.03)	1.12 (0.03)	1.14 (0.03)	1.12 (0.03)
log1p.teamsize	−3.44 (0.05)	−3.34 (0.05)	−3.23 (0.06)	−3.23 (0.05)	−2.86 (0.06)	−2.79 (0.06)	−2.93 (0.06)
log1p.#wiki.links	−0.81 (0.03)	−0.88 (0.04)	−0.91 (0.04)	−0.75 (0.03)	−0.80 (0.03)	−0.83 (0.03)	−0.88 (0.04)
log1p.#external.refs	−0.14 (0.02)	−0.15 (0.02)	−0.16 (0.02)	−0.14 (0.02)	−0.12 (0.02)	−0.11 (0.02)	−0.13 (0.02)
log1p.#lang.links	0.39 (0.02)	0.36 (0.02)	0.34 (0.02)	0.39 (0.02)	0.33 (0.02)	0.34 (0.02)	0.31 (0.02)
#level.1.sections	−0.33 (0.02)	−0.33 (0.02)	−0.32 (0.02)	−0.33 (0.02)	−0.32 (0.02)	−0.31 (0.02)	−0.32 (0.02)
#level.2.sections	−0.44 (0.01)	−0.43 (0.01)	−0.43 (0.01)	−0.44 (0.01)	−0.44 (0.01)	−0.44 (0.01)	−0.44 (0.01)
log1p.#images	0.16 (0.02)	0.16 (0.02)	0.16 (0.02)	0.17 (0.02)	0.15 (0.02)	0.14 (0.02)	0.14 (0.02)
log1p.#templates	0.60 (0.03)	0.59 (0.03)	0.59 (0.03)	0.61 (0.03)	0.56 (0.03)	0.56 (0.03)	0.55 (0.03)
#characters.per.word	−0.73 (0.03)	−0.71 (0.03)	−0.70 (0.03)	−0.74 (0.03)	−0.73 (0.03)	−0.71 (0.03)	−0.72 (0.03)
#words.per.sentence	−0.29 (0.05)	−0.28 (0.05)	−0.28 (0.05)	−0.29 (0.05)	−0.26 (0.05)	−0.26 (0.05)	−0.26 (0.05)
#categories	0.05 (0.01)	0.04 (0.01)	0.04 (0.01)	0.06 (0.01)	0.07 (0.01)	0.06 (0.01)	0.05 (0.01)
log1p.#avg.cat.size	−0.08 (0.03)	−0.09 (0.03)	−0.10 (0.03)	−0.07 (0.03)	−0.10 (0.03)	−0.10 (0.03)	−0.12 (0.03)
granularity	0.52 (0.03)	0.47 (0.04)	0.43 (0.04)	0.48 (0.03)	0.38 (0.04)	0.41 (0.04)	0.36 (0.04)
team.div.cosine	**0.39 (0.03)**		**−0.39 (0.05)**				
team.div.zscore		**0.52 (0.03)**	**0.83 (0.05)**				**0.40 (0.03)**
ind.div.cosine				**−0.17 (0.03)**		**0.84 (0.06)**	
ind.div.zscore					**−0.86 (0.03)**	**−1.71 (0.08)**	**−0.80 (0.03)**
AIC	38,162.48	37,950.46	37,902.57	38,303.15	37,579.22	37,338.96	37,351.17
BIC	38,414.44	38,202.43	38,167.80	38,555.11	37,831.19	37,604.19	37,616.40
Num. obs.	4,245,902	4,245,902	4,245,902	4,245,902	4,245,902	4,245,902	4,245,902

$p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE II

LOGISTIC REGRESSION FOR THE PROBABILITY THAT ARTICLES ARE FA OR GA. ALL PARAMETERS ARE SIGNIFICANTLY DIFFERENT FROM ZERO AT THE 0.1% LEVEL.

	Team.C	Team.Z	Team.C+Z	Ind.C	Ind.Z	Ind.C+Z	Team+Ind.Z
	(all control variables from Sect. III-F included)						
team.div.cosine	**0.10 (0.01)**		**−0.29 (0.02)**				
team.div.zscore		**0.20 (0.01)**	**0.42 (0.02)**				**0.10 (0.01)**
ind.div.cosine				**−0.08 (0.01)**		**0.53 (0.02)**	
ind.div.zscore					**−0.64 (0.01)**	**−1.13 (0.03)**	**−0.62 (0.01)**
AIC	192,292.62	191,961.57	191,753.70	192,322.90	189,480.96	188,643.04	189,381.69
BIC	192,544.59	192,213.54	192,018.92	192,574.87	189,732.93	188,908.27	189,646.92
Num. obs.	4,245,902	4,245,902	4,245,902	4,245,902	4,245,902	4,245,902	4,245,902

Considering top-level categories (TLC): Table III reports estimated parameters where we extend the rightmost model from Table I by 21 binary indicator variables encoding membership of articles in TLC. Our findings are robust so that team diversity is positive and individual diversity is negative for article quality. We further extended the model from Table III by adding all interaction effects of either team diversity or average individual diversity with all TLC indicators which introduces 42 additional effects. (Estimated parameters of that model are not reported in this paper.) Our relevant findings did not change qualitatively: team diversity has a positive effect and the average individual diversity has a negative effect on the FA-probability.

Interest diversity vs. role diversity: Table IV reports estimated parameters where we extend the rightmost model from Table I by six variables for the average composition and role diversity of teams, introduced in Sect. III-H. Our findings related to team diversity and individual diversity (where we consider diversity of interests, rather than role diversity) do not change.

V. CONCLUSION AND FUTURE WORK

One possible reason for the success of open peer production is that self-organizing teams of volunteers can draw from a large pool of potentially diverse contributors who can bring in complementary background knowledge and abilities. In this paper we perform a rigorous empirical analysis of the hypothesis that diverse teams of Wikipedia users tend to produce articles of higher quality.

We consider *task-oriented interest diversity* rather than other – not less relevant – variants, such as social or demographic diversity, tenure diversity, or role diversity. We define two users as having have high distance (diverse interests) to the extent that they contribute to different articles. In turn, the

TABLE III
Logit model for FA-probability with varying baseline probabilities for different top-level categories (TLC)

(all control variables from Sect. III-F included)	
team.div.zscore	$\mathbf{0.39}$ $(\mathbf{0.03})^{***}$
ind.div.zscore	$\mathbf{-0.81}$ $(\mathbf{0.03})^{***}$
Arts	0.28 $(0.05)^{***}$
Culture	-0.01 (0.05)
History	0.07 (0.05)
Humanities	0.39 $(0.05)^{***}$
Politics	-0.51 $(0.06)^{***}$
Geography	-0.40 $(0.06)^{***}$
World	0.12 $(0.06)^{*}$
Events	0.83 $(0.05)^{***}$
Life	0.70 $(0.07)^{***}$
Nature	0.17 $(0.07)^{*}$
Philosophy	-0.02 (0.17)
People	-0.10 $(0.05)^{*}$
Science_and_technology	-0.39 $(0.09)^{***}$
Sports	-0.55 $(0.07)^{***}$
Health	0.10 (0.08)
Society	-0.12 $(0.05)^{*}$
Law	0.13 (0.10)
Religion	0.06 (0.08)
Mathematics	-0.33 (0.17)
Matter	0.35 $(0.10)^{***}$
Reference_works	-1.42 $(0.38)^{***}$
AIC	36,435.93
BIC	36,979.65
Num. obs.	4,245,902

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE IV
Logit model for FA-probability including indicators of average composition and role diversity of teams, introduced in Sect. III-H. All parameters are significantly different from zero at the 0.1% level.

(all control variables from Sect. III-F included)	
team.div.zscore	$\mathbf{0.31}$ $(\mathbf{0.03})$
ind.div.zscore	$\mathbf{-1.10}$ $(\mathbf{0.04})$
avg.provide.content	-1.03 (0.06)
var.provide.content	-1.45 (0.09)
avg.edit.content	-1.41 (0.07)
var.edit.content	0.99 (0.05)
avg.coordinate	2.32 (0.05)
var.coordinate	-1.47 (0.05)
AIC	31,131.25
BIC	31,476.05
Num. obs.	4,245,902

team of users jointly writing an article has high diversity to the extent that it is composed of users with high pairwise distance. Thus, articles with high team diversity are written by atypical combinations of users who normally contribute to different articles.

A complementary measure of diversity used in this paper is the individual diversity of users. Two Wikipedia articles have high distance to the extent that they are written by different users. A user, in turn, is said to have high individual diversity (is a "jack of all trades") if she contributes to articles with high pairwise distance. Thus, users with high individual diversity contribute to atypical combinations of articles that are not normally co-edited by the same users.

Both indicators are defined as a function of the weighted 2-mode network connecting users to the articles they write. Thus, both indicators could, in principle, be computed for other production systems, for instance, open-source software communities – whenever we have actors connected to objects they work at. We adapt ideas to normalize diversity indicators via random graph models that control for the observed degrees of users and articles (i. e, their activity and popularity) but otherwise have no clustering into latent topics or knowledge disciplines. We show that these model-based indicators consistently outperform their respective counterparts obtained via an ad-hoc normalization (cosine similarity).

Based on previous related work we hypothesize that team diversity is positive for article quality, since diverse teams can draw from a larger pool of complementary background knowledge or experience – all other things being equal. On the other hand – drawing on the jacks-of-all-trades-are-masters-of-none argument [10] – we hypothesize that individual diversity of users is negative for article quality.

For both hypotheses we have found strong empirical support. According to these results, the best teams would be composed of specialists from different disciplines; the quality of the team output would deteriorate if most users belong to the same discipline but also if users are interdisciplinary "polymaths." These findings have been shown to be very robust. We get qualitatively the same results with models that control for many characteristics of the articles, for membership of articles in main topic areas, or for indicators of team composition or role diversity. Weakening the criteria for high-quality articles from featured to good also leaves the main findings unchanged.

It is noteworthy that our findings on the relation between individual diversity and quality is contrary to findings of [26]–[28] in the sense that we find individual diversity to be negative for article quality while [26]–[28] found a positive effect. However, we have to take into account several differences in the operationalization of the tests, where the most fundamental difference seems to be in the definition of diversity. *Editors' diversity* or *versatility* in [26]–[28] has been defined via the entropy of editors' contributions to top-level categories. In contrast, we defined *individual diversity* via the 2-mode user-article network, where a user is said to has diverse interests if she edits articles that are rarely co-edited by the same user.

Possible directions for future work include attempts to relate the organizational structure of the team of contributors to indicators of diversity. Individual users might be embedded in hierarchical structures [39] or might be members of factions having opposite opinions [23], and such characteristics of the collaboration network might interact with team diversity. Interaction between diversity and conflict in Wikipedia has been analyzed before [21] but not on the global scale as in this paper. It could also be that team diversity, or individual diversity, has varying benefits or drawbacks in different stages of article development. A dynamic analysis that considers diversity over time, relating it with indicators for the current state of the article, might shed light on this question.

References

[1] J. Lerner and J. Tirole, "The open source movement: Key research questions," *European economic review*, vol. 45, no. 4, pp. 819–826, 2001.

[2] E. von Hippel and G. von Krogh, "Open source software and the "private-collective" innovation model: Issues for organization science," *Organization science*, vol. 14, no. 2, pp. 209–223, 2003.

[3] ——, "Free revealing and the private-collective model for innovation incentives," *R&D Management*, vol. 36, no. 3, pp. 295–306, 2006.

[4] O. Arazy, W. Morgan, and R. Patterson, "Wisdom of the crowds: Decentralized knowledge construction in Wikipedia," in *Proc. 16th Workshop Information Technologies and Systems*, 2006, pp. 79–84.

[5] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz, "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie," in *Proc. 25th Ann. ACM Conf. Human Factors in Computing Systems*. ACM, 2007.

[6] L. Hong and S. E. Page, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 46, pp. 16 385–16 389, 2004.

[7] S. Wuchty, B. F. Jones, and B. Uzzi, "The increasing dominance of teams in production of knowledge," *Science*, vol. 316, no. 5827, pp. 1036–1039, 2007.

[8] J. Freeman and M. T. Hannan, "Niche width and the dynamics of organizational populations," *American Journal of Sociology*, vol. 88, no. 6, pp. 1116–1145, 1983.

[9] B. Kovács and R. Johnson, "Contrasting alternative explanations for the consequences of category spanning: A study of restaurant reviews and menus in san francisco," *Strategic Organization*, vol. 12, no. 1, pp. 7–37, 2014.

[10] G. Hsu, "Jacks of all trades and masters of none: Audiences' reactions to spanning genres in feature film production," *Administrative Science Quarterly*, vol. 51, no. 3, pp. 420–450, 2006.

[11] B. Kovács and M. T. Hannan, "The consequences of category spanning depend on contrast," in *Categories in markets: Origins and evolution*. Emerald Group Publishing Limited, 2010, pp. 175–201.

[12] L. Bromham, R. Dinnage, and X. Hua, "Interdisciplinary research has consistently lower funding success," *Nature*, vol. 534, no. 7609, p. 684, 2016.

[13] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones, "Atypical combinations and scientific impact," *Science*, vol. 342, no. 6157, pp. 468–472, 2013.

[14] S. K. Horwitz and I. B. Horwitz, "The effects of team diversity on team outcomes: A meta-analytic review of team demography," *Journal of management*, vol. 33, no. 6, pp. 987–1015, 2007.

[15] A. Joshi and H. Roh, "The role of context in work team diversity research: A meta-analytic review," *Academy of Management Journal*, vol. 52, no. 3, pp. 599–627, 2009.

[16] J. Pfeffer, "Organizational demography." *Research in organizational behavior*, 1983.

[17] K. A. Jehn, G. B. Northcraft, and M. A. Neale, "Why differences make a difference: A field study of diversity, conflict and performance in workgroups," *Administrative science quarterly*, vol. 44, no. 4, pp. 741–763, 1999.

[18] E. Mannix and M. A. Neale, "What differences make a difference? the promise and reality of diverse teams in organizations," *Psychological science in the public interest*, vol. 6, no. 2, pp. 31–55, 2005.

[19] F. Flöck, D. Vrandečić, and E. Simperl, "Towards a diversity-minded Wikipedia," in *Proc. 3rd Intl. Web Science Conference*. ACM, 2011, p. 5.

[20] L. P. Robert and D. M. Romero, "The influence of diversity and experience on the effects of crowd size," *Journal of the Association for Information Science and Technology*, vol. 68, no. 2, pp. 321–332, 2017.

[21] O. Arazy, O. Nov, R. Patterson, and L. Yeo, "Information quality in Wikipedia: The effects of group composition and task conflict," *Journal of Management Information Systems*, vol. 27, no. 4, pp. 71–98, 2011.

[22] J. Liu and S. Ram, "Who does what: Collaboration patterns in the Wikipedia and their impact on article quality," *ACM Transactions on Management Information Systems (TMIS)*, vol. 2, no. 2, p. 11, 2011.

[23] F. Shi, M. Teplitskiy, E. Duede, and J. Evans, "The wisdom of polarized crowds," 2017, arXiv:1712.06414.

[24] Y. Ren, J. Chen, and J. Riedl, "The impact and evolution of group diversity in online open collaboration," *Management Science*, vol. 62, no. 6, pp. 1668–1686, 2015.

[25] S. K. Lam, J. Karim, and J. Riedl, "The effects of group composition on decision quality in a social production community," in *Proceedings of the 16th ACM international conference on Supporting group work*. ACM, 2010, pp. 55–64.

[26] J. Szejda, M. Sydow, and D. Czerniawska, "Does a 'renaissance man' create good Wikipedia articles?" in *Proc. Intl. Conf. Knowledge Discovery and Information Retrieval (KDIR-2014)*, 2014, pp. 425–430.

[27] K. Baraniak, M. Sydow, J. Szejda, and D. Czerniawska, "Studying the role of diversity in open collaboration network: experiments on Wikipedia," in *International Conference and School on Network Science*. Springer, 2016, pp. 97–110.

[28] M. Sydow, K. Baraniak, and P. Teisseyre, "Diversity of editors and teams versus quality of cooperative work: experiments on Wikipedia," *Journal of Intelligent Information Systems*, vol. 48, no. 3, pp. 601–632, 2017.

[29] A. Kittur and R. E. Kraut, "Harnessing the wisdom of crowds in Wikipedia: quality through coordination," in *Proc. 2008 ACM conf. Computer Supported Cooperative Work*. New York, NY, USA: ACM, 2008, pp. 37–46.

[30] G. Wu, M. Harrigan, and P. Cunningham, "Characterizing Wikipedia pages using edit network motif profiles," in *Proc. 3rd intl. workshop Search and mining user-generated contents, Glasgow, Scotland, UK*. New York, NY, USA: ACM, 2011, pp. 45–52.

[31] S. Ransbotham and G. C. Kane, "Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia," *MIS Quarterly*, vol. 35, no. 3, pp. 613–627, 2011.

[32] J. Lerner and A. Lomi, "Knowledge categorization affects popularity and quality of Wikipedia articles," *PloS one*, vol. 13, no. 1, p. e0190674, 2018.

[33] M. Tsvetkova, R. García-Gavilanes, L. Floridi, and T. Yasseri, "Even good bots fight: The case of Wikipedia," *PloS one*, vol. 12, no. 2, p. e0171774, 2017.

[34] B. T. Adler and L. de Alfaro, "A content-driven reputation system for the Wikipedia," in *Proc. 16th Intl. Conf. WWW*. ACM, 2007, pp. 261–270.

[35] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij, "Network analysis of collaboration structure in Wikipedia," in *Proc. 18th Intl. Conf. WWW*. ACM, 2009, pp. 731–740.

[36] S. Javanmardi, C. Lopes, and P. Baldi, "Modeling user reputation in wikis," *Statistical Analysis and Data Mining*, vol. 3, no. 2, pp. 126–139, 2010.

[37] S. Maniu, B. Cautis, and T. Abdessalem, "Building a signed network from interactions in Wikipedia," in *Proc. Databases and Social Networks*. ACM, 2011, pp. 19–24.

[38] F. Flöck and M. Acosta, "Wikiwho: Precise and efficient attribution of authorship of revisioned content," in *Proc. 23rd Intl. Conf. WWW*. ACM, 2014, pp. 843–854.

[39] J. Lerner and A. Lomi, "The third man: Hierarchy formation in Wikipedia," *Applied Network Science*, vol. 2, no. 1, p. 24, 2017.

[40] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Munoz, "Scale-free networks from varying vertex intrinsic fitness," *Physical review letters*, vol. 89, no. 25, p. 258702, 2002.

[41] G. De Masi, G. Iori, and G. Caldarelli, "Fitness model for the italian interbank money market," *Physical Review E*, vol. 74, no. 6, p. 066112, 2006.

[42] J. E. Blumenstock, "Size matters: word count as a measure of quality on Wikipedia," in *Proc. 17th Intl. Conf. WWW*. ACM, 2008, pp. 1095–1096.

[43] E. H. Simpson, "The interpretation of interaction in contingency tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 238–241, 1951.

[44] C. R. Blyth, "On Simpson's paradox and the sure-thing principle," *Journal of the American Statistical Association*, vol. 67, no. 338, pp. 364–366, 1972.

[45] S. Barbosa, D. Cosley, A. Sharma, and R. M. Cesar Jr, "Averaging gone wrong: Using time-aware analyses to better understand behavior," in *Proc. 25th Intl. Conf. WWW*. ACM, 2016, pp. 829–841.

[46] A. Kittur, E. H. Chi, and B. Suh, "What's in Wikipedia? Mapping topics and conflict using socially annotated category structure," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2009, pp. 1509–1512.