**Social Network Analysis and Mining manuscript No.**
(will be inserted by the editor)

# Team diversity, polarization, and productivity in online peer-production

**Jürgen Lerner · Alessandro Lomi**

Accepted manuscript

**Abstract** We define network-based indicators to characterize diversity of Wikipedia teams and contributing users. A team of Wikipedia users is diverse to the extent that its members edit different articles. An individual user has diverse interests to the extent that she contributes to articles that are not normally co-edited by the same users, i. e., if she contributes to an atypical combination of articles. For both indicators we propose a model-based normalization by comparing observed and expected values computed on a reference random graph model that preserves expected degrees of users and articles. Using data on all articles of the English-language edition of Wikipedia, we show that diverse teams tend to produce high-quality (or "featured") articles. In contrast, teams of users that individually have diverse interests tend to produce articles of lower quality. These findings are robust with respect to several alternative explanations for article quality. We also show that the proposed model-based normalization of network indicators outperforms an adhoc normalization via more conventional cosine similarity measures. Finally, we analyze the interplay between team diversity and polarization sustained by adherence to behavioral norms predicted by balance theory. Results suggest that diversity can mitigate the – otherwise negative – effect of polarization on team productivity.

**Keywords** online peer-production · team diversity · team productivity · polarization · balance theory · Wikipedia · relational event models

J. Lerner
University of Konstanz, Germany
E-mail: juergen.lerner@uni-konstanz.de

A. Lomi
University of Lugano, Switzerland
University of Exeter Business School, GB
E-mail: alessandro.lomi@usi.ch

## 1 Introduction

Recent years have witnessed the emergence of open peer-production systems in which self-organizing teams of volunteers sustain private costs to provide public goods (Lerner and Tirole, 2001; von Hippel and von Krogh, 2003, 2006). Successful examples include open-source software communities (Conaldi and Lomi, 2013), large-scale collaborative open projects such as Polymath (Franzoni and Sauermann, 2014) or Linux (Lee and Cole, 2003), and Wikipedia (Lerner and Lomi, 2017) – the free online encyclopedia that provides the empirical setting for this paper. The success of open peer-production communities has sometimes been explained by the "wisdom of the crowds" argument (Arazy et al., 2006; Kittur et al., 2007). Self-organizing teams might be able to draw from large pools of contributors with potentially diverse and complementary background knowledge, experience, and capabilities. Building on a well-established empirical regularity in studies of teams in organizations (Horwitz and Horwitz, 2007), we would expect diversity within teams involved in peer-productions to be positively related to team performance expressed in terms of quality of team output (Hong and Page, 2004; Wuchty et al., 2007).

Such expectations about the virtue of diversity are somewhat at odds with recent findings in organizational sociology suggesting that diversity derived from membership in multiple categories, or engagements with multiple genres is detrimental to quality evaluations (Zuckerman, 1999; Zuckerman et al., 2003; Hsu, 2006; Kovács and Hannan, 2010; Phillips et al., 2013; Goldberg et al., 2016). Along a similar line, empirical studies have found that interdisciplinary research proposals (i. e., proposal that combine knowledge across conventional disciplinary boundaries) are typically discounted by funding agencies (Bromham et al., 2016). Clarifying the contentious role that diversity plays in team performance requires a careful distinction between the diversity of the *team* and the diversity of interests of the *individual* members of the team. In this paper we claim that these cross-level concepts of diversity have distinct empirical implications. Specifically, we predict that teams characterized by high levels of diversity will be systematically associated with collective outcomes of higher quality. However, teams containing members with disparate *individual interests* tend to produce outcomes of lower quality.

Both indicators of diversity (team diversity and individual diversity) are functions of the 2-mode network connecting Wikipedia users to the articles they write. We define that two users have high distance (i. e., different interests) if they contribute mostly to different articles. The group of users jointly writing 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 (Uzzi et al., 2013) of users who normally contribute to different articles, see Fig. 1 for illustration.

A complementary notion of diversity is the *individual diversity* of users – a term that we use as shorthand for the diversity of interests of individual users. Two Wikipedia articles have high distance if they are written mostly by different users. A user, in turn, has high individual diversity if she contributes

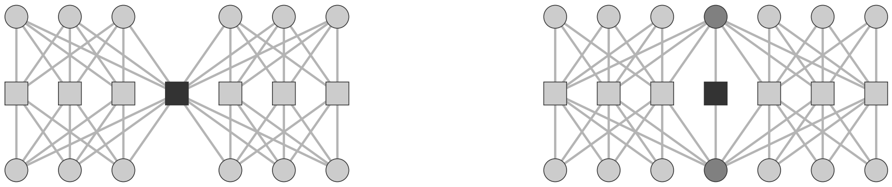

**Fig. 1** Two stylized bipartite collaboration networks. Wikipedia users (circles) contribute to articles (squares). Both networks have two distinct dense clusters that might represent latent topics or knowledge disciplines. *Left:* the black-colored article is written by members from different clusters and therefore has high team diversity; each of its contributors edits mostly within one cluster and therefore has a relatively low individual diversity. *Right:* the two dark-gray contributors in the middle edit articles from both clusters and therefore have high individual diversity (i. e., they are jacks 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.

to articles with high average pairwise distance. Thus, users with high individual diversity contribute to atypical combinations of articles that are not normally co-edited by the same users, see Fig. 1 for illustration.

In light of the illustrating example from Fig. 1 we can say that articles spanning latent topics have high team diversity, while users spanning latent topics have high individual diversity.

We normalize our diversity indicators via reference random graph models that control for the observed degrees of users and articles (i. e., their activity and popularity, respectively), 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.

As one of the main empirical results of this paper, we demonstrate, using data on all articles of the English-language edition of Wikipedia, that article quality is affected positively by diversity of the team, and negatively by individual diversity of team members. We show that these findings are robust with respect to a number of factors that might affect the quality of articles, with respect to controling for indicators of team composition and role diversity, and with respect to varying criteria for measuring article quality. We further analyze effects of diversity separately for articles in different top-level categories to assess how the topic area of the article affects the impact of diversity on quality.

Finally we analyze the interplay between diversity and the internal collaboration structure of the team on all featured articles and a sample of comparable non-featured articles. Specifically we are interested in how team diversity relates to polarization as expressed by adherence to behavioral norms predicted by balance theory (Heider, 1946; Cartwright and Harary, 1956). Findings obtained by extending models from Lerner and Lomi (2019) – who showed that polarization, in general, is associated with lower quality – suggest that the diversity of a team mitigates this negative effect of polarization, so that polarization seems to be less harmful for diverse teams.

In Sect. 2 we overview related work and formulate our hypotheses. Section 3 provides details on the empirical data, the definition of variables (including team diversity, individual diversity, and article quality), and a model used to assess the internal collaboration structure of a team. In Sect. 4 we present and discuss our results and Sect. 5 concludes with a summary of the main findings and indicates potential future work. This article is an extended version of the conference paper Lerner and Lomi (2018a).

## 2 Related Work and Hypotheses

The relation between team diversity and team performance has long been a central issue in the study of formal organizations (Ancona and Caldwell, 1992; Horwitz and Horwitz, 2007). Extant research demonstrates that this relation depends on the type of diversity as well as on a number of contextual or mediating factors (Jehn et al., 1999; Mannix and Neale, 2005; Joshi and Roh, 2009). Indeed, "diversity" itself is a diverse concept, as it can refer to heterogeneity in social or demographic characteristics, attitudes, opinions, interests, background, experience, or social positions, status, or roles (McPherson and Ranger-Moore, 1991). In this paper we focus on the effect of a network-based definition of diversity (Horwitz and Horwitz, 2007), according to which a team of Wikipedia editors is diverse to the extent that it contains members that typically contribute to different articles.

Effects of diversity in Wikipedia (Flöck et al., 2011) have been studied from different perspectives. For instance, it has been shown that diversity moderates the effect of team size on performance in the WikiProject "Film"(Robert and Romero, 2017). Liu and Ram (2011) analyzed how role diversity of Wikipedia teams affects output quality and Shi et al. (2017) 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. (2015) 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 afterward. Lam et al. (2010) also found that groups with intermediate tenure diversity make good decisions.

In a paper that inspired our analysis presented in Sect. 4.6, Arazy et al. (2011) showed that cognitive diversity moderates the effect of task conflict on article quality: the output quality of teams with high diversity benefits from the presence of task conflict, while conflict on less diverse teams is negative for article quality. Our work differs from Arazy et al. (2011) in a fundamental way. We are not analyzing the effect of the *amount* of conflict, but rather the effect of the *structure* of conflictual interaction resulting from edit actions in which users undo or redo the contributions of other users. We show that the negative effect of polarization – as expressed by adherence to the behavioral norms of structural balance (Cartwright and Harary, 1956) – on article quality (which has been demonstrated in Lerner and Lomi (2019)) is less severe in teams with high diversity. Additional differences from Arazy et al. (2011) are the much larger size of the data set we analyze and our operationalization of diversity and conflict.

Previous work that analyzed the effect of diversity of individual users on article quality includes Szejda et al. (2014); Baraniak et al. (2016); Sydow et al. (2017). *Editors' diversity* or *versatility* has been defined in these papers 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 is positive for 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 Szejda et al. (2014); Baraniak et al. (2016); Sydow et al. (2017) 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 writing a particular article has high diversity to the extent that 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" (Uzzi et al., 2013) of users, while a team with low diversity consists mostly of users that typically work together. Likewise, an individual user is said to have high diversity if she contributes to articles that are rarely co-edited by the same users. Thus, a diverse user contributes to atypical combinations of articles.

Based on previous 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, 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 jacks-of-all-trades-are-masters-of-none argument (Hsu, 2006), we hypothesize that individual diversity of users is negative for article quality:

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

Finally, extending ideas from Arazy et al. (2011) we hypothesize that the negative effect of polarization on article quality (Lerner and Lomi, 2019) is mitigated by the diversity of the team:

$H_3$ the higher the diversity of the *team* of users writting a Wikipedia article, the *lower* is the negative effect of polarization on article quality.

## 3 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.

### 3.1 Outcome Variable: Article Quality

The main outcome variable that we consider in this paper is the *quality* of an article, where we say that 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, implying an average rate of slightly less than one FA per 1,000 articles. FA is the highest in Wikipedia's assessment grades: featured articles (FA), A-class, good articles (GA), B-class, C-class, start-class, and stub-class. Wikipedia's article evaluations are often used in academic research (Kittur and Kraut, 2008; Wu et al., 2011; Ransbotham and Kane, 2011; Liu and Ram, 2011; Lerner and Lomi, 2018c) and have been found to be consistent with external evaluations (Kittur and Kraut, 2008). As a robustness check, we estimate models for article quality applying a weaker criterion in which an article is defined to be of high quality if it is featured or a good article (GA). There are 32,194 articles that are FA or GA.

The units of analysis in the first type of models analyzed in this paper are, thus, 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. 3.2 below), which are 4,245,902 articles.

Since the set of all articles is highly imbalanced with respect to the outcome variable (i. e., quality), we perform additional robustness checks on a sample of articles, defined in Lerner and Lomi (2019), that contains all featured articles and roughly the same number of *comparable* non-featured articles. The sampled non-featured articles are selected such that they have similar distributions as the featured articles in basic control variables such as length, age, number of edits, number of contributors, and many more, see details in Lerner and Lomi (2019).

Tests of Hypothesis $H_3$ follow the design of a case-control study (Borgan et al., 1995): we select articles based on the outcome variable split into featured articles ("cases") and comparable non-featured articles ("controls"),

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

mentioned in the previous paragraph. Then we analyze differences in the interplay between polarization and diversity between featured and non-featured articles, extending the models from Lerner and Lomi (2019). In this type of models, the outcome variables are the dynamically changing probabilities that specific users undo contributions of specific other users, modeled as a function of characteristics that indicate how the two users are embedded into the collaboration network, see further details in Sect. 3.10.

### 3.2 The Wikipedia Collaboration Network

The data structure we use to compute team diversity and individual diversity is the weighted 2-mode network connecting Wikipedia users to the articles they write. 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 (Sydow et al., 2017), that $u$ has added to $a$. (If $u$ has never contributed to $a$, we set $w(u, a) = 0$.) Thus, we measure contributions by the provision of content, rather than by editing activity which could be measured, for instance, by the number of edits. 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 (Tsvetkova et al., 2017).

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 (Adler and de Alfaro, 2007; Brandes et al., 2009; Javanmardi et al., 2010; Maniu et al., 2011; Flöck and Acosta, 2014; Lerner and Lomi, 2017) are not applied in this paper to define team diversity. However, survival of the contributions of individual users is analyzed with the relational event models described in Sect. 3.10.2.

As a matter of fact, 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 $u_1, \ldots, u_k$ such that

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

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 compute the index $i_0$ such that

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

which is the minimum index $i$ such that users $u_1, \ldots, u_i$ contribute at least 95% of the total contributions to article $a$. We keep only $u_1, \ldots, u_{i_0}$ as contributors of $a$, that is, we set the weights $w(u_i, a)$ for $i = i_0 + 1, \ldots, k$ 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), $A(u)$ for the set of articles to which user $u$ makes a nonzero contribution, $A$ for the set of all articles, and $U$ for the set of all users.

### 3.3 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

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

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 of interests, or overlap in interests:

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

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”

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

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

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

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

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

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

### 3.4 Model-based Normalization of User Distance

Even though the cosine similarity is an established measure 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 the sake of example, if there were two users contributing 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 in the following 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 Uzzi et al. (2013) who normalized the overlap $c$ in journal citations of scientific papers to the z-score measure

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

To estimate expectation and variance Uzzi et al. (2013) 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* (Caldarelli et al., 2002; De Masi et al., 2006), as summarized below.

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 (which is equal to the total sum of edge weights). Adapting ideas of Caldarelli et al. (2002); De Masi et al. (2006) to weighted 2-mode networks, we define that random weights $W_{ua}$ for $(u,a) \in U \times A$ are drawn independently from a distribution with expectation

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

This random graph has the property that the expected degrees of users and articles (i. e., their activity and popularity) are equal to their observed degrees (Caldarelli et al., 2002; De Masi et al., 2006). Besides this property, the random graph has no clustering into 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)

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

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

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

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*.

### 3.5 Individual Diversity (Extent of Being a Jack of All Trades)

The other main variables used in this paper are indicators for whether the contributors of an article individually 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" (Hsu, 2006), "polymaths", "Renaissance men" (Szejda et al., 2014), or interdisciplinary users with diverse interests.

Individual diversity of users is computed by the same formulas as team diversity of articles after transposing "users" and "articles" in the 2-mode network. More explicitly, if we are given a distance function $\mathit{dist}(a, b)$ for any two articles $a$ and $b$, the *individual diversity* of a user $u$ with $|A(u)| \geq 2$ is defined as the weighted average pairwise distance of its articles, where we set $w_{u;a,b} = w(u, a) + w(u, b)$ as the weight of the pair $(a, b)$ for user $u$. In formulas, individual diversity is defined by

$$\mathit{individual.diversity}(a) = \frac{\sum_{a \neq b \in A(u)} \mathit{dist}(a, b) \cdot w_{u;a,b}}{\sum_{a \neq b \in A(u)} w_{u;a,b}} \quad . \tag{5}$$

Similarity of articles $a$ and $b$ is defined via a measure of overlap of their sets of contributors:

$$c_{ab} = \sum_{u \in U} w(u, a) \cdot w(u, b) \quad , \tag{6}$$

leading to the "cosine similarity" of articles

$$\mathit{sim.cosine}(a, b) = \frac{c_{ab}}{\|a\| \cdot \|b\|} \quad ,$$

where the 2-norm $\|a'\|$ of an article $a'$ is defined by

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

The respective distance of two articles $a$ and $b$ that have at least one common collaborator is defined by

$$dist.cosine(a,b) = -\log(sim.cosine(a,b)) \ ,$$

and by substituting $dist.cosine$ for $dist$ in Eq. (5) we obtain the first indicator of individual diversity, denoted by $individual.diversity.cosine$.

Similar to user-user similarity, we define a model-based normalization for similarity between articles. The z-score of two articles $a$ and $b$, using the fitness model introduced above, is given by

$$z(c_{ab}) = \frac{c_{ab} - \frac{d_a \cdot d_b}{D^2} \cdot \sum_{u \in U} d_u^2}{\frac{d_a \cdot d_b}{D^2} \cdot \sqrt{\sum_{u \in U} d_u^4}} \ .$$

The respective measure of article-article distance is defined by

$$dist.zscore(a,b) = -\mathrm{sign}(z(c_{ab})) \cdot \log(1 + |z(c_{ab})|)$$

and by substituting $dist.zscore$ for $dist$ in Eq. (5) we obtain the second indicator of individual diversity denoted by $individual.diversity.zscore$ as an alternative to $individual.diversity.cosine$.

Finally, articles are assigned the weighted average individual diversity of their contributors. That is, for an article $a$ we define

$$individual.diversity(a) = \frac{\sum_{u \in U(a)} w(u,a) \cdot individual.diversity(u)}{\sum_{u \in U(a)} w(u,a)} \ ,$$

for both, the cosine and z-score variant of individual diversity.

Correlation among the various indicators of diversity over all articles is given in Table 1. We can see that the cosine-based indicator of team diversity and the z-score based indicator of team diversity have a relatively strong positive correlation (0.77). On the other hand, the correlation between the two measures of individual diversity (ind.div.cosine and ind.div.zscore) is much lower but still positive (0.39). For the cosine-based measures we observe a weakly positive correlation between team diversity and and individual diversity ($cor(team.div.cosine, ind.div.cosine) = 0.24$), while we observe for the respective z-score-based measures a weakly negative correlation ($cor(team.div.zscore, ind.div.zscore) = -0.38$).

**Table 1** Sample correlation among diversity indicators.

| | team.div.cosine | team.div.zscore | ind.div.cosine | ind.div.zscore |
|---|---|---|---|---|
| team.div.cosine | · | 0.77 | 0.24 | -0.04 |
| team.div.zscore | · | · | 0.06 | -0.38 |
| ind.div.cosine | · | · | · | 0.39 |

### 3.6 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. We include in our models the following control variables.

The variables that have the highest predictive power for the quality of articles are indicators of article size (Blumenstock, 2008) and the amount of work invested in writting 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*, where the last two variables capture the so-called reading complexity (Ransbotham and Kane, 2011). 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 (Lerner and Lomi, 2018c). Further, more specific control variables are introduced in the following two sections.

The correlation of the four diversity indicators with all control variables are given in Table 2. The strongest correlations (in absolute value) can be found between indicators of team diversity and the length and number of edits of the article. Since article length and number of edits have such a strong – and fairly obvious – influence on the probability of being featured (Blumenstock, 2008) it is necessary to control for these basic characteristics in a model explaining article quality by diversity.

### 3.7 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* (Simpson, 1951; Blyth, 1972; Barbosa et al., 2016). To test the robustness of our findings against this conjecture we assign articles

**Table 2** Sample correlation of diversity indicators with control variables.

| | len | age | edits | rvs | team | ch/w | w/sen | sect.1 | sect.2 |
|---|---|---|---|---|---|---|---|---|---|
| team.div.c | -0.46 | 0.03 | -0.29 | -0.18 | -0.21 | 0.00 | -0.02 | -0.28 | -0.18 |
| team.div.z | -0.50 | -0.07 | -0.47 | -0.36 | -0.39 | 0.04 | 0.01 | -0.36 | -0.22 |
| ind.div.c | 0.07 | 0.26 | 0.26 | 0.24 | 0.29 | -0.09 | -0.02 | 0.08 | 0.05 |
| ind.div.z | 0.17 | 0.23 | 0.37 | 0.35 | 0.38 | -0.10 | -0.03 | 0.16 | 0.07 |

| | cats | c.size | links | refs | imgs | templ | langs | gran |
|---|---|---|---|---|---|---|---|---|
| team.div.c | -0.04 | -0.05 | -0.29 | -0.35 | -0.14 | -0.21 | -0.04 | -0.23 |
| team.div.z | -0.00 | 0.02 | -0.24 | -0.37 | -0.10 | -0.10 | 0.00 | -0.00 |
| ind.div.c | 0.09 | -0.05 | 0.06 | 0.04 | 0.01 | -0.04 | 0.07 | -0.22 |
| ind.div.z | 0.04 | -0.00 | 0.02 | 0.16 | -0.03 | -0.08 | 0.00 | -0.21 |

to one or several of 21 top-level categories (TLC) (Kittur et al., 2009), as described in Lerner and Lomi (2018c). We then extend our models by 21 binary variables that indicate if the article is in the respective TLC. We also estimate more complex models in which we interact our variables of interest (team diversity and average individual diversity) with all 21 TLC indicators. These more complex models allow us to assess whether the impact of diversity on quality varies across topic areas.

### 3.8 Interest Diversity vs. Role Diversity

A further alternative explanation against which we test our findings is the conjecture that team diversity (as we define it in this paper) might just be a byproduct of varying *composition* of the team. As a matter of fact different users have very different levels of activity and tend to perform different *tasks* or play different *roles* (Liu and Ram, 2011). These differences could influence both, team diversity and FA-probability.

To assess varying composition of teams, we first define three numerical variables, assigned to users, that capture activity levels in three different tasks – or roles. 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 $avg.x(a) = 0$.

### 3.9 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 explanatory variable, except the binary TLC indicator variables, we subtract its mean and divide by its standard deviation. This normalization makes parameter sizes better comparable. More explicitly, if we estimate a parameter $\alpha$ for an explanatory variable $x$, then hypothetically increasing $x$ by one standard deviation (that is, by one) multiplies the predicted log-odds for being featured by $\exp(\alpha)$. For instance, a parameter $\alpha = 0.1$ implies an increase in the log-odds by 10.5%, a parameter $\alpha = 0.5$ implies an increase by 65%. Since baseline probabilities are very small, an increase in the log-odds is roughly equal to an increase in the probability.

### 3.10 Differences in the structure of collaboration networks: the interplay between diversity and polarization

In a different set of models we assess how the structure of collaboration networks *within* teams differs between high-quality and low-quality articles and how this relation is moderated by team diversity. The data used in this analysis is much more fine-grained than that used in the tests for the direct effects of diversity on article quality since we analyze individual edit actions in which users undo or redo contributions of other users at given points in time.

Previous work (Brandes et al., 2009; Lerner and Lomi, 2017) advocated the approach that *undo* events, that is, edits in which users make contributions of other users undone, are considered as negative interaction expressing disagreement and that *redo* events, that is, edits in which users restore contributions of other users, are considered as positive interaction expressing agreement. These signed edit events are weighted by the number of words undone or redone and induce an emergent network resulting from collaborative article writting in Wikipedia. Lerner and Lomi (2018b, 2019) suggested that the decisions to undo or redo content of others can in part be explained by the rules predicted by the theory of structural balance (Heider, 1946; Cartwright and Harary, 1956): actors are predicted to have a positive attitude towards the friends of their friends and towards the enemies of their enemies and a negative attitude towards the friends of their enemies and the enemies of their friends. (The terms "friend" and "enemy" have to be understood metaphorically as users who strongly argree or disagree, respectively.) These behavioral rules reduce cognitive dissonance (Festinger, 1962) and lead to a cognitively stable state (Heider, 1946). It is known that perfect agreement with these rules gives rise to signed networks that partition into two factions of nodes such that all positive ties are within factions and all negative ties are between factions (Cartwright and Harary, 1956). Therefore, adherence to the rules of balance theory can be considered as a network-based measure of polarization (Esteban and Ray, 1994).

Polarization – but also the desire to reduce cognitive dissonance or the tendency to consider the enemy of an enemy as a friend – often has negative consequences (Lord et al., 1979; Friedkin et al., 2016; Akerlof and Dickens, 1982; Saperstein, 2004). Based on these insights, Lerner and Lomi (2019) claimed and demonstrated that Wikipedia teams producing high-quality articles act in weaker agreement with the behavioral rules predicted by balance theory. In this paper we test Hypothesis $H_3$ stating that team diversity can mitigate this negative effect of polarization, since conflict – despite having a negative effect on output quality in general – can also induce actors to question their own beliefs and re-consider diverse viewpoints (Arazy et al., 2011). In the following we provide details on how we implement tests of Hypothesis $H_3$ in this paper.

### *3.10.1 Overview: relating team performance to network structure*

The analysis that relates the structure of the collaboration network to the quality of the resulting articles follows the design of a case-control study (Borgan et al., 1995): articles are selected and assigned to sub-samples dependent on the outcome variable (featured or not) as described in Lerner and Lomi (2019). We then analyze how the dynamic patterns explaining dyadic undo events (see Sect. 3.10.2 below) differ between featured and non-featured articles. In particular, we assess whether teams producing high-quality articles act more or less in accordance with balance theory and how this relation gets moderated by the diversity of the team and the average individual diversity of its members.

### *3.10.2 Relational event models explaining dyadic undo probabilities*

We apply relational event models for Wikipedia edit networks that fit into the framework proposed in Lerner and Lomi (2017). These models assume that for each pair of users $(u, v)$ contributing to a common article and for each point in time $t$ there is a latent probability $prob.undo(u, v; t)$ predicting how likely it is that user $u$ makes contributions of user $v$ undone at time $t$. When comparing the article's text from version to version we can determine the actual amount of undo. For the sake of example, assume that when user $u$ uploads a new version of the article at time $t$, she could potentially undo 100 words, contributed by user $v$ at some time before $t$. User $u$ may decide to undo a fraction of this text; for instance, if $u$ deleted 40 out of these 100 words, we would obtain an observed undo ratio of 0.4 for the dyad $(u, v)$ at time $t$. Models for the edit network specify the time-varying dyadic latent probabilities $prob.undo(u, v; t)$ by logistic regression such that they best explain the observed undo ratios.

The undo probability $prob.undo(u, v; t)$ is specified as a function of variables indicating how the dyad $(u, v)$ is embedded into the network of past positive and negative interaction that happened before time $t$. Of particular interest for this paper is a variable, denoted by $SB(u, v; t)$ – where $SB$ stands for "structural balance" – expressing to what extent $u$ and $v$ have common friends or common enemies and subtracting the extent to which $u$ and $v$ are connected to a third user who is a friend of $u$ and an enemy of $v$, or vice versa. Technically, $SB(u, v; t)$ is defined to be the sum of $friend.of.friend(u, v; t)$ and $enemy.of.enemy(u, v; t)$ minus the sum of $friend.of.enemy(u, v; t)$ and $enemy.of.friend(u, v; t)$, where the latter four variables are defined in Lerner and Lomi (2019). Structural balance theory predicts that the higher the value of $SB(u, v; t)$, the more friendly and less hostile user $u$ perceives user $v$. This in turn would lead to a decreased undo probability $prob.undo(u, v; t)$, so that balance theory predicts a negative parameter associated with $SB(u, v; t)$ in models explaining undo events. We include further network effects controling for past dyadic interaction, degree effects, and the *reputation* of users; see Lerner and Lomi (2019) and Tables 13 and 14.

To test the hypothesis that team diversity can mitigate the negative effect of polarization (as expressed by adherence to the rules of balance theory), we interact the variable $SB$ with the variable for team diversity and with a binary indicator for whether the edit event happens on a featured article or not. We hypothesize (1) that teams producing featured articles, in general, act in weaker agreement with balance theory (so that the parameter associated with the interaction effect $SB \times featured$ is expected to be positive) and (2) that this does not hold to the same extent for teams with high diversity (so that the parameter associated with the three-way interaction effect $SB \times featured \times team.div$ is expected to be negative).

## 4 Results and Discussion

### 4.1 Diverse teams vs. jacks of all trades

Table 3 reports estimated parameters of logit models for the probability that articles are featured (FA) as a function of *team.div.cosine*, *team.div.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.

**Table 3** Logistic regression modeling FA-probability dependent on team diversity. Estimated parameters and standard errors (in brackets). Effects related to our hypotheses are in **bold font**.

| | Team.C | Team.Z | Team.C+Z |
|---|---|---|---|
| (Intercept) | −11.23 (0.06)$^{***}$ | −11.28 (0.07)$^{***}$ | −11.31 (0.07)$^{***}$ |
| log1p.length | 2.70 (0.04)$^{***}$ | 2.72 (0.04)$^{***}$ | 2.67 (0.04)$^{***}$ |
| age | 1.06 (0.03)$^{***}$ | 1.01 (0.03)$^{***}$ | 0.99 (0.03)$^{***}$ |
| log1p.#edits | 2.18 (0.06)$^{***}$ | 2.26 (0.06)$^{***}$ | 2.25 (0.06)$^{***}$ |
| log1p.#reverts | 1.10 (0.03)$^{***}$ | 1.10 (0.03)$^{***}$ | 1.09 (0.03)$^{***}$ |
| log1p.#teamsize | −3.44 (0.05)$^{***}$ | −3.34 (0.05)$^{***}$ | −3.23 (0.06)$^{***}$ |
| log1p.#wiki.links | −0.81 (0.03)$^{***}$ | −0.88 (0.04)$^{***}$ | −0.91 (0.04)$^{***}$ |
| log1p.#external.refs | −0.14 (0.02)$^{***}$ | −0.15 (0.02)$^{***}$ | −0.16 (0.02)$^{***}$ |
| log1p.#lang.links | 0.39 (0.02)$^{***}$ | 0.36 (0.02)$^{***}$ | 0.34 (0.02)$^{***}$ |
| #level.1.sections | −0.33 (0.02)$^{***}$ | −0.33 (0.02)$^{***}$ | −0.32 (0.02)$^{***}$ |
| #level.2.sections | −0.44 (0.01)$^{***}$ | −0.43 (0.01)$^{***}$ | −0.43 (0.01)$^{***}$ |
| log1p.#images | 0.16 (0.02)$^{***}$ | 0.16 (0.02)$^{***}$ | 0.16 (0.02)$^{***}$ |
| log1p.#templates | 0.60 (0.03)$^{***}$ | 0.59 (0.03)$^{***}$ | 0.59 (0.03)$^{***}$ |
| #characters.per.word | −0.73 (0.03)$^{***}$ | −0.71 (0.03)$^{***}$ | −0.70 (0.03)$^{***}$ |
| #words.per.sentence | −0.29 (0.05)$^{***}$ | −0.28 (0.05)$^{***}$ | −0.28 (0.05)$^{***}$ |
| #categories | 0.05 (0.01)$^{***}$ | 0.04 (0.01)$^{***}$ | 0.04 (0.01)$^{***}$ |
| log1p.avg.cat.size | −0.08 (0.03)$^{**}$ | −0.09 (0.03)$^{***}$ | −0.10 (0.03)$^{***}$ |
| granularity | 0.52 (0.03)$^{***}$ | 0.47 (0.04)$^{***}$ | 0.43 (0.04)$^{***}$ |
| **team.div.cosine** | **0.39 (0.03)**$^{***}$ | | **−0.39 (0.05)**$^{***}$ |
| **team.div.zscore** | | **0.52 (0.03)**$^{***}$ | **0.83 (0.05)**$^{***}$ |
| AIC | 38,162.48 | 37,950.46 | 37,902.57 |
| BIC | 38,414.44 | 38,202.43 | 38,167.80 |
| Num. obs. | 4,245,902 | 4,245,902 | 4,245,902 |

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

The first three models in Table 4 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.

**Table 4** Logistic regression modeling FA-probability dependent on individual diversity (and both team and individual diversity). The same control variables as reported in Table 3 are included in the model but are not reported here. All parameters are significantly different from zero at the 0.1% level.

| | Ind.C | Ind.Z | Ind.C+Z | Team+Ind.Z |
|---|---|---|---|---|
| | *(all control variables from Sect. 3.6 included)* | | | |
| **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)** |
| **team.div.zscore** | | | | **0.40 (0.03)** |
| AIC | 38,303.15 | 37,579.22 | 37,338.96 | 37,351.17 |
| BIC | 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 |

The right-most model (*Team+Ind.Z*) in Table 4 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. To check whether non-independence of observations could bias our results, we also performed a robust parameter estimation (Carroll and Pederson, 1993) of the model *Team+Ind.Z.* Neither the directions (signs) of parameters nor their significance levels changed.

## 4.2 Defining FA and GA as high-quality articles

Tables 5 and 6 report findings for models that have exactly the same explanatory variables as those from Tables 3 and 4 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 5** Logistic regression modeling the probability that articles are FA or GA dependent on team diversity. All parameters are significantly different from zero at the 0.1% level.

| | Team.C | Team.Z | Team.C+Z |
|---|---|---|---|
| | *(all control variables from Sect. 3.6 included)* | | |
| **team.div.cosine** | **0.10 (0.01)** | | **−0.29 (0.02)** |
| **team.div.zscore** | | **0.20 (0.01)** | **0.42 (0.02)** |
| AIC | 192,292.62 | 191,961.57 | 191,753.70 |
| BIC | 192,544.59 | 192,213.54 | 192,018.92 |
| Num. obs. | 4,245,902 | 4,245,902 | 4,245,902 |

**Table 6** Logistic regression modeling the probability that articles are FA or GA dependent on individual diversity (and both, team and individual diversity). All parameters are significantly different from zero at the 0.1% level.

| | Ind.C | Ind.Z | Ind.C+Z | Team+Ind.Z |
|---|---|---|---|---|
| | *(all control variables from Sect. 3.6 included)* | | | |
| **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)** |
| **team.div.zscore** | | | | **0.10 (0.01)** |
| AIC | 192,322.90 | 189,480.96 | 188,643.04 | 189,381.69 |
| BIC | 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 |

## 4.3 Controling for composition and role diversity

Table 7 reports estimated parameters where we extend the rightmost model from Table 4 by six variables for the average composition and role diversity of teams, introduced in Sect. 3.8. Our findings related to team diversity and individual diversity (where we consider diversity of interests, rather than role diversity) do not change.

**Table 7** Logit model for FA-probability including indicators of average composition and role diversity of teams, introduced in Sect. 3.8. All parameters are significantly different from zero at the 0.1% level.

| | |
|---|---|
| *(all control variables from Sect. 3.6 included)* | |
| **team.div.zscore** | **0.31 (0.03)** |
| **ind.div.zscore** | **−1.10 (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 |

## 4.4 Effect of diversity across topical domains

Table 8 reports estimated parameters where we extend the rightmost model from Table 4 by 21 binary indicator variables encoding membership of articles in top-level categories (TLC). This model allows for varying baseline probabilities in the different TLC. Our main findings are robust in the sense that team diversity continues to have a positive effect and individual diversity continues

**Table 8** Logit model for FA-probability with varying baseline probabilities for different top-level categories (TLC)

| *(all control variables from Sect. 3.6 included)* | |
|---|---|
| **team.div.zscore** | **0.39 (0.03)**$^{***}$ |
| **ind.div.zscore** | **−0.81 (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 |

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

to have a negative effect on article quality. Articles belonging to some TLC indeed have significantly different probabilities to be of high quality. For instance, articles belonging to *Arts* or *Humanities* have higher FA-probabilities, while articles in *Politics*, *Geography*, or *Sports* have lower FA-probabilities.

We further estimated a model in which we include the interaction effects of team diversity with all TLC indicators, reported in Table 9. (A respective model interacting individual diversity with all TLC indicators is reported in Table 10.) We observe that the base effect of *team.diversity.zscore* is positive (parameter equal to 0.63) so that, in general, as diversity of teams increases, so does the likelihood of producing a high quality article. The effect of team diversity in some TLC is significantly different from this baseline effect. The topical domain in which team diversity has the weakest effect on quality is *Mathematics* for which the interaction effect *team.div.zscore:Mathematics* is equal to −0.62. Thus, for an article that is in *Mathematics* (but in no other TLC), a team whose diversity is by one standard deviation higher than the average produces a featured article with a probability that is $\exp(0.63-0.62) = 1.01$ times the baseline FA-probability – all other things being equal. Considering the associated standard errors we can conclude that team diversity has no significant

**Table 9** Effect of team diversity on FA probability separately by TLC.

| | |
|---|---|
| *(all control variables from Sect. 3.6 included)* | |
| *(all TLC base effects from Table. 8 included)* | |
| **team.div.zscore** | **0.63 (0.04)**$^{***}$ |
| **team.div.zscore:Arts** | **0.07 (0.06)** |
| **team.div.zscore:Culture** | **−0.04 (0.06)** |
| **team.div.zscore:History** | **−0.28 (0.06)**$^{***}$ |
| **team.div.zscore:Humanities** | **−0.05 (0.06)** |
| **team.div.zscore:Politics** | **−0.04 (0.09)** |
| **team.div.zscore:Geography** | **−0.28 (0.07)**$^{***}$ |
| **team.div.zscore:World** | **0.18 (0.08)**$^{*}$ |
| **team.div.zscore:Events** | **0.02 (0.07)** |
| **team.div.zscore:Life** | **−0.09 (0.09)** |
| **team.div.zscore:Nature** | **−0.12 (0.09)** |
| **team.div.zscore:Philosophy** | **0.31 (0.23)** |
| **team.div.zscore:People** | **−0.07 (0.05)** |
| **team.div.zscore:Science_and_technology** | **−0.07 (0.12)** |
| **team.div.zscore:Sports** | **−0.16 (0.09)** |
| **team.div.zscore:Health** | **−0.03 (0.10)** |
| **team.div.zscore:Society** | **0.11 (0.07)** |
| **team.div.zscore:Law** | **0.06 (0.12)** |
| **team.div.zscore:Religion** | **0.23 (0.10)**$^{*}$ |
| **team.div.zscore:Mathematics** | **−0.62 (0.23)**$^{**}$ |
| **team.div.zscore:Matter** | **−0.08 (0.15)** |
| **team.div.zscore:Reference_works** | **0.81 (0.49)** |
| AIC | 36,989.39 |
| BIC | 37,798.34 |
| Num. obs. | 4,245,902 |

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

effect on quality for articles in *Mathematics*. This, however, is an exception. The TLC with the second lowest effect of team diversity are *History* and *Geography*. Articles in these two categories have their FA-probabilities multiplied by $\exp(0.63 - 0.28) = 1.41$ if their team diversity increases by one standard deviation. This is lower than the overall effect of team diversity – but still positive. Articles in the TLC *Religion* seem to benefit even more than the average from team diversity. For an article in *Religion* the FA-probability gets multiplied by $\exp(0.63+0.23) = 2.36$ if the team diversity increases by one standard deviation. In conclusion, Hypothesis $H_1$ – claiming that diverse teams tend to produce Wikipedia articles of higher quality – can be validated in all TLC, except *Mathematics* (in which the effect of team diversity is near-absent).

Table 10 reports parameters of a model in which we include the interaction effects of individual diversity with all TLC indicator variables. We observe that the baseline effect of individual diversity in that model is negative (parameter equal to $-0.80$) so that jacks of all trades are found to do poor work, in general. The effect of individual diversity on quality is significantly different in some TLC. It seems to be most harmful in *Religion* (where the parameter is decreased by 0.45) and much less harmful in *Mathematics* (where the parameter is increased by 0.61). Individual diversity seems to have almost no effect

**Table 10** Effect of individual diversity separately by TLC.

| | |
|---|---|
| *(all control variables from Sect. 3.6 included)* | |
| *(all TLC base effects from Table. 8 included)* | |
| **ind.div.zscore** | **−0.80 (0.05)**$^{***}$ |
| **ind.div.zscore:Arts** | **−0.16 (0.07)**$^{*}$ |
| **ind.div.zscore:Culture** | **0.12 (0.07)** |
| **ind.div.zscore:History** | **0.02 (0.06)** |
| **ind.div.zscore:Humanities** | **−0.02 (0.06)** |
| **ind.div.zscore:Politics** | **0.41 (0.08)**$^{***}$ |
| **ind.div.zscore:Geography** | **0.27 (0.08)**$^{***}$ |
| **ind.div.zscore:World** | **−0.37 (0.08)**$^{***}$ |
| **ind.div.zscore:Events** | **−0.06 (0.07)** |
| **ind.div.zscore:Life** | **−0.35 (0.09)**$^{***}$ |
| **ind.div.zscore:Nature** | **0.15 (0.09)** |
| **ind.div.zscore:Philosophy** | **−0.00 (0.26)** |
| **ind.div.zscore:People** | **−0.36 (0.06)**$^{***}$ |
| **ind.div.zscore:Science_and_technology** | **−0.15 (0.12)** |
| **ind.div.zscore:Sports** | **0.79 (0.09)**$^{***}$ |
| **ind.div.zscore:Health** | **−0.26 (0.10)**$^{**}$ |
| **ind.div.zscore:Society** | **0.12 (0.07)** |
| **ind.div.zscore:Law** | **−0.27 (0.14)**$^{*}$ |
| **ind.div.zscore:Religion** | **−0.45 (0.11)**$^{***}$ |
| **ind.div.zscore:Mathematics** | **0.61 (0.24)**$^{*}$ |
| **ind.div.zscore:Matter** | **0.27 (0.15)** |
| **ind.div.zscore:Reference_works** | **−0.67 (0.59)** |
| AIC | 36,396.57 |
| BIC | 37,205.52 |
| Num. obs. | 4,245,902 |

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

on quality for articles that are in the TLC *Sports* (but in no other TLC). In conclusion, Hypothesis $H_2$ claiming that jacks of all trades tend to do poor work is supported in most TLC, with the exception of *Sports* (and perhaps *Mathematics*).

### 4.5 Analysis of a balanced sample of articles

As we have noted the baseline probability that a Wikipedia article is featured is very low. Less than one in 1,000 articles belongs to the FA category. The estimation of logistic regression models for a response variable that is so unbalanced might be problematic since even a small absolute increase in the probability can yield a significant relative increase. To perform further robustness checks we fit models for article quality on a balanced sample of articles, defined in Lerner and Lomi (2019), that contains all featured articles and roughly the same number of *comparable* non-featured articles. The sampled non-featured articles are selected such that they have similar distributions as the featured articles in the basic control variables introduced in Sect. 3.6 (see Lerner and Lomi (2019) for details on the selection process). Tables 11 and 12

**Table 11** Logistic regression for FA-probability estimated on the balanced sample of articles from Lerner and Lomi (2019). All parameters are significantly different from zero at the 0.1% level.

| | Team.Z | Ind.Z | Team+Ind.Z |
|---|---|---|---|
| *(all control variables from Sect. 3.6 included)* | | | |
| **team.div.zscore** | **0.48 (0.04)** | | **0.31 (0.04)** |
| **ind.div.zscore** | | **−0.86 (0.04)** | **−0.77 (0.05)** |
| AIC | 12,592.47 | 12,357.52 | 12,289.51 |
| BIC | 12,728.29 | 12,493.34 | 12,432.48 |
| Num. obs. | 9,401 | 9,401 | 9,401 |

**Table 12** Logistic regression for FA-probability estimated on the balanced sample of articles from Lerner and Lomi (2019). This model has no control variables (apart from an intercept). All parameters are significantly different from zero at the 0.1% level.

| | Team.Z | Ind.Z | Team+Ind.Z |
|---|---|---|---|
| **team.div.zscore** | **0.33 (0.03)** | | **0.16 (0.03)** |
| **ind.div.zscore** | | **−0.59 (0.03)** | **−0.55 (0.03)** |
| AIC | 12,909.19 | 12,619.75 | 12,594.94 |
| BIC | 12,923.48 | 12,634.05 | 12,616.39 |
| Num. obs. | 9,401 | 9,401 | 9,401 |

report parameters on the effects of team diversity and individual diversity estimated on this balanced sample. The two tables differ in that models reported in Table 11 include all control variables (which have also been used to generate the balanced sample) while Table 12 has no control variables (except an intercept). Results for both variants give further support for the main findings that team diversity is positive for article quality and individual diversity is negative.

## 4.6 Diversity as a moderator for the effect of polarization

We finally turn to the question if and how diversity moderates the effect of polarization on quality, related with Hypothesis $H_3$. We estimate relational event models explaining the probabilities that particular users undo the contributions of particular other users, introduced in Sect. 3.10, on the sample of articles from Lerner and Lomi (2019) which has also been used in Sect. 4.5. Units of analysis in these relational event models, however, are not articles but individual edit decisions: for each triplet $(u, v; t)$ where $u$ and $v$ are two different users and $t$ is a point in time in which $u$ could undo a certain amount of text contributed by $v$, we specify $prob.undo(u, v; t)$, that is, the probability that $u$ makes contributions of $v$ undone, by logistic regression. Further details on these models are given in Lerner and Lomi (2017, 2019).

The particular purpose of this paper is to assess how team diversity and individual diversity moderate the effect of polarization on quality. Polarization

**Table 13** Logit model for dyadic undo probabilities including effects of team diversity. All parameters are significant at the 0.1% level.

| | undo model | × featured |
|---|---:|---:|
| (Intercept) | −1.8870 (0.0004) | −1.8934 (0.0005) |
| number.of.users | 0.2157 (0.0001) | 0.2164 (0.0001) |
| reputation.of.source | −0.3515 (0.0002) | −0.3513 (0.0002) |
| reputation.of.target | −0.8331 (0.0001) | −0.8315 (0.0001) |
| undo.repetition | 0.1217 (0.0001) | 0.1230 (0.0001) |
| undo.reciprocation | 0.0421 (0.0001) | 0.0429 (0.0001) |
| redo.repetition | −0.0303 (0.0001) | −0.0285 (0.0001) |
| redo.reciprocation | −0.1848 (0.0001) | −0.1840 (0.0001) |
| undo.outdegree.source | 0.8813 (0.0009) | 0.8855 (0.0009) |
| undo.indegree.source | −0.6881 (0.0006) | −0.6945 (0.0006) |
| undo.outdegree.target | 0.1668 (0.0003) | 0.1655 (0.0003) |
| undo.indegree.target | 0.3480 (0.0002) | 0.3489 (0.0002) |
| redo.outdegree.source | 0.1941 (0.0007) | 0.1915 (0.0007) |
| redo.indegree.source | 0.1737 (0.0003) | 0.1758 (0.0003) |
| redo.outdegree.target | −0.1891 (0.0003) | −0.1926 (0.0003) |
| redo.indegree.target | −0.2778 (0.0002) | −0.2825 (0.0002) |
| SB | −0.3525 (0.0003) | −0.3739 (0.0005) |
| **team.div.zscore** | **−0.1635 (0.0002)** | **−0.1246 (0.0004)** |
| **SB:team.div.zscore** | **−0.0549 (0.0002)** | **−0.0365 (0.0003)** |
| featured | | 0.0120 (0.0007) |
| SB:featured | | 0.0339 (0.0006) |
| **team.div.zscore:featured** | | **−0.0851 (0.0005)** |
| **SB:team.div.zscore:featured** | | **−0.0330 (0.0004)** |
| AIC | 325,675,571.5082 | 325,354,414.7885 |
| BIC | 325,675,835.1820 | 325,354,733.9725 |
| Num. obs. | 7,862,108 | 7,862,108 |

is measured by adherence to the rules predicted by balance theory and is operationalized in the explanatory variable $SB$ (defined in Sect. 3.10) for which a negative parameter, indicating a decreased undo probability, that is, a more positive assessment, supports balance theory. It has been found in previous work that teams producing featured articles act in weaker accordance with balance theory and, thus, exhibit lower degrees of polarization, than teams producing articles of lower quality (Lerner and Lomi, 2019). In this section we assess the impact of diversity on this relation between polarization and diversity.

Table 13 reports estimated parameters in models that assess (among others) the effects of team diversity on dyadic undo probabilities. We find that $SB$ has a significantly negative effect on undo probabilities, supporting the predictions of balance theory. We also find that the interaction effect *SB:featured* is positive so that teams producing high-quality articles adhere less to the behavioral norms predicted by balance theory. We find that *team.div.zscore* has a decreasing baseline effect on dyadic undo probabilities, so that users in more diverse teams typically undo less. The interaction effect *SB:team.div.zscore* is negative implying that more diverse teams act in stronger agreement with balance theory. That is, they are more reluctant (than less diverse teams) to

**Table 14** Logit model for dyadic undo probabilities including effects of individual diversity. All parameters are significant at the 0.1% level.

| | undo model | × featured |
|---|---:|---:|
| (Intercept) | −1.8877 (0.0004) | −2.0399 (0.0005) |
| number.of.users | 0.1865 (0.0002) | 0.1904 (0.0002) |
| reputation.of.source | −0.3508 (0.0002) | −0.3513 (0.0002) |
| reputation.of.target | −0.8317 (0.0001) | −0.8302 (0.0001) |
| undo.repetition | 0.1219 (0.0001) | 0.1235 (0.0001) |
| undo.reciprocation | 0.0430 (0.0001) | 0.0439 (0.0001) |
| redo.repetition | −0.0277 (0.0001) | −0.0261 (0.0001) |
| redo.reciprocation | −0.1837 (0.0001) | −0.1832 (0.0001) |
| undo.outdegree.source | 0.8995 (0.0009) | 0.9035 (0.0009) |
| undo.indegree.source | −0.6878 (0.0006) | −0.6942 (0.0006) |
| undo.outdegree.target | 0.1808 (0.0003) | 0.1801 (0.0003) |
| undo.indegree.target | 0.3497 (0.0002) | 0.3499 (0.0002) |
| redo.outdegree.source | 0.1774 (0.0007) | 0.1747 (0.0007) |
| redo.indegree.source | 0.1783 (0.0003) | 0.1812 (0.0003) |
| redo.outdegree.target | −0.1949 (0.0003) | −0.2000 (0.0003) |
| redo.indegree.target | −0.2758 (0.0002) | −0.2802 (0.0002) |
| SB | −0.3368 (0.0003) | −0.3972 (0.0005) |
| **ind.div.zscore** | **0.1947 (0.0003)** | **0.2713 (0.0004)** |
| **SB:ind.div.zscore** | **0.0537 (0.0002)** | **0.0644 (0.0004)** |
| featured | | 0.2533 (0.0006) |
| SB:featured | | 0.0801 (0.0006) |
| **ind.div.zscore:featured** | | **−0.1044 (0.0005)** |
| **SB:ind.div.zscore:featured** | | **0.0031 (0.0005)** |
| AIC | 325,668,345.0008 | 325,271,439.6072 |
| BIC | 325,668,608.6746 | 325,271,758.7912 |
| Num. obs. | 7,862,108 | 7,862,108 |

undo contributions of the friends of their friends or the enemies of their enemies and more inclined to undo contributions of the enemies of their friends or the friends of their enemies. This implies that diversity seems to foster polarization. Last but not least we find a negative parameter associated with the three-way interaction effect *SB:team.div.zscore:featured* implying that diverse team that produce high-quality articles act in even stronger agreement with balance theory. Thus, the baseline finding that teams producing featured articles are less guided by balance processes (that is, they have weaker tendencies to polarize) gets moderated by team diversity: for diverse teams polarization does not seem to be associated with lower quality of the team output.

Table 14 reports estimated parameters in models that assess (among others) the effects of individual diversity (that is, the extent to which team members are jacks of all trades) on dyadic undo probabilities. As in Table 13 we find that *SB* has a significantly negative effect on undo probabilities – supporting the predictions of balance theory – and we also find that the interaction effect *SB:featured* is positive so that teams producing high-quality articles adhere less to the behavioral norms predicted by balance theory. We find that *ind.div.zscore* has – in contrast to team diversity – an increasing baseline effect on dyadic undo probabilities, so that users in teams composed of jacks of all trades typically undo more. The interaction effect *SB:ind.div.zscore* is positive implying that users with higher individual diversity act in weaker agreement with balance theory. Last but not least – and in contrast to the respective finding for team diversity – we find a positive parameter associated with the three-way interaction effect *SB:ind.div.zscore:featured* implying that teams that are composed of jacks of all trades and that produce high-quality articles act in weaker agreement with balance theory. Thus, the baseline finding that teams producing featured articles are less guided by balance processes (that is, they have weaker tendencies to polarize) gets even amplified individual diversity: for teams composed of jacks of all trades polarization seems to be even more harmful for the quality of the team output than for teams with an average level of individual diversity.

## 5 Conclusion and Future Work

One hypothetical 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 interest-based diversity rather than other – no less relevant – variants, such as social or demographic diversity, tenure diversity, or role diversity. We stipulate that two users have different interests (i. e., they are distant) if they contribute mostly to different articles. The team of users jointly writing an article, in turn, is said to be diverse if it is mostly composed by users with high pairwise distance. Thus, articles with high team diversity are written by atypical combinations of users who do not normally work together.

A complementary variable used in this paper is the individual diversity of users, that is, their extent of being a "jack of all trades". Two Wikipedia articles have high distance if they are written mostly by different users. A user, in turn, is said to have high individual diversity 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 (Hsu, 2006) – we hypothesize that individual diversity of users is detrimental for article quality.

We have found strong empirical support for both hypotheses. 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 obtain 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. With very few exceptions these findings are qualitatively the same if we analyze effects of team diversity or individual diversity separately for articles in 21 topic areas – although the strength of effects varies across topics.

It is noteworthy that our findings on the relation between individual diversity and quality is contrary to findings of Szejda et al. (2014); Baraniak et al. (2016); Sydow et al. (2017) in the sense that we find individual diversity to be detrimental for article quality while these authors 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 Szejda et al. (2014); Baraniak et al. (2016); Sydow et al. (2017) 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.

With a rather different set of models we analyzed the moderating effect that team diversity, or individual diversity, has on the relation between polarization and output quality. Interaction between diversity and conflict in Wikipedia, and its relation to team performance, has been analyzed before (Arazy et al., 2011) but not on the scale as in this paper. We found that the negative association between polarization and quality, which has been found in previous work (Lerner and Lomi, 2019), is mitigated by team diversity but amplified in teams composed of jacks-of-all-trades.

It could be that findings on the interaction between polarization and diversity are reflected in the separate analysis by topic areas (see Table 9). For instance, we have found that the quality-enhancing effect of diversity is strongest for articles in the top-level category *Religion* and weakest (or even absent) for articles in *Mathematics*. How the effect of diversity on quality changes with the position of the article in the knowledge hierarchy of Wikipedia is an issue that we offer to future research.

Promising directions for future work also include more detailed analyses of how, why, and under which circumstances team diversity is beneficial or detrimental for productivity. Do diverse teams have access to more knowledge and capabilities, or are diverse teams also better able to manage team processes and conflict resolution? It could also be that team diversity, or individual diversity, have 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.

**Acknowledgements** This work has been supported by Deutsche Forschungsgemeinschaft (DFG Grant Nr. LE 2237/2-1) and Swiss National Science Foundation (FNS Project Nr. 100018_150126 and 100018_165764/1).

## References

Adler BT, de Alfaro L (2007) A content-driven reputation system for the Wikipedia. In: Proc. 16th Intl. Conf. WWW, ACM, pp 261–270

Akerlof GA, Dickens WT (1982) The economic consequences of cognitive dissonance. The American economic review 72(3):307–319

Ancona DG, Caldwell DF (1992) Demography and design: Predictors of new product team performance. Organization science 3(3):321–341

Arazy O, Morgan W, Patterson R (2006) Wisdom of the crowds: Decentralized knowledge construction in Wikipedia. In: Proc. 16th Workshop Information Technologies and Systems, pp 79–84

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

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

Barbosa S, Cosley D, Sharma A, Cesar Jr RM (2016) Averaging gone wrong: Using time-aware analyses to better understand behavior. In: Proc. 25th Intl. Conf. WWW, ACM, pp 829–841

Blumenstock JE (2008) Size matters: word count as a measure of quality on Wikipedia. In: Proc. 17th Intl. Conf. WWW, ACM, pp 1095–1096

Blyth CR (1972) On Simpson's paradox and the sure-thing principle. Journal of the American Statistical Association 67(338):364–366

Borgan Ø, Goldstein L, Langholz B (1995) Methods for the analysis of sampled cohort data in the Cox proportional hazards model. The Annals of Statistics pp 1749–1778

Brandes U, Kenis P, Lerner J, van Raaij D (2009) Network analysis of collaboration structure in Wikipedia. In: Proc. 18th Intl. Conf. WWW, ACM, pp 731–740

Bromham L, Dinnage R, Hua X (2016) Interdisciplinary research has consistently lower funding success. Nature 534(7609):684

Caldarelli G, Capocci A, De Los Rios P, Munoz MA (2002) Scale-free networks from varying vertex intrinsic fitness. Physical review letters 89(25):258702

Carroll RJ, Pederson S (1993) On robustness in the logistic regression model. Journal of the Royal Statistical Society Series B (Methodological) pp 693–706

Cartwright D, Harary F (1956) Structural balance: A generalization of Heider's theory. The Psychological Review 63(5):277–293

Conaldi G, Lomi A (2013) The dual network structure of organizational problem solving: A case study on open source software development. Social Networks 35(2):237–250

De Masi G, Iori G, Caldarelli G (2006) Fitness model for the italian interbank money market. Physical Review E 74(6):066112

Esteban JM, Ray D (1994) On the measurement of polarization. Econometrica 62(4):819–851

Festinger L (1962) A Theory of Cognitive Dissonance, vol 2. Stanford University Press

Flöck F, Acosta M (2014) Wikiwho: Precise and efficient attribution of authorship of revisioned content. In: Proc. 23rd Intl. Conf. WWW, ACM, pp 843–854

Flöck F, Vrandečić D, Simperl E (2011) Towards a diversity-minded Wikipedia. In: Proc. 3rd Intl. Web Science Conference, ACM, p 5

Franzoni C, Sauermann H (2014) Crowd science: The organization of scientific research in open collaborative projects. Research policy 43(1):1–20

Friedkin NE, Proskurnikov AV, Tempo R, Parsegov SE (2016) Network science on belief system dynamics under logic constraints. Science 354(6310):321–326

Goldberg A, Hannan MT, Kovács B (2016) What does it mean to span cultural boundaries? variety and atypicality in cultural consumption. American Sociological Review 81(2):215–241

Heider F (1946) Attitudes and cognitive organization. The Journal of Psychology 21:107–112

von Hippel E, von Krogh G (2003) Open source software and the "private-collective" innovation model: Issues for organization science. Organization science 14(2):209–223

von Hippel E, von Krogh G (2006) Free revealing and the private-collective model for innovation incentives. R&D Management 36(3):295–306

Hong L, Page SE (2004) 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 101(46):16385–16389

Horwitz SK, Horwitz IB (2007) The effects of team diversity on team outcomes: A meta-analytic review of team demography. Journal of management 33(6):987–1015

Hsu G (2006) Jacks of all trades and masters of none: Audiences' reactions to spanning genres in feature film production. Administrative Science Quarterly 51(3):420–450

Javanmardi S, Lopes C, Baldi P (2010) Modeling user reputation in wikis. Statistical Analysis and Data Mining 3(2):126–139

Jehn KA, Northcraft GB, Neale MA (1999) Why differences make a difference: A field study of diversity, conflict and performance in workgroups. Administrative science quarterly 44(4):741–763

Joshi A, Roh H (2009) The role of context in work team diversity research: A meta-analytic review. Academy of Management Journal 52(3):599–627

Kittur A, Kraut RE (2008) Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In: Proc. 2008 ACM conf. Computer Supported Cooperative Work, ACM, New York, NY, USA, pp 37–46

Kittur A, Chi E, Pendleton BA, Suh B, Mytkowicz T (2007) 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

Kittur A, Chi EH, Suh B (2009) 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, pp 1509–1512

Kovács B, Hannan MT (2010) The consequences of category spanning depend on contrast. In: Categories in markets: Origins and evolution, Emerald Group Publishing Limited, pp 175–201

Lam SK, Karim J, Riedl J (2010) 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, pp 55–64

Lee GK, Cole RE (2003) From a firm-based to a community-based model of knowledge creation: The case of the linux kernel development. Organization science 14(6):633–649

Lerner J, Lomi A (2017) The third man: Hierarchy formation in Wikipedia. Applied Network Science 2(1):24

Lerner J, Lomi A (2018a) Diverse teams tend to do good work in Wikipedia (but jacks of all trades don't). In: Proc. 2018 Intl. Conf. Advances in Social Network Analysis and Mining (ASONAM 2018), IEEE Computer Society, pp 214–221

Lerner J, Lomi A (2018b) The free encyclopedia that anyone can dispute: an analysis of the micro-structural dynamics of positive and negative relations in the production of contentious Wikipedia articles. Social Networks Forthcoming `https://doi.org/10.1016/j.socnet.2018.12.003`

Lerner J, Lomi A (2018c) Knowledge categorization affects popularity and quality of Wikipedia articles. PloS one 13(1):e0190674

Lerner J, Lomi A (2019) The network structure of successful collaboration in Wikipedia. In: Proc. 52nd Hawaii Intl. Conf. System Sciences (HICSS 2019), IEEE Computer Society, pp 2622–2631

Lerner J, Tirole J (2001) The open source movement: Key research questions. European economic review 45(4):819–826

Liu J, 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

Lord CG, Ross L, Lepper MR (1979) Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of personality and social psychology 37(11):2098

Maniu S, Cautis B, Abdessalem T (2011) Building a signed network from interactions in Wikipedia. In: Proc. Databases and Social Networks, ACM, pp 19–24

Mannix E, Neale MA (2005) What differences make a difference? the promise and reality of diverse teams in organizations. Psychological science in the public interest 6(2):31–55

McPherson JM, Ranger-Moore JR (1991) Evolution on a dancing landscape: organizations and networks in dynamic blau space. Social Forces 70(1):19–42

Phillips DJ, Turco CJ, Zuckerman EW (2013) Betrayal as market barrier: Identity-based limits to diversification among high-status corporate law firms. American Journal of Sociology 118(4):1023–1054

Ransbotham S, Kane GC (2011) Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia. MIS Quarterly 35(3):613–627

Ren Y, Chen J, Riedl J (2015) The impact and evolution of group diversity in online open collaboration. Management Science 62(6):1668–1686

Robert LP, Romero DM (2017) The influence of diversity and experience on the effects of crowd size. Journal of the Association for Information Science and Technology 68(2):321–332

Saperstein AM (2004) 'The enemy of my enemy is my friend' is the enemy: Dealing with the war-provoking rules of intent. Conflict Management and Peace Science 21(4):287–296

Shi F, Teplitskiy M, Duede E, Evans J (2017) The wisdom of polarized crowds, arXiv:1712.06414

Simpson EH (1951) The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society Series B (Methodological) pp 238–241

Sydow M, Baraniak K, Teisseyre P (2017) Diversity of editors and teams versus quality of cooperative work: experiments on Wikipedia. Journal of Intelligent Information Systems 48(3):601–632

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

Tsvetkova M, García-Gavilanes R, Floridi L, Yasseri T (2017) Even good bots fight: The case of Wikipedia. PloS one 12(2):e0171774

Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. Science 342(6157):468–472

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

Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. Science 316(5827):1036–1039

Zuckerman EW (1999) The categorical imperative: Securities analysts and the illegitimacy discount. American journal of sociology 104(5):1398–1438

Zuckerman EW, Kim TY, Ukanwa K, Von Rittmann J (2003) Robust identities or nonentities? typecasting in the feature-film labor market. American Journal of Sociology 108(5):1018–1074