The social embeddedness of peer production: A comparative qualitative analysis of three Indian language Wikipedia editions

Sejal Khatri
sejal.khatri5@gmail.com
University of Washington
Seattle, Washington, USA

Aaron Shaw
aaronshaw@northwestern.edu
Northwestern University
Evanston, Illinois, USA

Sayamindu Dasgupta[*]
sayamindu@unc.edu
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA

Benjamin Mako Hill
makohill@uw.edu
University of Washington
Seattle, Washington, USA

ABSTRACT

Why do some peer production projects do a better job at engaging potential contributors than others? We address this question by comparing three Indian language Wikipedias, namely, —Malayalam, Marathi, and Kannada. We found that although the three projects share goals, technological infrastructure, and a similar set of challenges, Malayalam Wikipedia's community engages language speakers in contributing at a much higher rate than the others. Drawing from a grounded theory analysis of interviews with 18 community participants from the three projects, we found that experience with participatory governance and free/open-source software in the Malayalam community supported high engagement of contributors. Counterintuitively, we found that financial resources intended to increase participation in the Marathi and Kannada communities hindered the growth of these communities. Our findings underscore the importance of social and cultural context in the trajectories of peer production communities.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing**; **Collaborative and social computing systems and tools**; **Wikis**; **Empirical studies in HCI**.

KEYWORDS

online communities, peer production, community engagement, Wikipedia, Indian Wikipedias, social computing, knowledge equity, underrepresented languages, social embeddedness

ACM Reference Format:
Sejal Khatri, Aaron Shaw, Sayamindu Dasgupta, and Benjamin Mako Hill. 2022. The social embeddedness of peer production: A comparative qualitative analysis of three Indian language Wikipedia editions. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3491102.3501832

1 INTRODUCTION

Although open collaboration systems require a pool of volunteer contributors to remain productive [20, 30], most attempts to build these communities struggle to attract participants [32, 35, 45, 49]. While a niche topic is unlikely to attract a large group of would-be contributors, this is far from a complete explanation [41]. For example, despite the existence of Wikipedia in almost three hundred languages, participation and content creation are not proportionally distributed to the number of language speakers, or even the number of viewers. Why do Wikipedia language communities vary in their ability to engage potential contributors?

We seek answers to this question by comparing the experience of Malayalam (ML), Marathi (MR), and Kannada (KN) Wikipedias—three Indian language online communities that differ in the size of the community and the knowledge bases they have built. These three communities have similar resources and technological infrastructure and face a similar set of challenges. Despite numerous similarities, our data suggest that Malayalam Wikipedia has created a substantially more vibrant community with a higher rate of engaging potential contributors than the others. We use interview data with 18 members of the three communities to identify the reasons why Malayalam Wikipedia may have engaged potential contributors in ways that Marathi and Kannada have not.

Our findings indicate that even though internal differences between the three communities matter, the ultimate reasons for Malayalam's relative success in promoting engagement stem from broad social and cultural factors. Furthermore, we found that the introduction of resources designed to help struggling projects compounded problems in these communities. Our findings have several important implications. First, our case studies demonstrate how contextual, social, and cultural features play an underappreciated role in shaping online communities. Second, our work provides a cautionary tale for those seeking to expedite online community growth with external resources.

[*] Dasgupta was at the University of North Carolina at Chapel Hill when this work was submitted for review. Since then, he has moved to the University of Washington.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501832

This paper offers several contributions. First, we make an empirical contribution by providing a detailed description of social computing systems in underrepresented languages—a large and important group of systems that have received little attention. Our work identifies a range of challenges shared by Indian language Wikipedias that we suggest are likely general features of social computing systems in low-resource settings. Second, by systematically comparing similar communities in different cultures, we develop and present a theoretical framework that describes a set of mechanisms through which social and cultural context shapes collaborative activity. Third, we contribute to the literature on engaging contributors in peer production by highlighting the role of culture and by showing how monetary resources can be counterproductive.

2 BACKGROUND

Our work contributes to the social computing literatures on peer production, underrepresented languages, the role of culture, and Wikipedia. We review related work in each of these areas.

2.1 Engaging Contributors in Peer Production

Peer production is a term coined by Benkler to describe a collaborative production model that occurs through the mass aggregation of numerous small contributions from large groups of diversely and often intrinsically motivated individuals working over the Internet [8, 9, 11]. While Benkler's archetypes are free/libre open-source software (FLOSS) and Wikipedia, the peer production model extends to many of the most important social computing systems, including communities involved in knowledge aggregation, creative computing, and collaborative filtering [11]. A large portion of the scholarly attention on peer production has focused on English Wikipedia [38].

Although the excitement about peer production stems from its ability to produce knowledge commons with enormous social benefit [10, 11], peer production communities frequently struggle. While peer production depends on the recruitment and retention of new contributors [50, 72], most efforts fail to attract more than a handful of participants [11, 32], and most successful projects struggle to sustain participation over time [89, 96]. For example, English Wikipedias' contributor base peaked in March 2007 [29].

In response, an enormous body of research has sought to understand the motivation of peer production contributors [11, 76]. Much of this work has involved surveys of contributors that have typically found that motivations are diverse and predominantly intrinsic [25, 27, 54, 74]. A related body of empirical work has sought to understand the processes through which the motivation of contributors shifts over time [5, 6, 14, 50, 79, 84, 101] and have focused on issues of newcomer retention [29, 70–72].

Maximizing engagement in peer production systems involves increasing not only newcomer retention, but also the rate at which noncontributors seek to engage in the first place. As most studies of motivation in peer production rely on social psychological evidence drawn from existing contributors [7], this work is limited in its ability to speak to the complete set of dynamics that drive engagement. One alternative approach involves the extension of surveys to broad samples of nonparticipants to map participation "pipelines" [93]. Another approach involves conducting comparative studies across populations of peer production projects that vary in their demonstrated ability to attract new contributors [37]. This paper takes the latter approach.

2.2 Underrepresented Languages

Almost all peer production scholarships have focused on the largest English language communities [64]. This likely reflects linguistic bias among social computing researchers who read and write in English. It also likely reflects bias caused by the size of the English language peer production communities as larger communities make objects of study that are more attractive. Although non-English language peer production communities are common, most examples of research on non-English language communities are large European languages such as German [e.g., 24, 88, 103] or large east Asian languages such as Chinese [e.g., 104, 105]. A more fundamental issue is that peer production, at least within the social computing literature, has been defined largely in Western terms using Western analytical frameworks. The notion of free culture central to peer production has, on the other hand, a long and understudied history in non-Western contexts. For example, in the discussion of the history of "public communication" and "public reason," Amartya Sen pointed out that the introduction of the Chinese translation of the Indian Sanskrit treatise *Vajracchedikā Prajñāpāramitā Sūtra* that was done in 402 CE (printed in 868) carried a note explaining that it was made for "universal free distribution" [91, p. 82].

Many of the languages left out of peer production scholarship are what Besacier et al. [12] describes as "underresourced languages." In Besacier et al.'s definition, these languages have at least some of the following qualities: lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, and lack of electronic resources for speech and language processing such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries, and vocabulary lists. We prefer to call these languages "underrepresented" on the web, rather than "underresourced," for several reasons, including that underrepresentation is often a product of long-term racist and colonial efforts to suppress or obscure cultural traditions through political and economic domination. Furthermore, "resources" for a given language may represent attributes orthogonal to how well represented it is in digital media—a language rich in literary works may still be underrepresented in digital media due to various historical and technological reasons. It is also important to note that although languages spoken by a minority of the population of a territory are frequently underrepresented, many "majority" languages are as well [12].

Despite being largely ignored by scholars, peer production in underrepresented languages needs to be studied for several reasons. First, these communities face enormous technological and social challenges. Second, these communities are frequently the sites of interventions by governments and non-profit organizations that operate with little in the way of basic research to guide their actions. Third, bringing HCI and social computing's theoretical and analytical methods to bear on these issues takes steps toward addressing an important issue of knowledge equity in HCI. Although, Van Dijk [100] has argued that underrepresented languages reflect an important object of study for peer production scholars, very few authors have taken up his call.

2.3 Social Embeddedness

Extending the linguistic scope of peer production research also broadens its cultural scope. In general, *culture* refers to the ways that people think, feel, or act in a society [53]. We approach culture by drawing on the concept of *embeddedness* from sociology to understand how the social and cultural context relates to the different experiences of the communities in our study. In the sociological sense, all facets of society are *embedded* within social relationships and institutions. For example, economic markets are embedded in the social relationships, networks, and values that underpin transactions [28, 83]. Similarly, Evans [22] argued that the success of state efforts to develop computing industries in Brazil, India, and Korea in the 1970s and 1980s depended on whether state actors were embedded in surrounding social institutions, such as networks of experts and entrepreneurs. In these accounts, culture is a component of the broader social environment within which technical or technocratic interventions unfold.

With its strong basis in psychology, HCI has usually engaged in a more narrow and cognitive sense with culture. In a recent literature review, Kyriakoullis and Zaphiris [53] argued that HCI research typically uses culture to explain why users in different countries adopt an interface at differing rates, use a system in different ways, or find different interface configurations more or less usable. Kyriakoullis and Zaphiris agreed that HCI research on culture often seeks to reflect culture differences quantitatively using reductive systems for the categorization of cultures along a small number of dimensions, such as Hofstede's [39] influential but controversial six factors.

Culture frequently provides a "catch-all" explanation for why a particular technology might work differently among different groups. For example, two recent books discuss deployments of the One Laptop per Child (OLPC) project in South America.[1] Ames [3] described a deployment in Paraguay that she characterized as a failure. Ames attributed this failure to the differences between the cultural context of OLPC's designers and its users in Paraguay. She argued that OLPC struggled because MIT-based hackers made decisions that sought to serve the needs of users like themselves as youth—precocious, technophilic boys from the Global North. Ames demonstrated that OLPC struggled in Paraguay because the social context in its rural deployments bore little resemblance to the context imagined by OLPC's designers. Chan [15] described a different OLPC deployment in Peru whose success she attributed to the close involvement of volunteer indigenous leaders and open technology activists who provided a localized support ecosystem in conjunction with the national government and rural community members.

Differences in culture (in the sense of a bundle of cognitive attributes) might explain these differing outcomes. Maybe Peruvians are predisposed toward deploying, adopting, or using OLPC in some way that Paraguayans are not? However, the divergence of Ames and Chan's accounts raises a deeper empirical puzzle: what specific social processes created a context that was conducive to OLPC's success in Peru but not in Paraguay? Our paper proposes a broad framework that uses the idea of embeddedness to provide preliminary answers. We will return to the example of OLPC in the Discussion section (§6).

2.4 Comparing Wikipedia Language Editions

Because many peer production systems are deployed more-or-less identically across a range of social contexts, they provide opportunities to understand the specific ways that social computing systems are socially and culturally *embedded* and with what effect. Our work is conducted in exactly such a system: Wikipedia. Started in 2001 as an English website, Wikipedia quickly added a series of editions in different languages. As of September 2020, it includes more than 300 languages. These non-English Wikipedias share the goal of writing a free encyclopedia through peer production and use identical technological infrastructure. In most other respects, they operate as separate projects. Different Wikipedias are written by different users and governed by varying norms and rules [42]. Although articles can begin with translated content, translations are not typically kept in sync. While 16 languages on Wikipedia have more than one million articles, more than 90% of Wikipedia language editions have less than one hundred thousand.[2] Many smaller Wikipedias are in underrepresented languages. As a result, the content produced by these communities is often highly ranked in search engine results, serves as a source of data for natural language processing [61], and provides a source of data for knowledge brokers like Google and Facebook [23, 60].

Although most Wikipedia research has considered only English Wikipedia, some has looked at dynamics in other language editions [64]. For example, several comparative studies of Wikipedia language editions have performed high-level quantitative comparisons [55, 77, 78]. Other works have sought to employ comparisons of small numbers of Wikipedias to identify differences in the contribution patterns [31, 34, 82]. A third approach involves the measurement of the differences between societies by identifying coverage gaps across language editions [33, 67, 68].

Comparative research on Wikipedia has—with few exceptions [e.g., 68, 100]—almost never considered underrepresented language Wikipedias. The only comparative study of underrepresented Wikipedia language editions we are aware of makes the unsurprising point that underrepresented Wikipedia editions engage contributors more or less effectively based on the desire of language speakers for Wikipedia content in their language [100].

Explaining why some language Wikipedia editions have bigger communities and content bases than others is not as straightforward an exercise as it may seem. For example, research has demonstrated that differences in Internet connectivity is far from a complete explanation. Although both practitioners and scholars point to differences in society and culture as an explanation, we know of no work that has attempted to identify salient types of cultural differences, or to unpack the social mechanisms that connect these differences to engagement in peer production.

[1] Two authors of this paper, Dasgupta and Hill, worked for OLPC in early years of the project. That said, the authors have no specific knowledge of the deployments beyond what is described in the books by Ames and Chan.

[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias

For the reasons stated above in §2.2, underrepresented language Wikipedias reflect an excellent place to study the relationship between social and cultural embeddedness and engagement in peer production. In that community goals and technology are held constant, Wikipedia language editions provide an opportunity to study social and cultural variations. Moreover, understanding the low rate of content production in underrepresented Wikipedia language editions can answer an important empirical puzzle in social computing scholarship [85]. By helping understand the processes through which some underrepresented language Wikipedias engage contributors more effectively than others, our work can directly benefit speakers of underrepresented languages.

3 EMPIRICAL SETTING

Our empirical setting is underrepresented languages spoken in India. We selected this setting because there are an estimated 536 million people in India who prefer their primary language to English when reading on the Internet [48] and because most, if not all, Indian languages qualify as underrepresented. To scope our project, we considered all 22 official languages of India and narrowed down our list to languages that are officially recognized only in India and that have a Wikipedia language edition. Second, we categorized the communities into three buckets—relatively low, medium, and high engagement—based on differences in the number of articles divided by speakers.

Based on this metric, we categorized Malayalam (ML) and Punjabi under high engagement bucket, and Marathi (MR), Kannada (KN), and Hindi in the relatively low engagement bucket. We set aside Hindi as it is spoken in nine Indian states. In addition, we chose not to investigate Punjabi because the first three articles of Punjabi Wikipedia language edition's were created more than a year later than the other projects we were considering. The first ML article was created in December 2002 and the first MR and KN articles in May 2003 and June 2003, respectively. Although ML launched 6 months earlier from MR and KN, MR and KN had a higher number of articles in the early years. In April 2005, ML had 133 articles, whereas MR and KN had 662 and 241 articles, respectively.

Table 1 presents a range of quantitative measures drawn from readership and contributorship statistics made publicly available by WMF that attempt to reflect the potential contribution base for the three Wikipedias.[3] Details on these measures are presented in Table 3 in our appendix, and a subset of these data are visualized in Figure 1. In general, the pattern of results indicates that while MR has a much larger audience than ML—both potential and realized—ML has a community that is at least as engaged and productive. The audience of KN Wikipedia is more similar in size to that of ML but its engagement rate is more similar to that of MR. Along almost the full range of quantitative measures presented, the community of ML outperforms its peers.

As far as we can deduce, the success of ML in maintaining engagement does not appear to be attributable to factors such as Internet availability, freedom of speech, an established tradition of encyclopedias, and understanding of other languages—all factors described as important in previous work [100]. Although there is a higher literacy rate among Malayalam speakers (—93.1% in Kerala where Malayalam is spoken versus 82.34% in Maharashtra where Marathi is spoken [66]), along with a high Human Development Index rating (—0.782 in Kerala vs 0.697 in Maharashtra vs 0.683 in Karnataka where Kannada is spoken[4]),—this difference is overwhelmed by Kerala's much smaller population.

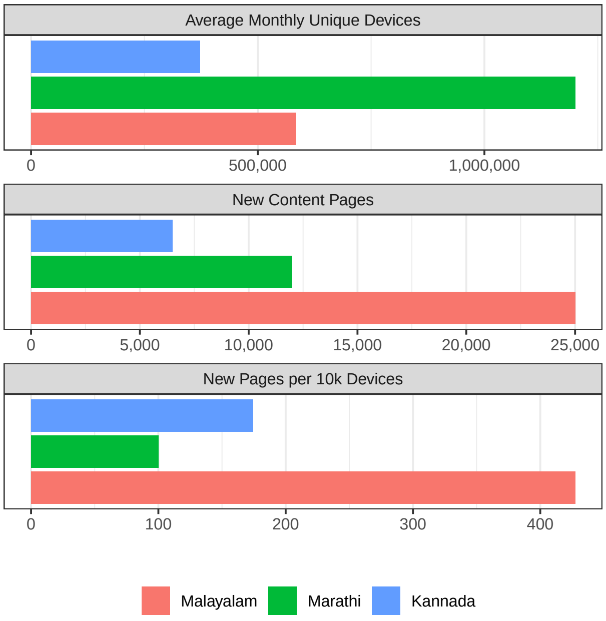

Figure 1: Average monthly unique devices represent an average monthly reader-base. New content pages represent the number of articles created, and New pages per 10k devices represent potential contributor engagement. Data is from January 2016–November 2019.

Before beginning our research project, we had familiarized ourselves with the data in Table 1 and several other sources of statistical data about both Wikipedia projects and the broader language speaking communities. As a result, we knew that ML had done a better job of engaging contributors than MR and KN along the range of metrics in Table 1. That said, we believed that the projects would be similar in many other respects. We did not believe, *ex ante*, that there was an obvious reason to think that one would have higher engagement rates than the others.

4 METHODS

To understand differences across the three language editions, we conducted a series of semistructured interviews with active participants of all three projects. We employed statistically nonrepresentative stratified sampling [99] to build a sample of adults with at least 1 year of experience in each of the Wikipedia language editions and made an effort to recruit both women and men as well as both administrators and non-administrators from each community. We

[3] https://stats.wikimedia.org/; https://meta.wikimedia.org/wiki/List_of_Wikipedias

[4] https://en.wikipedia.org/wiki/List_of_Indian_states_and_union_territories_by_Human_Development_Index

	Malayalam (ML)	Marathi (MR)	Kannada (KN)
Launch Date	2002	2003	2003
Number of Articles	71K	62K	27K
Native Speakers	35M	83M	44M
Article Depth (Collaborative Quality, Frequency Of Article Updates)	211	64	102
Contribution and Readership Statistics (Jan 2016–Nov 2019)			
Content Editors (Monthly Avg)	565	406	242
Active Content Editors (Monthly Avg)	86	55	41
New pages (content)	25K	12K	6.5K
New pages (content+non-content)	105K	41K	38K
Unique devices — readers (monthly avg)	585K	1.2M	373K
Unique desktop users (monthly avg)	135K	174K	83K
New registered users	15K	16K	8K

Table 1: Wikipedia Community Statistics

recruited the participants by posting screening surveys on each community's discussion channels including on each community's "Village Pump" community announcement page. We used snowball sampling to increase our reach and identify women participants who appear to be systematically underrepresented in our target languages, as they are in English Wikipedia [36].

As per common research ethics and the terms of the approved IRB-protocol governing this research, we have taken several steps to maintain the anonymity of our interviewees. First, we referred to subjects using their community acronym and a combination of a unique numerical label. Second, we included only summary-level participant information in Table 2. Although it is common to include demographic details on individual participants, the small size of these communities indicates that even minimal data on gender, tenure, and role could reveal participant identities.

We conducted interviews with 18 contributors—seven each from Malayalam and Marathi and four from Kannada. In each case, we sought long-term active contributors and interviewed every person that we could recruit. Although our focus on long-term contributors made the pool of potential interviewees smaller, it enabled us to recruit information rich cases for our study, which helped us achieve analytic saturation. Because the communities we studied are small, these relatively small samples within each wiki reflect a large portion of the most active contributors in each. These interviewees are described in summary form in Table 2. The first author of the paper traveled to Kerala, Maharashtra, and Karnataka between December 7, 2019, and January 7, 2020, to conduct face-to-face interviews. As a result, all but four subjects were interviewed in person in India—the remaining interviews were conducted over Zoom or phone from the United States and India. The interviews lasted between 39 min and more than 3 h for an average of 89 min and a total of 29 h. The interviews were conducted in English, Hindi, and Marathi or in a combination of those languages. All interviews were audio-recorded, fully transcribed, and translated into English by the first author. Furthermore, the interviews were conducted using a protocol that probed participants with open-ended questions about their personal experiences and motivation; their perceptions of dynamics, challenges, and goals within their primary language community; and their perceptions of other language Wikipedia editions. We have included a full copy of our interview protocol in the supplemental material.

Analysis was conducted by the first author following Charmaz's [17] approach to grounded theory. Although our codes were overwhelmingly inductive, we also included what Charmaz calls "sensitizing codes" derived from theoretical and empirical works that had influenced our research design. We conducted initial open coding in a line-by-line and incident-to-incident manner using the open source qualitative data analysis tool Taguette.[5] The first and final author discussed the codes and worked together to merge the codes into broader themes, write memos, and recode data in an iterative process. We conducted axial coding using Lucidchart, which is an online visual mapping tool. Finally, we synthesized and conceptualized the five-step theoretical model presented in §5 from the memos we generated about our themes. We have integrated *in vivo* codes into our theory to reflect perspectives and preserve the terms used by our interviewees.

5 FINDINGS

Our analysis revealed that the three underrepresented Wikipedia communities faced a number of common challenges. These included but were not limited to language localization on mobile devices, local language text entry, information retrieval, wikitext editing, Western influence on local language use, premature optimization, and limited resources for community outreach. As these were common across all projects, we describe these themes in Appendix A.3 but put them aside for now. The below findings focus on themes from our interviews that help explain why one underrepresented language Wikipedia might struggle relative to another. We organize these themes into explanations at three levels. In §5.1, we describe *micro-level* explanations focusing on differences in the day-to-day experience of participants attempting to contribute. In §5.2, we describe *meso-level* explanations, including differences in the norms and rules used within each community. Finally, we discuss *macro-level* explanations in §5.3, which focuses on differences in the broader social and cultural contexts in the societies of the three language-speaking communities.

[5] https://www.taguette.org/

Language Editions	Participant IDs	Gender	Tenure Range (years)	Roles
Malayalam	ML 8 - ML 14	1 F, 6 M	1 - 11	Editors, Organizers, Administrators
Marathi	MR 1 - MR 7	4 F, 3 M	3 - 8	Editors, Reviewers, Former administrator
Kannada	KN 15 - KN 18	1 F, 3 M	4 - 15	Editors, Administrators

Table 2: Participant Profile

The final theme described in §5.4 details the way that resources were deployed by NGOs and foundations to overcome challenges in both MR and KN—but not in ML. In both cases, this led to increased reliance on strong hierarchical community governance that contributed to increased barriers and, ultimately, to less contribution.

5.1 Direct Causes of Decreased Participation (Micro)

MR and KN faced a number of challenges that were less salient in ML. Many of these challenges reflected micro-level features of communities that directly deterred participation. In all cases, our interviewees found that these factors play a major role in Marathi and Kannada, but not in Malayalam. These included: low social support (§5.1.1), content disputes (§5.1.2), harassment (§5.1.3), poor conflict resolution (§5.1.4), a low sense of community identity (§5.1.5), and lack of technical resources (§5.1.6).

5.1.1 Low Social Support. Our interviewees suggested that senior MR editors invested few resources in the social support of either newcomers or new community members. MR4 explained that "...if we face some problems, then we did not understand who to ask for help. I had to study everything by myself." The members complained that senior editors were critical of newcomers and failed to follow Wikipedia's principle of assuming good faith.

5.1.2 Content Disputes. Interviewees from both MR and KN reported a large number of content disputes that they felt reduced their desire to contribute. For example, MR3, a skilled woman editor explained:

> Fights are about edits, some topics. It is not some personal fights. But that discourages people like me. If I am writing something and someone is questioning it, or I get some comments, then I feel like "Why am I contributing to Wikipedia?"

The demotivating nature of conflict was described as particularly salient by women like MR3, who may be more averse to conflict in general or more likely than men are to be burdened with professional and household duties [62]. Although content conflict surely occurred in ML as well, it appeared to be a more important factor in the experience of our MR and KN interviewees. In ways that highlight the interconnected nature of our themes, the interviewees suggested that content conflicts occurred for a series of organizational-level reasons, including restrictive norms (§5.2.1), the arbitrary use of policies in governance (§5.2.2), and patterns of gender inequality (§5.3.2).

5.1.3 Gender and Caste-Based Harassment. Extensive research has documented the low level of participation by women in Wikipedia [36, 47, 70, 93] as well harassment of women who do contribute [63]. A large body of work has indicated that the latter dynamic leads to decreased participation by women [62]. The interviewees from all three Wikipedias suggested that these dynamics extend to their projects as well—and to MR and KN in particular. They also reported an additional concern of harassment based on caste.

All five of the women editors we interviewed from MR and KN described having their contributions undone by other editors in ways they attributed to their gender. Furthermore, they expressed worry and confusion about not knowing what to do when this occurred. For example, one woman from MR described a pattern of harassment caused by an MR member who systematically reverted her edits:

> I had started an article and he comes every time and reverts my edits. I complained actually about him, but I did not get a good response. (MR2)

Similarly, KN18 told a story about a long pattern of incidents:

> I have been conducting diversity editathons (editing events) for the last 3 months. One of the major aims of these editathons is to bring more women editors. Before the new editor completes the article, an active editor jumps in and adds templates [to flag] that the article is not of good quality. The new editors, especially the women editors, are scared to continue editing.

Women editors in MR and KN explained that a pattern of hostile reactions singled out women, constituted harassment, and ultimately decreased retention of women newcomers.

Our interviewees suggested that harassment was not limited to gender. In particular, MR members described a pattern of caste-based harassment.[6] For example, when a MR community member began the process of deleting an article about a caste-based social reform movement due to Wikipedia's copyright policies, it led to caste-based conflict within MR and a sense of caste-based harassment among some editors. As the deletion was started by a senior editor whose Wikipedia's user name suggested an upper-caste identity,[7] their actions and subsequent justifications were imbued with a political agenda by some. The lack of trust within the MR community indicated that that the senior member's actions were understood to be caste-based harassment by some MR editors.

[6] The caste system in India [2] categorizes individuals based on heredity. A complex system and practice [18, 44], caste has been used for numerous discriminatory and exclusionary ends [43].

[7] South Asian names often code caste identities [81].

In another example that shows how harassment between gender and caste could be mixed, KN18 described a woman editor being "targeted" and subject to "abusive language" as well as being publicly called out, and effectively shamed, because of her lower caste status.

On the contrast, editors from ML explained that their community operated upon a strong norm of civility and respect and reported no examples of either gender or caste-based discrimination. ML11 explained that "we have a policy: Don't bash any editors. Never discourage an editor who is coming to edit Malayalam. If somebody who has access to technology and comes as a newcomer to Wikipedia, that person is [a] gem to us." Although our interviewees did not describe harassment in ML, this appeared to be at least partially explained by the fact that women editors on ML often choose to hide their gender. ML9 said, "I think that the gender gap is there. Women editors are not ready to reveal their identities or get included in community programs, events. They are participating in online events but they will not come to a meet up."

Although ML was not described as a bastion of gender and caste diversity, the kind of harassment and systematic discrimination reported in MR and KN were entirely absent from our interviews with ML participants. Once again, our interviewees attributed these more microlevel findings to differences in the macrolevel culture discussed in §5.3.2.

5.1.4 Poor Conflict Resolution. Both MR and KN editors reported poor conflict resolution processes that resulted in the banning of active editors and a demotivating environment. KN17 described an event when an active editor was blocked as part of the fallout from their addition of a "Reference Needed" template to an article written during one of the events to encourage women's participation in KN described by KN18 in the previous section. While some KN editors perceived the addition of the template as a good-faith suggestion for facilitating newcomer growth and improving content, others—including KN18—thought it constituted antisocial "newbie biting" that would deter future participation from women editors. Whatever the merits of either position, the conversation devolved into a long argument on the KN community discussion channel, which involved accusations of sexual harassment and an environment of confusion and frustration. Ultimately, the editor who added the template was banned, in large part for the way they conducted themselves in the community debate.

Our MR and KN interviewees explained that the way that the administrators and others in positions of authority wielded power and resolved conflicts in response to conflicts resulted in lower contributions for two reasons. First, it resulted in the banning of active users. Second, they felt that it reflected a low degree of coherence between decision-makers and community members and an inability to solve problems except through the blunt exercise of power. Our interviewees suggested that unsatisfactory conflict resolution contributed to the high attrition rates of active editors in both MR and KN.

5.1.5 Low Sense of Community Identity. Our interviewees describe MR and KN as having low degrees of community identity [50]. For example, MR Wikipedians reported not having a clear or measurable community goal. Even long-term contributor MR1 explained that "earlier there was this goal of 50k articles. Once it was achieved, I didn't see any goal." The interviewees reported no community discussion to identify new goals. When asked about community goals, KN15 reported, "the goal is to document stuff in Kannada. The goal started in the days of the previous leader. But now, we don't have a plan." Other KN members reported different goals. Administrators KN16 and KN18 each explained that the goal of their community was to increase the quantity and quality of scientific articles. That said, this goal appeared to not be shared by other members of the KN editor community. Indeed, KN community members express frustration with the current community administrators and their goals.

On the contrary, the ML community members we interviewed seemed completely content with the much more general goal of improving the quality of ML. For example, ML9 explained, "the community goal is to create more content-rich articles." In general, ML editors consistently expressed a strong sense of community identity. Our interviewees suggested that this shared goal and identity led to fewer conflicts and provided members with a productive and effective community experience. Once again, in ways that point to higher-level explanations, our interviewees attributed the low sense of community identity in MR and KN to behavior by centralized authority (§5.2.3).

5.1.6 Lack of Technical Resources. Interviewees from both MR and KN communities attributed the challenges they faced to a lack of technical resources. For example, MR1 invoked the absence of both the anti-vandalism tool *Twinkle*[8] and the editor engagement system *Wiki Love*[9] which are both available on English and many other larger Wikipedias saying, "this tool is absent from Marathi Wikipedia, and I don't think they will ever get that ... When I started the discussion, the point was who will do the translation [so] we don't have Twinkle [or] Wiki Loves." KN Wikipedians felt the same. For example, KN15 described a technical issue related to the Kannada script saying, "if you write a 15KB article [on KN],[10] a full formatted table, it counts as 0KB. I asked an active member why he doesn't fix it or allow us to do it. He said when I'll find the time, I will do it. It's been three years now." The issue experienced by KN15 is caused by a bug that means that the formatted text in Kannada wiki markup does not get counted in a widely viewed measure of Wikipedia users' total contributed text. KN15 is frustrated by this bug—and the KN's inability to fix this—because many Wikipedians are motivated to increase their total contribution and because he believes that the bug discourages the use of formatted text in KN.

On the other hand, ML has a balance between more and less technical contributors between both administrators and normal contributors. As a result, the Malayalam community is able to provide better technical support for its editors. This also reduces the technical participation barrier for newcomers, thus directly affecting and improving newcomer integration.

[8] https://en.wikipedia.org/wiki/Wikipedia:Twinkle

[9] https://en.wikipedia.org/wiki/Template:WikiLove_templates

[10] For reference, we estimate that 15 kilobytes would correspond to between 500–1,000 words.

5.2 Strong/hierarchical community governance (Meso)

Although our interviewees pointed to microlevel factors as leading directly to engagement, these explanations raise an obvious new question: *Why does the ML have relatively less conflict, better conflict resolution, stronger community identity, and more technical resources?* Our interviewees addressed this question by providing a series of what we refer to as "meso-level" explanations. We discussed four such themes that emerged from our analysis: territoriality (§5.2.1), restrictive content policies (§5.2.2), centralized authority (§5.2.3), and competition with other Wikipedias (§5.2.4).

5.2.1 Territoriality. Our interviewees from MR and KN repeatedly attributed the presence of conflicts (§5.1.2) to implicit community norms. In particular, our interviewees described extreme territoriality in MR where editors felt social ownership over the articles they worked on. Experienced editor MR2 explained that "one good thing is that [in the] Marathi community people have accepted that the subject that I am expert in belongs to only me. Yes!" MR1 after making substantial edits to an article explained, "on Marathi Wikipedia, you own that article." In English Wikipedia, this type of territorial behavior leads to resistance to improvements and increased conflict and can deter contributions from new contributors [35, 97].

5.2.2 Restrictive content policies. All Wikipedia editions create local policies in addition to Wikipedia's core policies. Our respondents suggested that the ability of MR to engage readers was impeded by a requirement that editors are expected to use Marathi with almost no words borrowed from other languages. Strict rules about pure language reflect a barrier to contribution because Marathi speakers typically mix languages. MR6 explained:

> The users are actually trying to make it a very pure kind of language. Now I say 'file' then you understand, but if I say 'dharidi' then it takes so much time to understand, as compared to 'file.' That happens for most of the articles. I know that if it is written in English then it is user friendly, and I would easily understand it. But going in a full Marathi way becomes a bit difficult.

Similarly, MR5 explained that "if we write a word in English and nobody knows its Marathi translation, still the user suggests to translate it in Marathi. Otherwise, that user comments that the content is wrong and merely copy-pasted."

Strict language rules create confusion and resistance from fluent Marathi contributors who want to contribute technical knowledge to MR but who have likely completed their higher education in English and may simply not know technical terms in Marathi. MR6 explained how this could represent a barrier, saying, "newcomers will have to be proficient in the use of language—the word that we are using is *shudh lekan* (pure writing)." Editors like MR6 acknowledged the difficulty of creating "pure" Marathi articles due to lack of language expertise and the relative absence of online sources of information in Marathi. The strict language policy of MR led to increased content disputes (§5.1.2), high rates of newcomer rejection, member frustration, and high attrition.

On the other hand, ML and KN adopted content policies that were more open to loan words. For example, ML14 explained that "...in the case of Malayalam, we are used to using technical terms in English, so we cannot completely rely on Malayalam." Interviewees in both ML and KN argued that these decisions increased the scope of participation in their projects and their processes of newcomer integration.[11]

In a related sense, KN used strict rules about citation that allowed only for the use of widely known reference sources and required that other sources be accepted only on a case-by-case basis. KN15 explained that on KN, "...everything you write has to be politically correct, it has to be signed off, it has to come through layers distilled, [the] essence gets lost." For example, one active KN contributor was not allowed to add information on local metro stations due to the lack of available references. Kannada editors like KN15 expressed distress about the lack of flexibility from KN administrators and authority figures whose decisions, in their mind, rendered important information off-limits to the readers of KN. Editors in both ML and MR reported much more open policies regarding references.

5.2.3 Strong centralized authority structures. Interviewees from both MR and KN reported power as being concentrated amongst a small number of community members. Interviewees from MR described their community's authority structure as highly centralized and out of touch. In an account that was largely repeated by MR7 in different words, MR3 explained that, "there are seven or eight admins, I guess, but only one is active. Actually, our admin stays in the US. So, that is one problem I think. Because we don't get to meet." In discussing issues in KN, KN17 explained that "in Kannada, community is balanced but power is not balanced." In further statements, KN17 attributed microlevel issues including ineffective social support (§5.1.1) to the concentration of power. Other interviewees explained how centralized authority led to unsatisfying resolution to conflicts (§5.1.4) and a low sense of community identity (§5.1.5). Furthermore, our interviewees attributed the lack of technical resources (§5.1.6) to the lack of representation of technical members among the administrators.

On the other hand, we found that the Malayalam community has larger and active adminship and a more decentralized power structure. ML12 explained that on ML "there are so many admins—nearly 21 admins—but most of the admins are not active in Malayalam Wikipedia in recent years. Only 4-5 admins are active in Malayalam Wikipedia projects." Although ML12 felt that "only" 4-5 active administrators was a problem for ML, this reflected more administrative activity and more distribution of power than in either MR or KN.

5.2.4 Competition with Other Wikipedias. One final challenge was largely unique to KN. Unlike the Marathi and Malayalam user groups, the Kannada Wikipedia user group—Karavali Wikimedians—works with three regional languages, namely, Tulu, Kannada, and Konkani. The result is a distribution of community resources and outreach that some users saw as a reason for the struggles of KN. One of the administrators, KN18, explained that "...users of Tulu Wikipedia are very less compared to Kannada Wikipedia ... so Tulu

[11] Centralized, and often government-run agencies often propose new terminology. However, readers' lack of familiarity with newly coined terms may lead to poor understanding [95]. On the other hand, terms borrowed from another language may be just as incomprehensible. Both alternatives represent difficult choices for language communities, as the example above shows.

Wikipedia needs more concentration." The distribution of time and resources of KN's administrators across the three languages was controversial among KN Wikipedians, and felt that KN's administrators' split allegiances directly harmed KN's efforts to engage users.

5.3 Less Supportive Social Environment (Macro)

Just as the micro-level explanations could be connected to meso-level themes, meso-level explanations also point to a "why" question: *Why does ML have less territoriality, less restrictive policies, and less centralized authority structures?* As before, each of the meso-level themes presented can be attributed to macro-level themes reflecting broader contextual differences that emerged as themes from our analysis. We discussed four such macro-level explanations: differences in government support for open-source and free knowledge (§5.3.1), gender/caste equity (§5.3.2), direct institutional support (§5.3.3), and attitudes toward volunteerism (§5.3.4).

5.3.1 Government support for open source and free knowledge. One clear macro-level explanation suggested that ML was helped by Kerala's long history of support for free/libre open source software (FLOSS). Originally dubbed "free software," FLOSS refers to a social movement promoting technology that users can study, modify, and distribute [94]. FLOSS and Wikipedia are closely linked: Wikipedia's model was inspired by FLOSS and both Wikipedia's technology and licenses are drawn from the FLOSS community [86].

The Indian state of Kerala, where most Malayalam-speakers live, has a long history of both top-down and bottom-up social reform movements as well as governance by left-wing political parties [52, 56], having switched historically between left-wing and center-left parties [51]. Furthermore, Kerala has a thriving FLOSS community that has engaged "with mainstream groups such as the government, media, and civil society at large" since the early 2000s [4, p. 109]. Kerala's embrace of FLOSS by government agencies and public entities is contrasted with experience in Maharashtra and Karnataka. A study of FLOSS usage in the Indian states' governments demonstrated that both the Maharashtra and Karnataka states did not adopt FLOSS and typically depended on proprietary software vendors like Microsoft [19].

Although expressing some reservations about the completeness of the story, KN16 pointed to the history of ruling political parties for ML's relative success in maintaining engagement while comparing the KN community to ML:

> [The] number of editors in Malayalam is far more than Kannada. If you see the highest number of Linux users are from Kerala. Next is Bengal. Somehow the open-source, free knowledge movements and the left... Somehow there's an equation, which I don't subscribe to, but it is there.

According to our interviewees, Malayalam speakers' history of involvement in open source has helped set the stage for eliciting participation in free knowledge communities such as Wikipedia. On the other hand, Marathi and Kannada language communities find less cultural resonance between Wikipedia and their readers. Like the "elective affinities" that Weber [102] identified between Protestant religious ethics and the spirit of capitalism, the Malayalam language community's long history of embracing an open source ethos may have created fertile ground for the growth of a vibrant Malayalam Wikipedia.

In a direct piece of evidence of this relationship, Kerala's government has taken steps to support Wikipedia-like free knowledge projects among children. Since 2009, the state has introduced Wikipedia's collaborative peer production model in the form of *School Wiki*[12]—a scaled down version of Malayalam Wikipedia—which has been introduced to young children with the help of the ML participants.[13] Although this project may have directly contributed to increased awareness of Wikipedia and increased contribution rates, its direct effects are difficult to ascertain. What is clear from our interviews is that the ML community feels that their work is more supported and in harmony with the cultural context of the Kerala society than does either the MR or KN communities.

5.3.2 Lower gender/caste-equity. A partial explanation for the different experiences of gender and caste-based harassment (§5.1.3) are differences in cultural attitudes. Kerala's government has emphasized the improvement of human development indicators (HDI) such as literacy and life expectancy whereas per capita income and gross domestic product (GDP) remained relatively low. This model of development is widely known as the "Kerala model" and has been extensively studied by development economists [52, 56, 90]. As a result, Kerala does well in terms of several measures of gender equity relative to other regions in India and has the highest literacy and educational achievement rate for women. The HDI measures for Maharashtra have consistently trailed Kerala.[14] Relatedly, a 2007 study observed a high degree of inequality in education across regions, gender, and caste groups in Maharashtra [80]. Similarly, research suggests relatively high degrees of gender inequity and regional disparities in Karnataka [46].

ML11, an administrator of ML, explained Kerala's unique context in this regard and its effect on ML:

> Kerala is a different land because of the land reformation, the caste reformation has only happened in Kerala. All other states in India are different, in case of caste problems and the caste discrimination and knowledge discrimination. In Kerala we are treating everybody equally nowadays.

We should be wary of taking these types of statements at face value. The Malayalam societal and cultural norms still dictate that women should be subservient to men both at home and in the labor market [69] and caste-based inequality is deeply rooted in Indian society in ways that are difficult for any government policies to eliminate. Neither Malayalam society nor its Wikipedia have completely eliminated harassment or discrimination based on gender or caste.

[12] https://schoolwiki.in

[13] https://www.deccanchronicle.com/nation/in-other-news/311016/school-wiki-to-link-15000-kerala-schools.html

[14] https://globaldatalab.org/shdi/shdi/

That said, it is also true that the three Wikipedias are situated in their specific societal and cultural contexts and that our interviewees repeatedly invoked Kerala's history of socio-religious reformation as a reason that the ML participants we talked to could not recall a single issue of caste-based discrimination, whereas Marathi and Kannada communities faced these types of challenge routinely. In this sense, the social embeddedness of peer production projects describes how the inclusiveness and participation in Wikipedia communities can be shaped by the societies that constitute their linguistic communities.

5.3.3 Direct institutional support. As a final macro-level explanation, our interviewees explained that the Malayalam community were able to promote their project using connections to the government and media institutions. Although direct support from the government for ML was minimal, School Wiki (discussed in §5.3.1) was a collaborative project between the Kerala state government and ML that provided contributors with direct access to students in public schools.

In addition, the ML members were able to successfully cultivate connections to the Kerala media that were absent in other projects. ML10 said, "Malayalam Wikipedia can successfully manipulate other media. We have notable articles from Malayalam Wikipedia in Malayalam dailies, newspapers, and there are some programs regarding Malayalam Wikipedia in [video] media also, and we get big media coverage." This point was echoed by a number of other ML participants.

Although MR received support from *Rajya Marathi Vikas Sanstha* (RMVS), a government agency in Maharashtra, the support was limited to outreach activities. MR1 explained that RMVS's "motive is to spread the word around Marathi Wikipedia and nothing else." Our interviewees from KN reported no government support for KN at all. KN18 explained that this was a source of frustration saying that "we wanted [the state] government to understand that we are doing so much for Kannada and you people are just not noticing at all." KN18 explained that the lack of support was not caused by the lack of effort or opportunity on the part of the community. They explained in detail how they had proposed several activities to the Karnataka government but that these overtures for collaboration had been rebuffed.

5.3.4 Attitudes toward volunteerism. In ways that echo the lack of an open source ethos described in §5.3.1, interviewees from MR and KN attributed low engagement to cultural attitudes toward volunteerism. For example, MR3 pointed out the lack of monetary benefits as the reason for low levels of engagement in MR saying, "people won't find time for [contributing to Wikipedia] because such things don't give you money." MR2 relied on broad cultural stereotypes to provide a similar explanation:

> Two of my friends questioned me about this [interview] meeting saying, 'Why are you going to meet her for an interview? What are you gaining from this?'

Many MR Wikipedians cited stereotypes about volunteerism among Marathi speakers to explain why engaging editors was an uphill battle.

KN Wikipedians reported similar dynamics. KN16 explained how his family seemed puzzled by his volunteer contributions to Wikipedia saying:

> I am very active. I am so active that in my house they criticize me for doing so much on that. It doesn't pay, right? Suppose at the same time I use [the time I spend on KN] somewhere else. I'm a freelancer. Suppose I use the same time for my own commercial or financial benefit? I can earn more, which I am not doing.

MR and KN editors repeatedly described being criticized by their communities for doing unpaid work in ways that caused them to reevaluate their participation. In this way, the lack of support for volunteerism in the Marathi and Kannada societies may have contributed to low editor engagement.

The difference in the situation reported by our ML informants was stark. ML13 explained that "I am a free software activist and enthusiast, and I do like volunteer things like that. Because I know this is needed for us. For the people." The Malayalam editors expressed deep support for volunteerism in the free software and free knowledge movement, and in general.

5.4 NGO Involvement

One final difference between the three Wikipedias we studied was the strong role that paid labor from NGOs played in MR and KN, but not in ML. This explanation does not fit neatly into our micro/meso/macro framework as these NGOs typically grant money designed to support the development of content in Marathi and Kannada precisely because MR and KN were already struggling.

Our interviewees suggested that monetary support did not lead to active community participation. Instead, members in funded communities described funding as ineffective at best. For example, when asked about community support, MR3 said:

> People get laptops and some people getting Internet connections with some schemes, through [NGOs] or Wikimedia Foundation. So that is good support. But I don't think that is translating to active contribution. Very few people who have got this support are contributing.

Because MR and KN were struggling to serve the needs of their communities, the NGOs stepped in with grant money. According to MR3, this money largely went into technology for people who ended up not substantially contributing. When money was effectively used, it typically went into supporting the labor of administrators who were already the most active members. Surprisingly, this backfired as well by increasing reliance on the work of these administrators (§5.2.3), aggravating issues of territoriality (§5.2.1), and providing resources to enforce strict policies (§5.2.2).

Our interviewees reported that although grants to non-profit organizations effectively supported community outreach efforts and training sessions, these efforts came with pressure for short-term achievements and provided little in the way of badly needed long-term support. Perhaps, more importantly, the introduction of paid labor played into the development of a project culture where would-be participants felt that it was only fair if they were paid to contribute too. Previous research on volunteer-run sports organizations has demonstrated that the introduction of paid labor can cause other volunteers to reduce their own contributions [21]. We heard evidence of similar dynamics in both MR and KN.

Beyond its effect on volunteers' motivation, the question of paid labor is a complex one in Wikipedia and peer production. On the one hand, peer production communities, such as Wikipedia, have a range of rules and norms about paid contributions designed to reduce reputational threats that might emerge if there was a sense that subjects could simply "buy" desirable coverage. Wikipedia is trustworthy precisely because it is a noncommercial third-party space. On the other hand, research has framed the voluntary work done toward this as emotional labor [58, 59, 62] and asked whether it is fair to expect uncompensated labor (including emotional labor) when underlying structural issues that cause underrepresentation persist. Our findings indicate that even when infusions of money and resources is an option, the existing dynamics and structure of the community need to be carefully considered.

6 DISCUSSION

Our findings propose more than a dozen answers to a single empirical puzzle: Why did Wikipedia in Malayalam engage potential contributors more effectively than Kannada and Marathi? We present many answers both because we believe that there is no single answer and because we view the answers as deeply intertwined. An analogy can be drawn to the "five whys" approach involving repeated asking "why?" to identify more fundamental causes [92]. The micro-level answers in §5.1 describe specific experiences that deter potential contributors. But why are these experiences more common in MR and KN than in ML? Our meso-level answers in §5.2 point to organizational structures that our interviewees identify as likely causes. Once again, why do these problematic governance structures exist in some Wikipedias but not others? Our macro-level answers in §5.3 provide the highest-level explanations and describe how the embeddedness of specific Wikipedia projects within the surrounding social and cultural contexts shaped the way each project unfolded.

We synthesize all themes into two explanatory maps in Figures 2 and 3. Figure 2 demonstrates how the relatively supportive macro-environment in Kerala led to a larger group of potential contributors to ML as well as a chain reaction of social processes that led to a Wikipedia that was better able to engage potential contributors. We visualize the very different dynamics in MR and KN in Figure 3. In both cases, the features of the social and cultural environment led to a reliance on a relatively small group of people for governance. This led, in turn, to barriers to entry that reduced contributions. The introduction of external financial support, shown in the red box, increased the reliance on centralized authority structures, aggravating the problem and introducing a negative feedback cycle.

While the details of our story are specific to the communities we studied, our multilevel approach provides a conceptual framework for understanding how social embeddedness may shape social computing systems. Our findings indicate how the embeddedness of sociotechnical systems and editor communities within their respective social and cultural environments—i.e., norms, values, relationships, hierarchies, organizing techniques, experiences, resources, and political traditions—have interacted in specific ways and resulted in Wikipedias that share some traits and not others. By saying that these communities are embedded, we are not reducing the outcomes we observe simply to "reflections" or "expressions" of social context [28]. The fact that Kerala has a tradition of egalitarian left-wing politics did not predetermine that Malayalam Wikipedia would have a more engaged editor base. Rather, the history and experiences of Kerala seem to have provided a set of techniques, logics, and shared values that Malayalam editors have drawn on in building their community.

To illustrate the broader usefulness of this embeddedness framework, we return to the puzzling divergent accounts of OLPC in Paraguay and Peru which are briefly introduced in §2.3. How might our approach explain why OLPC deployments were more successful in Peru than in Paraguay? Ames's [3] account of OLPC in Paraguay described a deployment where decisions on technology and organization were made centrally with less engagement from the local community. Deployments were coordinated by an NGO that relied almost exclusively on paid labor and had little experience in free software or free culture activism. Ames argued that OLPC in Paraguay often found itself in an adversarial relationship with existing power structures in Paraguay and its schools. In all of these ways, she describes how macro-level social factors led to meso-level decisions about organization which limited the effectiveness of the OLPC deployment. These patterns and outcomes resemble those we observed in the MR and KN Wikipedias. On the contrary, Chan's [15] description of OLPC in Peru resembles Wikipedia in Kerala more closely. Just as in ML, the deployment of OLPC in Peru was supported by volunteers from the local free software activist community and enjoyed support and collaboration from local government and schools. The result was a deployment that meshed with the social fabric of Peru's techno-culture more effectively than Paraguay's. Chan [15] argued, that this led to increased engagement by Peru's OLPC user community, better outcomes for the project, and a distinct local interpretation of what OLPC was about. In this way, variations in the embeddedness of each project along the lines described in Figures 2 and 3 can help explain divergent outcomes.

6.1 Alternative Explanations

While we conclude that variations in social embeddedness shapes engagement in peer production communities, our evidence also suggests some potential alternative explanations. One alternative attacks the fundamental design of our study. Perhaps the blunt quantitative measures presented in Table 1 and Figure 1 misrepresent the real state of these Wikipedias and ML is not as successful as it appears. For example, what if ML has more articles per user, but those articles are of lower quality?

While we cannot exclude this possibility, our interviewees claimed that ML articles are of higher quality. MR and KN Wikipedians largely described their respective Wikipedias' article quality as poor. KN16 said that "almost 50 to 60% articles are bad actually, not good." Similarly, MR1 stated that, "There are some 53k articles and I know 50k are flop! Definitely, 50k articles are unsourced, most of them must be copy violations. If you start actually cleaning the stuff, you'll say 'I want to start Marathi Wikipedia again'." On the

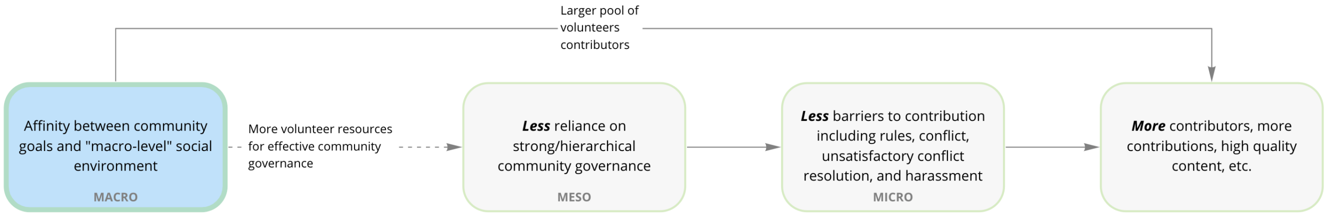
Larger pool of volunteers contributors
Affinity between community goals and "macro-level" social environment
MACRO
More volunteer resources for effective community governance
Less reliance on strong/hierarchical community governance
MESO
Less barriers to contribution including rules, conflict, unsatisfactory conflict resolution, and harassment
MICRO
More contributors, more contributions, high quality content, etc.

Figure 2: Virtuous Cycle: Explanatory mapping of the participation cycle in Malayalam Wikipedia.

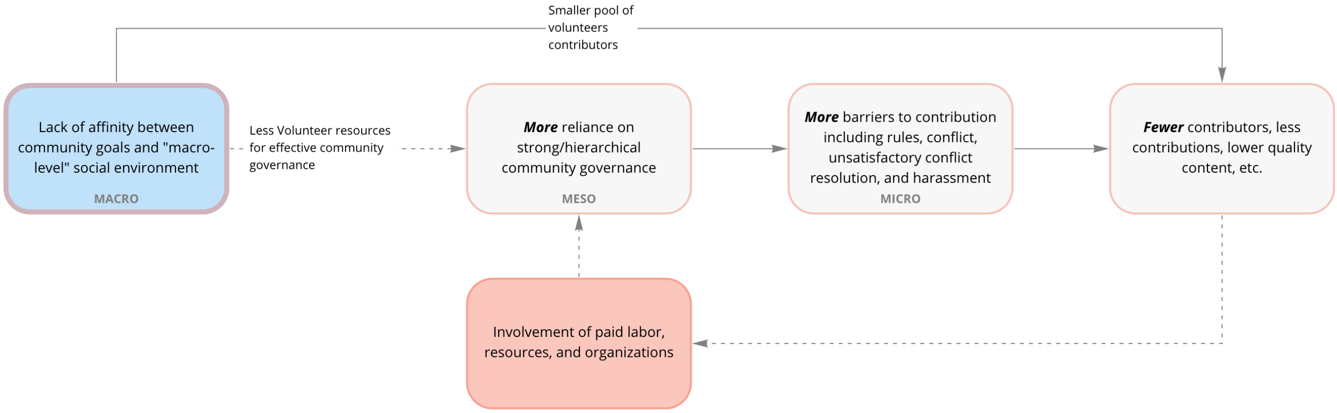
Smaller pool of volunteers contributors
Lack of affinity between community goals and "macro-level" social environment
MACRO
Less Volunteer resources for effective community governance
More reliance on strong/hierarchical community governance
MESO
More barriers to contribution including rules, conflict, unsatisfactory conflict resolution, and harassment
MICRO
Fewer contributors, less contributions, lower quality content, etc.
Involvement of paid labor, resources, and organizations

Figure 3: Vicious Cycle: Explanatory mapping of the feedback cycle in the Marathi and Kannada Wikipedias.

other hand, our ML interviewees perceived their respective language Wikipedias' article quality to be very good. ML14 said, "the articles are actually lengthier in the case of Malayalam, which I think is not present in many other languages." At a minimum, this evidence supports divergent perceptions of content quality.

Another reason for ML's high number of articles might be automatic content creation by bots—a common feature in many Wikipedia editions [57, 73]. We found little evidence for this either. ML11, a senior editor, explained:

> You cannot do a bot article in Malayalam Wikipedia, you will get banned instantly, but on other Wikipedias, basic [bot] editing is okay, and you can do a lot of articles. In Marathi or Kannada, they have already done it. 40% of articles are created from census articles, census data and that is not possible in Malayalam. That is the biggest difference.

Although we have not attempted to corroborate ML11's numbers for MR and KN, ML11's comments indicate that the difference in productivity between MR, KN, and ML may be even more stark than suggested by the quantitative measures presented in Table 1 and Figure 1.

Another possible alternative explanation is that collaborative projects such as peer-production initiatives come with a set of cultural constraints that arise out of the origins of the project but that these constraints apply unevenly across projects. By this, we can say that it may be possible that Western-origin peer production projects, such as Wikipedia, to come with a unique set of limitations that make it difficult to replicate the dynamics that makes the English language Wikipedia successful. Acey et al. [1] have described how some of the foundational principles of Wikipedia, such as an exclusive reliance on secondary, published sources are at odds with how knowledge is recorded and transmitted in most of the non-Western world. In an example of this dynamic in peer production projects other than Wikipedia, Ntabathia [75] found that the largely Western categorization system used by Open Street Maps—a peer produced geographic information system—rendered points of interest common in non-Western contexts invisible, and thus impossible to record. These barriers suggest that the peer production's value of "open" might apply more for Western ways of knowing. In a recent publication on decolonizing knowledge, Chan et al. [16] have urged for an expansion of the notion of "open" that allows for the inclusion of knowledge systems and epistemologies of marginalized people and communities that have been traditionally excluded from the canon of Western knowledge. Although we cannot rule out this critique, it is important to remember that the notion of "open" as being freely available to all (i.e., the current definition) is not an exclusively Western idea (as discussed in §2.2), and that it is possible that this alignment of values at a very fundamental level motivates contributors from non-Western contexts to contribute to the project, despite its Western origins. Furthermore, as support for this claim, our data includes no evidence that ML engaged contributors more effectively because Malayalam culture is aligned more with the Western values of Wikipedia. Finally, there are several examples of quite successful non-Western Wikipedia projects suggesting that these successful instances transcend any Western or Anglo-centric values that Wikipedia might embody.

Finally, we found some evidence that linguistic diversity within each project has an influence on engagement. It stands to reason that a high number of dialects might lead to disagreements, conflicts, and difficulties with collaboration. Anecdotally, this issue has been prominently reported in Hindi Wikipedia whose speaker base is geographically spread across India. Indeed, the Malayalam language has 18 dialects, whereas Marathi and Kannada languages have 42 and 20, respectively.[15] Although evidence in our data for this explanation was weak, both MR11 and MR7 described dialect-related content conflicts in MR and argued that it could be difficult to reach a consensus on dialect-based issues. The Malayalam and Kannada communities did not report dialect-based disputes.

6.2 Limitations

Our work has several methodological limitations. The lead author who conducted the interviews does not speak either Malayalam or Kannada, and this may have resulted in different interview qualities. We attempted to minimize this gap by using the same interview protocol across all three communities, but we cannot know how this limitation affected our results. In addition, our study only investigated the experiences of people who had joined and participated. While we believe that the comparisons we draw are fair, they do not reflect the experiences of those who had either abandoned or never joined or who wished to remain anonymous. The recruitment of interviewees from these groups poses challenges we were not able to overcome.

Like all grounded theory analyses, our goal is to generate new insights and we cannot know how our findings will generalize beyond our sample. We are most confident in the validity of the results among Indian language Wikipedias and have less confidence that our results will extend to other languages and geopolitical contexts, other peer production projects, and so on. Although we believe that our analysis is a valid interpretative account of the interviews we conducted, we cannot know whether or how the samples from each community in our study might be biased in ways that drive our findings. Although we used purposive theoretical sampling to mitigate this threat, it is possible that we drew idiosyncratic samples from one or more of the projects.

Finally, we recognize that our own interest in understanding contributor engagement might have led us to overlook trade-offs with other goals that communities might have. For example, although MR's strict language policies might have made the contribution more difficult, these policies might have also increased usability for Marathi speaking readers and provided better material for education. The relatively large readership of MR suggests that there might be an important tradeoff that underrepresented language Wikipedias must navigate between lowering barriers to contribution—e.g., by allowing lower quality content—and serving the (often difficult to ascertain) needs of an audience.

6.3 Design Implications

Our study suggests a number of implications for designers and contributors working on localized peer-production projects. In the most general sense, our work shows how new systems launched in underrepresented language contexts should employ design decisions that consider issues of social embeddedness.

We found that localized peer-production communities are intimately linked to local cultural and political dynamics (§5.3). This leads to our first suggestion of strategically selecting supporting organizations and partnerships. Relationships with organizations at all levels—schools and libraries, trade unions, non-profit organizations, and government agencies—can support peer-production projects in a variety of ways. That said, the outcomes of these relationships vary widely depending on social dynamics. Furthermore, the level of interventions and partnerships needs to be carefully considered. For example, in the case of the use of new or unfamiliar technical terms (§5.2.2), past scholarship has suggested that centralized agencies, such as language standard bodies, can support effective and widespread adoption [95]. On the contrary, the Marathi and Kannada Wikipedia communities sought to standardize terms themselves (§5.3.3). Government-controlled language standard body might more appropriately and effectively support familiarity with terminology through other content channels, such as mass media, publications, and textbooks.

A second suggestion pertains to localized peer production communities that attempt to follow an already established model (e.g., non-English Wikipedia editions following the English Wikipedia model). Such communities face the potential peril of "premature optimization" with regard to the creation and enforcement of norms, rules, and practices, which have real costs, before they are needed. All three communities did this and we discuss this in our appendix in A.3.6. Successful localization often results from a "best of both worlds" scenario where elements of an established initiative are carefully selected and combined with what the local community has to offer. The question of whether to use "pure" language (§5.2.2) is illustrative. While uniform language and style can benefit an encyclopedic project, the restriction on borrowing terms from English and the insistence on linguistic "purity" might also discourage participation. The use of terms that have local origins might be wise if the intended audience comprises of those who do not know English. In practice, the phrase "do not know English" often needs to be qualified when making arguments for "pure" language. On the other hand, the contributor pool of these communities, especially in their early stages, are more likely to know English loan words for technical terms [87]. Strict norms and rules might bar contributions from this group. A process of continuous engagement with local producers of content along with local consumers would help in the evaluation of what both groups' language practices and needs look like and support the creation of content standards on an incremental and ongoing basis.

7 CONCLUSION

Visited by more than a billion people each month, English Wikipedia is the fifth most popular website in the world and the most important website not created and managed by a for-profit company. Certainly, the large majority of people on earth do not benefit from

[15] Source: https://www.ethnologue.com/language/mal, https://www.ethnologue.com/language/mar, and https://www.ethnologue.com/language/kan.

English Wikipedia as they can not easily read reference material written in English. Moreover, many who visit English Wikipedia do so only because similar reference material is not available in the underrepresented languages they would prefer.

Frequently cited in fundraising pitches and TED talks, Wikipedia founder Jimmy Wale’s original vision for Wikipedia is “imagine a world in which every single person on the planet is given free access to the sum of all human knowledge.” To his credit, Wales has always understood that progress toward this vision would mean engaging volunteer speakers of underrepresented languages. Unfortunately, Wikipedia has been much less successful in these contexts than they have been among speakers in the world’s most highly resourced languages. Although some of the 300+ language Wikipedias targeting underrepresented language contexts have gained traction, most have struggled.

Our work is one more piece of evidence against what Toyama calls “packaged interventions” (i.e., purportedly replicable one-size-fits-all solutions that largely ignore context) [98]. Our work suggests that “English Wikipedia but in Language X” will rarely be an effective approach. Similarly, our explanation for Malayalam Wikipedia’s success relative to Marathi and Kannada Wikipedias is not in closer alignment with the English Wikipedia model.

That said, our work goes well beyond this. Our description of virtuous and vicious cycles that connect the micro-level features of peer production experience to meso-level organizational features to macro-level features of societies and cultures provides templates for a way to evaluate, critique, and even tailor interventions to build for virtuous cycles harnessing cultural resonance.

Our suggestion is that Malayalam’s relative success was due to specific resonances between the way that ML Wikipedia and Malayalam society are structured. Our work should not be understood as a blueprint for replicating Malayalam Wikipedia’s relative success at engaging contributors. Every context is different and resonance is a function of a wide spectrum of particulars. We believe that although the specifics will always vary, the broad structure of virtuous cycles will be similar among successful sociotechnical systems.

In addition, our description of vicious cycle that played out in MR and KN Wikipedias offers a specific warning to funders and others seeking to fix peer production projects through targeted interventions that, we show, can exacerbate fundamental underlying problems. Funding cannot fix a lack of cultural resonance, but it can make it worse.

Understanding the broad social dynamics that drive the relative success of contributor engagement among underrepresented language settings reflects a problem ideally suited to social computing research, an enormous and almost completely neglected challenge, and an opportunity to make progress on an important knowledge equity issue. We offer our paper as what we hope is a key first step toward making progress on this broader goal.

ACKNOWLEDGMENTS

The authors would like to thank the Wikipedia community members who participated in this study. The authors also thank the Community Data Science Collective, and especially Kaylea Champion, for their support and for providing feedback on the project. Financial support for this work came from the University of Washington Communication Department (grant FR0023) and the US National Science Foundation (award IIS-2045055).

REFERENCES

[1] Camille E Acey, Siko Bouterse, Sucheta Ghoshal, Amanda Menking, Anasuya Sengupta, and Adele G Vrana. 2021. Decolonizing the Internet by Decolonizing Ourselves: Challenging Epistemic Injustice through Feminist Practice. *Global Perspectives* 2, 1 (2021).

[2] Babasaheb Ambedkar. 2019. Castes in India: Their Mechanism, Genesis, and Development. In *Dr. Babasaheb Ambedkar : Writings and Speeches*. Vol. 1. Dr. Ambedkar Foundation, Ministry of Social Justice & Empowerment, Govt. of India, New Delhi, India.

[3] Morgan G. Ames. 2019. *The Charisma Machine: The Life, Death, and Legacy of One Laptop per Child*. MIT Press, Cambridge Mass.

[4] Satish Babu. 2011. FOSS as a Tool for Development: The Kerala Experience. In *2011 IEEE Global Humanitarian Technology Conference*. IEEE, Seattle, WA, USA, 108–110. https://doi.org/10.1109/GHTC.2011.82

[5] Martina Balestra, Coye Cheshire, Ofer Arazy, and Oded Nov. 2017. Investigating the Motivational Paths of Peer Production Newcomers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI ’17)*. ACM, New York, NY, USA, 6381–6385. https://doi.org/10.1145/3025453.3026057

[6] Martina Balestra, Lior Zalmanson, Coye Cheshire, Ofer Arazy, and Oded Nov. 2017. It Was Fun, but Did It Last?: The Dynamic Interplay between Fun Motives and Contributors’ Activity in Peer Production. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 1–13. https://doi.org/10.1145/3134656

[7] Gerard Beenen, Kimberly Ling, Xiaoqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, and Robert E. Kraut. 2004. Using Social Psychology to Motivate Contributions to Online Communities. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*. ACM, Chicago, Illinois, USA, 212–221. https://doi.org/10.1145/1031607.1031642

[8] Yochai Benkler. 2002. Coase’s Penguin, or, Linux and ‘The Nature of the Firm’. *The Yale Law Journal* 112, 3 (Dec. 2002), 369. https://doi.org/10.2307/1562247

[9] Yochai Benkler. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, New Haven, CT.

[10] Yochai Benkler. 2016. Peer Production and Cooperation. In *Handbook on the Economics of the Internet*, Johannes M. Bauer and Michael Latzer (Eds.). Edward Elgar, Cheltenham, UK, 91–119.

[11] Yochai Benkler, Aaron Shaw, and Benjamin Mako Hill. 2015. Peer Production: A Form of Collective Intelligence. In *Handbook of Collective Intelligence*, Thomas W. Malone and Michael S. Bernstein (Eds.). MIT Press, Cambridge, MA, 175–204.

[12] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic Speech Recognition for Under-Resourced Languages: A Survey. *Speech Communication* 56 (2014), 85–100.

[13] Pushpak Bhattacharyya, Hema Murthy, Surangika Ranathunga, and Ranjiva Munasinghe. 2019. Indic Language Computing. *Commun. ACM* 62, 11 (2019), 70–75.

[14] Susan L. Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work (GROUP ’05)*. ACM, New York, NY, 1–10. https://doi.org/10.1145/1099203.1099205

[15] Anita Say Chan. 2014. *Networking Peripheries: Technological Futures and the Myth of Digital Universalism* (illustrated edition ed.). The MIT Press, Cambridge, Mass.

[16] Leslie Chan, Budd Hall, Florence Piron, Rajesh Tandon, and Wanósts’a7 Lorna Williams. 2020. Open Science Beyond Open Access: For and with Communities, A Step towards the Decolonization of Knowledge. (July 2020). https://doi.org/10.5281/ZENODO.3946773

[17] Kathy Charmaz. 2014. *Constructing grounded theory*. sage.

[18] Partha Chatterjee. 1996. The Manifold Uses of Jati. In *Region, Religion, Caste, Gender and Culture in Contemporary India*, T. V. Sathyamurthy (Ed.). Number v. 3 in Social Change and Political Discourse in India. Oxford University Press, Delhi ; New York, 281–292.

[19] Rahul De, Lewin Sivamala, Ravi A. Rao, Sharmila Chakravarty, Supriya Dey, and Uma Bharath. 2015. *Economic Impact of Free and Open Source Software Usage in Government: Final Report*. Technical Report.

[20] Nicolas Ducheneaut. 2005. Socialization in an Open Source Software Community: A Socio-Technical Analysis. *Computer Supported Cooperative Work (CSCW)* 14, 4 (2005), 323–368. https://doi.org/10.1007/s10606-005-9000-1

[21] Bernard Enjolras. 2002. The commercialization of voluntary sport organizations in Norway. *Nonprofit and voluntary sector quarterly* 31, 3 (2002), 352–376.

[22] Peter B. Evans. 1995. *Embedded Autonomy: States and Industrial Transformation*. Princeton University Press, Princeton, NJ.

[23] Heather Ford, Mark Graham, and Eric Meyer. 2015. *Fact Factories: Wikipedia and the Power to Represent*. Ph.D. Dissertation. University of Oxford.

[24] Jana Gallus. 2016. Fostering Public Good Contributions with Symbolic Awards: A Large-Scale Natural Field Experiment at Wikipedia. *Management Science* 63, 12 (Sept. 2016), 3999–4015. https://doi.org/10.1287/mnsc.2016.2540

[25] Rishab Aiyer Ghosh. 2005. Understanding Free Software Developers: Understandings from the FLOSS Study. In *Perspectives on Free and Open Source Software*, Joseph Feller, Brian Fitzgerald, Scott A. Hissam, and Karim R. Lakhani (Eds.). MIT Press, Cambridge, MA, 23–36.

[26] Sanjay Ghosh and Anirudha Joshi. 2014. Text Entry in Indian Languages on Mobile: User Perspectives. In *Proceedings of the India HCI 2014 Conference on Human Computer Interaction - IHCI '14*. ACM Press, New Delhi, India, 55–63. https://doi.org/10.1145/2676702.2676710

[27] Ruediger Glott, Rishab Ghosh, and Philipp Schmidt. 2010. *Wikipedia Survey.* Technical Report. UNU-MERIT, Maastricht, Netherlands.

[28] Mark Granovetter. 1985. Economic Action and Social Structure: The Problem of Embeddedness. *Amer. J. Sociology* 91, 3 (Nov. 1985), 481–510. https://doi.org/10.1086/228311

[29] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist* 57, 5 (May 2013), 664–688. https://doi.org/10.1177/0002764212469365

[30] Aaron Halfaker, Oliver Keyes, and Dario Taraborelli. 2013. Making Peripheral Participation Legitimate: Reader Engagement Experiments in Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13).* ACM, New York, NY, USA, 849–860. https://doi.org/10.1145/2441776.2441872

[31] Noriko Hara, Pnina Shachaf, and Khe Foon Hew. 2010. Cross-Cultural Analysis of the Wikipedia Community. *Journal of the American Society for Information Science and Technology* 61, 10 (2010), 2097–2108. https://doi.org/10.1002/asi.21373

[32] Kieran Healy and Alan Schussman. 2003. The Ecology of Open-Source Software Development. (2003).

[33] Brent Hecht and Darren Gergle. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10).* ACM, Atlanta, Georgia, USA, 291–300. https://doi.org/10.1145/1753326.1753370

[34] Molly G Hickman, Viral Pasad, Harsh Kamalesh Sanghavi, Jacob Thebault-Spieker, and Sang Won Lee. 2021. Understanding Wikipedia Practices Through Hindi, Urdu, and English Takes on an Evolving Regional Conflict. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–31.

[35] Benjamin Mako Hill. 2013. Almost Wikipedia: What Eight Early Online Collaborative Encyclopedia Projects Reveal about the Mechanisms of Collective Action. In *Essays on Volunteer Mobilization in Peer Production.* Massachusetts Institute of Technology, Cambridge, Massachusetts.

[36] Benjamin Mako Hill and Aaron Shaw. 2013. The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation. *PLoS ONE* 8, 6 (June 2013), e65782. https://doi.org/10.1371/journal.pone.0065782

[37] Benjamin Mako Hill and Aaron Shaw. 2019. Studying Populations of Online Communities. In *The Oxford Handbook of Networked Communication*, Brooke Foucault Welles and Sandra González-Bailón (Eds.). Oxford University Press, Oxford, UK, 173–193.

[38] Benjamin Mako Hill and Aaron D. Shaw. 2020. The Most Important Laboratory for Social Scientific and Computing Research in History. In *Wikipedia @ 20: Stories of an Incomplete Revolution*, Joseph M. Jr. Reagle and Jackie L. Koerner (Eds.). MIT Press, Cambridge, Massachusetts.

[39] Geert Hofstede. 1984. *Culture's Consequences: International Differences in Work-Related Values.* SAGE.

[40] Joan E. Hughes and Ravi Narayan. 2009. Collaboration and Learning with Wikis in Post-Secondary Classrooms. *Journal of Interactive Online Learning* 8, 1 (2009), 63–82.

[41] Sohyeon Hwang and Jeremy D Foote. 2021. Why do people participate in small online communities? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[42] Sohyeon Hwang and Aaron Shaw. [n.d.]. Heterogeneous Practices in Collective Governance. ([n. d.]).

[43] Kancha Ilaiah. 2006. Merit of Reservations. *Economic and Political Weekly* 41, 24 (2006), 2447–2449.

[44] Aniket Jaaware. 2018. *Practicing Caste: On Touching and Not Touching* (first ed.). Fordham University Press. https://doi.org/10.5422/fordham/9780823282265.001.0001

[45] Steven L. Johnson, Samer Faraj, and Srinivas Kudaravalli. 2014. Emergence of Power Laws in Online Communities: The Role of Social Mechanisms and Preferential Attachment. *Management Information Systems Quarterly* 38, 3 (2014), 795–808.

[46] Gopal Kadekodi, Ravi Kanbur, and Vijayendra Rao. 2007. Governance and the'Karnataka Model of Development'. *Economic and Political Weekly* (2007).

[47] Bhuvana Meenakshi Koteeswaran. 2021. Bridging the Gender Gap. (2021).

[48] KPMG. 2017. Indian Languages –Defining India's Internet. https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf.

[49] Robert E. Kraut and Andrew T. Fiore. 2014. The Role of Founders in Building Online Groups. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14).* ACM, Baltimore, Maryland, USA, 722–732. https://doi.org/10.1145/2531602.2531648

[50] Robert E. Kraut, Paul Resnick, and Sara Kiesler. 2012. *Building Successful Online Communities: Evidence-Based Social Design.* MIT Press, Cambridge, MA.

[51] G Gopa Kumar. 2003. Kerala: Stable Bipolar Alignment. *Journal of Indian School of Political Economy* 15, 1-2 (2003), 79–95.

[52] John Kurien. 1995. The Kerala Model: Its Central Tendency and the Outlier. *Social Scientist* 23, 1/3 (1995), 70–90. https://doi.org/10.2307/3517892

[53] Leantros Kyriakoullis and Panayiotis Zaphiris. 2016. Culture and HCI: A Review of Recent Cultural Studies in HCI and Social Networks. *Universal Access in the Information Society* 15, 4 (Nov. 2016), 629–642. https://doi.org/10.1007/s10209-015-0445-9

[54] Karim R. Lakhani and B. Wolf. 2005. Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects. In *Perspectives on Free and Open Source Software*, Joseph Feller, Brian Fitzgerald, Scott A. Hissam, and Karim R. Lakhani (Eds.). MIT Press, 3–22.

[55] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the World Reads Wikipedia: Beyond English Speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19).* ACM, New York, NY, USA, 618–626. https://doi.org/10.1145/3289600.3291021

[56] GK Lieten. 2002. Human Development in Kerala: Structure and Agency in History. *Economic and Political Weekly* (2002).

[57] Randall M. Livingstone. 2016. Population Automation: An Interview with Wikipedia Bot Pioneer Ram-Man. *First Monday* (Jan. 2016). https://doi.org/10.5210/fm.v21i1.6027

[58] Maggie MacAulay and Rebecca Visser. 2016. Editing Diversity In: Reading Diversity Discourses on Wikipedia. *Ada: A Journal of Gender, New Media, and Technology* 9 (May 2016). https://doi.org/10.7264/N36M3541

[59] Laura March and Sayamindu Dasgupta. 2020. Wikipedia Edit-a-Thons as Sites of Public Pedagogy. *Proc. ACM Hum.-Comput. Interact.* 100:1–100:26, CSCW2 (2020). https://doi.org/10.1145/3415171

[60] Lousie Matsakis. 2019. Google Gives Wikimedia Millions—Plus Machine Learning Tools. *Wired* (Jan. 2019).

[61] Mohamad Mehdi, Chitu Okoli, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. 2017. Excavating the Mother Lode of Human-Generated Text: A Systematic Review of Research That Uses the Wikipedia Corpus. *Information Processing & Management* 53, 2 (March 2017), 505–529. https://doi.org/10.1016/j.ipm.2016.07.003

[62] Amanda Menking and Ingrid Erickson. 2015. The Heart Work of Wikipedia: Gendered, Emotional Labor in the World's Largest Online Encyclopedia. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15).* ACM, New York, NY, USA, 207–210. https://doi.org/10.1145/2702123.2702514

[63] Amanda Menking, Ingrid Erickson, and Wanda Pratt. 2019. People Who Can Take It: How Women Wikipedians Negotiate and Navigate Safety. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* Association for Computing Machinery, Glasgow, Scotland, UK, 472:1–472:14. https://doi.org/10.1145/3290605.3300702

[64] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. 2015. "The Sum of All Human Knowledge": A Systematic Review of Scholarly Research on the Content of Wikipedia: "The Sum of All Human Knowledge": A Systematic Review of Scholarly Research on the Content of Wikipedia. *Journal of the Association for Information Science and Technology* 66, 2 (Feb. 2015), 219–245. https://doi.org/10.1002/asi.23172

[65] Ministry of Home Affairs, Government of India. 2011. *Census.* Technical Report.

[66] Ministry of Home Affairs, Government of India. 2011. State of Literacy. https://censusindia.gov.in/2011-prov-results/data_files/india/Final_PPT_2011_chapter6.pdf.

[67] Marc Miquel-Ribé and David Laniado. 2018. Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions. *Frontiers in Physics* 6 (2018). https://doi.org/10.3389/fphy.2018.00054

[68] Marc Miquel-Ribé and David Laniado. 2019. Wikipedia Cultural Diversity Dataset: A Complete Cartography for 300 Language Editions. *Proceedings of the International AAAI Conference on Web and Social Media* 13 (July 2019), 620–629.

[69] Aparna Mitra and Pooja Singh. 2007. Human Capital Attainment and Gender Empowerment: The Kerala Paradox*. *Social Science Quarterly* 88, 5 (2007), 1227–1242. https://doi.org/10.1111/j.1540-6237.2007.00500.x

[70] Jonathan T. Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. 2013. Tea and Sympathy: Crafting Positive New User Experiences on Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13).* ACM, New York, NY, USA, 839–848. https://doi.org/10.1145/2441776.2441871

[71] Jonathan T. Morgan and Aaron Halfaker. 2018. Evaluating the Impact of the Wikipedia Teahouse on Newcomer Socialization and Retention. In *Proceedings*

of the 14th International Symposium on Open Collaboration (OpenSym '18). ACM, New York, NY, 20:1–20:7. https://doi.org/10.1145/3233391.3233544

[72] Sneha Narayan, Jake Orlowitz, Jonathan Morgan, Benjamin Mako Hill, and Aaron Shaw. 2017. The Wikipedia Adventure: Field Evaluation of an Interactive Tutorial for New Users. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1785–1799. https://doi.org/10.1145/2998181.2998307

[73] Sabine Niederer and José van Dijck. 2010. Wisdom of the Crowd or Technicity of Content? Wikipedia as a Sociotechnical System. *New Media & Society* 12, 8 (Dec. 2010), 1368–1387. https://doi.org/10.1177/1461444810365297

[74] Oded Nov. 2007. What Motivates Wikipedians? *Commun. ACM* 50, 11 (2007), 60–64. https://doi.org/10.1145/1297797.1297798

[75] Jude Ntabathia. 2019. Categories of control and visibility in mapping infrastructures. In *Proceedings of the Conference on Computing & Sustainable Societies - COMPASS 19*. ACM Press, Accra, Ghana, 174–183. https://doi.org/10.1145/3314344.3332494

[76] Chitu Okoli, Mohamad Mehdi, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. 2012. *The People's Encyclopedia under the Gaze of the Sages: A Systematic Review of Scholarly Research on Wikipedia*. SKEMA Business School. http://dx.doi.org/10.2139/ssrn.2021326

[77] Felipe Ortega. 2009. *Wikipedia: A Quantitative Analysis*. Ph.D. Dissertation. Universidad Rey Juan Carlos, Madrid, Spain.

[78] Felipe Ortega and Jesus M. Gonzalez Barahona. 2007. Quantitative Analysis of the Wikipedia Community of Users. In *Proceedings of the 2007 International Symposium on Wikis (WikiSym '07)*. ACM, New York, NY, USA, 75–86. https://doi.org/10.1145/1296951.1296960

[79] Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians Are Born, Not Made: A Study of Power Editors on Wikipedia. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work (GROUP '09)*. ACM, New York, NY, 51–60. https://doi.org/10.1145/1531674.1531682

[80] Madhu S Paranjape and Madhu S. 2007. Uneven Distribution of Education in Maharashtra: Rural-Urban, Gender and Caste Inequalities. *Economic and Political Weekly* (2007).

[81] Kamna Patel. 2017. What Is in a Name? How Caste Names Affect the Production of Situated Knowledge. *Gender, Place & Culture* 24, 7 (July 2017), 1011–1030. https://doi.org/10.1080/0966369X.2017.1372385

[82] Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. 2006. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication* 12, 1 (2006), 88–113. https://doi.org/10.1111/j.1083-6101.2006.00316.x

[83] Karl Polanyi. 1958. The Economy as Instituted Process. In *Trade and Market in the Early Empires*, Karl Polanyi, Conrad M. Arensberg, and Harry W. Pearson (Eds.). Free Press, New York, NY, 243–270.

[84] Jennifer Preece and Ben Shneiderman. 2009. The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation. *AIS Transactions on Human-Computer Interaction* 1, 1 (2009), 13–32.

[85] Morten Rask. 2008. The Reach and Richness of Wikipedia: Is Wikinomics Only for Rich Countries? *First Monday* (May 2008). https://doi.org/10.5210/fm.v13i6.2046

[86] Joseph Reagle. 2010. *Good Faith Collaboration: The Culture of Wikipedia*. MIT Press, Cambridge, MA.

[87] Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating Code-Switching on Twitter with a Novel Generalized Word-Level Language Detection Technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1971–1982. https://doi.org/10.18653/v1/P17-1180

[88] Joachim Schroer and Guido Hertel. 2009. Voluntary Engagement in an Open Web-Based Encyclopedia: Wikipedians and Why They Do It. *Media Psychology* 12 (Jan. 2009), 96–120(25). https://doi.org/10.1080/15213260802669466

[89] Charles M. Schweik and Robert C. English. 2012. *Internet Success: A Study of Open-Source Software Commons*. MIT Press, Cambridge, MA.

[90] Amartya Sen. 1998. Mortality as an Indicator of Economic Success and Failure. *The Economic Journal* 108, 446 (Jan. 1998), 1–25. https://doi.org/10.1111/1468-0297.00270

[91] Amartya Sen. 2013. *The Argumentative Indian: Writings on Indian History, Culture and Identity*. Farrar, Straus and Giroux, New York.

[92] Olivier Serrat. 2017. The Five Whys Technique. In *Knowledge Solutions: Tools, Methods, and Approaches to Drive Organizational Performance*, Olivier Serrat (Ed.). Springer, Singapore, 307–310. https://doi.org/10.1007/978-981-10-0983-9_32

[93] Aaron Shaw and Eszter Hargittai. 2018. The Pipeline of Online Participation Inequalities: The Case of Wikipedia Editing. *Journal of Communication* 68, 1 (Feb. 2018), 143–168. https://doi.org/10.1093/joc/jqx003

[94] Richard M. Stallman. 2002. *Free Software, Free Society: Selected Essays of Richard M. Stallman*. Lulu. com, Morrisville.

[95] M. Reza Talebinejad, Hossein Vahid Dastjerdi, and Ra'na Mahmoodi. 2012. Barriers to Technical Terms in Translation: Borrowings or Neologisms. *Terminology* 18, 2 (2012), 167–187. https://doi.org/10.1075/term.18.2.02tal

[96] Nathan TeBlunthuis, Aaron Shaw, and Benjamin Mako Hill. 2018. Revisiting "The Rise and Decline" in a Population of Peer Production Projects. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 355:1–355:7. https://doi.org/10.1145/3173574.3173929

[97] Jennifer Thom-Santelli, Dan R. Cosley, and Geri Gay. 2009. What's Mine Is Mine: Territoriality in Collaborative Authoring. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1481–1484. https://doi.org/10.1145/1518701.1518925

[98] Kentaro Toyama. 2015. *Geek heresy: Rescuing social change from the cult of technology*. PublicAffairs, New York, NY.

[99] Jan E. Trost. 1986. Statistically Nonrepresentative Stratified Sampling: A Sampling Technique for Qualitative Studies. *Qualitative Sociology* 9, 1 (March 1986), 54–57. https://doi.org/10.1007/BF00988249

[100] Ziko Van Dijk. 2009. Wikipedia and Lesser-Resourced Languages. *Language Problems and Language Planning* 33, 3 (2009), 234–250.

[101] Georg von Krogh, Sebastian Spaeth, and Karim R. Lakhani. 2003. Community, Joining, and Specialization in Open Source Software Innovation: A Case Study. *Research Policy* 32, 7 (2003), 1217–1241. https://doi.org/10.1016/S0048-7333(03)00050-7

[102] Max Weber. 2003. *The Protestant Ethic and the Spirit of Capitalism*. Dover Publications, Mineola, NY.

[103] Olga Zagovora, Fabian Flöck, and Claudia Wagner. 2017. "(Weitergeleitet von Journalistin)": The Endered Presentation of Professions on Wikipedia. In *Proceedings of the 2017 ACM on Web Science Conference (WebSci '17)*. ACM, New York, NY, 83–92. https://doi.org/10.1145/3091478.3091488

[104] Xiaoquan Zhang and Chong Wang. 2012. Network Positions and Contributions to Online Public Goods: The Case of Chinese Wikipedia. *Journal of Management Information Systems* 29, 2 (Oct. 2012), 11–40. https://doi.org/10.2753/MIS0742-1222290202

[105] Xiaoquan Michael Zhang and Feng Zhu. 2011. Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia. *American Economic Review* 101, 4 (June 2011), 1601–1615. https://doi.org/10.1257/aer.101.4.1601

A APPENDIX

A.1 Background Information on Wikipedias

A.1.1 Malayalam Wikipedia. *Malayalam Wikipedia (ML)* was launched in December 2002.[16] Malayalam is a Dravidian language spoken in the Indian state of Kerala and the union territories of Lakshadweep and Puducherry by the Malayali people.[17] It is written in Malayalam script and spoken by around 35 million native speakers in India [65]. Malayalam Wikipedia is formally supported by Wikimedians of Kerala user group established on October 12, 2018.[18]

A.1.2 Marathi Wikipedia. *Marathi Wikipedia (MR)* was launched in May 2003.[19] Marathi is an Indo-Aryan language spoken predominantly by around 83 million native speakers of Maharashtra in India [65]. It is written in the Devanagari script, which is also used by Hindi, and has the third-largest number of native speakers in India.[20] Marathi Wikipedia is formally supported by the Marathi Wikimedians user group established on March 15, 2019.[21]

A.1.3 Kannada Wikipedia. *Kannada Wikipedia (KN)* was launched in June 2003.[22] Kannada is a Dravidian language spoken predominantly by the people of Karnataka.[23] It is written in the Kannada script and is spoken by around 44 million native speakers in India [65]. Kannada Wikipedia is formally supported by the Karavali Wikimedians user group established on February 2017 which also supports the Tulu and Konkani language Wikipedias.[24]

[16] https://en.wikipedia.org/wiki/Malayalam_Wikipedia
[17] https://www.ethnologue.com/language/mal
[18] https://meta.wikimedia.org/wiki/Wikimedians_of_Kerala
[19] https://en.wikipedia.org/wiki/Marathi_Wikipedia
[20] https://www.ethnologue.com/language/mar
[21] https://meta.wikimedia.org/wiki/Affiliations_Committee/Resolutions/Recognition_of_Marathi_Wikimedians_User_Group
[22] https://en.wikipedia.org/wiki/Kannada_Wikipedia
[23] https://www.ethnologue.com/language/kan
[24] https://meta.wikimedia.org/wiki/Karavali_Wikimedians/Reports

A.2 Descriptive Statistics

The definitions of descriptive statistics from the Wikimedia Foundation and used in our paper are presented in Table 3.

A.3 Common Challenges

A.3.1 Language Localization and Wikipedia Mobile Editing. Most of the Indian language Wikipedia editors in our study created new articles by translating content from English Wikipedia. These editors found content translation on mobile devices to be challenging as it required them to switch between different tabs on mobile. In addition, many systems that make editing easier are not localized for smaller communities. For example, editors complained about the lack of localized templates that enable the addition of new citations by filling a simple form instead of manually coding markup.

A.3.2 Challenges in text entry. Participants from all three projects complained that the Wikipedia mobile interface frequently does not include the ability to input Indian language text by default—either because of the limitations of users’ phones or the Wikipedia mobile interface itself. Even when technical issues could be overcome, the increased script and code complexity in Indic languages mean that many Indian language users may still face challenges in contributing [13, 26]. One editor from MR7 explained that “English grammar can be edited fast, Marathi and Hindi takes time to write.”

A.3.3 Challenges in information retrieval. In addition, users conducting Internet searches in an underrepresented language face problems getting accurate results. One Malayalam Wikipedian explained that “if they make spelling mistakes ... then the search will not show the article, unlike in English. So people will think that there is no Malayalam content for that” (ML12). Our interviewees explained that this reduced ability to gather information affects both contributors’ ability to find references and source material and readers’ ability to find Wikipedia articles for the topics they search for.

A.3.4 Wikitext editing. All three Wikipedias struggled with poor Internet connections that were common among would-be contributors. Although Wikipedia is relatively easy to read over a slow connection, WMF’s rich text visual editor that enables users to edit Wikipedia without having to learn the wiki markup requires a high-bandwidth and low-latency pipe. Reflecting on 2G Internet connections common in rural India, one ML editor said that “the visual editor is a biggest disaster to the whole Mediawiki world” (ML11). Editors like ML11 were frustrated with the high bandwidth requirements of WMF’s visual editor and explained that many Indian contributors were forced to resort to “wiki markup”—a famously difficult form of text-based code that has been shown by previous research to provide a major barrier to contributing to Wikipedia [40].

A.3.5 Western influence on local language digital use. Contributors from all three Wikipedias argued that their projects struggled with the perception that their languages were of lower level than English. For example, MR7 said that “I see that there is no craze for [Marathi] as the generation is shifting towards international languages.” English is the *de facto* business language of India. Like many other subjects across the three efforts, MR7 felt that underrepresented languages suffered because young people were increasingly interested in contributing to English language content to build skills in order to prepare them for new opportunities in business.

A.3.6 Premature Optimization: High Standards and Expectations from English Wikipedia. All three communities struggled with a sense that they were forced to live up to the standards of larger Wikipedia editions due to expectations from both readers and other Wikipedians. For example, KN17 explained that new editors to KN “are thinking that Wikipedia has a big name and people who edit Wikipedia are amazing, and whatever I will create, will be read by some millions of users” (KN17). KN17 explained that readers’ expectations about Wikipedia could be paralyzing and demotivating for many newcomers. In addition, KN 17 explained that senior editors felt responsible for maintaining the credibility of Wikipedia due to the high reputation of English Wikipedia. For example, all three communities pointed out that rigid policies about content quality established on English Wikipedia as that community matured were imported to other Wikipedia communities in ways that hindered the type of growth and experimentation that supported the growth of English Wikipedia.

A.3.7 Limited resources for community outreach. Finally, all three language editions faced challenges related to the high cost of conducting effective user outreach. All three communities felt that would-be contributors required training and guidance to get started. As a result, all projects directed energy to increase Wikipedia awareness by providing in-person wiki editing training and outreach workshops as well as events at an enormous cost in terms of volunteer time and other resources. For example, MR4, KN16, and ML9 cited workshops and tutorials as the single most important model for increasing engagement. Similarly, editors like MR2 and MR3 cited the lack of resources for conducting effective follow-ups from workshops as a cause of low engagement rates. Members of all three communities felt that the technical difficulties described in the previous sections could most effectively be overcome through careful in-person training that walked would-be contributors through the processes of doing effective search, inputting language in non-Latin scripts, writing wikitext on low bandwidth connections, and so on. Members of all three communities recognized that doing so would require costly human labor and volunteer efforts that were also in short supply.

Metric	Description
Article "depth"	Defined as ([Edits/Articles] × [Non-Articles/Articles] × [1- Stub-ratio]) is a rough indicator of a Wikipedia's quality, showing how frequently its articles are updated. It does not refer to academic quality.
Content editors	The count of editors with one or more edits, including on redirect pages, with content page type.
Active content editors	The count of editors with five or more edits in a given month, including on redirect pages, with content page type.
New pages	The count of new pages created, excluding pages being redirects. We measure this by counting page creations and ignoring any page deletions or restores.
Unique devices	A key content consumption metric is unique devices; how many distinct devices we have visiting our web properties in a given time period.
Newly registered user	Newly registered user is a standardized user class used to measure the number of new users signing up on a given wiki project for the first time. It is used as a proxy for user acquisition.

Table 3: Definitions of metrics drawn from the Wikimedia Foundation.