

Determinants of Wikipedia Quality: the Roles of Global and Local Contribution Inequality

Ofer Arazy

University of Alberta
ofer.arazy@ualberta.ca

Oded Nov

New York University Polytechnic Institute
onov@poly.edu

ABSTRACT

The success of Wikipedia and the relative high quality of its articles seem to contradict conventional wisdom. Recent studies have begun shedding light on the processes contributing to Wikipedia's success, highlighting the role of coordination and contribution inequality. In this study, we expand on these works in two ways. First, we make a distinction between global (Wikipedia-wide) and local (article-specific) inequality and investigate both constructs. Second, we explore both direct and indirect effects of these inequalities, exposing the intricate relationships between global inequality, local inequality, coordination, and article quality. We tested our hypotheses on a sample of a Wikipedia articles using structural equation modeling and found that global inequality exerts significant positive impact on article quality, while the effect of local inequality is indirect and is mediated by coordination.

Author Keywords

Wikipedia, quality, contributing inequality, coordination.

ACM Classification Keywords

H5.m. Information interfaces and presentation: Misc.

General Terms

Management, Human Factors, Theory.

INTRODUCTION

Recent years have seen the emergence of a community-based model for creation of knowledge-based goods [3, 19] by large, distributed, self governing communities of volunteers. A prominent example of the community-based production model is Wikipedia [1, 12, 15]. Given its lack of clear monetary incentives for participation, and its loose governance, the quality of Wikipedia seems counter-intuitive. While the wiki collective knowledge production model lacks traditional quality control procedures and therefore susceptible to risks [7], recent studies demonstrate that the quality of Wikipedia is comparable to that of traditional encyclopedias [9]. Further, recent studies on Wikipedia suggest that errors and vandalism are quickly

corrected [17, 18], and that expert reviewers determine the credibility of the articles to be higher than non-expert reviewers [5].

How, then, does Wikipedia succeed in developing high-quality content? In recent years, researchers have studied various aspects of quality in Wikipedia, including topical scope and coverage [2, 10], articles' rigor and diversity [14], and metrics for quality [17]. A few studies identified factors that are related to Wikipedia articles' quality: Anthony et al [1] found that individual contributions' quality is positively related with the number of contributions, and that contributions of highest quality come from a large number of anonymous, infrequent editors. Wilkinson and Huberman's [21] found that high-quality articles were edited by a larger number of editors, while [17] showed that the level of explicit coordination in an article's discussion page is associated with high quality (i.e. featured) articles. In a recent paper, Kittur and Kraut [12] built on these earlier studies to develop a model for the factors that determine change in article quality. They demonstrated the role contribution inequality (referred to as "implicit coordination") in moderating the impact of group size on quality. Their findings highlight the importance of contribution inequality as a key to understanding the mechanism for creating high quality articles.

Some questions, however, regarding the factors driving Wikipedia articles' quality, and in particular the effect of contribution inequality, remain open. In this study, we focus our attention on the distinction between the activity of a set of editors in a specific article (i.e. local activity) as opposed to the activity of this set of editors across Wikipedia at large (i.e. global activity). Heavy contributors to Wikipedia at large may be individuals with high commitment to Wikipedia and its principal tenets (i.e. "information should be free"), whereas individuals who contribute intensively to a specific article but relatively little to others may be more concerned with the specific article topic rather than Wikipedia at large. Thus, we distinguish between *local inequality* - inequality of editors' contribution in a particular article, and *global inequality* - inequality in overall Wikipedia activity levels for the same set of editors. To illustrate this, consider an article that was written by ten editors, where each has made the same number of edits to the article, in which case the local inequality would be low. If, however, two of these ten editors contributed heavily to other Wikipedia articles and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CSCW 2010, February 6–10, 2010, Savannah, Georgia, USA.

Copyright 2009 ACM 978-1-60558-246-7/09/04...\$5.00.

the remaining eight editors contributed very little to Wikipedia at large, global inequality (of the same group of editors) would be high. Kittur and Kraut [12] have argued that local inequality reduces the need for coordination (since most work is done by a few heavy contributors) and enables the development of a shared mental model, which results in quality output. We believe that global inequality, too – plays an important role in determining article quality. Furthermore, we argue that both global and local inequalities can impact coordination levels. The objective of this paper is, thus, to explore the relationships between editors' global contribution inequality, local contribution inequality, coordination levels, and the resulting article quality. We employ a multi-stage statistical model, which allows us to examine both direct and indirect effects.

RESEARCH MODEL

In Wikipedia, global inequality of contribution is likely to represent differences of commitment, interest and involvement in Wikipedia at large. When examined within the context of any given article, these differences are likely to translate into differences in the commitment and involvement in the edit work of the specific article. Furthermore, since high global inequality is associated with diversity of interests and commitment, which are likely to require more coordination, we propose:

H1. Global inequality is positively related to local inequality.

H2. Global inequality is positively related to coordination.

People who are highly active in Wikipedia at large are more likely to have an official role in Wikipedia, such as administrators. During the ongoing process article editing, the presence of administrators or other devoted Wikipedians in a group of editors is helpful in conflict resolution and breaking deadlocks. However, the group should not include too many such administrators, as “too many cooks spoil the broth”. Such a diverse group composition will be manifested in high global inequality, and we therefore propose:

H3. Global inequality is positively related to article quality.

Local inequality in contribution level among editors of a given article is likely to represent – at least to some extent – diversity of knowledge bases and points of view related to the topic of the article. Functional diversity of team members was found to be positively related to a number of performance indicators in teams [4], such as external knowledge sharing [6] and sustained performance [11]. This positive relation between local inequality and the performance of groups of editors (i.e. article quality) is supported by the findings of [12], and therefore we propose:

H4. Local inequality is positively related to article quality.

Kittur and Kraut [12] argued that local inequality have argued that local inequality improves coordination, since when most work is done by a few “heavy” contributors,

these editors develop shared mental models, whereby they can maintain a common view of the article content and structure without the need for explicit communication. These focal editors play a leadership role and direct the other editors, thus enabling indirect coordination. In line with these arguments, we expect that:

H5. Local inequality is positively related to coordination.

Finally, in line with the findings of [17], who studied how activity in Wikipedia articles' discussion pages - where coordination takes place - is related to information quality, we argue that coordination is important for enabling editors to hash out differences of opinion and synthesize distinct editors' knowledge into a coherent article. We thus expect:

H6. Coordination is positively related to article quality.

RESEARCH METHOD

In order to test the proposed hypotheses, we conducted an empirical study of the English version of Wikipedia. The unit of analysis is an article, where the editing of each article represents a distinct group project, and the group is defined by the set of editors who contributed to the article. Previous research on article quality employed Wikipedia's internal rating mechanism, and specifically the ‘Feature Articles’ category. This approach is not suitable for studying the impact of coordination, since processes for becoming a featured article explicitly require additional coordination activities¹. To the best of our knowledge, there exists no dataset of Wikipedia articles with reliable quality ratings. Thus, we sought an alternative sample. In an unpublished survey [16] that was advertised at the AISWorld mailing list (operated by the Association for Information Systems and serving information systems researchers), fifty scholars answering an advertisement rated the **accuracy** and **completeness** of 50 Wikipedia articles of their choice (the topic of these articles was science and technology) on a 5-point Likert scale. We used these ratings as metrics of article quality.

The additional variables were measured by ‘harvesting’ Wikipedia. Since the contents of Wikipedia changes continually, the estimates for all constructs were based on the article version on which quality assessments were made. Coordination was estimated based on discussion page activity - in line with [12] – and we used the number of words as a metric. Local Inequality was calculated based on the number of edits made to the article, while Global Inequality was estimated based on the number of edits

¹ We found that the Wikipedia's article rating (as indicated by articles' category, e.g. ‘featured’, ‘A’, etc.) is highly correlated with the coordination levels on the discussion page (correlations were in the 0.8-0.9 range). Moreover, Wikipedia's article rating is also highly correlated with the number of contributing editors (i.e. group size). This suggests that Wikipedia's internal rating process may be biased, such that editors are involved in influencing the rating of their own article, and the more editors the article has the more it is likely to be rated positively.

made by the article's editor group to Wikipedia at large. Inequality (both local and global) was calculated using the Gini coefficient [8]. The Gini coefficient can range between 0 and 1: with low inequality (i.e., all editors contribute fairly evenly), the Gini coefficient approaches 0; and with high inequality (i.e., very few editors contribute), the Gini coefficient approaches 1. This measure has been used to measure contribution inequality in open source software projects [13] as well as Wikipedia [12]. Since the number of group members (i.e. article editors) may affect inequality levels, the extent of coordination, or even article quality, we controlled for the effect of group size.

RESULTS

Descriptive statistics are presented in Table 1. We notice that groups are large, with an average of close to 250 contributors per article. Global inequality is greater than the average local inequality. The extent of knowledge sharing is high, with an average of 3500 words in each article's discussion page. The quality of the articles averages at 3.7 and 3.0 for Accuracy and Completeness respectively.

Construct	Measure	Mean	Med	SD
Group size	Number if editors	249	40	777
Global inequality	Gini coefficient of editor's overall edits	.79	0.82	.12
Local inequality	Gini coefficient. of editor's article edits	.34	0.32	.15
Coordination	Discussion Page Words	3502	172	13807
Product quality (1)	Accuracy (1-5)	3.7	4	1.1
Product quality (2)	Completeness (1-5)	3.0	3	1.2

Table 1. Constructs' mean, median, and standard deviation

Data analysis was conducted with structural equation modeling, using *Partial Least Squares* (PLS) algorithm, which estimates multi-stage path models.

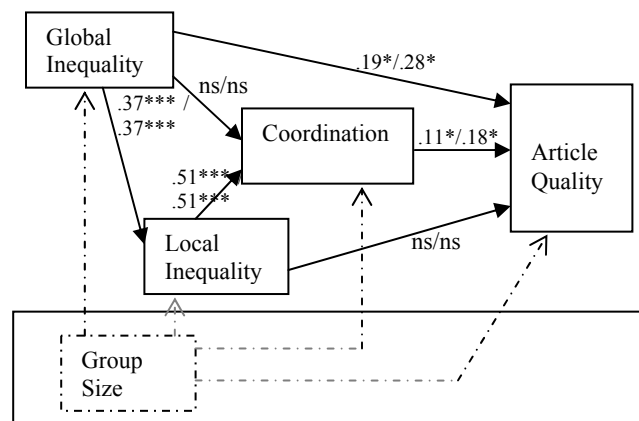


Figure 1. PLS results for two alternative models (when using Accuracy / Completeness as the outcome variable). * = P<0.05; ** = P<0.01; * = P<0.001; ns = non-significant**

PLS requires few assumptions about distribution of variables (i.e. it does not assume Normal distribution), and thus is suitable for analyzing variables with power-law distribution (in our data, *group size* and *coordination*

exhibit such power law distribution). Discriminant validity was examined by analyzing the correlations between constructs, which did not exceed the recommended threshold of 0.5. The significance of structural path estimates was computed using the bootstrapping re-sampling method, and the structural model was evaluated based on both the R^2 for each variable and the statistical significance of structural paths. Figure 1 presents the results for both output variables (accuracy / completeness).

DISCUSSION AND CONCLUSION

The success of Wikipedia and the relative high quality of its articles seem to contradict conventional wisdom. Recent studies have begun shedding light on the processes contributing to Wikipedia's success, highlighting the role of coordination [12, 17], and contribution inequality [17]. In this preliminary study, we expand on these works in two ways. First, we make a distinction between global and local inequality, and explore both constructs. Second, we explore both direct and indirect effects, exposing the intricate relationships between global inequality, local inequality, coordination, and article quality. We tested our model on a set of Wikipedia articles where quality assessment was independent of internal Wikipedia processes. Some of the hypothesized relationships were supported, and the results were consistent for both two outcome variables (Accuracy and Completeness).

We found that – in line with prior results [17] – coordination has a significant positive impact on article quality, suggesting that communication and coordination between contributors is essential for developing high quality content. We also found that global inequality impacts local inequality significantly, demonstrating that editors' activity in Wikipedia at large is associated with their activity in the specific article.

Interestingly, we found that while *local* inequality has a significant positive impact on coordination (contrary to H5), it does not exert a significant direct effect on quality (therefore H4 is not supported). In contrast, *global* inequality does not have a significant impact on coordination (contrary to H2), but does exert a significant direct positive effect on quality (H3) Thus, it seems that global inequality impacts quality directly, while local inequality works indirectly, by facilitating coordination, which in turn contributes to quality. These findings suggest that both inequalities are important, although they influence article quality in different ways.

The results regarding the effects of local inequality stand in contrast to the findings of [12], who (a) suggested that local inequality and coordination are complementary, such that high inequality would result in low coordination, and (b) showed that local inequality impacts article quality. While these discrepancies may stem from the different sample we have used, we suspect that they may be due to the fact that [12] did not test for indirect effects, suggesting that the effect of inequality is moderated by coordination. The

positive relation between inequality and coordination may be explained by the long tail associated with high inequality, such that when the edit work is divided between many people more coordination is required.

Another novel finding relates to the control variable. We found that group size exerted a positive impact on all constructs, but none of these effects were statistically significant, in contrast to what was reported in prior studies. We believe that this is because the effect of group size on quality was moderated by various (insignificant) effects on inequality and coordination.

Therefore, the results presented here offer a novel and refined perspective on the relationship between inequality, coordination, and Wikipedia quality. In addition to the theoretical contribution, our findings have important implications for practice. We believe that the insights from this study extend beyond the specific context of Wikipedia and are relevant for other types of computer-supported cooperative work. First, managers and administrators of large-scale computer-supported collaborations should strive to achieve a team composition that is unequal, i.e. includes few highly active individuals and many relatively inactive outsiders. This suggests that a 2-tier incentive scheme be devised, as different members are driven by different motivations (as was illustrated by varying activity profiles in discussion forums; [20]). Second, since it is often difficult to attract a large number of outsiders, entry barriers should be kept as low as possible. Third, our results regarding the effects of coordination suggest that mechanisms for enabling coordination for unequal teams be provided. For example, it may be useful to designate highly active users as article-specific moderators, so that while all users will be authorized to contribute, only the moderators will perform higher-level tasks (e.g. forking, conflict resolution).

In conclusion, our results shed new light on the factors contributing to Wikipedia success. However, these results could be biased by articles' topics and the method for estimating quality. Further research is required in order to (a) validate the findings across alternative samples on different topics, (b) test the model when using more robust estimates of article quality, and (c) expand the model to include interaction effects and additional constructs.

REFERENCES

1. Anthony, D., Smith, S. W., & Williamson, T. The quality of open source production: Zealots and good samaritans in the case of Wikipedia. *Rationality and Society*, forthcoming.
2. Arazy O., Morgan W., and Patterson R. Wisdom of the crowds: decentralized knowledge construction in Wikipedia. *Proc. of the 16th Workshop on Information Technologies & Systems (WITS'06)*, December 2006.
3. Benkler, Y. *The wealth of networks*. (2006). Yale University Press.
4. Bunderson, J. S., & Sutcliffe, K. Comparing alternative conceptualizations of functional diversity in management teams: Process and performance effects. *Academy of Management Journal*, 45, (2002), 875-893.
5. Chesney, T. An empirical examination of Wikipedia's credibility. *First Monday*, 11 (2006) (11).
6. Cummings, J. N. Work group, structural diversity, and knowledge sharing in a global organization. *Management Science*, 50 (3), (2004), 352-364.
7. Denning, P., J. Horning, D. Parnas, and L. Weinstein. Inside risks: Wikipedia risks. *Communications of the ACM* 48, 12, (2005), p.152.
8. Dixon, P. M., J. Weiner, T. Mitchell-Olds, R. Woodley. Bootstrapping the Gini coefficient of inequality. *Ecology* 68, 1987, 1548-1551.
9. Giles, G. Internet encyclopaedias go head to head. *Nature*, 438 (2005), 900-901.
10. Halavais A. and Lackaff D. An analysis of topical coverage of Wikipedia. *Journal of Computer-Mediated Communication*, 2008, 429-440.
11. Keck, S., & Tushman, M. Environmental and organizational context and executive team structure. *Academy of Management Journal* 36, (1993), 1314-1344.
12. Kittur, A., Kraut, R. E. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. *Proceedings of CSCW 2008*, ACM Press (2008).
13. Kuk G. Strategic Interaction and knowledge sharing in the KDE Developer Mailing List, *Management Science*, 52, 7, (2006), 1031-1042
14. Lih, A. Wikipedia as participatory journalism: Reliable sources? *Fifth International Symposium on Online Journalism*, (2004), Austin, TX.
15. Nov, O. What motivates Wikipedians. *Communications of the ACM* 50, 11, (2007), 60-64.
16. Press, L. Unpublished Wikipedia web survey results. *Personal communication*. (2006).
17. Stvilia, B., Twidale, M., Smith, L. C., Gasser, L. Information quality work organization in Wikipedia. *JASIST*, 59, 6, (2008), 983-1001.
18. Viegas, F., Wattenberg, M., & Dave, K. Studying cooperation and conflict between editors with history flow visualizations. *Proc. of CHI 2004*, ACM Press (2004).
19. von Krogh, G., & von Hippel, E. The promise of research on open source software. *Management Science*, 52, 7, (2006), 975-983.
20. Welser T., Gleave E., Fisher D., and Smith M. (2007). Visualizing the signatures of social roles in online discussion groups, *Journal of social Structure*, 8(2).
21. Wilkinson, D., & Huberman, B. Assessing the value of cooperation in Wikipedia. *First Monday*, 12, 4, (2007).