

## Network Structure, Efficiency, and Performance in WikiProjects

Anonymous

*Keywords: peer-production, networks, collaboration, wikipedia, social learning*

### Extended Abstract

Large-scale collaboration in the absence of top-down control is notoriously difficult[2], yet the internet has enabled decentralized collaborations at a global scale. Wikipedia is one of the most visible examples, with millions of self-organized editors collaboratively producing a high-quality encyclopedia [3, 8]. A better theoretical understanding of such projects is desirable, in order to identify best-practices and inform the design of new projects. We focus specifically on how the network structure [10] formed by editor interactions within a WikiProject relate to two variables: *efficiency* (how much is accomplished by a unit of work) and *performance* (the best quality typically reached). We confirm behaviors observed in numerical and small-scale lab studies [7, 4, 9, 5], such as an overall trade-off between performance and efficiency. However, we find evidence that some variables correlate positively with both variables.

We focus on WikiProjects on the English-language Wikipedia. WikiProjects are groups of related articles, and form distinct communities, with distinct norms and processes. One of the main roles of a WikiProject is to assess the quality of articles within that project, which we use to quantify article progress. We also rely on designations of featured article (FA) or good article (GA) status, which are decided independently of WikiProject assessments. We combine multiple data sets from the English-language Wikipedia. For information about edit history, we used a publicly-available data set containing metadata about all edits from July 12, 2006 to December 2, 2015. To get the rating history of each article, we used a custom script to scrape the daily logs produced by WP 1.0 Bot for all unique WikiProjects (2279) between May 4, 2006 and December 2, 2015. Finally, we used a publicly-available log of page events (including rename events) to reconstruct the unique identifier for each article title mentioned in the rating history logs.

We define efficiency as the mean number of articles that reach a grade level, divided by the number of edits since the previous grade level. Let  $T(W, G)$  be the set of article assessment transitions from below grade  $G$  to grade  $G$  or higher in project  $W$ . Let  $N(W, G)$  be the number of articles in project  $W$  which ever transition from below grade  $G$  to grade  $G$  (or higher). Let  $r(t)$  be the number of article revisions since its previous grade transition. The grade  $G$  efficiency is:

$$E(W, G) = \left[ \frac{1}{N(W, G)} \sum_{t \in T(W, G)} \frac{r(t)}{g(t)} \right]^{-1}, \quad (1)$$

where  $g(t)$  (the number of grade levels crossed by transition  $t$ ) divides edits evenly when transitions span multiple grade levels. Our performance measure  $P(W)$  is defined simply as the percentage of articles in project  $W$  which have reached good or featured status:

$$P(W) = \frac{f(W) + g(W)}{n(W)}, \quad (2)$$

where  $f(W)$  and  $g(W)$  are the number of featured and good articles respectively, and  $n(W)$  is the total number of articles.

For each WikiProject, we construct and analyze the coeditor network: with editors being connected if they have edited at least one article in common. These relationships represent possible flows of information. We focus specifically on mean degree, degree skewness, and path length. High-degree editors have more collaborators, which can increase diversity and access to information at the possible expense of higher coordination costs [6, 4]. Highly skewed degree distributions can amplify the biases of high-degree editors while reducing the need for explicit coordination [7]. Networks with shorter path lengths allow information to travel more quickly at the possible expense of less localized diversity [9, 1]. The mean path length is computationally intensive, so we employ stratified sampling, choosing sample sizes such that the relative error is less than 10, and exclude the 183 largest projects.

Our main findings (Figure 1, Table 1) are:

- Despite an overall performance/efficiency trade-off ( $r = -0.12$ ,  $p < 0.001$ ), Low-degree coeditor networks tend to have both higher performance and higher efficiency;
- Short path lengths tend to be associated with higher performance, consistent with a conformity-based learning strategy;
- Structural inequality (degree skewness) is associated with lower performance.

Our findings shed light on the importance of network structure for successful collaboration at very large scales. Our empirical results are strictly correlative, not causal, and further work is necessary to determine causality. However, the consistency of our findings with those in numerical simulations and small-scale lab studies helps solidify a picture of the role of network structure in collaboration. The higher efficiency and performance observed for low-degree projects might suggest benefits of collaborations of small closely-knit teams, while the correlations between path length and performance might suggest overlapping team membership in order to connect distant parts of the network. And the correlation of low degree skewness with performance may suggest benefits of decentralized team structure. These findings might be informative for developing collaboration policies, for designing interventions to recommend tasks based on how they will influence network structure, or for creating interventions which seek to encourage the behavioral norms most effective for existing network structure.

## Figures and Tables

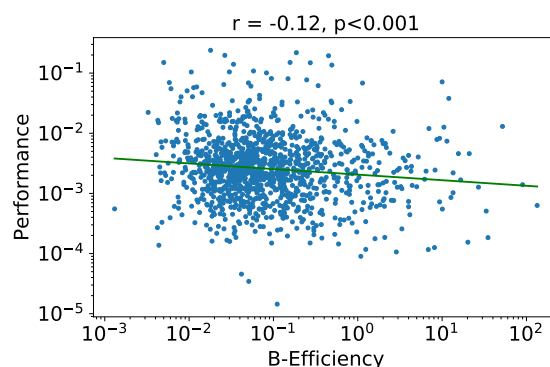


Figure 1: WikiProject performance is anticorrelated with efficiency.

	Perf <sup>†</sup>	A-Eff <sup>†</sup>	B-Eff <sup>†</sup>	C-Eff <sup>†</sup>
Mean degree <sup>†</sup>	-0.7***	-0.8***	-0.6***	-0.3*
Out degree skew <sup>†</sup>	-0.4***	-0.5**	-0.3*	-0.06
Mean path length <sup>†</sup>	-0.33***	-0.09	-0.05	-0.09
C-Efficiency <sup>†</sup>	-0.08*	—	—	—
Connected frac.	0.01	0.09*	0.15***	0.06
Talk fraction <sup>†</sup>	0	-0.02	-0.03	0.01
Mean similarity <sup>†</sup>	0.06**	-0.03	0.01	0.02
Mean editors/art. <sup>†</sup>	0.3**	0.3	0.2*	0.09
Article count <sup>†</sup>	-0.4	0.7*	0.8**	0.7**
Editor count <sup>†</sup>	0.4	0.9**	0.8**	0.5*
Revision count <sup>†</sup>	0.6*	-1**	-1.1***	-1***
First assessment	0.05	0.11**	0.31***	0.43***
Mean article age	-0.03	-0.04	-0.01	-0.05*
N	1179	966	1260	1415
R <sup>2</sup> <sub>adj</sub>	0.37	0.17	0.30	0.43

<sup>†</sup> Log-transformed. \*  $p < 0.05$ . \*\*  $p < 0.01$ . \*\*\*  $p < 0.001$ .

Table 1: Standardized coefficients for OLS models.

## References

- [1] Daniel Barkoczi and Mirta Galesic. Social learning strategies modify the effect of network structure on group performance. *Nature*, 7, 2016.
- [2] Jo Freeman. The tyranny of structurelessness. *Berkeley Journal of Sociology*, 17:151–164, 1972.
- [3] Jim Giles. *Internet encyclopaedias go head to head*. Nature Publishing Group, 2005.
- [4] Benjamin Golub and Matthew O Jackson. Naive learning in social networks and the wisdom of crowds. *Am Econ J-Microecon*, 2(1):112–149, 2010.
- [5] Patrick Grim, Daniel J Singer, Steven Fisher, Aaron Bramson, William J Berger, Christopher Reade, Carissa Flocken, and Adam Sales. scientific networks on data landscapes: question difficulty, epistemic success, and convergence. *Episteme*, 10(04):441–464, 2013.
- [6] Lu Hong and Scott E Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *PNAS*, 101(46):16385–16389, 2004.
- [7] Michael Kearns. Experiments in social computation. *Communications of the ACM*, 55(10):56–67, 2012.
- [8] Brian Keegan and Casey Fiesler. The Evolution and Consequences of Peer Producing Wikipedia’s Rules. In *ICWSM*, 2017.
- [9] Winter A Mason, Andy Jones, and Robert L Goldstone. Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, 137(3):422, 2008.
- [10] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.