Network Structure, Efficiency, and Performance in WikiProjects

Edward L. Platt
University of Michigan
Ann Arbor, Michigan
elplatt@umich.edu

August 7, 2017

Abstract

TODO

1 Introduction

> The problem with Wikipedia is that it only works in practice. In theory, it can never work.
>
> Miikka Ryokas [4]

Wikipedia successful decentralized.
Efficiency Performance
Coeditor network
NK Simulations
Contributions

2 Background and Related Work

TODO

3 WikiProjects

3.1 Data

Our analysis combines multiple datasets from the English-language Wikipedia. For information about edit history, we used a publicly-available dataset containing metadata about all edits between TODO. To get the rating history of each article, we wrote a script to scrape the daily logs produced by WP 1.0 Bot for each WikiProject between TODO. 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.

3.2 Efficiency and Performance

To model the relationship between performance, efficiency, and network structure, we must have a way to quantify performance and efficiency. We define these quantities on a per-WikiProject basis to enable comparisons across different projects. Our definitions are inpsired by work modeling collective problem-solving as an optimization problem [6, 8, 7, 5, 2]. The efficiency quantifies how quikly a solution is reached, while the performance quantifies how good the eventual solution is.

For a WikiProject, efficiency quantifies how quickly project participants can improve the assessed quality of an article. Quality assessments are made through consensus of the project participants themselves, so different projects can have different standards and practices for assessing article quality. So the efficiency is not a measure of how quickly some objective measure of quality improves, but rather of how quickly the project participants can reach consensus on the improvments that need to be made and make those improvements. Because our definition relies on assessment transitions, we define efficiency variables for each of the project-level quality assessments:

A, B, and C. If $T(W, G)$ is the set of transitions in project W from below grade G to grade G (or higher), then we quantify the efficiency $E(W, G)$ as:

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

where $r(t)$ is the number of revisions since the previous grade transition, and $g(t)$ is the number of grade levels crossed by transition t. The $g(t)$ term is added because an assessments often raise article quality by several grades, in which case the revisions are divided evenly between all grade levels achieved. It is also worth noting that we measure efficiency in terms of revisions made, rather than time passed. We focus on revisions because the amount of work done on an article varies widely from day to day.

For performance, we wish to quantify how good articles tend to be when they reach a stable state. Measuring performance is difficult for several reasons: there is no objective meaasure of article quality avaialbe, and articles are always changing, making it difficult to know which articles should be considered complete or stable. We use an extremely simple performance measure that gives surprisingly consistent results. In addition to per-project quality assessment, articles can be given "featured article" or "good article" status. The criteria for these statuses are consistent across all of Wikipedia, and any editor can participate in the discussion and decision to award good or featured status. In other words, the good and featured statuses are more objective than per-project assessments. Our performance measure $P(W)$ is just 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 numbered of featured and good articles respectively, and $n(W)$ is the total number of articles.

3.3 Coeditor Networks

For each WikiProject, we compare the efficiency and performance measures to the structural properties of its coeditor network. The *coeditor network* of a WikiProject consists of nodes representing editors. Two editors are connected when they have both edited the same article or talk page. The edges are directed, with the direction representing the direction of *plausible information flow*; an edge from editor A to editor B exists if A edited an article and then B edited the same article at a later time. Edges can exist in both directions e.g., if an article was edited first by A, then by B, and again by A. For simplicity, we assign all edges unit weight. We focus on three structural properties: degree, characteristic path length, and min-cut.

The node degree distribution is the simplest structural property we analyze for WikiProject coeditor networks. The in-degree (out-degree) of a node is the number of edges to (from) that node. Taking the average of either in-degree or out-degree gives the same value: the *mean degree* of the network. In our context, the mean degree represents how many others a given editor has collaborated with. We also consider the *skewness* of the in-degree and out-degree distributions. A large positive degree skewness value for a WikiProject coeditor network implies that a small number of editors have a very large number of collaborators, while a small positive value implies that the editors having the most collaborators don't have many more than a typical editor.

We also calculate the characteristic path length for each WikiProject coeditor network. The *distance* from editor A to editor B is the length of the shortest path from A to B. The *characteristic path length* is the mean distance between all editor pairs. If no path exists between two editors, we exclude that pair from the mean. For brevity, we will simply refer to this quantity as the *path length*. The path length represents how quickly information can move through the network. Networks with longer paths require more interactions for information to propagate through the network, which has been shown to reduce efficiency in some settings [8, 2].

Our final network measure quantifies the connectivy of a project's coeditor network using min-cut size. The minimum st-cut between nodes s and t is the set of edges that must be removed in order that no path exists from s to t. The minimum cut

(min-cut) of a graph is the smallest minimum st-cut over all node pairs st. The size of the graph min-cut quantifies the connectivity of a graph, but only incorporates information about edges lying on paths crossing the min-cut. Instead, we use the mean size of all minimum st-cuts, which we refer to as the *mean min-cut*. This measure quantifies the number of redundant paths information can take through the network. Networks with higher redundancy are more resilient to errors on one path [1] and allow innovations to propagate through complex contagion, in which innovations are only adopted after multiple exposures thorugh different sources [3].

The mean path and min-cut network measures are computationally intensive, requiring distance and minimum st-cut calculations for each all node pairs. For larger projects, these calculations are impractical. When coeditor networks were large, we employed sampling to determine mean path length and mean min-cut. For mean path length, source nodes were sampled, and path length was calculated to all destination nodes from each of these. For min-cut, node pairs were sampled. In both cases, stratification was used to ensure the same number of nodes were were sampled from each of 12 node degree quantiles. We estimated the error due to sampling by determining true values for a medium-sized project, and calculating error as a function of sample-size. Sample sizes were chosen such that relative error was below 10%. Even with sampling, however, it was impractical to calculate these properties for the largest projects, so we exclude them from the analysis. To control for bias in project size, we include several size-related variables in our models.

3.4 Model

We model performance and efficiency in WikiProjects using ordinary least-squares linear regression. Each WikiProject is taken as a single observation. The models include each project's coeditor network properties as independent variables. We also include several project-level variables to control for confounding factors. The C-efficiency is included in the model for performance to control for the presence of articles that are actively being improved. Projects with lower efficiency will have more works-in-progress and could have an artifically low performance without controlling for efficiency. Some variables are log-transformed to reduce heteroscedasticity.

Our models are summarized in Table 1. Mean min-cut was found to be highly correlated with degree (see Figure 1), so we exclude min-cut to prevent collinearity. The high correlation between mean degree and min-cut implies that in most cases the minimum st-cut is simply the either set of edges from s or the set of edges to t. The rarity of non-trivial min-cuts suggests that WikiProject coeditor networks have very few central bottlenecks and are thus highly decentralized. In-degree and out-degree skewness were also highly correlated, so each dependent variable was modelled twice: once with in-degree skewness and once with out-degree skewness.

The regression results are consistent whether in-degree or out-degree skewness is used, although the out-degree models are slightly more significant when modelling efficiency. We also see that B-efficiency and C-efficiency have very similar models, but that A-efficiency behaves differently in its dependence on degree skewness and connectivity. The different behavior of A-efficiency is likely explained by the observation that the A-Class quality is infrequently used in practice, meaning that the quality level is usually achieved when an article is rated as a good or featured article, which involves a different consensus process than lower ratings.

The negative dependence of performance on C-efficiency suggests there is generally a tradeoff between performance and efficiency. However, low degree is correlated with both higher efficiency and higher performance, suggesting that it is sometimes possible to improve both simultaneously. Much of the existing numerical work on networked social learning focuses on path length rather than degree, so we explore this result further using simulations in the next section.

For path length, we find that longer lengths correspond to lower performance, contrary to the conjecture that longer path lengths allow more exploration [8] but consistent with a conformity-based social learning strategy [2].

We also observe that high degree skewness is corre-

Figure 1: TODO

lated with lower performance and lower A-efficiency, suggesting that articles in projects with decentralized coeditor networks reach featured or good status more efficiently, and reach higher quality ratings in general.

4 Numerical Simulations

Intro

4.1 Learning Strategies

Individual learning
- Social learning and iteration
- Best neighbor
- Conformity
- Consensus

4.2 Network family

TODO
- Base network
- Duplication and rewiring

4.3 Simulation results

TODO

5 Discussion

TODO

6 Conclusion

TODO

7 Acknowledgements

I would like to thank Daniel Romero for valuable guidance and feedback; Danielle Livneh and Karthik Ramanathan for help collecting the data sets; Yan Chen and Tanya Rosenblat for feedback on the methodology; and the atendees of the May 25, 2017 MIT Center for Civic Media lab meeting and the Berkman-Klein Center's Cooperation Working Group for helpful feedback on preliminary results. This research was funded by the University of Michigan School of Information.

References

[1] Albert, R., Jeong, H., and Barabsi, A.-L. Error and attack tolerance of complex networks. *Nature 406*, 6794 (2000), 378–382.

[2] Barkoczi, D., and Galesic, M. Social learning strategies modify the effect of network structure on group performance. *Nature 7* (2016).

[3] Centola, D., and Macy, M. Complex contagions and the weakness of long ties. *American journal of Sociology 113*, 3 (2007), 702–734.

[4] Cohen, N. The Latest on Virginia Tech, from Wikipedia. *New York Times 23* (2007).

[5] Grim, P., Singer, D. J., Fisher, S., Bramson, A., Berger, W. J., Reade, C., Flocken, C., and Sales, A. scientific networks on data landscapes: question difficulty, epistemic success, and convergence. *Episteme 10*, 04 (2013), 441–464.

[6] Lazer, D., and Friedman, A. The network structure of exploration and exploitation. *Administrative Science Quarterly 52*, 4 (2007), 667–694.

[7] Mason, W., and Watts, D. J. Collaborative learning in networks. *Proceedings of the National Academy of Sciences 109*, 3 (2012), 764–769.

[8] Mason, W. A., Jones, A., and Goldstone, R. L. Propagation of innovations in networked groups. *Journal of Experimental Psychology: General 137*, 3 (2008), 422.

	Perf†	Perf†	A-Eff†	A-Eff†	B-Eff†	B-Eff†	C-Eff†	C-Eff†
Mean degree†	-0.84***	-0.78***	-0.61**	-0.84***	-0.50***	-0.61***	-0.34**	-0.31*
In-degree skewness†	-0.58***	—	-0.43**	—	-0.19	—	-0.1	—
Out-degree skewness†	—	-0.48***	—	-0.57***	—	-0.26*	—	-0.066
Mean path†	-0.357***	-0.356***	-0.026	-0.096	-0.020	-0.045	-0.092**	-0.086
C-eff†	-0.083**	-0.085**	—	—	—	—	—	—
Connectivity	0.018	0.016	0.081	0.088	0.126***	0.142***	0.080**	0.081**
Mean editors/article†	0.37***	0.40***	0.22	0.27	0.20	0.22*	0.076	0.075
Article count†	-0.32	-0.27	0.63*	0.70*	0.76**	0.78**	0.67***	0.67***
Editor count†	0.68*	0.48	0.74*	0.97**	0.60*	0.72**	0.52*	0.46*
Revision count†	0.37	0.39	-0.84**	-0.88**	-1.05***	-1.05***	-1.00***	-1.00***
First assessment	0.028	0.058	0.086*	0.101**	0.309***	0.321***	0.456***	0.460***
Mean article age	-0.040	-0.029	-0.049	-0.037	-0.019	-0.014	-0.046*	-0.045*
DoF	1278	1278	1048	1048	1371	1371	1540	1540
R^2_{adj}	0.37	0.37	0.15	0.16	0.30	0.30	0.43	0.43

Table 1: Standardized coefficients for OLS models.
† Log-transformed. * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.