Network Structure, Efficiency, and Performance in WikiProjects

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

August 3, 2017

Abstract

TODO

1 Introduction

> The problem with Wikipedia is that it only works in practice. In theory, it can never work.
>
> Miikka Ryokas Cohen (2007)

Wikipedia successful decentralized.
Efficiency Performance
Coeditor network
NK Simulations
Contributions

2 Background and Related Work

TODO

3 Empirical Methods

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 WPBot1.0 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 . The effieincy 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.

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 and allow innovation to propagate through complex contagion, in which innovations are only adopted after multiple exposures thorugh different sources .

3.4 Model

OLS, controls

3.5 Empirical Results

TODO

4 Numerical Simulation

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

Daniel Romero. Danielle Livneh, Karthik Ramanathan. Yan Chen, Tanya Rosenblat. MIT Center for Civic Media. Cooperation Working Group at the Harvard Berkman-Klein Center. School of Information.

References

Noam Cohen. 2007. The Latest on Virginia Tech, from Wikipedia. *New York Times* 23 (2007).