

Characterizing Contribution Strategies and Quality with Bi-Partite Networks in Wikipedia

order of preference : (1) talk, (2) ignite talk, (3) poster

Maximilian Klein, Thomas Maillart, John Chuang

According to the adage “*given enough eyeballs, all bugs are shallow*” [1], open source projects achieve high quality from the variety of edits performed by many contributors. We aim to test this adage by characterizing the contribution strategies by editors in 12 categories of Wikipedia articles, using networks models with two node types: articles and editors. The *biased random walk* method in bi-partite networks [2] is a reflexive method, in which produced entities are ranked as a function of producing entities, and conversely in a recursive way [3]. This approach is a two node-type version of the *pageRank* algorithm [4] and compares to a random walker jumping between editors and articles (see Figure 1A). Jumps are controlled by parameters α and β , which determine the probability to jump from editors to articles with more links (resp. from articles to editors with more links). When α or β equal zero, this indicates an equal probability to jump to a neighboring node. However, larger α (resp. β) indicates a lower probability to jump to a node with more links. In other words, for a node of a given type, α (or β depending on the node type) controls the importance of the connectivity of neighboring nodes with nodes of same type as the original one. α or β therefore provide direct information on the structure of influence between articles and editors.

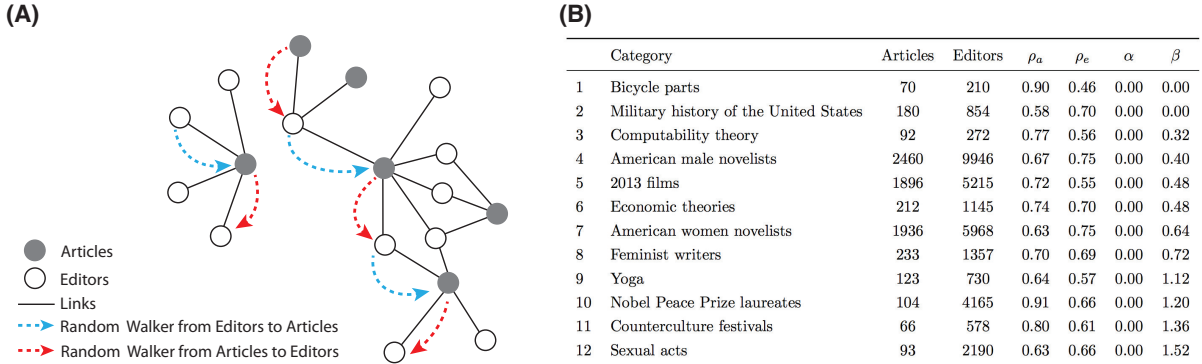


Figure 1: **(A)** Bi-partite network with Article and Editor nodes. Dashed arrows show how the random walker jumping between articles and editors with probability controlled by the appropriately biased connectivity of the each node. **(B)** Table shows the best rank-correlation ρ_a and ρ_e of the algorithm with the ground truth for each Wikipedia category, as well the value of the divestment bias β .

Calibration of α and β by maximizing rank-correlation with ground-truth quality metrics for editors [5] and articles [6] shows that the bi-partite model of influence accounts particularly well for both quality of articles ($0.58 < \rho_a < 0.91$) and expertise of editors ($0.46 < \rho_e < 0.75$). The best rank-correlation is obtained for $\alpha = 0$ and $0 \leq \beta \leq 1.52$ (c.f. Figure 1B). Since $\alpha = 0$, articles always benefit from scrutiny by more editors, hence confirming the famous adage. However, for categories with $\beta > 0$ the quality of an article is also negatively influenced by the portfolio size of its editors. In other words, when β is large, the more articles an editor edits, the lower his or her contributed value to one single article. β also characterizes each category and the way quality is achieved through the cumulative contribution of information. For β small, all editors have a fairly equal chance to contribute positively to any article, while for β large, an editor can only contribute positively to a subset of articles (the larger β the smaller the set). Typically, it is unlikely for a single person to have visited all *counterculture festivals*, performed all *sexual acts* or *Yoga* practices, while it is easier for anyone to gather relevant information on *bicycle parts* or the *US military history*. In conclusion for $\beta > 0$, Raymond’s adage should be appended in the following way : “*given enough eyeballs, all bugs are shallow. But each pair of eyeballs cannot kill all bugs*”.

References

- [1] E. Raymond, *Knowledge, Technology & Policy* **12**, 23 (1999).
- [2] G. Caldarelli, *et al.*, *PloS one* **7** (2012).
- [3] C. A. Hidalgo, R. Hausmann, *Proceedings of the National Academy of Sciences* **106**, 10570 (2009).
- [4] L. Page, S. Brin, R. Motwani, T. Winograd (1999).
- [5] S. R. Geiger, A. Halfaker, *Computer supported cooperative work (CSCW '13)* (2013).
- [6] C. D. Warncke-Wang, M., J. Riedl, *Proceedings of WikiSym '13* (2013).