The Article/Editor Ranking : Bringing Order to Wikipedia

Maximilian Klein
OCLC Research
777 Mariners Island Blvd
San Mateo, CA, 94404
kleinm@oclc.org

Thomas Maillart
School of Information
University of California,
Berkeley, 102 South Hall
Berkeley, CA 94720
thomas.maillart@ischool.berkeley.edu

John Chuang
School of Information
University of California,
Berkeley, 102 South Hall
Berkeley, CA 94720
chuang@ischool.berkeley.edu

ABSTRACT

We introduce a new method to jointly rank the quality of articles and the expertise of editors in Categories of Wikipedia, based on the bi-partite network information of who has contributed at least once to an article on the one hand, and what are the articles that have edited at least once by a given author on the other hand. We show that this "reflexive" ranking method exhibits high correlations with usual article quality and user expertise metrics, which account for quality on Wikipedia (that we assume to be a grand truth here). In particular, we find that the quality of an article can be captured very well by our method right after a few edits, while the expertise of editors is captured increasingly better over time. Our results suggest that it is easier to predict the quality of an article from the editors who touched it, rather than editor expertise from articles they have edited.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

to be completed

Keywords

to be completed, if necessary

1. INTRODUCTION

From product reviews to online collaboration platforms, the World Wide Web is increasingly populated with knowledge contributed by the crowds. One positive aspect is the immediate sharing of information, which in turn helps others take more informed decisions about the quality of a product, the cleanliness of a restaurant, whether it is worth buying a book. But the reliability of the contribution is limited by the level of expertise of the person who made the contribution. In many case, repeated contributions of the same type by several individuals (e.g., reviews) average out idiosyncracies (assuming that individual do not influence each others). Many crowd sourcing mechanisms have been designed in this way []. The reliability/quality of a contribution is however more critical when the same contribution is not necessarily repeated, or repeatable. Open source collaboration projects face such an issue: rewriting several times Wikipedia or the whole open source software codebase would be simply be impossible. Yet these open collaboration projects can achieve remarkable quality [**?**]. This is achieved through *peer-production* a labor organization, which mainly relies on two main ingredients : (i) task self-selection and (ii) peer-review : individuals choose the tasks they believe they are more qualified for, and they review the work of others also according to their skills [**?**]. These rules have been initially used for open source software development [**?**], for which it is possible to ultimately test the quality of the work by executing the code. However, for natural language knowledge, like Wikipedia, there is no "machine" to execute the code. Therefore the quality of knowledge remains subjective, and it is mainly tied to the expertise of the person in the domain of contribution, and to the number of persons (and their respective expertise) who have contributed to a piece of knowledge (e.g., an article). The level of expertise in one domain can in turn be approximated by the number of related articles contributed by the same individual, and so on.

On Wikipedia, is there a way to assess relative quality of both articles and editors by only considering whose editor has touched which article ? This is the problem we are addressing in this paper.

The rest of this paper is organized as follows. With begin with discussion of related work, followed by a description of the method. We then present the results and conclude.

2. RELATED WORK

In this study, we build on a growing stream of literature that aims to see economies a complex network of entities (e.g., firms, countries) gaining a competitive advantage from a set of abilities, which in turn allow them sell products to other entities [**?**]. Capabilities cannot be observed, and the approach assumes that products are a proxy of each entity's capabilities. As result, if they have overlapping sets of capabilities, entities may compete for selling similar products. From an economic perspective, the way entities compete on similar (resp. dissimilar) market segments allows comparing the structure, and arguably, the competitiveness of entities' economy. The relation between products and entities can be modeled by so-called *bi-partite networks* with remarkable properties [**?**] that we leverage in this paper, for the sake of understanding better how content quality and editors' expertise emerge in open collaboration. The core idea is to introduce a reflexive metric, which helps understand the value of entities (i.e., *fitness*) from the products they sell, as well as the *fitness* of a product from the number of entities, which have the capabilities to produce and sell it. This is in fact the first step of an iterative method called *reflexivity*, in which at each step the value of an entity (resp. a product) can be evaluated from the previous fitness (resp. ubiquity) at the previous step. The reflexivity method is explained in much more details in the Method section, for Wikipedia articles and editors. However, in its initial formulation [**?**], the algorithm suffers a number of pitfalls, among which the most important one is its convergence to a fixed point. Indeed, after a sufficient number of iterations, all entities have the same fitness, and all products have the same ubiquity, while the algorithm should on the contrary further discriminate entities and products as the number of iterations increases.

Several alternative methods have been proposed to ensure non-convergence of the algorithm[**?**, **?**, **?**, **?**]. In particular, Caldarelli et al. [**?**] have explained in details the nature of the problem and proposed an alternative method, based on biased Markov chains, which allow get rid of the convergence problem on the on hand, and further understand the nature of the bi-partite network structure on the other hand. This reformulation and extension of the reflexivity algorithm initially proposed is comparable to the pageRank algorithm developed to rank web pages based on the number of time they are linked by other pages [**?**], only that there are two types of nodes (entity, product) instead of only one (webpage).

The problem of ranking entities and their respective production is relevant to the flourishing production of knowledge on the Web, and is directly related to two outstanding problems, which have been previously debated. First, how do we gauge the quality (resp. reliability) of blog posts, book reviews (e.g., on Amazon), or restaurant reviews (e.g., on Yelp) ? Second, how to grant editing and administrative privileges on community networks (e.g., Slashdot) and on online collaborative platforms (e.g. Wikipedia) [**?**].

Some websites (e.g., ebay? Amazon?) allow rate the rater. This approach is similar to the very first steps of the reflexivity method ! [**?**]

more citations needed here

In this paper, we investigate the properties of the biased Markov chains formulation of the reflexivity method, as proposed by Caldarelli et al. [**?**], in the case of online collaboration, namely Wikipedia.

3. METHOD

While it relies on the same principles as in [**?**], our proposed model is conceptually different : our "products" are Wikipedia articles, which are not the basis for competitions but rather for cooperation. Editors enrich the articles together to make best articles. However, like countries, editors have limited capabilities and limited resources (e.g., time), which force them make choices on their contributions.

The matrix $\mathbf{M}$ shown in Figure **??** shows when an article has changed at some point by a given editor. The matrix is ordered on both dimensions by decreasing order of editors who have changed more articles (vertical axis) and by decreasing order of articles that have been changed by most editors (horizontal axis) for a category of Wikipedia articles (here Feminist Writers). Although it is a rough count, the matrix tells already about the experience of an editor in the given category, and the attention an article has gotten from editors, which is an implicit quality measure according to the second principle of peer-production : peer-review. This count is the zero order of the Article/Editor ranking algorithm, and thus the initiation step is given by

$$\begin{cases} d_e^{(0)} = \sum_{a=1}^{N_a} \\ u_a^{(0)} = \sum_{e=1}^{N_{N_e}} \end{cases} \tag{1}$$

Now, let's consider the second step : if an article has been changed by editors who edited more articles, then the quality of the article should be higher. Similarly, if an editor has edited articles that have been edited by more editors, then the expertise of the editor should be higher (the only reason for making this claim is the collaborative nature of Wikipedia, and learning by imitation). Accordingly, the third step is the following : if an article has been changed by editors who edited more articles that have edited by more editors, then the quality of the article should be higher. Similarly, if an editor has edited articles that have been edited by more editors that have edited more articles, then the expertise of the editor should be higher. The algorithm goes on recursively, incorporating the quality (resp. expertise) information of the article (resp. editor) at the previous step.

tell about the stochastic process here [**?**]

say that it is a ranking algorithm. We care only about the ranking of articles and editors

The algorithm at step n then writes,

$$\begin{cases} w_c^{(n+1)}(\alpha,\beta) = \sum_{p=1}^{N_p} \mathbf{G}_{cp}(\beta)\mathbf{w}_p^{(n)}(\alpha,\beta) \\ w_p^{(n+1)}(\alpha,\beta) = \sum_{c=1}^{N_c} \mathbf{G}_{pc}(\beta)\mathbf{w}_c^{(n)}(\alpha,\beta) \end{cases} \tag{2}$$

where the Markov transition matrix $\hat{\mathbf{G}}$ is given by

$$\begin{cases} \mathbf{G}_{cp}(\beta) = \frac{\mathbf{M}_{cp}\mathbf{k}_c^{-\beta}}{\sum_{c'=1}^{N_c} \mathbf{M}_{c'p}\mathbf{k}_{c'}^{-\beta}} \\ \mathbf{G}_{cp}(\beta) = \frac{\mathbf{M}_{cp}\mathbf{k}_c^{-\beta}}{\sum_{c'=1}^{N_c} \mathbf{M}_{c'p}\mathbf{k}_{c'}^{-\beta}} \end{cases} \tag{3}$$

The random walkers starting from editors jump on articles at odd steps, and back on editors from article at even steps. Figure **??** shows how the ranking typically converges over

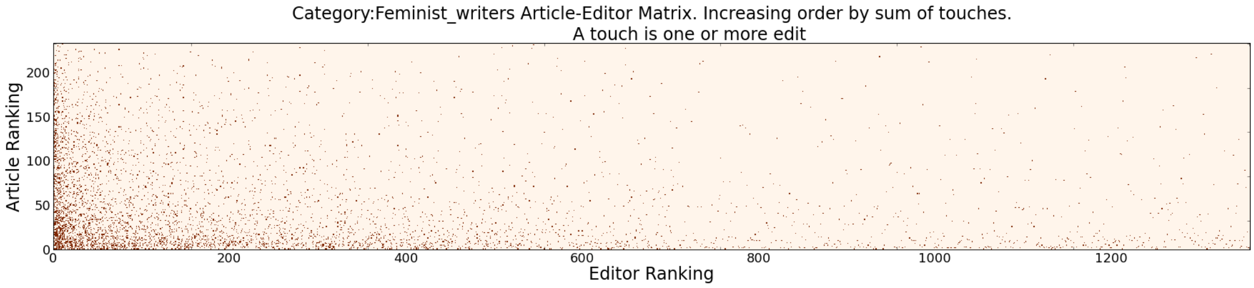

Figure 1: Typical M matrix for a Wikipedia category (here, *Feminist Writers*) ordered on both dimensions by descending order of number of articles modified by an editor (horizontal axis) and of number editors who have modified an article (vertical axis). The structure of M is triangular and shows that some editors have a pervasive activity over articles, while most editors edit only a few. Similarly, some articles receive widespread attention by editors, while most articles are modified only by a few editors.

iterations. As shown on Figure 1, the algorithm converges in a non-trivial way **(explanation for this?)**. In the iterative solution we see how certain editors start low, but then climb in rankings. This means that they are editing few articles, but those articles are of higher quality. Likewise certain articles climb over iterations, they are edited by relatively few editors, but those editors are fitter.

WIth the following balance condition,

$$\mathbf{G}_{pc}\mathbf{w}_c^* = \mathbf{G}_{cp}\mathbf{w}_p^* \tag{4}$$

An analytical solution can be found [**?**]

$$\begin{cases} w_c^* = A(\sum_{p=1}^{N_p} M_{cp} k_p^{-\alpha}) k_c^{-\beta} \\ w_p^* = B(\sum_{c=1}^{N_c} M_{cp} k_c^{-\beta}) k_p^{-\alpha} \end{cases} \tag{5}$$

that we use onwards. Note that in the case $\alpha = \beta = 0$, the analytical solution is nothing else that the zeroth..

4. DATA

The main input of the Article/Editors ranking algorithm is the matrix M which determines, for a category on Wikipedia, which articles have modified at least once by each editor. We collected contribution data for articles in 10 categories (c.f. table 4 for summary statistics on the categories). In addition, we made 10 snapshots of equal number of edits to account for the evolution over time of each category (see Figure **??**).

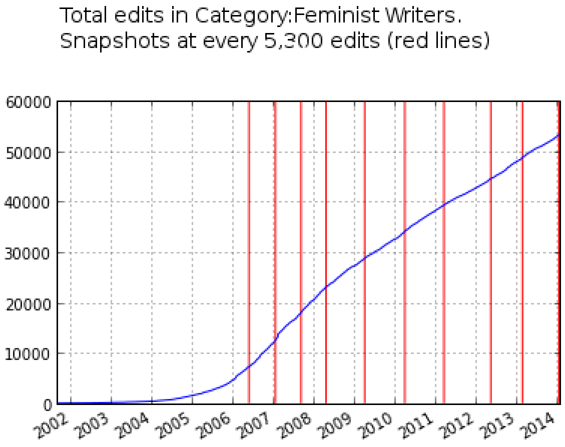

Figure 2: cumulative snapshots Feminist Writers eps

Category	Users	Articles	Edits
2013 films	5215	1896	150956
American male novelists	9946	2460	224783
American women novelists	5968	1936	138716
Bicycle parts	210	70	4981
Computability theory	272	92	7117
Counterculture festivals	578	66	10515
Economic theories	1145	212	28658
Feminist writers	1357	233	25738
Military history of the US	854	180	20172
Nobel Peace Prize laureates	4165	104	91522
Sexual acts	2190	93	45901
Yoga	730	123	25315

For each snapshot, we constructed the matrix $\mathbf{M_{e,a}}$ of contributors versus edited articles, similar to the country versus products matrix of the Economics Domain. For each snapshot, the values in $\mathbf{M_{e,a}}$ are defined as the number of edits made by editor e on article a in the category occurring in the snapshot time. Note that the final snapshot represents the entire history of the category up to the present date.

The matrix $\hat{\mathbf{M}}_{\mathbf{e,a}}$ is a binary representation of $\mathbf{M_{e,a}}$ where each nonzero entry is replaced with 1. This represents if editors have touched which articles rather than how much they have touched each article. In the economics domain, the distinction of making a binary matrix out of the data is interpreted an alternative metric to GDP per capita, rather than GDP. Here we could also see the distinction as normalized editor fitness. It has a typical triangular structure as shown on Figure **??**. The matrix $\hat{\mathbf{M}}_{\mathbf{e,a}}$ constitutes the basic input for implementing the Biased Markov Chain Approach, which we will call the $\mathbf{w}^*$ algorithm, which is an analytic solution to the iterative $\mathbf{W}$ algorithm. [**?**]

The exogenous metric for editors v_e we take is *labourhours*. For each editor the contribution history upto the snapshot point, divide into strings of *editsessions*, edits that occur within 1 hour of the previous edit. Then *labourhours* are determined by subtracting the looking at the total time between the first and last edit in each edit session, and then summing the labours of each edit session. [**?**, **?**].

For an exogenous measure of article quality, v_a, we use a group of 5 text analysis metrics performed on Wikipedia articles at the lastest time in the snapshot. These are ratio of mark-up to readable text, number of headings, article length, citations per article length, and outgoing intrawiki links. To reduce the dimensionality of these 5 metrics, we perform Principal Component Analysis, and accept the principal component. Variance explained by the first principal component, was as high as .7 and never below .5 http://www-users.cs.umn.edu/ morten/publications/wikisym2013-tellmemore.pdf, http://mailer.fsu.edu/ bstvilia/papers/quantWiki.pdf [**?**].

5. RESULTS

5.1 Correlations with exogenous variables

The results of calibrating our model are encouraging as we find high correlations between the results of our w_e^* algorithm and our exogenous variables v.

We define ρ at the maximum achievable spearman rank correlation between w^* and v for editors or articles, by category, and over time. The variation of ρ, by for editor for any category ranges from 0.75 to 0.46 with a mean 0.64. That same statistic for articles is articles from 0.91 to 0.57 with a mean of 0.72, which is overall higher.

From snapshotting we see view ρ as a function of time. In the case of editors we see increasing trends in all categories over time. This means that w_e^* benefits from incorporating more contribution history. As for articles, from the start of a category's history the correlations remain stable. 3

5.2 Negative values of α and β

Another surprising result is that we find at times, negative values for α and β.

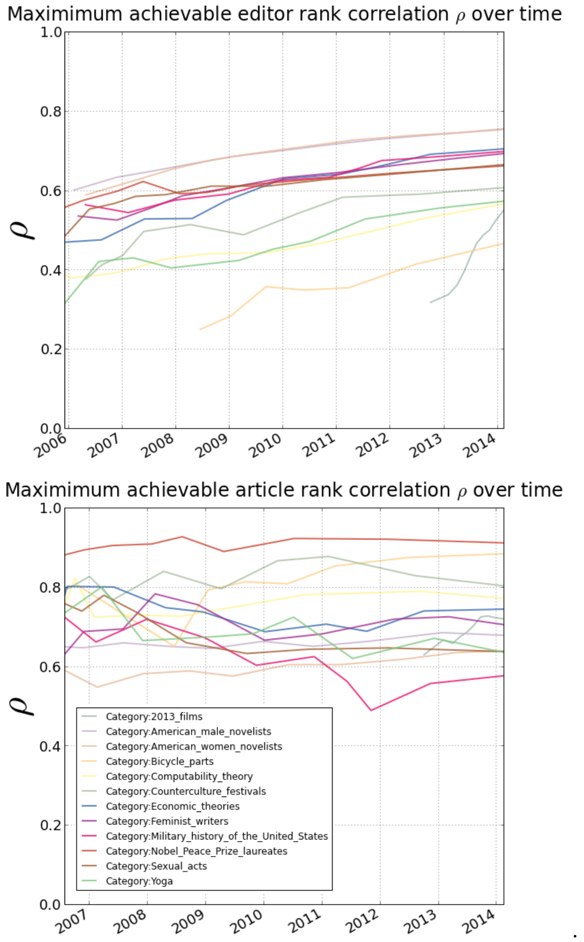

Figure 3: ρ over time, by category and type

For editors, maximum ρ always occurs strictly within a radius of 0.01 around the origin on the α-β plane. The $(0, 0)$ solution is significant in that it represents unbiased arithmetic averages. 4

For articles, the solutions have varied solutions, possibly including a negative alpha. For instance in Category Economic Theories we find a family of maximizing solutions with a negative values of α and β. 4 Why there should be this family is due optimizing solutions is due to a number of reasons. One is our use of the ranking correlation. Since we are comparing only ranks positions, there are many ways to achieve our best-prediction list and perturbing α and β do not affect the rank being produced. This also means that the behaviour displayed by the landscape indicates that one of several possibilities: That either α is positive but small, in which case β is bounded tightly, which means that better editors are producing more obscure articles. Or that α is negative and that β is more loosely bounded. This indicates that α is dominating, as is the case in equations (2). In our particular example we also find that this family is shifted towards the positive β axis. So even as the trade off occurs, its favours β which interpreted through equations (2) means that editors who edit relatively more obscure articles are important to success of articles in Category Economic

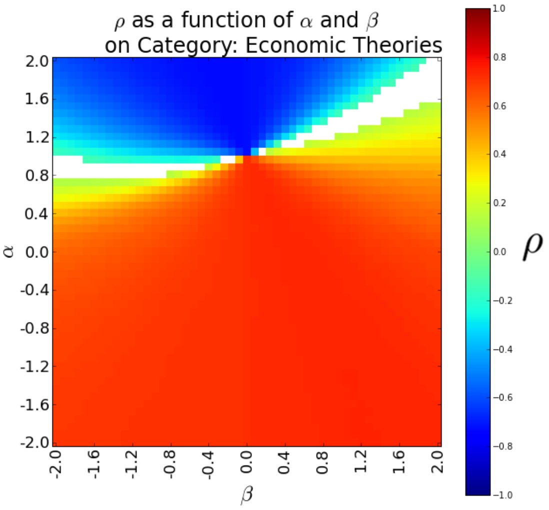

Figure 4: Landscape

Theories.

6. DISCUSSION

One of the brightest results here is that in using the method described for predicting world economies by GDP, we not only can predict Wikipedia editor and article rankings, but outperform the original application. Whereas in Caldarelli the achieve a correlation of 0.4, [?] that is on the lower end of our results. A theory to explain this behaviour goes back to the original motivation in creating this method, that of trying to find "capabilities". In Economics there are circumstantial factors that influence a countries capabilities, e.g. Geography or Politics. However in Wikipedia all outputs are only due to the editors' true underlying capabilities. There are no commodities and articles have no intrinsic value until they are written. If this explained why our correlations were higher, it might also be true for other online collaboration sites, where users operate in a within a greater meritocracy.

In Wikipedia there has been discussion about the importance of super-users who represent a small fraction of editors but contribute a majority of content. [?]. It is possible now to take α as a measure of importance of superuser contributions. Since different categories we correlate more highly for different ranges of α it is possible to compare the success of super-users in different categories. Moreover we can also find which categories are most closely modelled by low or negative values of α which represents better articles are made by less fit editors . Ways to determine this pheonomena could be two or more power users fighting over a page can leave it worse than not being touched at all. Another is that perhaps less fit editors have a greater tendency to collaborate. Or that less fit, "newbie" editors start editing quite obvious, and there for ubiquitous articles, which have high quality already. Whatever the reason, it shows an anti-competitiveness. It means the contributions of less fit authors are important. This is a departure from the economics domain, where the best fits for GDP are only in the positive / positive α-β quadrant.

7. LIMITATIONS

We do not directly measure the *capabilities* of editors. For instance, what does one editor do best to improve an article, among the five metrics (ratio of mark-up to readable text, number of headings, article length, citations per article length, and outgoing intrawiki links) we have used to assess the quality of an article ?

We have not looked at the evolution of the quality of correlation as a function of iteration steps.

"less is more" : Why is it that when we incorporate more information in the matrix M, the results are not better ? Here we took the most simple metric (unlike HH method).

Quality of the exogenous metrics, especially for editors. It would be definitely make to look at real capabilities.

We only look at the ranking not at the real quality/expertise values ? Can we learn more the real values about the gap between articles/editors ?

8. DISCUSSION

One of the brightest results here is that in using the method described for predicting world economies by GDP, we not only can predict Wikipedia editor and article rankings, but outperform the original application. Whereas in Caldarelli the achieve a correlation of 0.4, [?] that is on the lower end of our results. A theory to explain this behaviour goes back to the original motivation in creating this method, that of trying to find "capabilities". In Economics there are circumstantial factors that influence a countries capabilities, e.g. Geography or Politics. However in Wikipedia all outputs are only due to the editors' true underlying capabilities. There are no commodities and articles have no intrinsic value until they are written. If this explained why our correlations were higher, it might also be true for other online collaboration sites, where users operate in a within a greater meritocracy.

In Wikipedia there has been discussion about the importance of super-users who represent a small fraction of editors but contribute a majority of content. [?]. It is possible now to take α as a measure of importance of superuser contributions. Since different categories we correlate more highly for different ranges of α it is possible to compare the success of super-users in different categories. Moreover we can also find which categories are most closely modelled by low or negative values of α which represents better articles are made by less fit editors . Ways to determine this pheonomena could be two or more power users fighting over a page can leave it worse than not being touched at all. Another is that perhaps less fit editors have a greater tendency to collaborate. Or that less fit, "newbie" editors start editing quite obvious, and there for ubiquitous articles, which have high quality already. Whatever the reason, it shows an anti-competitiveness. It means the contributions of less fit authors are important. This is a departure from the economics domain, where the best fits for GDP are only in the positive / positive α-β quadrant.

9. LIMITATIONS

We do not directly measure the *capabilities* of editors. For instance, what does one editor do best to improve an ar-

ticle, among the five metrics (ratio of mark-up to readable text, number of headings, article length, citations per article length, and outgoing intrawiki links) we have used to assess the quality of an article ?

We have not looked at the evolution of the quality of correlation as a function of iteration steps.

“less is more” : Why is it that when we incorporate more information in the matrix M, the results are not better ? Here we took the most simple metric (unlike HH method).

Quality of the exogenous metrics, especially for editors. It would be definitely make to look at real capabilities.

We only look at the ranking not at the real quality/expertise values ? Can we learn more the real values about the gap between articles/editors ?

10. CONCLUSION