Title

Maximilian Klein
confusing@notconfusing.com

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
Abstract

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

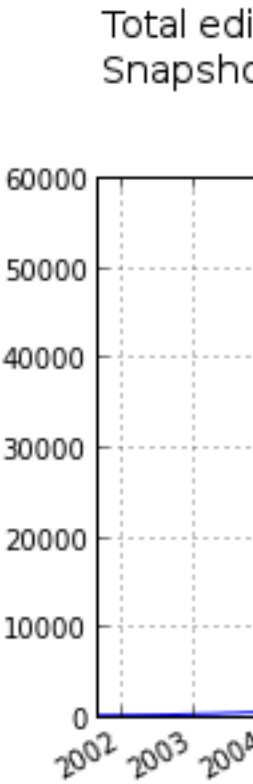
1. INTRODUCTION
3 elements. Literature. Question. Data. The need for a new metric.

2. We also inadvertently develop a new measure for the convtroversialness of articles, and the collaborativeness of a group of editors.

2. METHOD
1. The current investigation involved collecting historical data of edition and quality metrics, from 10 categories of articles in English Wikipedia, with focus on fine-grained edits by contributors to articles.

2. The chosen categories contain between 50 and 4000 articles, and between 50 and 5000 contributors have edited at least 100'000 times all the articles over their history. (c.f. table 1 for summary statistics on the categories).

3. For each category, we constructed 10 accumulative snapshots, that is each starting from the first edit in the category until 10% more of the categories total edits have occurred.



??snapshot points for Feminist Writers.png
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. [**?**]

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.

The output of $\mathbf{w}^*$ are a pair of rankings w_e^* and w_a^* for editors and articles respectively.

$w_e^* = A(\sum_{a=1}^{N_a} M_{e,a} k_a^{-\alpha}) k_e^{-\beta}$

$w_a^* = B(\sum_{e=1}^{N_e} M_{e,a} k_e^{-\beta}) k_a^{-\alpha}]$

Next we collect exogenous metrics as comparison for both w_e^* and w_a^*, which we call v_e and v_a.

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 [**?**].

We then applied the most general implementation of the **FQ** algorithm as developed for modelling the economy and competitiveness of countries. The **FQ** is a nonlinear generalization of the Hidalgo Hausman "Reflections Method". [**?**]. The algorithm has both a stochastic, iterative implementation, and an analytic solution. We demonstrate the iterative solution, to gain some intuition for the algorithm.

$w_c^* = A(\sum_{p=1}^{N_p} M_{cp} k_p^{-\alpha}) k_c^{-\beta}$

$w_p^* = B(\sum_{c=1}^{N_c} M_{cp} k_c^{-\beta}) k_p^{-\alpha}$

Table 1: Summary statistics for each category

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

Figure 1: Matrix M ordered by decreasing order of edits on both contributors and articles dimensions.

At each iterative step we simultaneously rank editor "fitness", and article "ubiquity". In the linear model, the first iteration of "fitness" is the sum of articles to which that editor has contributed, and the "ubiquity" is the sum of editor who have contributed to that article. In the second iteration, say a user is as fit as the average the ubiquities of the of the pages edited. But this is all things being equal.

In the economic domain, the best products are those that are made by the fewest countries. Therefore in our average we want to give more weight to those best producing countries. This measure of good contributors being more important to success, is measured by alpha. A higher alpha means that a good product needs to be exported by the best countries. In Caldarelli, to correlate best with GDP rankings alpha = 1.5 Our result we find the opposite - negative values of alpha. in the not competitive but collaborative wikipedia, where the best articles are produced by the highest number of editors - negative alpha.

2.1 Calibrating

5. While the implementation presented here is strictly similar to **??**, the interpretation is slightly different in the context of group collaboration. Indeed, while countries competes for selling products, the hypothesis here is that Wikipedia contributors cooperate, at least in a very informal way, for improving the quality of articles.

3. RESULTS

4. DISCUSSION

6. *Refer to problems here, if any.*

5. CONCLUSIONS