Assessing value of cooperation in Wikipedia

A STRONG CORRELATION BETWEEN QUALITY AND NUMBER OF EDITS IN WIKIPEDIA ARTICLES

Alessio Devoto
Giacomo Acitelli

What is our goal: show correlation between article quality and number of edits in ca.wiki

How are we going to get there?

1. Step 1. How Wikipedia grows: the more an article has been edited, the more it will be edited in the future. There will be **a few highly edited articles**, and **a sheer number of not popular articles, with a few edits.** This produces a **lognormal distribution** in the number of edits per article.

2. Step 2. Compare number of edits of high quality articles to other articles, showing **a strong correlation between number of edits and quality**. We show that edits correspond on average to an increase in article quality.

Step 1: how Wikipedia grows

The **math model** that shows how likely wiki articles are going to be edited is a **lognormal distribution**.

$$\Delta n(t) = [a + \xi(t)]\, n(t)$$

$$P\big(n(t)\big) = \frac{1}{n\sqrt{2\pi s^2 t}} exp[-\frac{(log(n) - at)^2}{2s^2 t}]$$

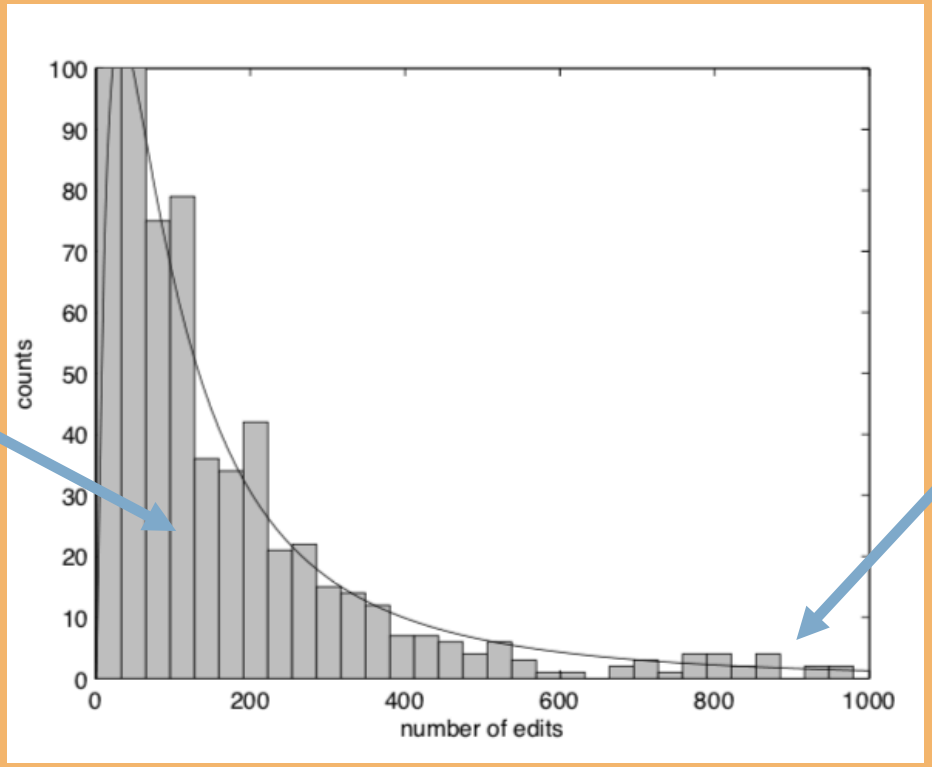
- *n(t)* is the total number of edits up until time *t*
- *a* is a constant that accounts for the rate of accretion
- *ξ(t)* is a random term accounting for flutuations

- $\mu = at$
- $\sigma = s^2t$
- Both are **linearly related** to the age *t of the article*

Step 1: how Wikipedia grows

Lognormal distribution for the number of edits per article for articles of age t = 240 weeks

A lot of articles that are rarely edited.



Few highly edited articles.

Step 2: compare high quality articles with other articles

- But first: What do we mean by "**high quality**"? There have been several attempts to assess the quality of an article in Wikipedia. Here we consider a list of so called "**featured articles**" voted by Wikipedia users as "the best articles in Wikipedia".
- We have to sets of articles to compare: high quality vs others. Though care must be taken in comparing them. The number of edits is influenced by:

1. Visibility ➡ Group article based on PageRank (We used views)
2. Age ➡ Normalize number of edits of article of age t with mean and variance for all articles of that age. Formally:

$$x(A) = \frac{\log n(A) - \mu(\tau)}{\sigma(t)}$$

Datasets

We had to use (and merge) info from two sources:

1. Compressed XML (huge 3 GB) dump that contains **info about pages and all edits history**

```
<page>
<title>Main Page</title>
<id>1269</id>
<revision>
<id>1269</id>
<timestamp>2005-07-07T15:31:37Z</timestamp>
<contributor>      <username>Arde~aawiki</username>      </contributor>
</revision> ….
```

2. Wikipedia's REST api to retrieve info about Pageviews (number of views for each page)

{"project":"ca.wikipedia","article":"Barcelona","granularity":"monthly","timestamp":"2015100100" …

Methods: what did we do?

1. Download list of **featured articles.** (using BeautifoulSoup)
2. Download list of **Wikipedia Bots.** (using BeautifoulSoup)
3. Parse the XML.gzip dump (600 K articles + edits for each article). We had **to filter out edits made by Wikipedia BOTS.** (Took a while)
4. Send HTTP requests to Wiki API to get info about Pageviews. (Took a lot)
5. Join the previously computed dataframes
6. Group them based on number of views and plot results (normalized and non normalized)

name	views
Front Revolucionari Antifeixista i Patriòtic	168.00000
Front d'Alliberament del Ogaden	113.00000
Front d'Alliberament Animal	778.00000
Front Unit de Salvació Democràtica	113.00000
Front Unit Democràtic Popular de Benishangul-Gumaz	181.00000
Front Revolucionari Antifeixista i Patriota	7187.00000
Front Marxista Valencià	702.00000
Front Oriental de la Segona Guerra Mundial	8719.00000
Front Navarrès Independent	414.00000
Front d'Alliberament Nacional de Tripura	336.00000
Front d'Alliberament de les Açores	697.00000
Front d'Alliberament de la Terra	788.00000
Front d'Alliberament Nacional de Jammu i Caixmir	271.00000
Front d'Alliberament Africà del Sudan	215.00000
Front Unit de Moçambic	305.00000
Front Unit Bengalí d'Alliberament	53.00000
Front Republicà	98905.00000
Front Oriental de la II Guerra Mundial	270.00000

DataFrame that holds pageviews for each page retrieved through wiki API

Methods: what did we do?

Unnamed: 0	edits	editors	inception date
Àbac	289.00000	87.00000	2001-03-17
Abadia	61.00000	44.00000	2001-03-17
Adagi	55.00000	34.00000	2001-03-17
Adam	174.00000	71.00000	2001-03-17
Addicció	166.00000	71.00000	2001-03-17
Addicte	4.00000	3.00000	2001-03-17
Astronomia	295.00000	106.00000	2003-10-25
AIX	49.00000	31.00000	2002-01-06
Acampada	61.00000	37.00000	2003-01-20
Alpinisme	138.00000	61.00000	2003-08-31
Aeròbic	100.00000	55.00000	2003-01-20
Aeròbic karate	12.00000	9.00000	2003-02-15
Aikido	242.00000	86.00000	2003-02-15
Aixecament de pesos	4.00000	2.00000	2003-02-15
Atletisme	693.00000	238.00000	2002-02-24
Arquitectura	352.00000	129.00000	2003-09-27
Sometent	120.00000	67.00000	2003-03-03
Arqueologia	187.00000	77.00000	2003-10-20

DataFrame resulting from the parsing of the XML

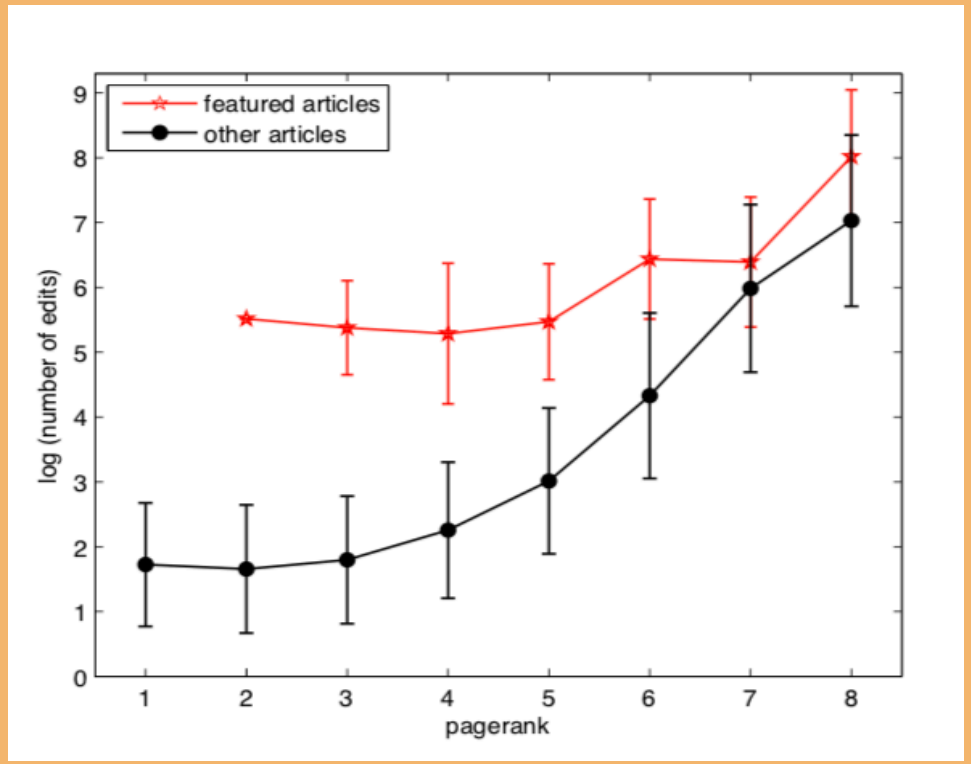
JOIN

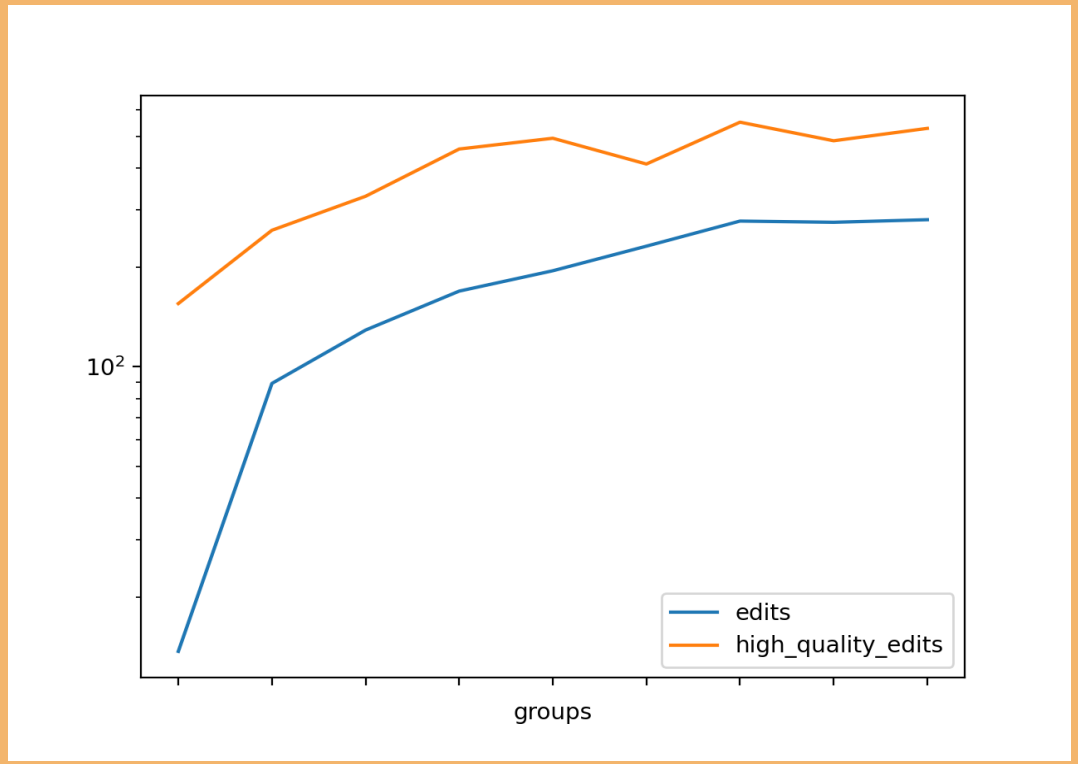
name	views
Front Revolucionari Antifeixista i Patriòtic	168.00000
Front d'Alliberament del Ogaden	113.00000
Front d'Alliberament Animal	778.00000
Front Unit de Salvació Democràtica	113.00000
Front Unit Democràtic Popular de Benishangul-Gumaz	181.00000
Front Revolucionari Antifeixista i Patriota	7187.00000
Front Marxista Valencià	702.00000
Front Oriental de la Segona Guerra Mundial	8719.00000
Front Navarrès Independent	414.00000
Front d'Alliberament Nacional de Tripura	336.00000
Front d'Alliberament de les Açores	697.00000
Front d'Alliberament de la Terra	788.00000
Front d'Alliberament Nacional de Jammu i Caixmir	271.00000
Front d'Alliberament Africà del Sudan	215.00000
Front Unit de Moçambic	305.00000
Front Unit Bengalí d'Alliberament	53.00000
Front Republicà	98905.00000
Front Oriental de la II Guerra Mundial	270.00000

DataFrame that holds pageviews

Results (non normalized): sursprinsinlgy similiar to the paper's results

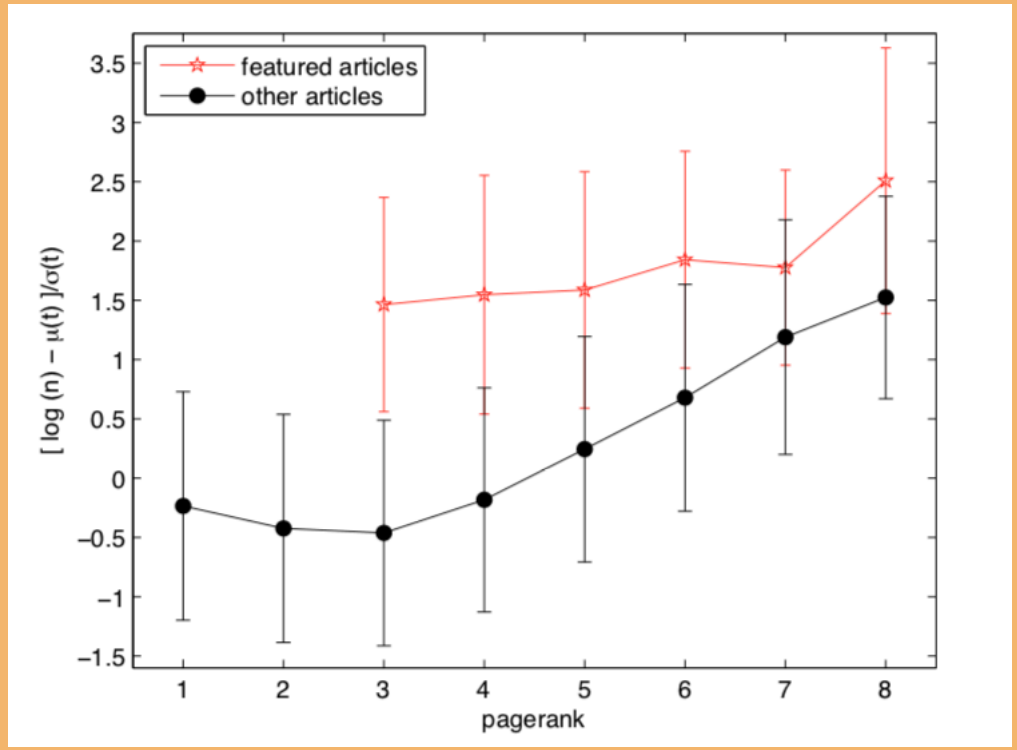
Pictures of plots NON NORMALIZED of quality vs non quality articles : Paper's vs ours

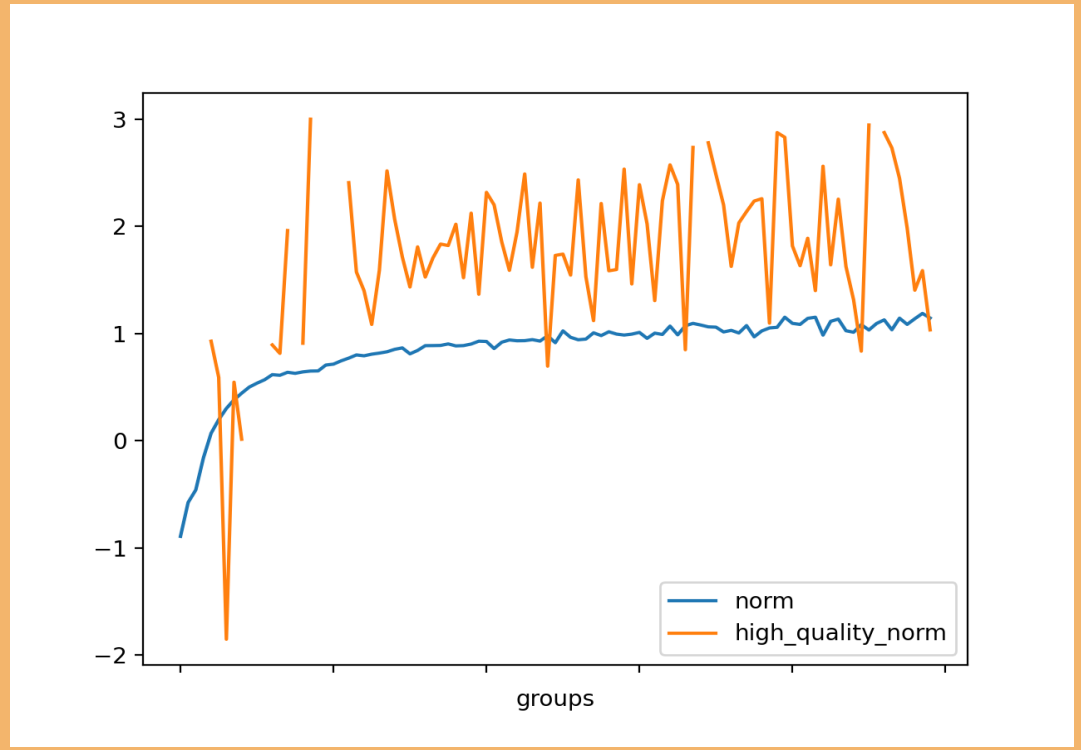




Results (normalized) :

Pictures of plots of NORMALIZED quality vs non quality articles : Paper's vs ours





References:

Assessing the value of cooperation in Wikipedia Dennis M. Wilkinson and Bernardo A. Huberman HP Labs, Palo Alto, CA 94304 February 1, 2008

- T. Chesney. An empirical examination of Wikipedias credibility. First Monday, 11(11), 2006.
- A. Capocci, V. Servidio, F. Colaiori, L. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: the case of Wikipedia. Phys. Rev. E, 74:036116, 2006.
- T. Chesney. An empirical examination of Wikipedias credibility. First Monday, 11(11), 2006.
- A. Lih. Wikipedia as participatory journalism. In Proc. 5th International Symposium on Online Journalism, austin, TX, 2004.