DO WIKI-PAGES HAVE PARENTS? AN ARTICLE-LEVEL INQUIRY INTO WIKIPEDIA'S INEQUALITIES

Abhishek Nagaraj, Priya Seetharaman, Rahul Roy, Amitava Dutta
Indian Institute of Management Calcutta, Indian Institute of Management Calcutta, Indian Institute of Management Calcutta, George Mason University
abhishekn2010@email.iimcal.ac.in, priyas@iimcal.ac.in, rahul@iimcal.ac.in, adutta@gmu.edu

Abstract

We hypothesize that articles on Wikipedia have "parents" who contribute a significant portion of their edits. We establish a notion of inequality based on the Gini Co-efficient for articles on Wikipedia and find support for the existence of this phenomenon of parenting. We base our study on data collected from the Tagalog and Croatian Wikipedias. Ultimately we claim that our research has significant implications for policy for both Corporate Wikis as also for Wikipedia. We state these implications and also suggest directions for future research.

Keywords: Wikipedia, Inequality, Parenting, Knowledge Management, Corporate Wikis

1. Introduction

Wikipedia the online collaborative encyclopedia has captured the attention of not only scholars from a variety of fields but also from mainstream media. One of the fundamental objectives of these investigations has been to determine the reasons for Wikipedia's ability to nearly match other respected publications such as the Encyclopedia Britannica in terms of article quality (Giles 2005). A variety of parameters based on page characteristics have been used to explain differences in article quality. These range from simple parameters like word count (Blumenstock 2008) to more complex models linking article quality to author authority and peer reviews (Hu et al. 2007).

Another important line of investigation has been to look at contributors themselves and explain their behavior. At the very basic level authors have been classified based on simple properties like edit counts and the period for which they have been active. A study by Kittur et al. (2007) for example uses this distinction to examine the changing influence of "elite" and "common" users over time in Wikipedia. An important study in this category has been the one by Anthony et al. (2005) which contends that two types of users contribute significantly to article quality – the "Good Samaritans", one time users who make high quality contributions and the "Zealots", committed users who have been contributing significantly over the past.

The present study lies primarily in that class of papers which tries to identify a particular category of contributors and links them to article quality. We call this category "parents". In the following sections we define what we mean by "parents" and list a few of the characteristics that parents demonstrate. We then describe our methodology and use inequality measures to find support for the phenomenon of "parenting". In the concluding section we make suggestions about the implications of such a finding and directions for future research.

2. Objectives and Hypothesis

In order to look for evidence of parenting in Wikipedia we draw from economics literature to apply the concept of the "Gini Coefficient" introduced by Corrado Gini to measure the inequality of wealth distribution in a population (Gini 1936). We use this parameter to define and measure the inequality in contributions for a particular article. We define the number of contributions made by a particular contributor to a particular article as his "wealth" and the total number of contributions to a particular article as the "total wealth" of that particular article. We apply the Gini Coefficient defined in this manner

to calculate inequality over a particular article and contend that a high degree of inequality for a given article signifies that it is being "parented" by a few users.

To calculate the Gini Coefficient we first plot the Lorenz curve for a particular article. The Lorenz curve is defined as "a graphical representation of the cumulative sum of contributions where we sort contributors on the horizontal axis by their amount of contribution" (Ortega et al. 2008). For a population with zero inequality i.e. one where all contributors have an equal number of edits, the Lorenz curve is a straight line – the diagonal in a unit square with side equal to the total number of contributions. For any other value of user contributions, the Lorenz curve will be convex and will lie below the imaginary line of perfect equality. The area between these two figures is the Gini co-efficient. Thus for a perfectly unequal situation (i.e. where one user makes all the contributions) the Gini co-efficient is one while in the cases of perfect equality it is zero. In other cases it will lie between these two values.

The Gini Coefficient has so far been rarely applied to look at user contributions on Wikipedia. A recent study by Ortega et al (Ortega et al. 2008) investigated over 10 different editions of Wikipedia looking at the inequality in the distribution of the sum total of contributions for each edition. They find that Wikipedia as a whole demonstrates a large degree of inequality which remains stable over time. There is however nothing to be said about the differences in inequality among different articles and the inequality in contributions for a given article, both factors important to make conclusions about the existence of parenting.

The contribution of this paper is to use this inequality effect to look at article level inequality to find support for the phenomenon of parenting.

3. Parenting in Wikipedia

In this section we shall describe our methodology and describe our results. We conducted our studies in two stages. We first used the Tagalog[1] Wikipedia, a small-medium sized Wikipedia for our studies. Once we were reasonably sure of our claims we conducted further analyses on the Croatian Wikipedia, a much larger edition. Our choice of Wikipedia editions was based on a variety of factors including the total number of articles, the total number of edits, the total number of users, the total number of "active users" and the "depth" of the edition. A latest estimate of these figures and their definition can be obtained via the Wikimedia foundation[2].

The entire dump of these versions of Wikipedia was downloaded as on 30th July 2009. The dump is provided by the Wikimedia foundation[3] and it lists each article and the history of edits made by all users to each article along with other details like a timestamp of the edit and the username or the IP address of the user if he is not registered. Once this dump was obtained we cleaned it to remove entries made by bots. Bots are automated programs which troll Wikipedia performing a variety of functions like adding missing reference sections and reverting vandalism. We also deleted non-article pages like discussion pages or categorization pages. Further analyses were performed on these cleaned versions.

The Tagalog Wikipedia contained 29089 unique pages and 13859 unique user ids. While there were ~420k total revisions this number reduced to ~120k after the data was cleaned. This formed the dataset for our analyses. On initial analyses we were able to verify a well known fact about Wikipedia – most users would contribute only one edit. This is shown in Figure 1 where we plot the number of contributors on the Y axis and frequency on the X axis.

[1] Tagalog is a language mainly spoken in the Philippines

[2] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[3] http://download.wikimedia.org/hrwiki

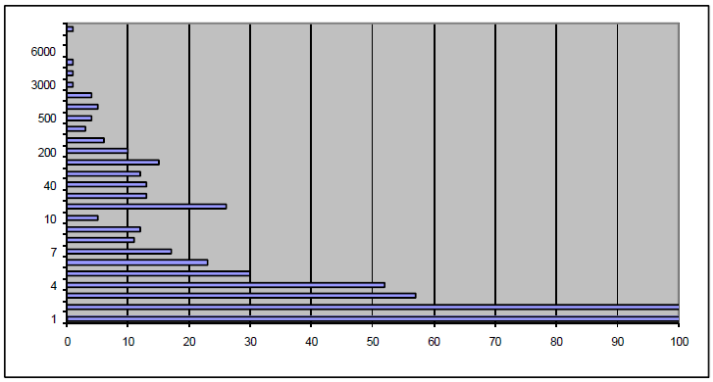

Figure 1. Frequency Distribution of User Contributions

Apart from this, the Tagalog Wikipedia also contains a large number of articles which have only one edit. 80.99% of the articles have two edits or less. This is shown in Figure 2.

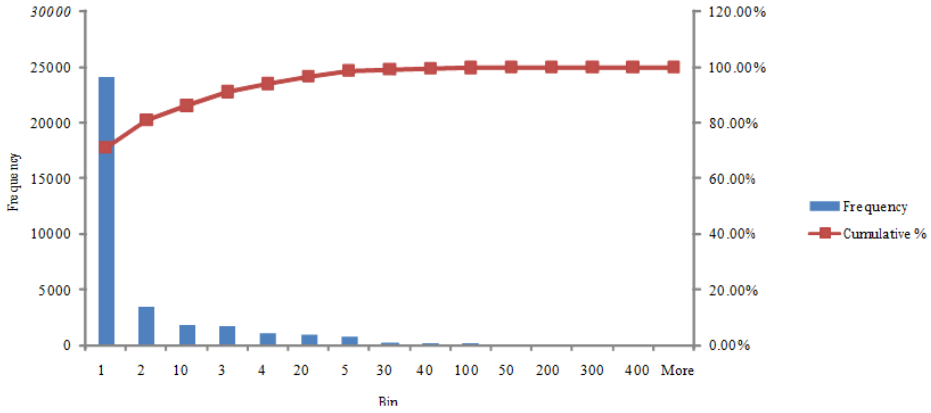

Figure 2. Histogram of Number of Article Edits

Once we had established this overall level of inequality in the Tagalog Wikipedia we then turned to looking for traces of parenting. As an initial investigation we looked at the top 4 articles by edits on the Tagalog Wikipedia. These are as evidenced by their titles the most popular articles on this Wikipedia. Topics like the country of the Wikipedia "Philippines" and its capital "Manila" are bound to attract editor attention. Yet as shown in Table 1 parents are able to capture these pages and contribute to them in a significant way. This gives us a starting point to trace such parenting features in a larger Wikipedia on a more systematic basis.

Table 1. Top Pages by Total Edits in the Tagalog Wikipedia

Rank	Page Title	Parent	Parent Edits	Total Edits	% by Parent
1.	Unaang	Kampfgruppe	288	297	96.97%
2.	Maynila	DragosteaDinTei	165	234	70.51%
3.	talaan mga bansa	AnakngAraw	201	231	87.01%
4.	Pilipinas	Bluemask	79	196	40.31%

In order to not restrict our investigation to finding the "top parent" we turn to the Gini Coefficient as described above. By using "inequality" as a measure of parenting instead of simply the top user by edits as shown above we are now able to capture a variety of cases where a few users might parent a single page or where parenting as a feature is simply absent. We use this newly introduced measure to look at the Croatian Wikipedia. The Croatian Wikipedia after cleaning was found to contain approximately 1.5M user revisions and about 138k unique articles.

We now turn to measuring the inequality of pages in this Wikipedia. Wikipedia defines "Featured Articles" to be the best quality articles on Wikipedia[4]. This classification is based on decisions taken by contributors and must satisfy a stringent list of criteria[5]. We use "Featured Articles" to be a convenient proxy for "high quality articles" and use the list of Featured Articles on the Croatian Wikipedia for further analyses. As of July 30, 2009 there were 223 Featured Articles on the Croatian Wikipedia. We sampled 25 articles out of this list randomly. We also sampled a list of 25 non-featured articles randomly controlling for mean number of article edits in the second case.

Now in order to calculate the Gini Co-efficient for these two sets we used the formula proposed by Angus Deaton (1997) as given in Figure 3.

$$G = \frac{N+1}{N-1} - \frac{2}{N(N-1)u}(\Sigma_{i=1}^{n} P_i X_i)$$

Figure 3. Formula to calculate the Gini Coefficient

In the above formula, G represents the Gini Coefficient for a particular page, N represents the total number of edits for a particular article, u represents the mean number of edits, P_i represents the rank of a particular contributor where Rank 1 is held by the "richest" contributor in terms of edits and X_i represents the total number of edits or the "wealth" of a particular contributor. Using the above formula we are now in a position to calculate the Gini Co-efficient for the two sets of randomly selected 25 articles. These articles have an average of about 183 edits per article. The results are shown in Table 2 and Table 3. Figure 4 shows how a graphical representation of this calculation for the article "Zagreb".

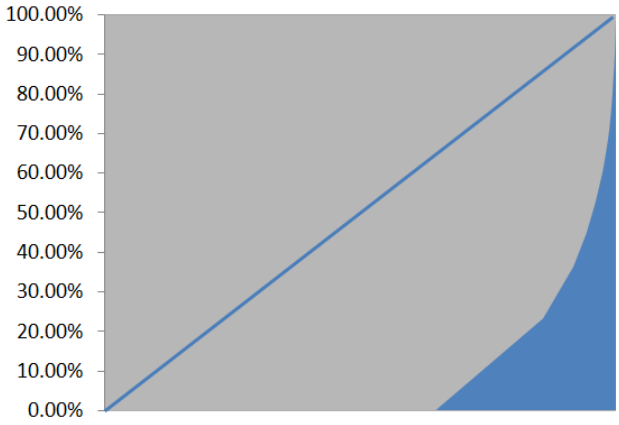

Figure 4. Lorenz Curve for the article "Zagreb". The area between the line and the curve is the Gini Co-efficient, in this case 67.44%

[4] http://en.wikipedia.org/wiki/Featured_Article

[5] http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

Table 2. Featured Articles	
Page Title	**Gini Coefficient**
Arnold Schanberg	72.86%
Asirija	77.01%
Autizam	73.36%
Dioniz	59.13%
Filozofija	68.75%
Francisco Franco	68.08%
Gabriel Garcia Marquez	60.73%
Gospodar prstenova	65.97%
Gospodarstvo Bocvane	85.97%
Harry Potter i Darovi smrti	66.82%
Holokaust	59.58%
Hrvatski jezik	70.27%
Indijanci	88.99%
Jean-Paul Sartre	67.97%
Jezik	65.89%
Nizozemski jezik	66.74%
Nordijska mitologija	81.46%
Odisej	66.85%
Orgazam	72.43%
Sherlock Holmes	60.98%
Staroslavenski jezik	67.65%
Ukrajinci	78.14%
Vikinzi	61.21%
William Shakespeare	60.75%
Zagreb	67.44%

Table 3. Non-Featured Articles	
Page Title	**Gini Coefficient**
Adolf Hitler	68.22%
Borema (nogomet)	74.71%
Bosna i Hercegovina	62.65%
Britney Spears	67.02%
Crna Gora	62.56%
Donji Miholjac	59.07%
Grude	61.56%
HNK Hajduk Split	81.99%
Hrvati	67.63%
Hrvatska nogometna reprezentacija	72.31%
Hrvatska Republika Herceg-Bosna	78.89%
Hrvatski demo sastavi	65.50%
Kosovo	65.28%
Livno	65.80%
Nezavisna Drint	63.75%
NK Dinamo Zagreb	79.10%
Nordijska mitologija	65.43%
Osijek	67.76%
Predlo	74.06%
Rijeka	66.78%
RNK Split	76.08%
Slavonski Brod	65.95%
Slovenija	59.19%
Split	69.47%
Srbi	68.17%
Srbija	62.11%

The average Gini Co-efficient for Featured Articles is 61.44% (std. deviation 7.61%) while that for Non-Featured articles is 68.12% (std. deviation 6.15%). This finding strongly suggests that there is a high degree of inequality in Wiki-pages; that is there is strong evidence for the presence of a small group of users who "parent" articles. For our finding we have used pages of high quality (Featured Articles) and pages with a high number of revisions. These are pages that are in some sense "popular" and would intuitively be the hardest for a particular group of "parents" to dominate. Yet, we see that this is exactly

what happens. We find support for the proposition that even in Wikipedia's most popular parts parenting exists. This is our contribution to the existing understanding of Wikipedia.

4. Implications and suggestions for future research

The major implication of our study is the fact that for well-established mature pages to develop the phenomenon of parenting is inevitable. Pages develop and mature when then find a few people ready to nurture it. This leads to two concrete suggestions, one for the business world and the other for Wikipedia administrators. For corporate wikis there is undoubtedly a case to be made for an explicit allotment of people to pages, hoping that such attention would cause the pages to mature. Wikipedia policy on the other hand should look at easing this process of pages finding parents either via explicitly creating "parent" roles or by "become-a-parent" suggestion boxes based on browsing history.

As for future research, there is obviously the need to further investigate the importance of parenting in featured pages achieving their status. This would strengthen the idea that parenting leads to higher quality. Our study also opens up the possibility of using hybrid parameters to test page quality which include the Gini Coefficient. Another interesting thing to do would be to conduct this study in an orthogonal manner i.e. look at the inequalities among user contributions across pages. That could further strengthen our hypothesis of contributors being parents of some articles and fleeting editors on others. Lastly it would be interesting to test for a notion of "good" and "bad" parenting – parents who nurture pages and moderate discussion as opposed to parents who impose their view on an article. Tying in this perspective of parenting to previous literature like Anthony et al. (2005) could lead to interesting results.

By proposing and finding support for this phenomenon of parenting we have thus opened up an interesting perspective which could lead to interesting results in the future.

References

Anthony D., Smith S. and Williamson, T., "Explaining quality in internet collective goods: Zealots and good samaritans in the case of Wikipedia", November 2005. Retrieved online. *http://web.mit.edu/iandeseminar/Papers/Fall2005/anthony.pdf*.

Blumenstock, J.E. "Size matters: word count as a measure of quality on wikipedia", in *Proceedings of the 17th international conference on World Wide Web*, ACM, New York, USA (Pub.), April 2008, pp. 1095-1096.

Deaton A, "The analysis of household surveys: a microeconometric approach to Development Policy", The John Hopkins University Press, Baltimore, 1997, pp. 137.

Giles, J. "Special Report: Internet encyclopaedias go head to head", *Nature* 438, December 2005, pp. 900-901.

Gini C., "On the Measure of Concentration with Special Reference to Income and Wealth," in *Abstracts of Papers Presented at the Cowles Commission Research Conference on Economics and Statistics*, Colorado College Publications, 1936.

Hu M., Lim E., Sun A., Lauw H.W. and Vuong B., "Measuring article quality in wikipedia: models and evaluation", in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, ACM, New York, USA (Pub.), November 2007, pp. 243-252.

Kittur A., Chi E. H., Pendleton B. A., Suh B., Mytkowicz T., "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisi", in *Proceedings of the 25th Annual ACM Conference on Human Factors in Computing Systems*, ACM, New York, USA (Pub.), April 2007.

Ortega F., Gonzalez-Barahona J.M., Robles G., "On the Inequality of Contributions to Wikipedia", in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, IEEE Computer Society Press, Big Island, Hawaii, January 2008, pp. 308.