{{Citation needed}}:
The dynamics of referencing in Wikipedia

[Short Paper]

Chih-Chun Chen
CAMS
CNRS & EHESS
190 ave de France
Paris, France
c.chen@abmcet.net

Camille Roth
CAMS
CNRS & EHESS
190 ave de France
Paris, France
roth@ehess.fr

ABSTRACT
The extent to which a Wikipedia article refers to external sources to substantiate its content can be seen as a measure of its externally invoked authority. We introduce a protocol for characterising the referencing process in the context of general article editing. With a sample of relatively mature articles, we show that referencing does not occur regularly through an article's lifetime but is associated with periods of more substantial editing, when the article has reached a certain level of maturity (in terms of the number of times it has been revised and its length). References also tend to be contributed by editors who have contributed more frequently and more substantially to an article, suggesting that a subset of more qualified or committed editors may exist for each article.

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

General Terms
Theory

Keywords
Wikipedia, Collaborative systems, Authority

1. INTRODUCTION
The reliability of Wikipedia as an information source has always been a subject of controversy, with some arguing that it is comparable to encyclopedias that centrally monitor and curate their content [5], and others casting doubt on its trustworthiness (see [3] for a more detailed discussion). At the same time, significant effort has been made to understand the processes and mechanisms that underlie

the editing of Wikipedia articles, in particular the conditions under which collective editing is most productive [8], [11] and likely to lead to higher quality articles [13].

Within Wikipedia itself, different measures have been used to compare the quality, reliability, and trustworthiness of articles, editors and edits. For example, one approach is to use the proportion of the edit retained in the article and then to cast editor reputation in terms of the retention rate of his/her edits [3]. Article-centric approaches can be qualitative and based on article content, or quantitative and based on certain article features, such as length or review history [2, 6]. Editor-centric approaches define quality in terms of the composition (e.g. number, diversity) of contributors and/or their reputation [7] (where editor reputation might be defined in terms of network position e.g. [9]). There have also been studies addressing the dynamics of editing and article construction [12, 10]. These analyses have been invaluable in giving us insight into how the range of editing behaviour might give rise to articles that differ vastly in terms of quality.

While it is possible to frame quality and reliability in terms of only article content and consider articles' trustworthiness only in relation to other Wikipedia articles, this neglects the invocation of external sources by articles to substantiate content. The extent to which an article does this can be treated as a measure of its externally invoked authority (an even more intricate measure would also take into account the reliability of the external sources referred to). The goal of our work is to better understand the dynamics of referencing with respect to an article's maturity and its editors.

1.1 Dataset
In this initial study, we extracted a sample of articles from the entire English Wikipedia as of 5th April, 2011. Because the timescales of articles can vary greatly (some may grow quickly and/or be frequently edited, while others may see very slow growth or little editing attention over the same real time period), real time would be unlikely to reflect the maturity of an article. We therefore chose to take number of revisions rather than time stamp as the age indicator. In this initial study, we considered only a small sample of 137,104 Wikipedia articles which were randomly selected from the $\approx$3.6 million articles in the English Language Wikipedia
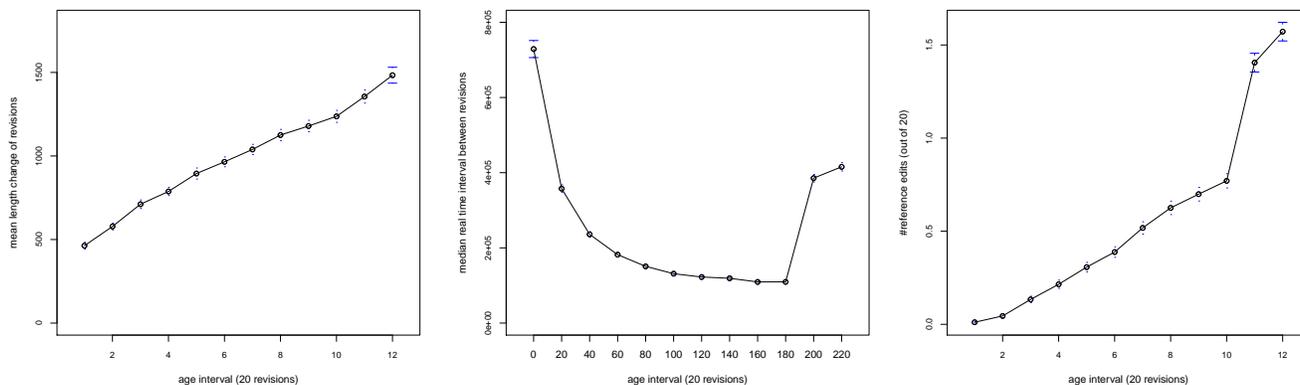
Figure 1: Left: Mean length change with article age; Middle: Inter-revision delay (median real time interval between edits) with article age; Right: Proportion of reference revisions with article age. Article age measured in 20-revision intervals.

and included only the most mature articles (roughly the top decile in terms of number of revisions) with 240 or more revisions with at least one reference, giving us only 5,434 articles (10,930 of the 137,104 articles had 240 or more revisions, but only 5,434 of these had one or more references at the end of the evaluated period).[1] Although we acknowledge that this is only a small fraction of the entirety of Wikipedia, the main contribution of this paper is not to provide an empirically exhaustive analysis of referencing in Wikipedia but to demonstrate a protocol for characterising the referencing process. Nevertheless, we expect the main premises of our findings to generalise.[2]

2. REFERENCING AND ARTICLE MATURITY

The editing dynamics of Wikipedia articles can differ enormously; some articles may have intensive periods of high activity but remain largely untouched outside of these (e.g. event-based or media-related articles) while others may grow with more regularly distributed contributions. Considering this, we take the number of revisions as a proxy for article age, the real time between revisions as a measure of activity intensity, and the length of a revision as a measure of its substantialness. The findings reported here refer to the states of articles at the same age, 240 revisions old. We did not control for reverts since we found their frequency to be negligible in our dataset (this may have been due to the fact that our sample did not include many controversial articles).

2.1 Article growth and referencing

The growth rate of articles appears to change through time. Initially, revisions are fairly insubstantial, with small changes in article length (see Figure 1: Left). They are also sparse in time, with large intervals between each edit (see Figure

---

[1]References were extracted using the "<ref>" tag (depending on the article format, these might be represented to the user in different ways in different articles, e.g. as footnotes or in a References section at the end of the article).

[2]Since we only included articles that had reached a certain level of maturity, we did not consider the effect that referencing (or lack of referencing) might have on the survival rate of an article at different stages.
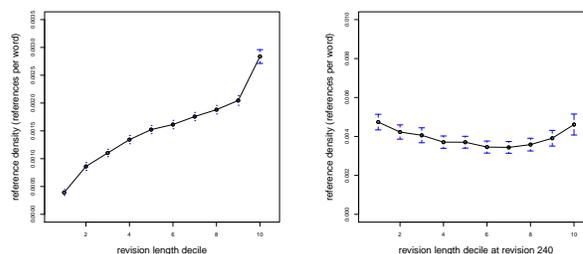
Figure 2: Left: Mean plot of page reference densities and article length for all revision intervals; Right: Mean plot of page reference densities and article length at 240 revisions.

1: Middle). During this period, there appears to be very little referencing. After a certain number of revisions, the article goes through a period of higher activity when edits happen at greater frequency (more revisions within a given time) and referencing starts to happen with more regularity (see Figure 1: Right). After the highly active period, there is a phase of more lengthy edits and referencing.

2.2 Reference density and length

The reference density (number of references per unit length) of an article is a measure of the degree to which external sources are used to support and substantiate the article's content. Not taking into account the number of revisions that articles have gone through, longer articles tend to have a higher reference density (see Figure 2: Left). This seems to suggest a state of article maturity in which content becomes better substantiated. However, the relationship between reference density and length is not straightforward. For any given revision period, length on its own does not appear predictive of reference density: see e.g. Figure 2 (Right) where reference density is plotted against article length at an age of 240 revisions, without demonstrating any clear correlation. This suggests some non-trivial interaction between length, age (in terms of number of revisions) and reference density. The section that follows considers the dynamics of reference editing to try to identify some of the underlying processes.

3. UNDERLYING PROCESSES

Figure 1 (Right) suggests that referencing only starts after a critical number of revisions, and Figure 2 (Left) suggests that this might be due to the fact that referencing only occurs when articles reach a critical length. However, for any given article, it still remains an open question what initiates this process of referencing and what leads to subsequent referencing.

In the previous section, we already saw that longer articles tend to have higher reference densities and that periods during which more substantial edits are made tend to be better referenced. This may be due to the fact that as articles become longer and more substantial (containing more assertions or 'points'), they require more external substantiation.

We find support for two underlying processes (not mutually exclusive):

1. Substantiation of articles reinforces itself so that better referenced edits provoke further better referenced edits. In this case, there should be an auto-correlation for reference density of contributions throughout the lifetime of the article.

2. Referencing occurs when a set of committed and qualified editors are attracted to the article and start to make more substantial, referenced edits, in which case we would expect editors making reference edits to also make longer edits.

3.1 Referencing as substantiation

Figure 3 suggests that periods in which the article grows in more substantial chunks (with lengthier revisions) are also those in which referencing is more frequent, i.e. where the reference density of revisions (number of revisions per unit of length change) is higher.

We also found an auto-correlation for the reference density of revisions. To illustrate, the heatmap in Figure 4 (Left) shows the correlations between the reference density of re-
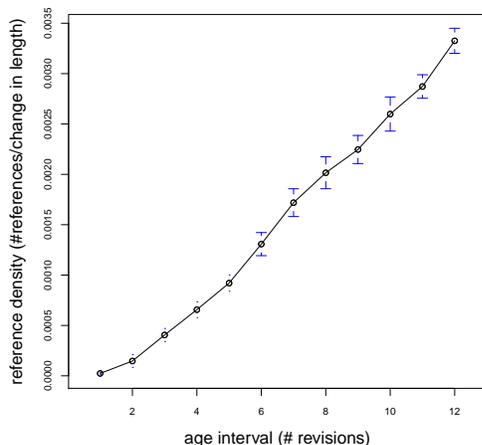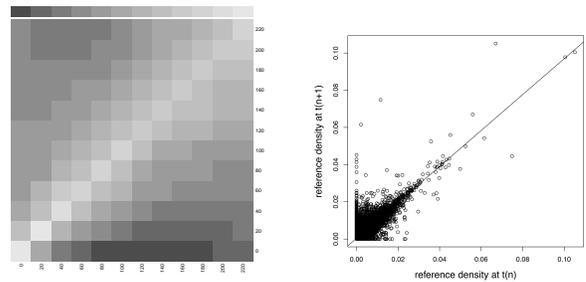
Figure 4: Left: Heatmap of the correlations between the reference densities of articles at different ages (20-revision intervals). Right: Scatter plot of reference densities at revision interval t and reference densities at revision interval $t + 20$

visions (number of revisions per unit of length change) of different revision intervals with the reference density of the final revision interval. The lighter regions indicate increasing Pearson correlation coefficients values towards cells corresponding to successive revision intervals $(t, t+20)$. Figure 4 (Right) plots the reference density at t revisions against the reference density at $t + 20$ revision, with a Pearson correlation of 0.877.

3.2 Editors contributing references

One means by which reference density becomes dependent on article maturity is that once an article has reached a certain state, it begins to inspire more more conscientious editing behaviour by a set of qualified editors or attract more 'serious' editors [3]. Figure 5 shows that editors who add references tend to be those who edit both more substantially and more frequently and also that those who have contributed more than 2 references edit more (both in terms of contribution length and in terms of frequency) than those who only contribute one or two references. Analyses of variance confirmed these differences were significant at $p = 0.001$ between reference contributors and non-contributors, and between reference contributors contributing 2 or more references and those contributing fewer than 2 references (in terms of both length and frequency).

It should be emphasised that although those who contribute references tend to be those who contribute more often and longer revisions, this does not necessarily say anything about the quality of their contributions (e.g. [1], who found that shorter contributions tend to have higher retention). Rather, we can say that editors who bother to substantiate their contributions by invoking the authority of external sources are those whose contributions tend to be longer and more frequent with respect to that article. It is also an open question whether editors contributing references tend to do so in support of their own content contributions and assertions, or whether they tend to add references in response to dissent by other editors (e.g. in response to debate to reinforce their position or to address requests for substantiation).

Figure 3: Reference density of revisions with article age (20-revision intervals)

---

[3] Although bots also play a role in the editing process, the proportion of edits they were responsible for in our sample was less than 1%. We therefore retained their edits in the analyses. It is also not clear that removing bot edits from the analyses would be valid since it might obscure the responses of human editors to bot edits.
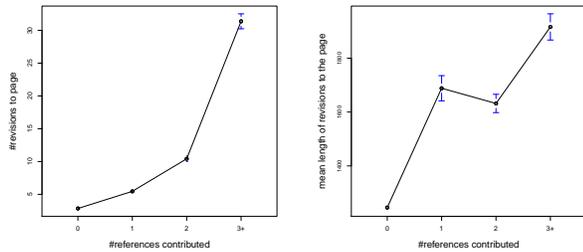
Figure 5: Meanplots comparing editors who have made 0, 1, 2, and > 2 reference edits in terms of (Left) median length of contributions (with respect to an article) and (Right) number of contributions to the article.

4. CONCLUSIONS AND FURTHER WORK

In this paper, we introduced a protocol for characterising the referencing process in the context of general Wikipedia article editing. Although we applied this to only a small sample of articles, we expect the findings to generalise. Our findings suggest that support and substantiation of articles by referencing external sources only occurs after articles have reached a certain level of maturity. Furthermore, referencing tends to occur during periods in which edits are more substantial. We also found that the reference density of edits is auto-correlated and that editors who contribute references are those who have contributed more frequently and more substantially to the article.

Further work. As already mentioned, a limitation of this preliminary study is that it only relates to a subset of Wikipedia articles, and up to a certain level of article maturity. For a more comprehensive analysis, our first goal would be to determine how well the findings generalise to other articles, whether there are differences between subgroups of articles (e.g. featured vs. non-featured, different topics), and what happens later on in articles' lifetimes. Other extensions relate to the two key findings on auto-correlation of reference density and the editor characteristics.

Although we found reference density to be auto-correlated, there still remain many questions about how/if it changes through the lifetime of an article (whether it remains constant or whether it is self-amplifying) and, related to this, what the precise relationship between assertion and substantiation is. Does reference density increase linearly with number of assertions, or do both increase with age and/or length (e.g. later on in the article's lifetime, there may be dissent against particular, more succinct points, but for which substantiation would still be sought)? Addressing this would require us to analyse the content of articles.

Similarly, although we found that editors who contribute references tend to be those who contribute more frequently and more substantially, we did not identify the motivations or individual-level mechanisms underlying these contributions. Do editors tend to add references to support their own content contributions, or do they only do so in response to dissent or requests (either explicitly with the "{{*Citation*

needed}}" tag, or implicitly in another editor's edit)? To better understand the triggers for referencing, it may be necessary to consider the article Talk pages or to combine our analyses with survey/interview data and previous findings on editor motivation (e.g. [4]).

5. ACKNOWLEDGEMENTS

6. REFERENCES

[1] D. Anthony, S. Smith, and T. Williamson. Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of Wikipedia. Electronically, Fall 2005.

[2] J. E. Blumenstock. Size matters: word count as a measure of quality on wikipedia. In *Proc. 17th Intl. Conf. on World Wide Web WWW '08*, pages 1095–1096, New York, NY, USA, 2008. ACM.

[3] T. Chesney. An empirical examination of Wikipedia's credibility. *First Monday*, 11(11), 2006.

[4] A. Forte and A. Bruckman. Why do people write for wikipedia? Incentives to contribute to open-content publishing. In *GROUP 05 Workshop on sustaining community: The role and design of incentive mechanisms in online systems*, 2005.

[5] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, Dec. 2005.

[6] M. Hu, E. P. Lim, A. Sun, H. W. Lauw, and B. Q. Vuong. Measuring article quality in wikipedia: models and evaluation. In *Proc. 16th ACM Conf. on information and knowledge management*, CIKM '07, pages 243–252, New York, NY, USA, 2007. ACM.

[7] S. Javanmardi and C. Lopes. Statistical Measure of Quality in Wikipedia. In *1st Workshop on Social Media Analytics (SOMA '10)*, July 2010.

[8] A. Kittur and R. E. Kraut. Beyond Wikipedia: coordination and conflict in online production groups. In *Proc. 2010 ACM Conf. on Computer supported cooperative work*, CSCW '10, pages 215–224, New York, NY, USA, 2010. ACM.

[9] N. Korfiatis, M. Poulos, and G. Bokos. Evaluating authoritative sources using social networks: An insight from wikipedia. *Online Information Review*, 30(3):252–262, 2006.

[10] D. Laniado and R. Tasso. Co-authorship 2.0: patterns of collaboration in Wikipedia. In *Proc. 22nd ACM Conf. on Hypertext and hypermedia*, HT '11, pages 201–210, New York, NY, USA, 2011. ACM.

[11] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Information quality work organization in Wikipedia. *JASIST (J. Am. Soc. Inf. Sci. Tech.)*, 2008.

[12] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with *history flow* visualizations. In *Proc. SIGCHI Conf. on Human factors in computing systems*, CHI '04, pages 575–582, New York, NY, USA, 2004. ACM.

[13] D. M. Wilkinson and B. A. Huberman. Cooperation and quality in wikipedia. In *Proc. WikiSym 2007 Intl. Symp. on Wikis*, WikiSym '07, pages 157–164, New York, NY, USA, 2007. ACM.