How Long Do Wikipedia Editors Keep Active?

Dell Zhang
DCSIS
Birkbeck, University of London
London WC1E 7HX, UK
dell.z@ieee.org

Karl Prior
DCSIS
Birkbeck, University of London
London WC1E 7HX, UK
kprior01@dcs.bbk.ac.uk

Mark Levene
DCSIS
Birkbeck, University of London
London WC1E 7HX, UK
mark@dcs.bbk.ac.uk

ABSTRACT

In this paper, we use the technique of survival analysis to investigate how long Wikipedia editors remain active in editing. Our results show that although the survival function of occasional editors roughly follows a lognormal distribution, the survival function of customary editors can be better described by a Weibull distribution (with the median lifetime of about 53 days). Furthermore, for customary editors, there are two critical phases (0-2 weeks and 8-20 weeks) when the hazard rate of becoming inactive increases. Finally, customary editors who are more active in editing are likely to keep active in editing for longer time.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Survival Analysis; H.1.2 [**Models and Principles**]: User/Machine Systems—*human factors*; H.2.8 [**Database Management**]: Database Applications—*data mining*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*user issues*

General Terms

Human Factors, Measurement

Keywords

Social Media, User Modelling, Behaviour Mining, Survival Analysis.

1. INTRODUCTION

It has recently been observed that the growth of Wikipedia has slowed down significantly. In particular, Wikimedia Foundation (WMF) has reported that[1]: "Between 2005 and 2007, newbies started having real trouble successfully joining the Wikimedia community. Before 2005 in the English Wikipedia, nearly 40% of new editors would still be active a year after their first edit. After 2007, only about 12-15% of new editors were still active a year after their first edit."

In this paper, we focus on investigating how long Wikipedia editors remain active in editing through *survival analysis* [5, 6] — a branch of statistics which is widely applied to modelling death in biological organisms and failure in mechanical systems — in order to obtain useful insights into Wikipedia's sustainable growth. Making an analogy to the modelling of people's expected lifetime, an editor is considered to be "born" when he starts editing (i.e., joins the community) and to "dies" when he stops editing (i.e., leaves the community). Specifically, we consider an editor to be "dead" or inactive if he did not make any edit for a certain period of time. Here we set the threshold of inactivity to be 5 months, since it reflects WMF's concern as demonstrated in the recent Wikipedia Participation Challenge[2].

The rest of this paper is organised as follows. In Section 2, we review the related work. In Section 3, we present and discuss the results of our analysis. In Section 4, we make conclusions.

2. RELATED WORK

The global slowdown of Wikipedia's growth rate (both in the number of editors and the number of edits per month) has been investigated by Suh et al. [12]. It is found that medium-frequency editors now cover a lower percentage of the total population while high-frequency editors continue to increase the number of their edits. Moreover, there are increased patterns of conflict and dominance (e.g., greater resistance to new edits in particular those from occasional editors), which may be the consequence of the increasingly limited opportunities in making novel contributions. An ecology inspired population model that assumes a resource limitation has been proposed to characterise the overall growth of Wikipedia. In this paper, we approach the problem from a different angle and arrive at conclusions which complement theirs.

The significant differences in Wikipedia editors' predispositions and patterns of contribution have been observed by researchers before [1, 9, 11]. In this paper, we also notice such a phenomenon and report new discoveries about the contrast between Wikipedia editors with different editing frequencies.

[1] `http://strategy.wikimedia.org/wiki/March_2011_Update`

[2] `http://www.kaggle.com/c/wikichallenge`

WikiSym'12, Aug 27–29, 2012, Linz, Austria.

The technique of survival analysis [5, 6] has been shown to be very useful in analysing information systems. For example, the estimated lifetime of a webpage could reflect its desirability [10]. It has recently been applied to a couple of studies on Wikipedia editors' behaviour. In one work [7, 8], the survival function for all Wikipedia editors is empirically estimated, but no parametric model has been produced. In another work [2], the survival function for all Wikipedia editors is fit by a mixture of two truncated lognormal distributions, but our work has revealed that the survival function for customary editors is better described by a Weibull distribution. Furthermore, those previous studies have not looked into the hazard function for Wikipedia editors.

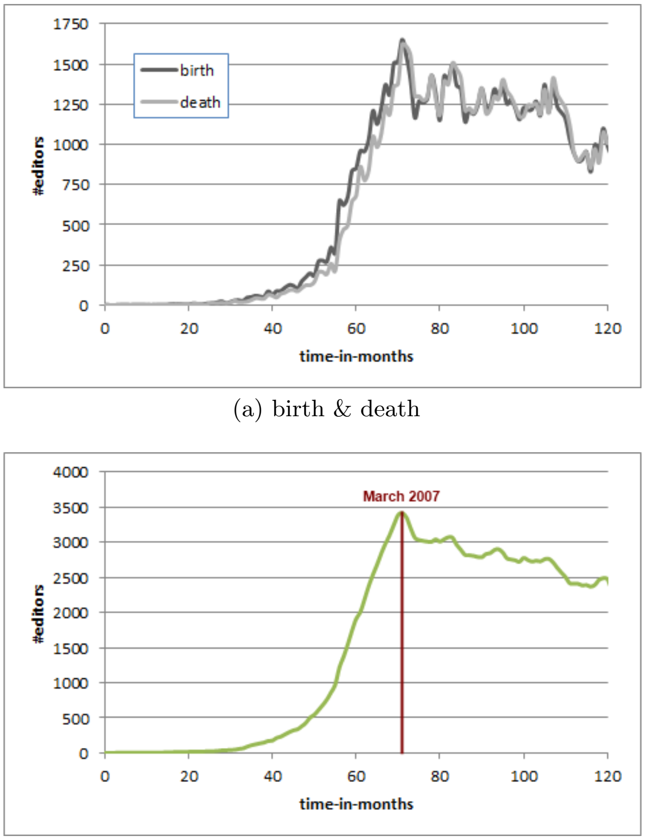
3. OUR RESULTS

The dataset for this study consists of 110,383 registered editors of (English) Wikipedia that were randomly sampled. The bots (i.e., automatic agents that do maintenance task on Wikipedia) were excluded from data collection. Moreover, we removed 38,348 one-timers who had only made one edit so far, because their contribution and influence to Wikipedia are known to be negligible [11, 12]. Finally the complete edit history of those remaining 72,035 editors were processed to extract a population of 86,468 "lives". If an editor started editing again after being "dead" (inactive), it would be considered as a new life instance, because in this study we are more interested in the continuous active period of an editor rather than his overall time in the community, and the same editor could exhibit quite different behaviour patterns when he came back after a very long break (e.g., due to the change of motivation). One prominent characteristic of lifetime data is that many samples may be *censored* [5, 6] — they were still alive at the end of data collection therefore their lifetime values are only known to be longer than a certain duration. Such a censoring problem requires special treatment in probability estimation etc. when performing data analysis.

The evolution of Wikipedia editors' community along with time is shown in Figure 1. It can be seen that although there were always many new editors joining the community, there were more editors leaving the community since the 71st month (March 2007), thus the accumulated number of active editors reached the peak at that time and then continued to decrease.

The histogram of Wikipedia editors' lifetime is shown in Figure 2, where the lifetime is measured in days and scaled logarithmically (using natural logarithm). The lifetime distribution clearly consists of two distinct regimes separated roughly at the point of 8 hours: the left regime corresponds to occasional editors who fail to find interest in editing Wikipedia articles after the first few attempts; the right regime corresponds to customary editors who stay in the community editing Wikipedia articles until they lose interest because of some reason. Let's focus on analysing the behaviour patterns of customary editors, as it is them who constitute the backbone of the community.

The objects of primary interest in survival analysis — *survival function* [5, 6] and *hazard function* [5, 6] — for customary editors are presented in Figure 3a and Figure 3b respectively.



(a) birth & death

(b) active editors

Figure 1: The evolution of Wikipedia editors' community.

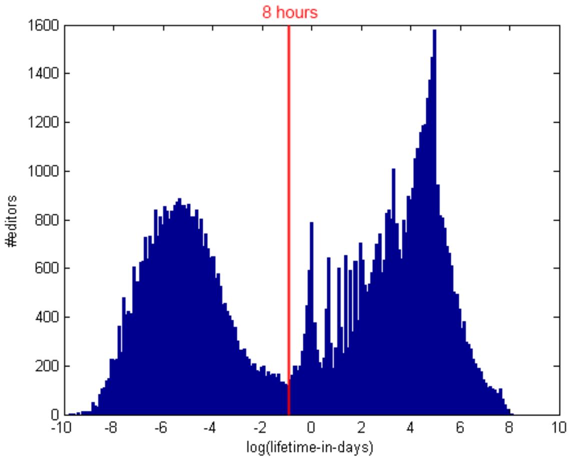
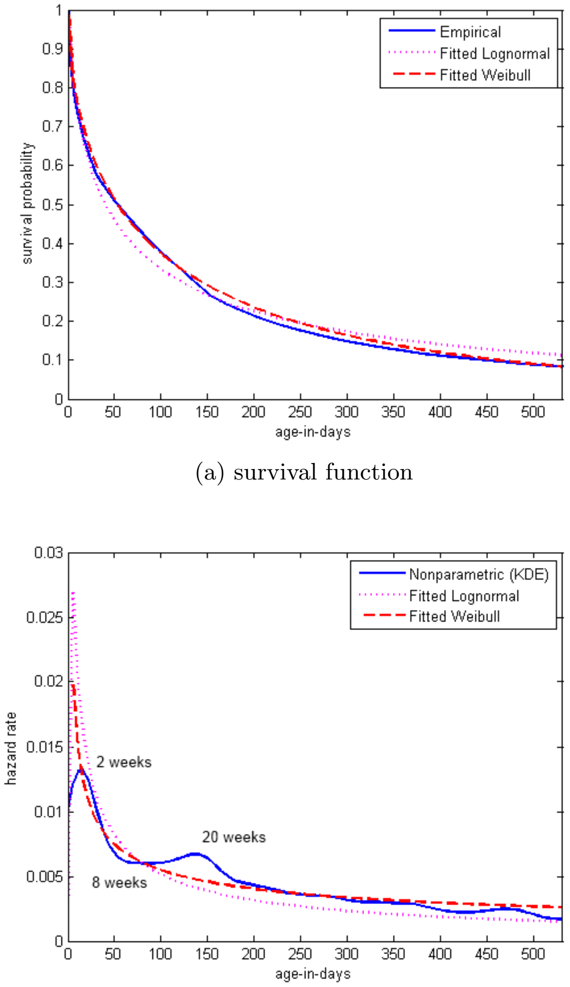


Figure 2: The histogram of Wikipedia editors' lifetime.

The survival function, conventionally denoted S, is the probability that the time of death T is later than some specified time t:

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(t)\,\mathrm{d}t\ . \qquad (1)$$

The empirical survival function for customary editors is calculated using the *Kaplan-Meier estimator* [5, 6] which can handle censored data. To project out and compute editors' departure probabilities at times beyond the end of the study,



(a) survival function

(b) hazard function

Figure 3: The survival analysis results for customary editors of Wikipedia.

we need to fit a parametric survival function to the empirical data. After trying out a number of popular probability distributions (including exponential, extreme value, lognormal, normal, Rayleigh, and Weibull), we have found that although the survival function of occasional editors roughly follows a (truncated) lognormal distribution (which confirms the finding in [2]), the survival function of customary editors can be better described by a Weibull distribution

$$f(t) = \begin{cases} \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-(t/\eta)^{\beta}} & t \geq 0 \ , \\ 0 & t < 0 \ , \end{cases} \qquad (2)$$

with the scale parameter $\eta = 102.68$ and the shape parameter $\beta = 0.55$. As shown in Figure 3a, the Weibull distribution curve clearly matches the customary editors' lifetime data better than the lognormal distribution curve does. The shape parameter of the fitted Weibull distribution is less than 1, which indicates that the overall departure rate decreases over time, i.e., those who leave the community tend to leave early, and those who stay in the community become less likely to leave over time. Given the parametric survival

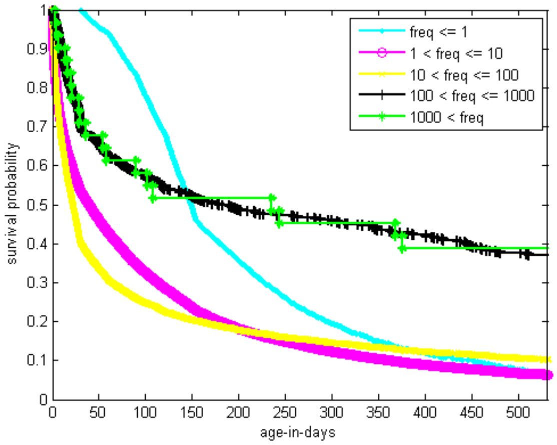


Figure 4: Comparison of survival functions between customary editors of Wikipedia with different monthly editing frequencies ($freq$).

function, we are able to make inference about the *expected future lifetime* of an editor who has stayed in the community for t_0 days:

$$\int_0^\infty t \frac{f(t+t_0)}{S(t_0)}\,\mathrm{d}t = \frac{1}{S(t_0)} \int_{t_0}^\infty S(t)\,\mathrm{d}t \ . \qquad (3)$$

This reduces to the expected lifetime (a.k.a. mean time to failure) at birth for $t_0 = 0$. Furthermore, the age at which a specified proportion q of editors will remain can be found by solving the equation $S(t) = q$ for t. Using the fitted Weibull distribution, we estimate the *median lifetime* (at which half of the customary Wikipedia editors leave the community) to be about 53 days.

The hazard function, conventionally denoted λ, is defined as the event ("death") rate at time t conditional on survival until time t or later (that is, $T \geq t$):

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0} \frac{\Pr(t \leq T \leq T + \Delta t | T \geq t)}{\Delta t} \\ &= -\frac{\mathrm{d}S(t)/\,\mathrm{d}t}{S(t)} \ . \end{aligned} \qquad (4)$$

The empirical hazard function for customary editors is obtained through the non-parametric method *kernel density estimation* based on the empirical survival function given above. It can be seen from Figure 3b that the empirical hazard function is closer to the parametric hazard function derived from the fitted Weibull distribution than that derived from the fitted lognormal distribution. Moreover, the empirical hazard rate curve is in general decreasing along with the editor's "age" in the community, except for the periods of 0-2 weeks and 8-20 weeks, which suggest that these are the two critical phases to retain Wikipedia editors in the community.

In order to further understand the survival patterns of customary editors, we group them into five classes according to their monthly editing frequencies ($freq$): (i) $freq <= 1$, (ii) $1 < freq <= 10$, (iii) $10 < freq <= 100$, (iv) $100 < freq <= 1000$, (v) $1000 < freq$. The sizes of those classes are 6744, 32583, 8818, 675, and 31 respectively. The exponential scale using the powers of 10 is chosen to define the above classes because the monthly editing frequencies roughly follow the *power law* [3]. This is the same editor classification criterion used in [12], except that the classification is not recalculated every month as we would like to analyse the relationship between an individual editor's monthly editing frequency and his whole lifetime.

Figure 4 plots the survival function for each class of customary editors separately. It is clear that low-frequency editors (class i and ii) are more likely to have a short lifetime than medium-frequency editors (class ii and class iii), and similarly we can say that medium-frequency editors (class ii and class iii) are less likely to have a long lifetime than high-frequency editors (class v). In other words, higher editing frequency implies longer lifetime — the more active an editor is, the longer he will keep active in editing.

4. CONCLUSIONS

The major contribution of this paper is to show that for customary Wikipedia editors,

- the survival function can be well described by a Weibull distribution (with the median lifetime of about 53 days);
- there are two critical phases (0-2 weeks and 8-20 weeks) when the hazard rate of becoming inactive increases;
- more active editors tend to keep active in editing for longer time.

Why Wikipedia editors become inactive is still largely an open research problem. As one would expect, reverts demotivate Wikipedia editors and drive newcomers away [4]. It will be promising to further use survival analysis methods such as the *Cox proportional hazards model* [5, 6] to understand how underlying factors like reverting determine the departure dynamics of Wikipedia editors, which has been left for future work.

5. ACKNOWLEDGMENT

We thank the anonymous reviewers for their helpful comments.

6. REFERENCES

[1] S. L. Bryant, A. Forte, and A. Bruckman. Becoming wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work (GROUP)*, pages 1–10, Sanibel Island, FL, USA, 2005.

[2] G. L. Ciampaglia and A. Vancheri. Empirical analysis of user participation in online communities: the case of Wikipedia. In *Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM)*, pages 219–222, Washington, DC, USA, 2010.

[3] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[4] A. Halfaker, A. Kittur, and J. Riedl. Don't bite the newbies: How reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym)*, pages 163–172, Mountain View, CA, USA, 2011.

[5] D. W. Hosmer, S. Lemeshow, and S. May. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley-Interscience, 2nd edition, 2008.

[6] D. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text*. Springer, 3rd edition, 2011.

[7] F. Ortega. *Wikipedia: A Quantitative Analysis*. PhD thesis, Universidad Rey Juan Carlos, 2009.

[8] F. Ortega and D. Izquierdo-Cortazar. Survival analysis in open development projects. In *Proceedings of the 2nd International Workshop on Emerging Trends in Free/Libre/Open Source Software Research and Development (FLOSS)*, pages 7–12, Vancouver, Canada, 2009.

[9] K. A. Panciera, A. Halfaker, and L. G. Terveen. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the 2009 International ACM SIGGROUP Conference on Supporting Group Work (GROUP)*, pages 51–60, Sanibel Island, FL, USA, 2009.

[10] J. E. Pitkow and P. Pirolli. Life, death, and lawfulness on the electronic frontier. In *CHI*, pages 383–390, Atlanta, GA, USA,, 1997.

[11] R. Priedhorsky, J. Chen, S. K. Lam, K. A. Panciera, L. G. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In *Proceedings of the 2007 International ACM SIGGROUP Conference on Supporting Group Work (GROUP)*, pages 259–268, Sanibel Island, FL, USA, 2007.

[12] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. The singularity is not near: Slowing growth of Wikipedia. In *Proceedings of the 2009 International Symposium on Wikis (WikiSym)*, Orlando, FL, USA, 2009.