Empirical Analysis of User Participation in Online Communities: The Case of Wikipedia

Giovanni Luca Ciampaglia*
University of Lugano
Faculty of Informatics
Via G. Buffi 13, 6900 Lugano, Switzerland

Alberto Vancheri
University of Lugano
Academy of Architecture
Largo Bernasconi 3, 6850 Mendrisio, Switzerland

Abstract

We study the distribution of the activity period of users in five of the largest localized versions of the free, online encyclopedia Wikipedia. We find it to be consistent with a mixture of two truncated log-normal distributions. Using this model, the temporal evolution of these systems can be analyzed, showing that the statistical description is consistent over time.

Introduction

An interesting open question in the study of online communities is whether these systems evolve over time, what the causes for this evolution are, and what may happen to these platforms in case the current state of affairs gives way to an unfavorable one. Among the various features that are likely to evolve over time, the balance between green and expert users plays an important role in the economy of any community where content is user-generated (UGC), as in the well-known free, online encyclopedia called Wikipedia.

Here, we present a framework for measuring properties of a virtual community by looking at its temporal evolution. We focus on a single quantity: the time period of participation of an individual to the process of creation of the content. We measure the time between the first and the last edit of a user, which we call the *lifetime* of a user account, and perform a statistical testing of hypotheses on what model best describes it.

Related work: recently, a survival analysis of the activation and de-activation of users has been performed (Ortega 2009). Our approach is different as our model is essentially a clustering technique, so that we can directly evaluate the statistical properties of the classes of the short and long-lived users.

Another work (Wilkinson 2008) is also interested in the same 'deactivation' phenomenon in users-generated content communities. It deals however with a different quantity (the number of edits) in order to characterize the ranking of the level of contributions by users.

Besides these empirical works, studies on network growth, that feature the 'death' of nodes, account for an heavy-tailed statistics too (Lehmann, Jackson, and Lautrup 2005): the present work is in part motivated by the fact that these generative models also need to be compared with empirical data.

Our contribution: in this paper, we report the following contributions:

1. We find all datasets to be consistent with the hypothesis that the lifetime of an user account is described by the superposition of two truncated log-normal distributions. An interpretation for this phenomenon is that two different regimes govern the participation of individuals to these versions of the Wikipedia project: the occasional users, who fail to find interest in the project after the first few attempts to contribute, and the long-term users, whose withdrawal is probably more related to external factors like the loss of personal incentives in contributing and similar causes.
2. Using our model, we characterize how the participation of users over time evolves, as the system ages. We find that the statistical description of the one-timers is stable over time, while the properties of the group of long-term users change as a consequence of the aging of the system.
3. We find evidence that the inter-edit time distribution decays with an heavy tail. In view of this finding, we check that our analysis is not affected by the choice of the parameter used for determining when an user is to be considered "inactive"; we find that for the one-timers it has no quantitative effect. For the statistics of the long-lived users we find instead a very good qualitative agreement.

The Model

A *truncated* distribution in $(a, b) \in \mathbb{R} \cup \{\pm\infty\}$ is specified by considering the original distribution conditional on being in (a, b). For the normal distribution with location μ and scale σ the probability density of its truncated version is thus defined as:

$$p(x) = \frac{1}{\Phi(b') - \Phi(a')} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \quad \text{if } x \in (a, b) \qquad (1)$$

and 0 elsewhere. Φ denotes the cumulative distribution function of the standard Gaussian and the extremes are standardized, ie. $a' = (a - \mu)/\sigma$, (similarly for b').

*G.L.C. acknowledges the support of the Swiss National Science Foundation under project no. 200020-125128
Copyright © 2010, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

A *mixture* model is the superposition of two or more distributions, so that the resulting density is a weighted average from several components: $p(x) = \sum_{k=1}^{K} \pi_k p_k(x; \boldsymbol{\theta}_k)$. Here $p_k(x; \boldsymbol{\theta}_k)$ is the density of the k-th component, with vectors of parameters $\boldsymbol{\theta}_k$, evaluated at x.

Inference for mixture models is, however, often infeasible using constrained maximization methods because of the peculiar behavior of the log-likelihood function (Bishop 2006). A solution to this problems comes under the form of the *expectation-maximization* (EM) algorithm (Dempster, Laird, and Rubin 1977).

In our case, we have to take into account the truncation from equation 1 when computing the new estimate $\mu^{(s+1)}$, $\sigma^{(s+1)}$ of the parameters for each component in the M-step of the algorithm (for a description of EM refer to (Bishop 2006)). To do this, we do employ the following approximation, based on the moments of the truncated normal (Johnson, Kotz, and Balakrishnan 1994):

$$\sigma^{(s+1)} = \left(\hat{\sigma}^{(s)}\right)^2 \left[1 - \frac{b'\phi(b') - a'\phi(a')}{\Phi(b') - \Phi(a')} + - \left(\frac{\phi(a') - \phi(b')}{\Phi(b') - \Phi(a')}\right)^2\right] \quad (2)$$

$$\mu^{(s+1)} = \hat{\mu}^{(s)} - \sigma^{(s+1)} \frac{\phi(b') - \phi(a')}{\Phi(a') - \Phi(b')} \quad (3)$$

here ϕ refers to the density of the standard Gaussian distribution, while $\hat{\mu}^{(s)}$ and $\hat{\sigma}^{(s)}$ refer to the estimates of the non truncated distribution based on the old weights.

In most non-pathological cases,[1] we find that this approximation procedure produces asymptotically unbiased estimators.

Results

The data we have analyzed comprise the meta-data of the whole history of revisions (in all namespaces) for five localized versions of the Wikipedia project: the English, the German, the Italian, the French and the Portuguese versions.

We consider only revisions from non-robot, registered users. We define the lifetime τ of an user account as the period between the times τ_i and τ_f of the first and last revision performed by the user, respectively.

In order to consider only those users who have ceased their activity, which we call *inactive*, we define as inactive all users whose last revision τ_f dates back more than τ_{max} given days from the date of the most recent revision recorded in the data. This of course does not guarantee that a user classified as inactive would be permanently so in the future. Moreover, the analysis might be influenced by the choice of τ_{max}. We carefully address both issues later in this section.

Lifetime distribution: we can infer the parameters of our truncated mixture of log-normals by applying our custom EM technique to $u = \ln(\tau)$. Table 1 and Figure 1 display the results of the parameter estimation. In all cases, a K-S test does not reject the hypothesis that the data are drawn from the same distribution.

Table 1: Estimated parameters for the truncated normal model. In parenthesis the significant digit of the standard error of the estimator. p-values for statistically significant estimates (≥ 0.1) are quoted in bold.

language	μ_1	μ_2	σ_1	σ_2	p-val.
Italian	-5.4(3)	4.3(3)	1.7(3)	1.9(3)	**0.688**
German	-2.2(2)	5.6(2)	3.8(2)	1.1 (3)	**0.632**
French	-5.5(2)	4.6(3)	1.8(2)	1.8(3)	**0.464**
Portuguese	-5.5(4)	3.5(3)	1.5(4)	2.2(4)	**0.612**
English	-5.3(5)	3.2(4)	1.6(5)	2.2(5)	**0.54**

On the other hand, table 2 shows there is no support for a power law behavior in the data, as the p-values from a K-S test reject the hypothesis that the data are drawn from this distribution in all cases except for the German. Moreover, the power law decay is found only in the very tail of the distribution, thus this second model fails completely to characterize the structure in the data for all but the largest values of τ. The exponent α and the starting point τ_{min} of the power law decay are estimated with an MLE method (Clauset, Shalizi, and Newman 2009).

Table 2: Power law fit. p-values for statistically significant estimates (≥ 0.1) are quoted in bold.

language	α	τ_{min} $n(> \tau_{\text{min}})$	p-value
Italian	3.99 ± 0.14	688.525 (461)	0.09
German	4.84 ± 0.1	1013.89 (1342)	**0.1**
French	3.58 ± 0.14	681.01 (351)	0.09
Portuguese	3.91 ± 0.11	619.37 (693)	0.08
English	6.95 ± 0.08	1119.43 (5376)	0.04

We believe there are two reasons for the fact the power law is not a satisfactory model of these data while ours is. First, since there cannot be any user whose participation in the project is longer than the age of the project itself, the data must be naturally bounded, which explains the sharp cutoff for high values of τ. Second, the structure for values of $\tau < \tau_{\text{min}}$ might be due to the superposition of two different regimes: one comprising users whose interest fades away quickly, as a result of the few interactions they have with the rest of the community, and the long-term users, whose motivation for participating are not anymore affected by the daily outcomes of their editing actions, but rather by some stronger form of incentives (e.g. ideology in free, open projects).

We would like to note that we find an equally good fit for a mixture of 3 or 4 components, but given the empirical

[1] The 'troublesome' cases arise when multiple components overlap significantly, so to make them undistinguishable from a single component, or when one (or both) of the truncation extremes fall very shortly (ie. less than one standard deviation) from one or more components' center. We don't think this problem endangers too much the quality of our results, as our data don't seem to fall in any of these cases (see table 1).

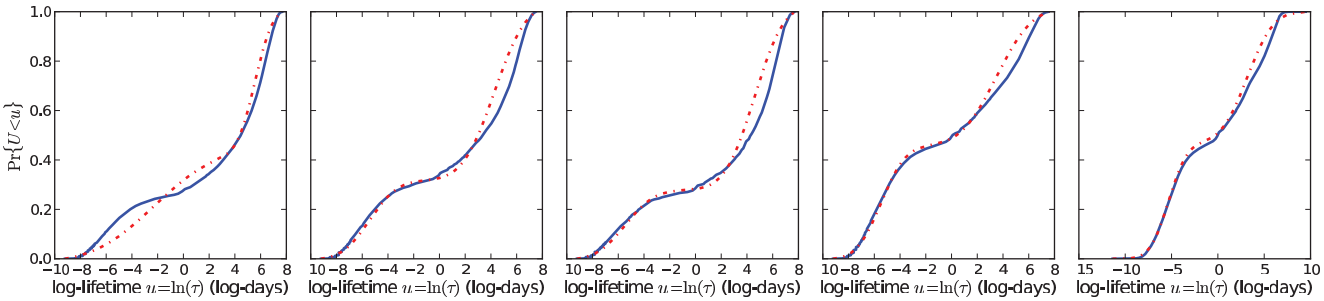

Figure 1: Truncated normal model fit. (a) German, (b) Italian, (c) French, (d) Portuguese, (e), English Wikipedia. Solid blue line, empirical CDF; dot-dashed red line, model prediction.

evidence for the existence of two distinct classes of users, we preferred to opt for the hypothesis of a model with 2 components.

We note that the fit of the German data is not in perfect agreement with the others. Indeed, the EM algorithm is only guaranteed to converge to a local maximum of the likelihood, but this discrepancy might be also due to the choice of $\tau_{\max}$. [2] In the following, we address the general problem of evaluating how much our analysis depends on the choice of the parameter $\tau_{\max}$.

Inactivity periods: let us denote with t the time of the most recent revision. A simple criterion for selecting the inactive users is to consider the time from the last revision $\tau_f - t$ and classify as inactive those user for which $\tau_f - t > \tau_{\max}$, where $\tau_{\max}$ is some threshold past which an inactive account can be safely considered as being permanently so.

A better understanding of the implication of taking this criterion can be achieved by looking at the statistics of the maximum $Y_N = \max\{S_0, \ldots, S_{N-1}\}$ of the inter-edit time S of the users with N edits, for any given N. Figure 2 show how $< Y_N >$ grows as a function of N.

If the underlying distribution of the inter-edit times had a tail decaying faster than a power law (ie. exponentially or like a Gaussian), the maximum would be a slowly increasing function of N (Sornette 2004). Here, instead, we see that for small values of N the maximum is growing faster than that, hinting at the fact that the inter-edit time distribution must be decaying with an heavy tail.

Figure 2 also illustrates that a fixed threshold classification might be skewed towards the high-activity users.

Temporal evolution: given all these questions it is natural to ask whether the results we get differ if a different value of $\tau_{\max}$ is chosen. Figure 3 depicts a comparison of the results of our analysis for several choices of the value of $\tau_{\max}$. For each pair of parameters $\pi_{1,2}$, $\mu_{1,2}$, $\sigma_{1,2}$, we plot their values estimated with EM at various ages of the system. More specifically, we took yearly-spaced points in the period that goes from an initial timestamp to the last recorded

[2] In order to increase the chances of hitting a good maximum of the log-likelihood function, our estimation procedure is repeated 25 times for each data set.

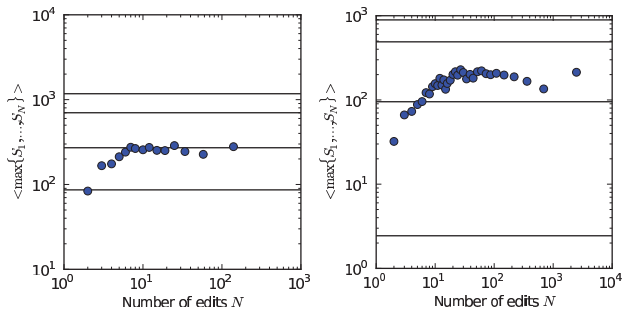

Figure 2: Analysis of the inter-edit time distribution. (a) French, (b) Portuguese Wikipedia. Filled blue circles: scaling of the average maximum statistics as a function of the number of edits (i.e. sample size). Axes are on the log-log scale. The grouping is count dependent so that in each bin there are at least 100 observations. The relative standard error bars are all smaller than the data points. Horizontal solid lines: from bottom to top, median, 75th, 95th and 99th percentile of the full distribution of the maxima, ie. not grouped by number of edits.

timestamp in our datasets. The initial timestamp is chosen so that the system contained at least 100 users at that time. At each of this dates, we restrict our datasets to the users whose last revision is precedent to such date.

The plots reveal very interesting patterns in the evolution of the systems: first, μ_1 is quite insensitive to the choice of $\tau_{\max}$; second, it stabilizes to the value of ≈ -5.5 after some time, which means that this is possibly the characteristic time scale of involvement of occasional contributors. The average log-lifetime μ_2 has instead a qualitatively similar behavior across different choices of $\tau_{\max}$, exemplifying the dynamics of aging of the system (from approximately one month at the beginning of the project, up to over one year for recent measurements). We note that occasionally the estimates sharply diverge from this common trend which might explain the disagreement for the German case at $\tau_{\max} = 180$.

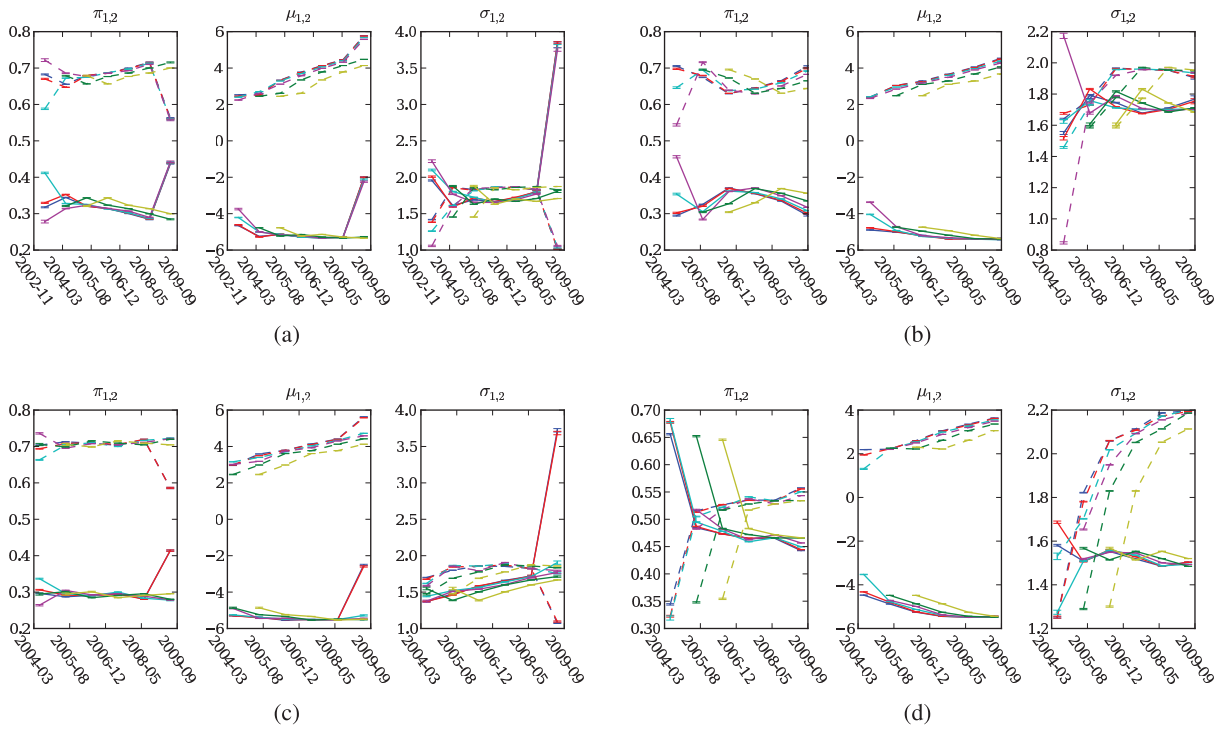

Figure 3: Evolution of the fitted parameters. (a) German, (b) Italian, (c) French, (d) Portuguese. Colors correspond to different values of the time $\tau_{\max}$ used to classify an user as inactive: 7 (blue), 30 (red), 90 (cyan), 180 (magenta), 365 (green), 730 (yellow). Solid lines: π_1, μ_1, σ_1 (short-lived users). Dashed lines: π_2, μ_2, σ_2 (long-lived users). Error bars are the 95% confidence intervals of the estimator.

Further investigation will be needed to better understand the source of these divergences.

Discussion

In this paper we analyzed the evolution of five of the largest language versions of Wikipedia over their history. We found strong support for the fact that the period of participation of users to the community is distributed according to the mixture of two log-normal laws. A possible criticism to our methodology could be that the comparison with a simple power law without any form of cutoff was not totally fair. Still, what we think to be truly interesting in this study is that all versions of the wikis we analyzed show the presence of some form bi-modality in the lifetime statistics: the mere empirical observation that some users live more than others would not be sufficient to explain this fact, so we believe that is far from being a trivial finding. In principle, our analysis depends on the arbitrary choice of one parameter, namely the threshold of inactivity $\tau_{\max}$. In practice the estimation is stable enough to discern clear temporal trends in the evolution of these systems. An explanation for this might be that we are estimating a property of the whole community of users, so that a robust notion of "inactivity" emerges from our statistical description, no matter how problematic it might be to define it at the level of the single individual.

References

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer Science, New York, USA.

Clauset, A.; Shalizi, C. R.; and Newman, M. E. J. 2009. Power-law distributions in empirical data. *SIAM Review*.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B* 39(1):1–38.

Johnson, N. L.; Kotz, S.; and Balakrishnan, N. 1994. *Continuous Univariate Distributions*, volume 1. Wiley Series in Probability and Statistics.

Lehmann, S.; Jackson, A. D.; and Lautrup, B. 2005. Life, death and preferential attachment. *Europhys. Lett.* 69:298–303.

Ortega, F. 2009. *Wikipedia: A Quantitative Analysis*. Ph.D. Dissertation, Universidad Rey Juan Carlos.

Sornette, D. 2004. *Critical Phenomena in Natural Sciences: Chaos, Fractals, Self-organization and Disorder: Concepts and Tools*. Heidelberg: Springer-Verlag.

Wilkinson, D. M. 2008. Strong regularities in online peer production. In *Proceedings of the 9th ACM conference on Electronic commerce*.