Wikipedia editing dynamics

Y. Gandica[*]
*Centre for Computational Physics, Department of Physics, University of Coimbra, 3004-516 Coimbra, Portugal
and Department of Mathematics and Namur Center for Complex Systems—naXys, University of Namur,
rempart de la Vierge 8, B 5000 Namur, Belgium.*

J. Carvalho and F. Sampaio dos Aidos
Centre for Computational Physics, Department of Physics, University of Coimbra, 3004-516 Coimbra, Portugal

(Received 7 November 2014; published 29 January 2015)

A model for the probabilistic function followed in editing Wikipedia is presented and compared with simulations and real data. It is argued that the probability of editing is proportional to the editor's number of previous edits (preferential attachment), to the editor's fitness, and to an aging factor. Using these simple ingredients, it is possible to reproduce the results obtained for Wikipedia editing dynamics for a collection of single pages as well as the averaged results. Using a stochastic process framework, a recursive equation was obtained for the average of the number of edits per editor that seems to describe the editing behavior in Wikipedia.

DOI: 10.1103/PhysRevE.91.012824 PACS number(s): 89.75.Fb, 87.23.Ge, 89.20.Hh, 87.64.Aa

I. INTRODUCTION

The extensive monitoring of people's daily activities, in particular their online actions, yields a large amount of information which, by adopting stochastic techniques, can be used to determine some of the probability distributions that govern social interactions. Statistical physics uses a probabilistic description [1] to obtain useful information about the general behavior of many-particle systems. In this way, it is possible to find some properties of macrosystems, regardless of the complex individual behavior of each particle, or, in the case of social systems, of each human being.

Human activity is very far from being a random process, in the sense of anybody being able to do anything at any time. Although each person has, at each time, the opportunity to choose from different options, actually one is immersed in an intricate net of social rules, schedules, socially accepted manners, power and economic constraints, etc., which end up determining the available usual options. The probability distribution for each type of human behavior greatly depends on the specific human trait that is being studied. Understanding the laws that govern human activity is a great challenge for science as it may arguably be considered the most complex stochastic process in nature.

The editing of Wikipedia (WP) is one of these sources of probabilistic outcome [2]. Some effort has already been put into attempts to understand the evolution of WP as a network, with pages or topics as nodes and links between them as edges [3–6]. Several models have been proposed to describe the activity patterns of editors over different pages [7,8]. For a single page, Wilkinson and Huberman proposed a simple stochastic mechanism to obtain the probability distribution for that page's number of edits as a function of time, based on the simple rule that "edits beget edits" [9]. This case of preferential attachment belongs to the elemental process introduced by Simon [10] in 1955, in an early observation of universal patterns in linguistic, sociological, and biological data.

The high quality of the WP encyclopedia is the result of a collective effort by millions of volunteers, in an apparently disorganized process of editing, acceptance, and rejection, which works, in fact, as an effective and robust peer review procedure. Halfaker *et al.* studied some WP editor characteristics that lead to the process of selection of high-quality contributions [11].

In agreement with the concept of universality, in the statistical physics meaning, we propose in this paper a statistical approach for the probability of each editor to interact with the WP page, based on three principles already shown to be essential in human interaction. We use the preferential attachment mechanism to represent the strong tendency of users to improve and defend their previous contributions [11]. This ownership feeling competes with the authority of users that are experts on the page topic, which is expressed in our model by an increased value of a parameter that we associate with each editor and that is usually called fitness [12–14]. Fitness may also describe the different drive that different persons have to push forward their opinions. Finally, an aging factor [15,16] is proposed, associating the time-dependent behavior with an initial high motivation to edit, followed by a tendency to decrease the editing activity by theme completeness, personal saturation, blockage [17], and/or any other possible personal reason. The analytical calculations used to describe the editing dynamics with the three above-mentioned ingredients have produced a recursive equation for the average number of edits per editor that describes qualitative and quantitatively the real behavior displayed by the WP editing dynamics.

This article is structured as follows: In Sec. II, the real data sample is presented. Section III discusses the choice of the ingredients used to describe real data while Sec. IV explains in detail the model used to represent the editing process. In Sec. V, the analytical treatment is developed and the conclusions are finally presented in Sec. VI.

II. REAL DATA

The real data results shown in this work were obtained from a January 2010 dump of the English WP, containing 4.64×10^6 pages, accessible at the WikiWarMonitor web page [18].

[*] ygandica@gmail.com

The data sample used, a "light dump," contains a reduced information listing of all the page edits (such as the edit number and the editor identification). Only pages with more than 2000 edits were analyzed. They were divided into five ranges of R (the ratio between the number of different editors involved in the editing of one page and the total number of edits of that page), from 0.1 to 0.6, in bins of 0.1. Only the first 2000 edits of each page were analyzed and compared with the simulation results.

III. EDITING PROBABILITY

It was argued in a previous paper [19] that one of the main characteristics of WP editing is an approximately constant rate R of incoming new editors, as illustrated in Fig. 1. This figure shows the editors' activity, by plotting a symbol for each edit of the article made by each specific editor, where editors are numbered according to the chronological order of their debut in the article.

The real WP page Jesus (left) is compared with a simulation (right) for some chosen components and parameters, as explained later. The plot for the real page shows clearly an initial intensive activity for each editor, as can be seen by the high density of symbols near the diagonal (the line that corresponds to the first edit made by each editor), followed, for most editors, by a clear decay. This indicates that most editors have an initial high motivation to edit which, in time, just fades away. However, it is also clear from the thick long horizontal lines in the same plot that there are some supereditors (editors with far more edits than the average) who actually manage to maintain the editing drive.

It seems clear that we need three ingredients to allow for a good qualitative description of this behavior. Preferential attachment is surely an essential part [20]. The presence of hubs (the supereditors in WP editing) in a network is a common signature for preferential attachment. A fitness function is also required to enhance the editing probability of some thereafter supereditors. Preferential attachment alone cannot explain the supereditor distribution in Fig. 1, as all supereditors would then enter the editing process very early. Fitness allows for the possibility of supereditors to start at any later time. Finally, we must also include an aging function to decrease the editing probability as time progresses. This effect is displayed in Fig. 1 by the dampening of most editors' editing frequency as time elapses.

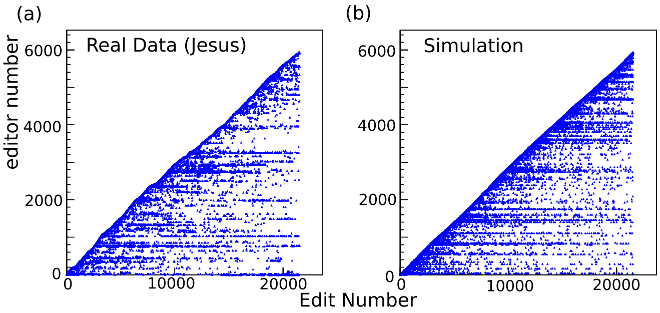

FIG. 1. (Color online) Editors' activity as a function of edit number for (a) the WP page Jesus and (b) the simulation with preferential attachment, an aging function, and editor's fitness, as explained in the text. In both cases, a symbol is plotted for each edit made by each editor.

The study of the editing process requires some definitions. The successive article edits are numbered in chronological order and the variable that refers to a specific edit number is denoted by e where, in our universe of 2000 edits, $1 \leqslant e \leqslant 2000$. The editors are also assigned a number in chronological order of their inception and the variable that refers to editor E_i is i, where $1 \leqslant i \leqslant 2000R$. e_i is defined as the value of e when editor E_i starts to edit and $\varepsilon_i = e - e_i$ is defined as the number of edits after editor E_i has started to edit. We shall refer to an editor by E_i, by editor number i, or as the editor who started at e_i. (Note that $1 \leqslant e_i \leqslant 2000$ but that, for each page, there are only $2000R$ different values of e_i. Note also that $\varepsilon_i = 0$ when editor E_i starts to edit.)

For each range of values of R, all the pages with more than 2000 edits were selected. Let N_p be the number of selected pages. We took the first 2000 edits from each of these pages and measured the number of edits done by each editor. Let $k(e_i,e)$ be the number of edits done by editor E_i up to edit number e. Using $k(e_i,e)$, for $e = 2000$, we made two different calculations. In one calculation, we present the results for the collection of N_p pages. We took all the $N_p \times 2000R$ values of $k(e_i,e)$ and inserted them in bins with equal length (we used 1). Then we measured the number of editors whose number of edits lay inside each bin, thus obtaining the probability distribution for the number of edits per editor for that collection of pages. In the other calculation, we evaluated a simple average of $k(e_i,e)$ [which we denote by $\overline{k}(e_i,e)$] for all the editors who started to edit at edit number e_i over all N_p different WP pages. This is done for each of the e possible values of e_i (note that if, in a specific page, no editor starts to edit at edit number e_i, that one counts as zero edits for the calculation, which will always average for N_p terms). Finally, these e values of $\overline{k}(e_i,e)$ were inserted in bins of equal length (0.2) and the probability distribution for the average number of edits per editor was obtained.

Several simulations of the WP editing dynamics were performed with all these ingredients. Two different absolutely continuous random variables were tried to describe fitness: a random variable ζ with a uniform distribution in the interval (0,1) and a random variable which follows the power law

$$\xi = (0.01 + \zeta)^{-\gamma}. \qquad (1)$$

We found that the uniform random variable was not powerful enough to create the supereditors who appear late in the editing process. The power law, on the contrary, seemed capable of providing very few editors with a very high editing proficiency. For the aging mechanism, we tested the inactivation suggested in Ref. [19] and an exponential form of aging, $e^{-q\varepsilon_i}$, which is, for a sufficiently large pool of pages, equivalent to the inactivation procedure in the calculation of the average number of edits per editor. We found that this form of aging destroys the editor's contribution too quickly, thus hindering a long editorial history for the editors. Therefore, we tried a power law function $\varepsilon_i^{-\alpha}$, which has already been used in Ref. [16], and which allowed for a smoother inhibition to the continued editorial activity of the editors.

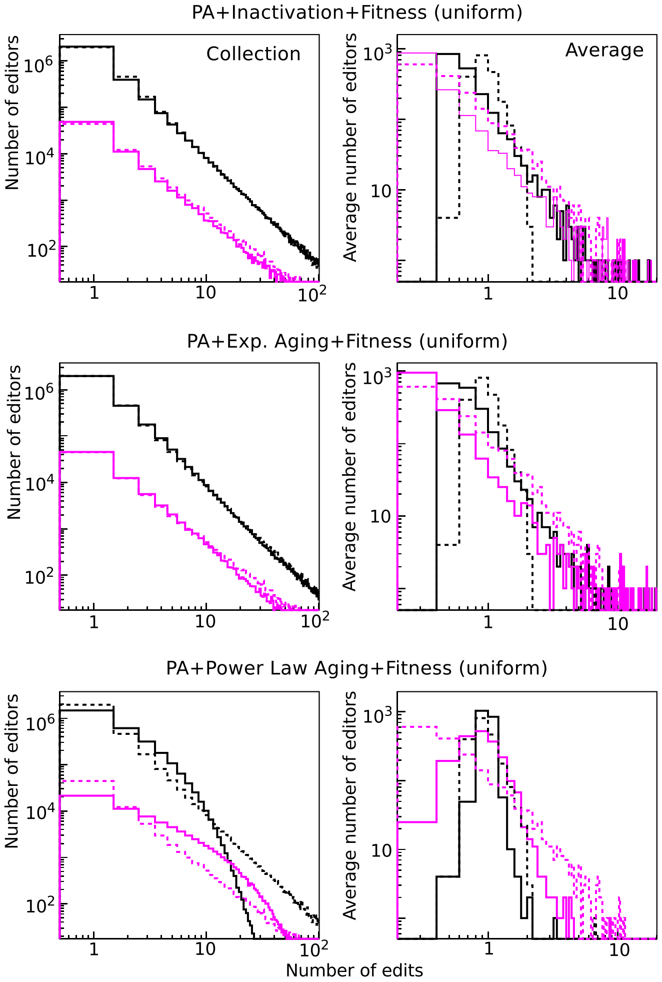

FIG. 2. (Color online) Comparison between the number of edits per editor obtained from real data (dashed lines) and from simulations (full lines), for two ranges of R (ratio between the number of editors and the number of edits for each page): [0.1,0.2] pink (gray) and [0.4,0.5] black. The left column shows the results for a collection of pages and the right column the results for the average over all the pages (the number of real and simulated WP pages is the same). The top row displays the results for the simulation with preferential attachment, uniform fitness, and inactivation (see the text for the explanation of the different components). The middle row shows the results for preferential attachment, uniform fitness, and exponential aging, and the bottom row the results for preferential attachment, uniform fitness, and power law aging. The agreement is very poor for all combinations.

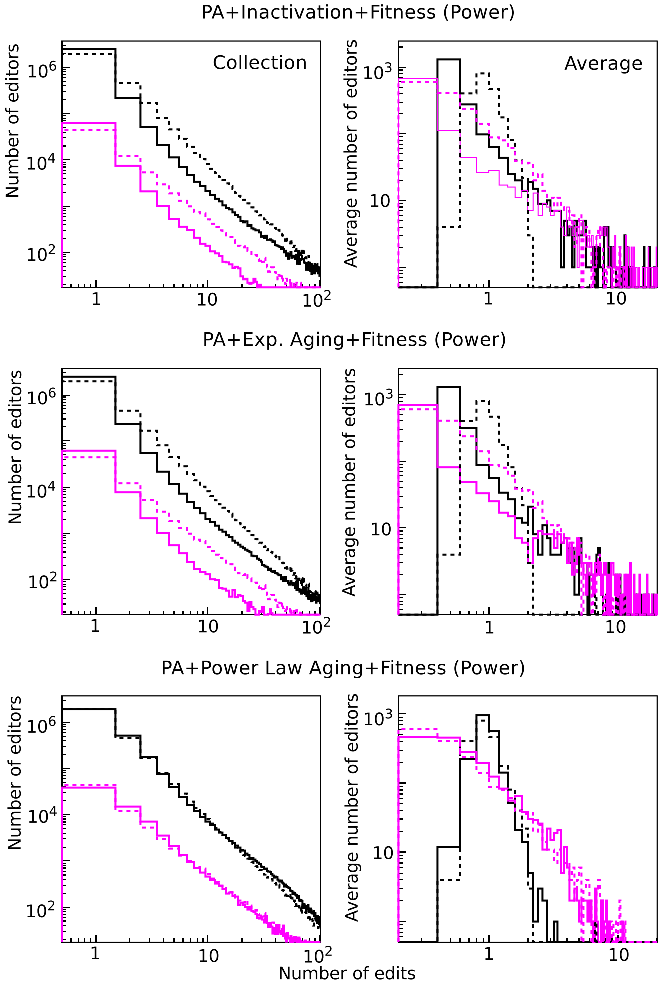

FIG. 3. (Color online) Comparison between the number of edits per editor obtained from real data (dashed lines) and from simulations (full lines), for two ranges of R (ratio between the number of editors and the number of edits in each page): [0.1,0.2] pink (gray) and [0.4,0.5] black. The left column shows the results for a collection of pages and the right column the results for the average over all the pages (the number of real and simulated WP pages is the same). The top row displays the results for the simulation with preferential attachment, power law fitness, and inactivation (see the text for the explanation of the different components). The middle row shows the results for preferential attachment, power law fitness, and exponential aging, and the bottom row the results for preferential attachment, power law fitness, and power law aging. Only the last combination shows a good agreement between real data and simulation for all the distributions.

The model parameters were chosen from a comparison between the WP real data and the simulation, using a two-sample Kolmogorov-Smirnov test [21]. The values obtained are $\gamma = 0.90$ for the fitness parameter, $q = 0.0005$ for the aging exponential parameter, and $\alpha = 1.25$ for the aging power law parameter.

In Figs. 2 and 3, we show a comparison between real data and simulations. The same number of pages was used in simulations and in real data calculations. For two different ranges of R, we show both the number of edits per editor for a collection of pages and the average number of edits per editor (as explained above). In Fig. 2, the simulations were done with uniform fitness and each of the three kinds of aging, and in Fig. 3 the simulations were run with power law fitness and again each of the three kinds of aging. The results indicate that the best fit is obtained with the power law fitness and the power law aging mechanism.

The exponential and the power law forms of aging are further compared in Fig. 4, which shows the average number of edits $\overline{k}(e_i,e)$ as a function of e_i for $e = 2000$, for real data and for two simulations. Both simulations were performed with preferential attachment and the power law form of fitness. One uses the exponential aging mechanism and the other uses the power law. The comparison is striking.

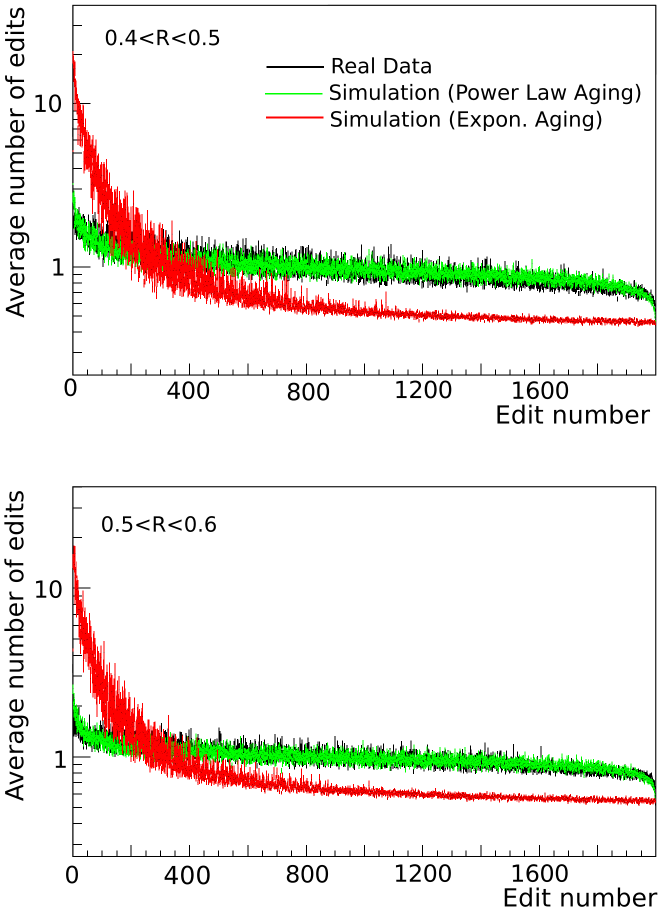

FIG. 4. (Color online) Average number of edits per editor $\overline{k}(e_i,e)$ as a function of e_i for $e = 2000$ for two ranges of R (ratio between the number of editors and the number of edits in each page): top [0.4,0.5] and bottom [0.5,0.6]. The real data are shown in black, the simulation with preferential attachment, power law fitness, and exponential aging in red (dark gray), and the simulation with preferential attachment, power law fitness, and power law aging in green (light gray, over the real data). The generally good description of real data by the latter simulation is clear.

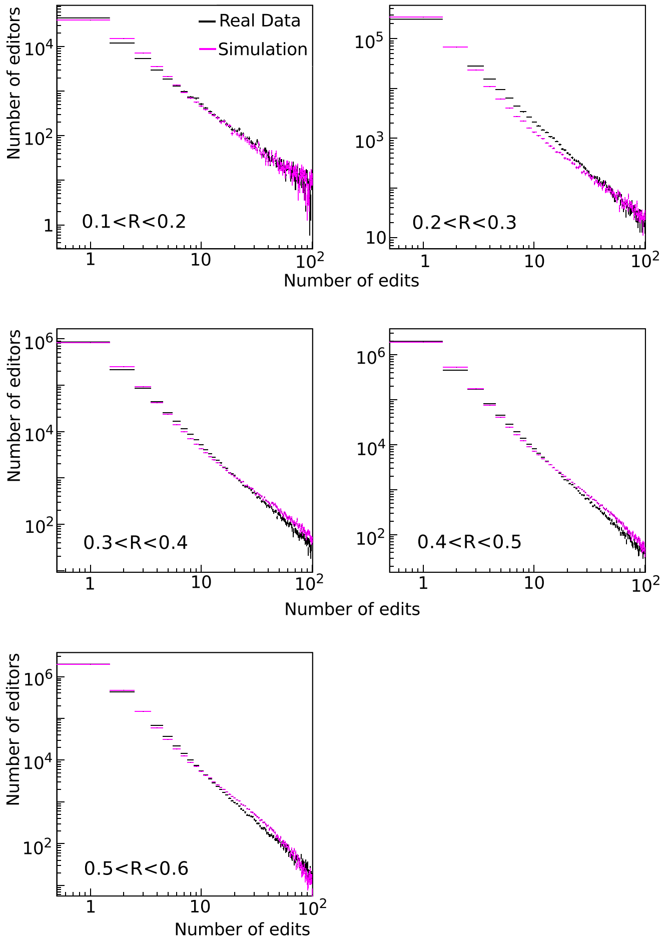

FIG. 5. (Color online) Comparison between the number of edits per editor obtained from real data, in black, and simulation, in pink (gray), for a collection of pages and five ranges of R (the ratio between the number of editors and the number of edits in each page). (The number of real and of simulated WP pages is the same.) The simulation was obtained with preferential attachment, power law fitness (with $\gamma = 0.90$), and power law aging (with $\alpha = 1.25$) (see the text for the explanation of the different components). The agreement between real data and simulation for all the distributions is, in general, good.

IV. FITTING MODEL

The agent-based model used to obtain the simulations in this paper is based on the ingredients discussed in the previous section. The model is described in detail in [19] but the only parts required here are the elements of the editing dynamics.

A computer simulation run starts with one editor and the choice of a value of R. At each dynamical step e, a new editor comes in and edits the page for the first time with probability R. Then, at each time step, an old editor E_i will edit the article with probability $1 - R$. The probability of choosing a particular editor E_i among all the old editors will be

$$\Pi_i = \frac{k(e_i,e-1)x_i\varepsilon_i^{-\alpha}}{\sum_j k(e_j,e-1)x_j\varepsilon_j^{-\alpha}}, \qquad (2)$$

where x_j is the fitness parameter of the editor who started at e_j, which is initially chosen following the power law mentioned in Eq. (1), and is maintained during the whole run. The sum in the denominator is over all editors who have edited the article before edit number e.

The qualitative agreement between real data and simulation is good for all the ranges of R, as shown in Fig. 5 for the collection of pages and in Fig. 6 for the average, taking into account the reduced number of ingredients and parameters used.

V. ANALYTIC CALCULATION

In this section, an analytical approach for the problem of finding the average number of edits per editor is developed. The mathematical problem of obtaining the number of edits by an editor at each edit number e is similar to the problem of finding the degree of a node in a network. It has been mentioned before that the editors play the role of nodes and the edits the role of connections, the difference being, of course, that in this

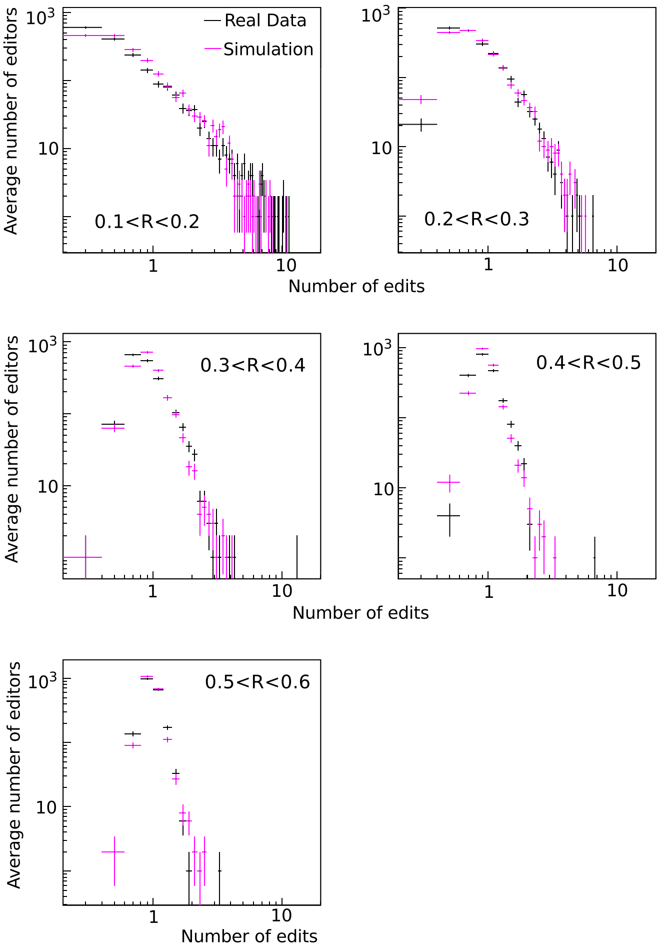

FIG. 6. (Color online) Comparison between the number of edits per editor obtained from real data, in black, and simulation, in pink (gray), for the average over all the pages and five ranges of R (the ratio between the number of editors and the number of edits in each page). (The number of real and of simulated WP pages is the same.) The simulation was obtained with preferential attachment, power law fitness (with $\gamma = 0.90$) and power law aging (with $\alpha = 1.25$) (see the text for the explanation of the different components). The agreement between real data and simulation for all the distributions is, in general, good.

case the editing does not connect one editor to another. Instead, it connects an editor to the WP page that is being edited, which is outside the network.

This problem is often studied under the continuum approximation [13,14,16,20,22,23] although some authors have performed exact calculations [24,25]. A network with preferential attachment and a power law aging has already been studied by Dorogovtsev and Mendes [16]. Using a continuum approach, they obtained an equation for $\overline{k}(e_i,e)$ (in the following discussion, their notation will be adapted to this paper's specific problem):

$$\frac{\partial \overline{k}(e_i,e)}{\partial e} = \frac{\overline{k}(e_i,e)(e-e_i)^{-\alpha}}{\int_0^e du\, \overline{k}(u,e)(e-u)^{-\alpha}}, \quad \overline{k}(e,e) = 1. \tag{3}$$

In the continuum approach it is assumed that e, e_i, and k are continuous real variables. Dorogovtsev and Mendes proceeded to show that, for $\alpha < 1$, the solution can be written in terms of a hypergeometric function, and that the distribution of the number of edits per editor asymptotically follows a power law. For $\alpha \geqslant 1$, however, the problem is more complicated. It is easy to see that Eq. (3) fails for this range of values of α, due to a divergence of the integral in the denominator. According to Dorogovstsev and Mendes, the number of edits per editor should now follow an exponential behavior, but the statistics is too poor for this statement to be verified by simulations.

The power law aging effect described in this paper is the same as the one Dorogovtsev and Mendes use in the above formalism. However, as the value obtained for the parameter α is larger than 1, their results cannot be used. Clearly, the divergence of the integral in Eq. (3) is caused by the continuum approach, as the divergent term $(e-u)^{-\alpha}$ would never be larger than 1 in the discrete approach for $\alpha \geqslant 1$.

Therefore it is necessary to go back to the discrete formalism. Assuming an ensemble of similar articles, the average number of edits $\overline{k}(e_i,e)$ made by the editors who started at e_i will now be

$$\overline{k}(e_i,e) = \lim_{N\to\infty} \sum_{j=1}^{N} \frac{k_j(e,e_i)}{N}, \tag{4}$$

where $k_j(e,e_i)$ is the value of $k(e_i,e)$ for article j of the ensemble (it can be zero, if no editor started at e_i in that particular article). Now, let $\kappa(e,e_i)$ and $\xi(e_i)$ be the random variables "number of edits of the editor who started at e_i at edit number e" and "fitness of the editor who started at e_i," respectively. Then

$$\overline{k}(e_i,e) = \sum_{k=0}^{e-e_i+1} k\, P\{\kappa(e,e_i) = k\}, \quad 1 \leqslant e_i \leqslant e, \tag{5}$$

where P is the probability function. Naturally, the term $k = 0$ does not contribute to the summation in Eq. (5). However, it was left there to stress the nonzero probability for the editor who started at e_i not to edit a specific WP page $[P\{\kappa(e,e_i) = 0\} \neq 0$ for $i > 1]$. For $e_i = e$, we obviously have

$$\overline{k}(1,1) = 1 \quad \text{and} \quad \overline{k}(e,e) = R, \quad \forall\, e > 1. \tag{6}$$

The change of $\overline{k}(e_i,e)$ in one step, from e to $e+1$, is given by

$$\begin{aligned} \Delta\overline{k}(e_i,e+1) &= \overline{k(e_i,e+1) - k(e_i,e)} \\ &= \sum_{k=0}^{1} k P\{\kappa(e+1,e_i) - \kappa(e,e_i) = k\}, \\ 1 \leqslant e_i \leqslant e, \end{aligned} \tag{7}$$

and

$$\Delta\overline{k}(e+1,e+1) = R. \tag{8}$$

Π_i has already been defined by Eq. (2) as the probability for choosing editor E_i to edit the article at edit number e, assuming that, at e, no new editor comes in. In this case, Π_i is a random variable that, besides depending on e and e_i, is proportional to the editor's number of edits $\kappa(e,e_i)$ and to its fitness $\xi(e_i)$. However, this implies that Π_i must also depend on the number of edits, the edit number, and the fitness of all the other editors (because of the normalization constant for the probability), and the simplification of Eq. (7) requires some care.

We define the random vectors $\boldsymbol{\kappa}(e)$ and $\boldsymbol{\xi}(e)$ and the integer and real vectors $\boldsymbol{k}(e)$ and $\boldsymbol{x}(e)$ respectively as vectors with e components given by

$$\begin{aligned} [\boldsymbol{\kappa}(e)]_i &= \kappa(e,e_i), \quad [\boldsymbol{\xi}(e)]_i = \xi(e_i), \\ [\boldsymbol{k}(e)]_i &= k_i, \quad [\boldsymbol{x}(e)]_i = x_i \end{aligned} \tag{9}$$

for $i = 1,\dots,e$, where k_i is a variable that can take integer values from zero up to e and x_i is a variable that takes values in the fitness set of values. $\Pi_i(\boldsymbol{k}(e),\boldsymbol{x}(e))$ (we shall omit the dependence on e and e_ℓ, for $1 \leqslant \ell < e$) is then related to the conditional probability

$$\begin{aligned} P\{\kappa(e+1,e_i) - \kappa(e,e_i) = 1 | \boldsymbol{\kappa}(e) = \boldsymbol{k}(e), \boldsymbol{\xi}(e) = \boldsymbol{x}(e)\} \\ = (1-R)\Pi_i(\boldsymbol{k}(e),\boldsymbol{x}(e)). \end{aligned} \tag{10}$$

In Eq. (7) only the term $k = 1$ survives and we can write

$$\begin{aligned} \Delta\bar{k}(e_i,e+1) = \sum_{\boldsymbol{k}(e)} \sum_{\boldsymbol{x}(e)} P\{\kappa(e+1,e_i) - \kappa(e,e_i) = 1, \boldsymbol{\kappa}(e) \\ = \boldsymbol{k}(e), \boldsymbol{\xi}(e) = \boldsymbol{x}(e)\} \end{aligned} \tag{11}$$

$$\begin{aligned} = (1-R) \sum_{\boldsymbol{k}(e)} \sum_{\boldsymbol{x}(e)} \Pi_i(\boldsymbol{k}(e),\boldsymbol{x}(e))\, P\{\boldsymbol{\kappa}(e) \\ = \boldsymbol{k}(e), \boldsymbol{\xi}(e) = \boldsymbol{x}(e)\}, \end{aligned} \tag{12}$$

where the sum over $\boldsymbol{k}(e)$ runs over all sets of integers for which there is a positive contribution to the sum. The sum over $\boldsymbol{x}(e)$ will be an integral if the fitness turns out to be a continuous variable (as is the case in our model). In order to continue, a specific expression for Π_i must be chosen. Equation (2) yields

$$\Pi_i(\boldsymbol{k}(e),\boldsymbol{x}(e)) = \frac{k_i\, x_i\, (e+1-e_i)^{-\alpha}}{\sum_{\ell=1}^{e} k_\ell\, x_\ell\, (e+1-\ell)^{-\alpha}}. \tag{13}$$

Unfortunately, this expression does not allow for the simplification of Eq. (12). One way to solve this problem is to make the approximation that the denominator is approximately a function of e alone,

$$F(e+1) \simeq \sum_{\ell=1}^{e} k_\ell\, x_\ell\, (e+1-\ell)^{-\alpha}. \tag{14}$$

This approximation amounts to assuming that the probability distribution $P\{\boldsymbol{\kappa}(e) = \boldsymbol{k}(e), \boldsymbol{\xi}(e) = \boldsymbol{x}(e)\}$ will cause a tight spread of the summation in Eq. (14) around an e-dependent value. Using this expression in Eq. (12), we obtain

$$\begin{aligned} \Delta\bar{k}(e_i,e+1) = \frac{1}{F(e+1)}(1-R)(e+1-e_i)^{-\alpha} \\ \times \sum_{k_i,x_i} k_i\, x_i\, P\{\kappa(e,e_i) \\ = k_i, \xi_i = x_i\}. \end{aligned} \tag{15}$$

As the random variables $\kappa(e,e_i)$ and ξ_i are not independent, we cannot simplify this equation by factorizing the probability into two terms (the probability for κ and the probability for ξ), and Eq. (15) becomes

$$\begin{aligned} \Delta\bar{k}(e_i,e+1) = \frac{1}{F(e+1)}(1-R)(e+1-e_i)^{-\alpha} \\ \times E\{\xi_i E[\kappa(e,e_i)|\xi_i]\}, \end{aligned} \tag{16}$$

where $E[\kappa(e,e_i)|\xi_i]$ is the conditional expectation value for $\kappa(e,e_i)$, assuming ξ_i. However, this expression suggests that we can obtain a more tractable result if we start with the conditional average value $\bar{k}(e_i,e|x_i)$, assuming a specific value x_i for the fitness of editor E_i, instead of starting with the nonconditional average. Then we do not get the final sum in x_i and the sum in k_i just produces the conditional average again. Similar calculations to the ones above lead to the following result, after using approximation (14):

$$\begin{aligned} \bar{k}(e_i,e+1|x_i) = \bar{k}(e_i,e|x_i) + \frac{1}{F(e+1)}(1-R) \\ \times (e+1-e_i)^{-\alpha}\, x_i\, \bar{k}(e_i,e|x_i). \end{aligned} \tag{17}$$

This equation, together with Eq. (6), can be solved by recurrence, and the result for the nonconditional average can then be obtained by

$$\bar{k}(e_i,e) = \sum_{x_i} \bar{k}(e_i,e|x_i)\, P\{\xi_i = x_i\}. \tag{18}$$

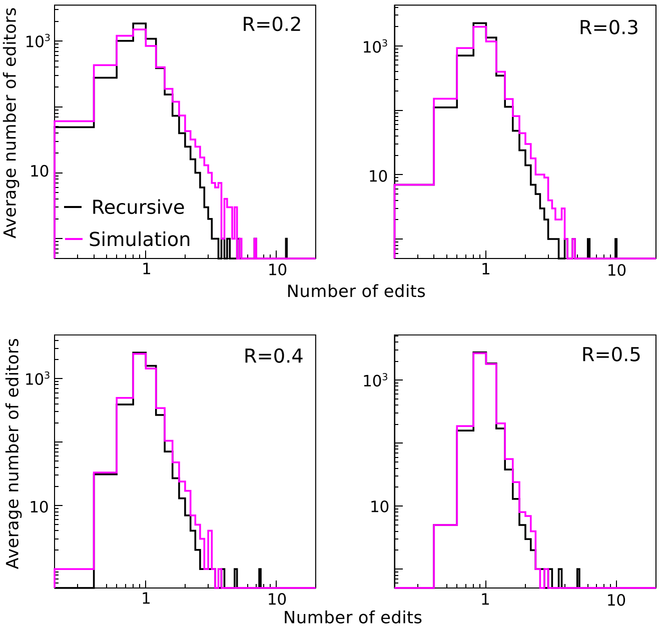

FIG. 7. (Color online) Comparison between the average number of edits per editor obtained with the recursive approach, in black, and simulation, in pink (gray), for four values of R (the ratio between the number of editors and the number of edits in each page, 0.2, 0.3, 0.4, and 0.5). The simulation was obtained with the same ingredients as before: preferential attachment, power law fitness (with $\gamma = 0.90$), and power law aging (with $\alpha = 1.25$) (see the text for the explanation of the different components). The agreement between the recursive calculation and simulation is, in general, good.

$F(e+1)$ can be evaluated by summing Eq. (17) for all values of e_i from 1 to e and using Eqs. (6) and (18). The final result is

$$F(e+1) = \sum_{e_i=1}^{e} \sum_{x_i} (e+1-e_i)^{-\alpha}\, x_i\, \bar{k}(e_i,e|x_i)\, P\{\xi_i = x_i\}. \tag{19}$$

We choose the values of x_i that allow for the calculation of the integral in Eq. (18) by a Gauss-Legendre quadrature. Equation (17) is then solved recursively for those values of x_i, thus providing a value for the average number of edits made by the editor who starts to edit at step e_i, and who has fitness x_i [in the approximation Eq. (19)]. Finally, Eq. (18) provides the results for $\bar{k}(e_i,e)$.

Several numbers of points were tried for the Gauss-Legendre quadrature. It was found that 40 points were enough to obtain convergence, as the results almost did not change with a larger number (we went up to 100 points). The recursive calculation is compared with the simulation in Fig. 7. Both calculations were performed with 5000 edits and the simulation was obtained with an average over 50 000 pages. The agreement is, in general, good for all values of R. The discrepancies are not due to statistical error, but to the approximation in Eq. (14). This approximation seems to be reasonable.

VI. SUMMARY

In this paper, the editing of Wikipedia, which is an available free source of human knowledge, was analyzed. It was shown that this process, far from being random, is instead well reproduced in terms of previous activity, capability to edit, and an aging effect. We successfully reproduced the distribution of both the number of edits per editor for a collection of single pages and its average as shown by real data. It is proposed that the essential ingredients for the probabilistic function are preferential attachment, a power law aging function, and an editor's random fitness, also following a power law distribution. The comparison of real data with the results obtained by simulations with different ingredients, previously found as emergent from human interactions, was shown. An agent-based model was developed, using the best-fitting ingredients, and it successfully reproduces real data, both for the editing of a collection of single pages and for the average over many pages. A recursive expression of the probabilistic function was achieved for the average number of edits for a WP page. The agreement of the analytical approach with the simulation results is good, and is also in accordance with the real data.

ACKNOWLEDGMENTS

Y.G. acknowledges financial support from Grant No. PEst-C/FIS/UI0405/2011, Projecto Unidade I&D 405, Centre for Computational Physics of the University of Coimbra. The work of Y.G. presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Program, initiated by the Belgian State, Science Policy Office.

[1] M. Kardar, *Statistical Physics of Particles* (Cambridge University Press, Cambridge, 2007).

[2] T. Yasseri and J. Kertész, Value production in a collaborative environment sociophysical studies of Wikipedia, J. Stat. Phys. **151**, 414 (2013).

[3] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet, Wikipedias: Collaborative web-based encyclopedias as complex networks, Phys. Rev. E **74**, 016115 (2006).

[4] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia, Phys. Rev. E **74**, 036116 (2006).

[5] V. Zlatić and H. Štefančić, Model of Wikipedia growth based on information exchange via reciprocal arcs, Europhys. Lett. **93**, 58005 (2011).

[6] L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi, Temporal analysis of the Wikigraph, in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, edited by T. Nishida, Z. Shi, U. Visser, X. Wu, J. Liu, B. Wah, W. Cheung, and Y.-M. Cheun (IEEE, New York, 2006).

[7] M. Wattenberg, F. B. Viégas, and K. Hollenbach, Visualizing activity on wikipedia with chromograms, in *Human-Computer Interaction INTERACT 2007*, Lecture Notes in Computer Science Vol. 4663, edited by Baranauskas, P. Palanque, J. Abascal, and S. Barbosa (Springer, Berlin, 2007), p. 272.

[8] D. Laniado and R. Tasso, Patterns of collaboration in Wikipedia, in *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, edited by P. le Bra (ACM, New York, 2011), pp. 201–210.

[9] D. M. Wilkinson and B. A. Huberman, Assessing the value of cooperation in Wikipedia, First Monday **12**, 4 (2007).

[10] H. Simon, On a class of skew distribution functions, Biometrika **42**, 425 (1955).

[11] A. Halfaker, A. Kittur, R. Kraut, and J. Riedl, A Jury of Your Peers: Quality, Experience and ownership in Wikipedia, in *WikiSym'09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, edited by D. Riehle (ACM, New York, 2009), Article no. 15.

[12] R. M. Kimmons, Understanding collaboration in Wikipedia, First Monday **16**, 12 (2011).

[13] G. Bianconi and A.-L. Barabási, Competition and multiscaling in evolving networks, Europhys. Lett. **54**, 436 (2001).

[14] S. N. Dorogovtsev and J. F. F. Mendes, Scaling properties of scale-free evolving networks: Continuous approach, Phys. Rev. E **63**, 056125 (2001).

[15] N. L. Vicent Gómez, Hilbert J. Kappen, and A. Kaltenbrunner, A likelihood-based framework for the analysis of discussion threads, World Wide Web **16**, 645 (2013).

[16] S. N. Dorogovtsev and J. F. F. Mendes, Evolution of networks with aging of sites, Phys. Rev. E **62**, 1842 (2000).

[17] S. Javanmardi, C. Lopes, and P. Baldi, Modeling User Reputation in Wikis, Statistical Analysis and Data Mining: The ASA Data Science Journal **3**, 126 (2010).

[18] http://wwm.phy.bme.hu/.

[19] Y. Gandica, F. S. dos Aidos, and J. Carvalho, The dynamic nature of conflict in Wikipedia, Europhys. Lett. **108**, 18003 (2014).

[20] A.-L. Barabási and R. Albert, Emergence of scaling in random Networks, Science **286**, 509 (1999).

[21] W. Eadie, D. Drijard, F. James, M. Roos, and B. Sadoulet, *Statistical Methods in Experimental Physics* (North-Holland, Amsterdam, 1971), pp. 269–271.

[22] P. L. Krapivsky, S. Redner, and F. Leyvraz, Connectivity of growing random networks, Phys. Rev. Lett. **85**, 4629 (2000).

[23] A.-L. Barabási, R. Albert, and H. Jeong, Mean-field theory for scale-free random networks, Physica A **272**, 173 (1999).

[24] L. Kullmann and J. Kertész, Preferential growth: Exact solution of the time-dependent distributions, Phys. Rev. E **63**, 051112 (2001).

[25] S. Dorogovtsev, J. Mendes, and A. Samukhin, Structure of growing networks with preferential linking, Phys. Rev. Lett. **85**, 4633 (2000).