PHYSICAL REVIEW E **93**, 042130 (2016)

# Growth fluctuation in preferential attachment dynamics

Yasuhiro Hashimoto*
*Division of Information Engineering, Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Ibaraki Prefecture, Japan*

(Received 30 September 2015; published 25 April 2016)

In the Yule-Simon process, creation and selection of words follows the *preferential attachment* mechanism, resulting in a power-law growth in the cumulative number of individual word occurrences as well as the power-law population distribution of the vocabulary. This is derived using mean-field approximation, assuming a continuum limit of both the time and number of word occurrences. However, time and word occurrences are inherently discrete in the process, and it is natural to assume that the cumulative number of word occurrences has a certain fluctuation around the average behavior predicted by the mean-field approximation. We derive the exact and approximate forms of the probability distribution of such fluctuation analytically, and confirm that those probability distributions are well supported by the numerical experiments.

DOI: 10.1103/PhysRevE.93.042130

## I. INTRODUCTION

The Yule-Simon process is a classical mathematical model that describes a branching process in discrete time and state space. It was originally introduced by Yule to explain the population dynamics of biological species in continuous time and discrete state space [1–3], and later modified by Simon into the discrete time and state model [3,4]. In Simon's scheme, the process yields a *word* sequence. At each time step, there is a probability $\alpha$ that the next word in the sequence will be a new vocabulary, and a complementary probability $1-\alpha$, or $\bar{\alpha}$, that it will be selected from the words already existing in the sequence. This process is analogous to that of book reading, where novel or known words appear one after another sequentially. One of the significant results of Yule's and Simon's works is the derivation of the population distribution that follows the power-law form, also known as Zipf's law in the rank-frequency distribution [5].

Now let us denote $i$ as the index of distinct words sorted in the ascending order of time when they are created. The probability of word $i$ being chosen among the existing words is proportional to the number of occurrences of word $i$ in the sequence [6]. This is defined as follows:

$$P(i,t) = n_i(t)/N(t), \tag{1}$$

where $n_i(t)$ is the number of occurrences of word $i$ until time step $t$ and $N(t)$ is the length of the sequence at $t$, that is, the total number of word occurrences until $t$—$N(t) = t$ from the definition. The name of the *preferential attachment* mechanism derives from this proportionality in the word selection, sharing the same idea inherent in the well-known *urn* models [7].

The Yule-Simon process has been used as an archetype of various other dynamic processes such as the Barabási-Albert (BA) graph model [8], which describes the growth of the web. The BA graph grows by adding nodes (web pages) to the graph one by one, resulting in a certain number of edges (hyperlinks) connected to the existing nodes in proportion to their degree (the number of edges belonging to the target node). The BA graph can be thought of as a limiting case of the Yule-Simon process where $\alpha = 1/2$. Indeed, the "node" and "degree" appearing in the BA graph are analogous to the "word" and "word occurrence," respectively, in the Yule-Simon process [9]. Barabási and others analyzed how the node gathers the number of edges in the evolution of the graph and showed that the degree grows in a power-law fashion in the continuum limit of time and degree as follows:

$$k_i(t) \propto (t/t_i)^{1/2}, \tag{2}$$

where $k_i(t)$ is the expected degree of node $i$ at time $t$ and $t_i$ is the time when node $i$ joined the graph. Following the same logic, the expected value of the cumulative number of occurrences of word $i$ at time $t$, denoted by $n_i^*(t)$, is derived as follows:

$$n_i^*(t+\Delta t) = n_i^*(t) + \bar{\alpha} P(i,t)\Delta t. \tag{3}$$

Then, via the integral form

$$\int \frac{dn_i^*}{n_i^*} = \bar{\alpha} \int \frac{dt}{t}, \tag{4}$$

we obtain

$$n_i^*(t) = (t/t_i)^{\bar{\alpha}}, \tag{5}$$

using the initial condition $n_i^*(t_i) = 1$. The homology between Eqs. (2) and (5) implies that the BA graph is actually a particular case of the Yule-Simon process with $\alpha = 1/2$.

The mean-field estimation elucidates the expected behavior of the increase in the cumulative number of word occurrences under the preferential attachment mechanism, as shown above. Even so, we can assume that the individual word occurrence will deviate from the expected value under a certain period of observation. There might be words that occur more frequently than expected and others that appear less frequently. As anomalous behavior often attracts our interest more than ordinary behavior [10], these deviations are the focus of our study. In other words, we are interested in the *individuality* that the growth of an element in a system exhibits. We can attribute such individuality to factors such as the so-called *fitness* [11] of each element, environmental contingency, or the inherent dynamics of the system, including the preferential attachment principle. If we work further on this course—aiming at

*hashi@cs.tsukuba.ac.jp

 

revealing the origin of anomaly in individuality—one of the fruitful approaches would be to look into the correlation latent in the time series of individuals [12,13] or further to consider the intercorrelation between them.

In this study, we aim at revealing the characteristics of the fluctuation which the individual element in the system exhibits under the preferential attachment mechanism, and which the mean-field approximation can hardly capture. We note that there exist many preceding studies related to the fluctuation observed in growing systems—for example, networks [14] and various social systems such as cities, scientific output, human communication, and so on [15–18]. However, in that context, the fluctuation is captured for a *class* of system elements, and the focus is on the growth of the system itself, not on individuals. Our goal is to provide a theoretical baseline against which the magnitude of the individual fluctuation is evaluated. In order to do that, we identify the shape of the probability distribution of the deviation scale that an individual element has in the preferential attachment dynamics.

In the following sections, first we analytically derive the probability distribution of the growth fluctuation that the individual words exhibit in the Yule-Simon process. Then, we check the validity of the formula through a comparison with the results from numerical experiments.

## II. DERIVATION

Let us denote $P[n_i(t) = n]$ as the probability that the cumulative number of occurrences of word $i$ at step $t$ [denoted by $n_i(t)$] equals $n$, and $P[n_i(t) \to n]$ as the probability that $n_i(t)$ becomes $n$ from $n-1$ right at $t$. Introducing $\tau$ to denote the elapsed time from $t_i$, and $s_i = t_i + \tau$ as the time to measure the probabilities, $P[n_i(s_i) = n]$ and $P[n_i(s_i) \to n]$ can be written recursively as follows: For $n = 1$,

$$P[n_i(s_i) = 1] = \prod_{t=t_i}^{s_i-1}\left(\alpha + \bar{\alpha}\frac{t-1}{t}\right) = \frac{\Gamma(t_i)\Gamma(s_i-\bar{\alpha})}{\Gamma(s_i)\Gamma(t_i-\bar{\alpha})}, \tag{6}$$

and for $n = 2$,

$$\begin{aligned} P[n_i(s_i) \to 2] &= P[n_i(s_i-1) = 1]\frac{\bar{\alpha}}{s_i-1} \\ &= \bar{\alpha}\frac{\Gamma(t_i)\Gamma(s_i-1-\bar{\alpha})}{\Gamma(s_i)\Gamma(t_i-\bar{\alpha})}, \\ P[n_i(s_i) = 2] &= \sum_{u=t_i+1}^{s_i}\left\{P[n_i(u)\to 2]\prod_{t=u}^{s_i-1}\left(\alpha+\bar{\alpha}\frac{t-2}{t}\right)\right\} \\ &= \bar{\alpha}\frac{\Gamma(t_i)\Gamma(s_i-2\bar{\alpha})}{\Gamma(s_i)\Gamma(t_i-\bar{\alpha})}\sum_{u=t_i+1}^{s_i}\frac{\Gamma(u-1-\bar{\alpha})}{\Gamma(u-2\bar{\alpha})}. \end{aligned} \tag{7}$$

Equation (6) means that word $i$ is not chosen for an interval $\tau$ after its first appearance. Equation (7) means that at a certain time point $u$ in the interval $[t_i + 1 : s_i]$, word $i$ is chosen only once, and after that it can never be chosen until $s_i$. Further, for $n > 2$, the form of the probabilities becomes more complicated because it has the term of weighted and nested sums of the ratios of Gamma functions in it. However, let us write down a few more values one by one: For $n = 3$,

$$\begin{aligned} P[n_i(s_i) \to 3] &= P[n_i(s_i-1) = 2]\frac{2\bar{\alpha}}{s_i-1} = 2\bar{\alpha}^2\frac{\Gamma(t_i)\Gamma(s_i-1-2\bar{\alpha})}{\Gamma(s_i)\Gamma(t_i-\bar{\alpha})}\sum_{u=t_i+1}^{s_i-1}\frac{\Gamma(u-1-\bar{\alpha})}{\Gamma(u-2\bar{\alpha})}, \\ P[n_i(s_i) = 3] &= \sum_{u=t_i+2}^{s_i}\left\{P[n_i(u)\to 3]\prod_{t=u}^{s_i-1}\left(\alpha+\bar{\alpha}\frac{t-3}{t}\right)\right\} \\ &= 2\bar{\alpha}^2\frac{\Gamma(t_i)\Gamma(s_i-3\bar{\alpha})}{\Gamma(s_i)\Gamma(t_i-\bar{\alpha})}\sum_{u=t_i+2}^{s_i}\left[\frac{\Gamma(u-1-2\bar{\alpha})}{\Gamma(u-3\bar{\alpha})}\sum_{v=t_i+1}^{u-1}\frac{\Gamma(v-1-\bar{\alpha})}{\Gamma(v-2\bar{\alpha})}\right], \end{aligned} \tag{8}$$

and for $n = 4$,

$$\begin{aligned} P[n_i(s_i) \to 4] &= P[n_i(s_i-1) = 3]\frac{3\bar{\alpha}}{s_i-1} \\ &= 6\bar{\alpha}^3\frac{\Gamma(t_i)\Gamma(s_i-1-3\bar{\alpha})}{\Gamma(s_i)\Gamma(t_i-\bar{\alpha})}\sum_{u=t_i+2}^{s_i-1}\left[\frac{\Gamma(u-1-2\bar{\alpha})}{\Gamma(u-3\bar{\alpha})}\sum_{v=t_i+1}^{u-1}\frac{\Gamma(v-1-\bar{\alpha})}{\Gamma(v-2\bar{\alpha})}\right], \\ P[n_i(s_i) = 4] &= \sum_{u=t_i+3}^{s_i}\left\{P[n_i(u)\to 4]\prod_{t=u}^{s_i-1}\left(\alpha+\bar{\alpha}\frac{t-4}{t}\right)\right\} \\ &= 6\bar{\alpha}^3\frac{\Gamma(t_i)\Gamma(s_i-4\bar{\alpha})}{\Gamma(s_i)\Gamma(t_i-\bar{\alpha})}\sum_{u=t_i+3}^{s_i}\left\{\frac{\Gamma(u-1-3\bar{\alpha})}{\Gamma(u-4\bar{\alpha})}\sum_{v=t_i+2}^{u-1}\left[\frac{\Gamma(v-1-2\bar{\alpha})}{\Gamma(v-3\bar{\alpha})}\sum_{w=t_i+1}^{v-1}\frac{\Gamma(w-1-\bar{\alpha})}{\Gamma(w-2\bar{\alpha})}\right]\right\}. \end{aligned} \tag{9}$$

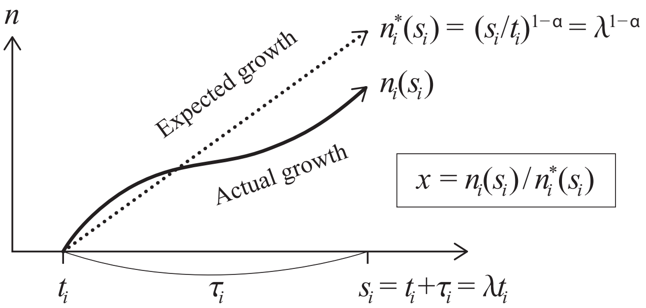


FIG. 1. A diagram of the relationship between the variables depicting the growth of the cumulative number of word occurrences.

Looking at Eqs. (6), (7), (8), and (9) deliberately, we can inductively infer their general form as follows:

$$P[n_i(s_i) = n] = \begin{cases} \frac{\Gamma(t_i)\Gamma(s_i-\bar{\alpha})}{\Gamma(s_i)\Gamma(t_i-\bar{\alpha})} & \text{if } n = 1, \\ (n-1)!\bar{\alpha}^{n-1}\frac{\Gamma(t_i)\Gamma(s_i-n\bar{\alpha})}{\Gamma(s_i)\Gamma(t_i-\bar{\alpha})}\sum_{\phi=t_i+n-1}^{s_i}\mathcal{S}_n(\phi) & \text{if } n > 1. \end{cases} \tag{10}$$

The term $\mathcal{S}_n(\phi)$ is defined as the following recursive function with a depth of $n-1$:

$$\mathcal{S}_n(\phi) = \begin{cases} \frac{\Gamma(\phi-1-\bar{\alpha})}{\Gamma(\phi-2\bar{\alpha})} & \text{if } n = 2, \\ \frac{\Gamma[\phi-1-(n-1)\bar{\alpha}]}{\Gamma(\phi-n\bar{\alpha})}\sum_{\psi=t_i+n-2}^{\phi-1}\mathcal{S}_{n-1}(\psi) & \text{if } n > 2. \end{cases} \tag{11}$$

This is the exact form of the probability distribution wherein the cumulative number of occurrences of word $i$ at time $s_i$ will be $n$. For sufficiently large values of $t_i$ and $s_i$, these equations can be transformed as follows:

$$P[n_i(s_i) = n] \sim \begin{cases} t_i^{\bar{\alpha}}s_i^{-\bar{\alpha}} & \text{if } n = 1, \\ (n-1)!\bar{\alpha}^{n-1}t_i^{\bar{\alpha}}s_i^{-n\bar{\alpha}}\sum_{\phi=t_i+n-1}^{s_i}\mathcal{S}_n(\phi) & \text{if } n > 1, \end{cases} \tag{12}$$

and

$$\mathcal{S}_n(\phi) \sim \begin{cases} \phi^{-\alpha} & \text{if } n = 2, \\ \phi^{-\alpha}\sum_{\psi=t_i+n-2}^{\phi-1}\mathcal{S}_{n-1}(\psi) & \text{if } n > 2, \end{cases} \tag{13}$$

where we use the asymptotic approximation of the ratio of Gamma functions for large $t_i$; $\Gamma(t-a)/\Gamma(t) \sim t^{-a}$. The general solution given by Eqs. (12) and (13) represent one of the principal results of this article.

If $\alpha \to 0$, or $\bar{\alpha} \to 1$, all weighting factors $\phi^{-\alpha}$ in Eq. (13), or all ratios of Gamma functions in Eq. (11), become exactly equal to 1. Consequently, we obtain a specific value of the sum part of Eqs. (10) and (12) as follows:

$$\lim_{\alpha\to 0}\sum_{\phi=t_i+n-1}^{s_i}\mathcal{S}_n(\phi) = \frac{(\tau-n+2)^{n-1}}{(n-1)!}, \tag{14}$$

which is the volume of an $(n-1)$-dimensional triangular pyramid where all of the edges aligned to a corresponding basis vector have the length $\tau - n + 2$. Substituting Eq. (14) into Eq. (12), we obtain a relatively simple form, as follows:

$$\lim_{\alpha\to 0} P[n_i(s_i) = n] = t_i s_i^{-n}(\tau-n+2)^{n-1}. \tag{15}$$

Alternatively, in the case of larger $\alpha$ such as 1/2 in the BA graph, it is unclear whether a simple form like Eq. (15) is available, in which case we have to numerically calculate Eqs. (12) and (13) directly, if needed. A naive calculation of all terms in the nested sum would require approximately $\tau^{n-1}$ operations, which is practically infeasible. However, by caching and reusing the values $n$ and $\phi$ generated with each calculation of $\mathcal{S}_n(\phi)$ the total number of calculations is drastically reduced. This is how we numerically calculate the solution to Eqs. (12) and (13) in the next section.

The following discussion is based on Eq. (15), which is the particular solution for a sufficiently small $\alpha$. What we ultimately want to know is the scale of the deviation of the cumulative number of word occurrences from their expected value. The absolute size of the deviation depends on $t_i$ as well as $\tau$. Equation (5) indicates that the cumulative number of word occurrences increases more slowly as $t_i$ is larger. Therefore, for a given $\tau$, the deviation of such words should be smaller relative to words with smaller $t_i$. For this reason, the size of the deviation should be normalized depending on $t_i$ using different values of $\tau$. To do this, we introduce a scale factor $\lambda$ as follows:

$$s_i = t_i + \tau_i = \lambda t_i. \tag{16}$$

Here $\tau$, the observation period of the deviation, varies word by word, and $\lambda$ is constant for every word and greater than one

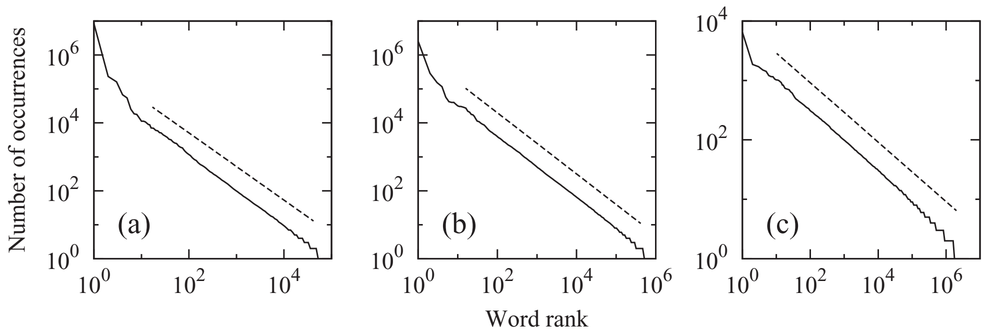


FIG. 2. The rank-frequency distribution in the case of (a) $\alpha = 0.01$, (b) 0.1, and (c) 0.5. Solid and dotted lines show, respectively, the simulation results and theoretical curves as an eye guide, which is given by Simon's rate equation approach [4] and proportional to [word rank]$^{\bar{\alpha}}$.

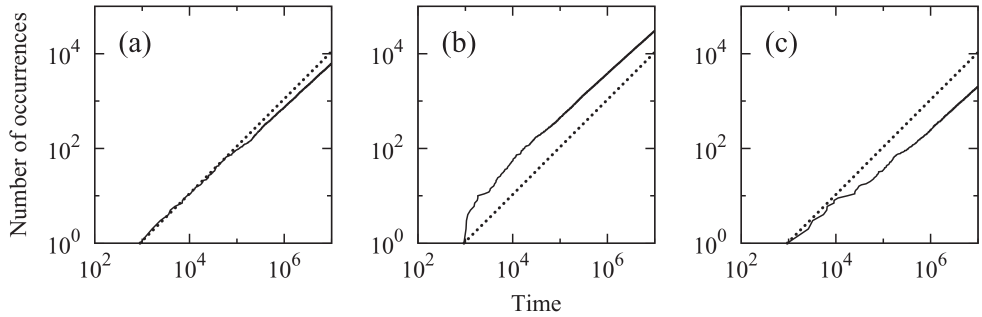


FIG. 3. The growth of the cumulative number of occurrences of three sampled words created at the (a) 89th, (b) 90th, and (c) 91st orders in the case of $\alpha = 0.1$. Solid and dotted lines show, respectively, the simulation results and the corresponding expected growth curves given by the mean-field estimation Eq. (5).

by definition. Substituting Eq. (16) into Eq. (5), we obtain

$$n_i^*(s_i) = \left(\frac{\lambda t_i}{t_i}\right)^{\bar{\alpha}} = \lambda^{\bar{\alpha}}. \tag{17}$$

This temporally normalized expected value of the cumulative number of word occurrences, $\lambda^{\bar{\alpha}}$, is used as a *reference* value to measure the scale of the deviation for each word. Replacing $n$ in $P[n_i(s_i) = n]$ with $x\lambda^{\bar{\alpha}}$, we define the probability

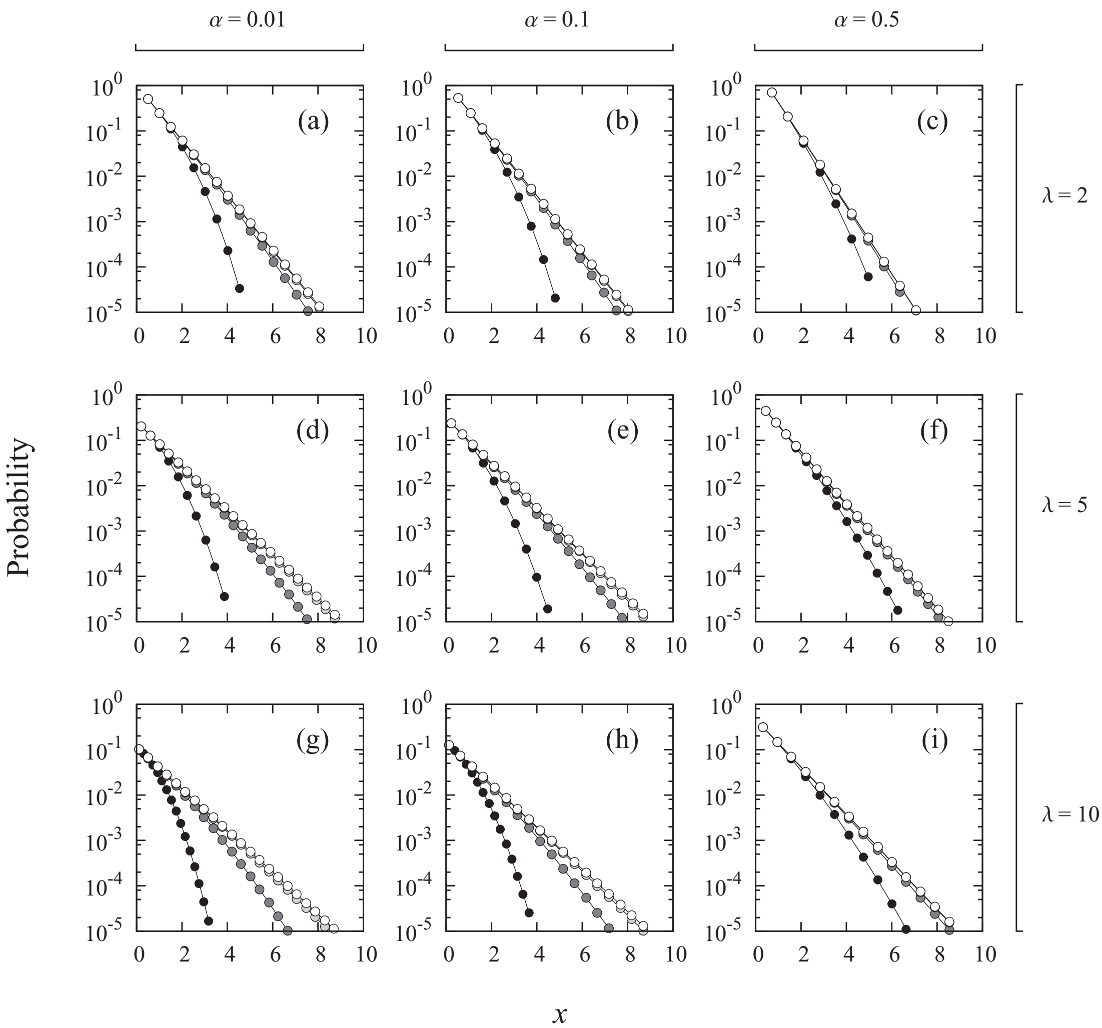


FIG. 4. Numerical solutions of the general form Eq. (12) for $t_i = 10^1$ (black circles), $10^2$ (dark gray), $10^3$ (light gray), and $10^4$ (white) under the condition of different $\lambda$ and $\alpha$ values: $\lambda = 2$, 5, 10 and $\alpha = 0.01$, 0.1, 0.5. For example, (a) shows the case of $\alpha = 0.01$ and $\lambda = 2$. The horizontal and vertical axes are the value of $x$ and $P(x)$, respectively. Note that the vertical axis is in logarithmic scale, whereas the horizontal axis is in linear, so that exponential decay draws a descending straight line.

distribution of the deviation scale as

$$P(x) \equiv P[n_i(s_i) = x\lambda^{\bar{\alpha}}]. \tag{18}$$

The idea of the deviation scale factor $x$ is depicted in Fig. 1. Thus, the particular solution Eq. (15) is transformed as follows:

$$\lim_{\alpha\to 0} P(x) = t_i(\lambda t_i)^{-x\lambda}[(\lambda-1)t_i - x\lambda + 2]^{x\lambda-1}$$
$$= \lambda^{-1}\left(1 - \frac{1}{\lambda} - \frac{x-2/\lambda}{t_i}\right)^{x\lambda-1}. \tag{19}$$

For a large majority of words, supposing $t_i \gg x \sim 1$, Eq. (19) is approximated as

$$\lim_{\alpha\to 0} P(x) = \frac{1}{\lambda-1}\left(1-\frac{1}{\lambda}\right)^{x\lambda}, \tag{20}$$

which is independent of $t_i$. Hence, this formula represents the probability distribution of the fluctuation for all words. This concise relationship is the other principal result of this article. The particular solution Eq. (20) clearly shows that the probability distribution of the deviation scale decays exponentially.

## III. NUMERICAL VALIDATION

We confirm that the general form Eq. (12) and the particular form for a sufficiently small $\alpha$ Eq. (20) are good predictors of the actual behavior of the growth fluctuation in the Yule-Simon process. First, we ran the numerical simulation of the Yule-Simon process for different $\alpha$ values of 0.01, 0.1, and 0.5, where the total number of word occurrences is $10^7$. Consequently, the final vocabulary sizes are approximately $10^5$, $10^6$, and $5 \times 10^6$, respectively. Figure 2 shows the rank-frequency distribution for each $\alpha$ value, and we see that Zipf's law actually holds in every case with the power exponent $\bar{\alpha}$ as predicted by Simon's rate equation approach. There is no discrepancy between the simulation result and theoretical expectation in the population distribution. However, we see a

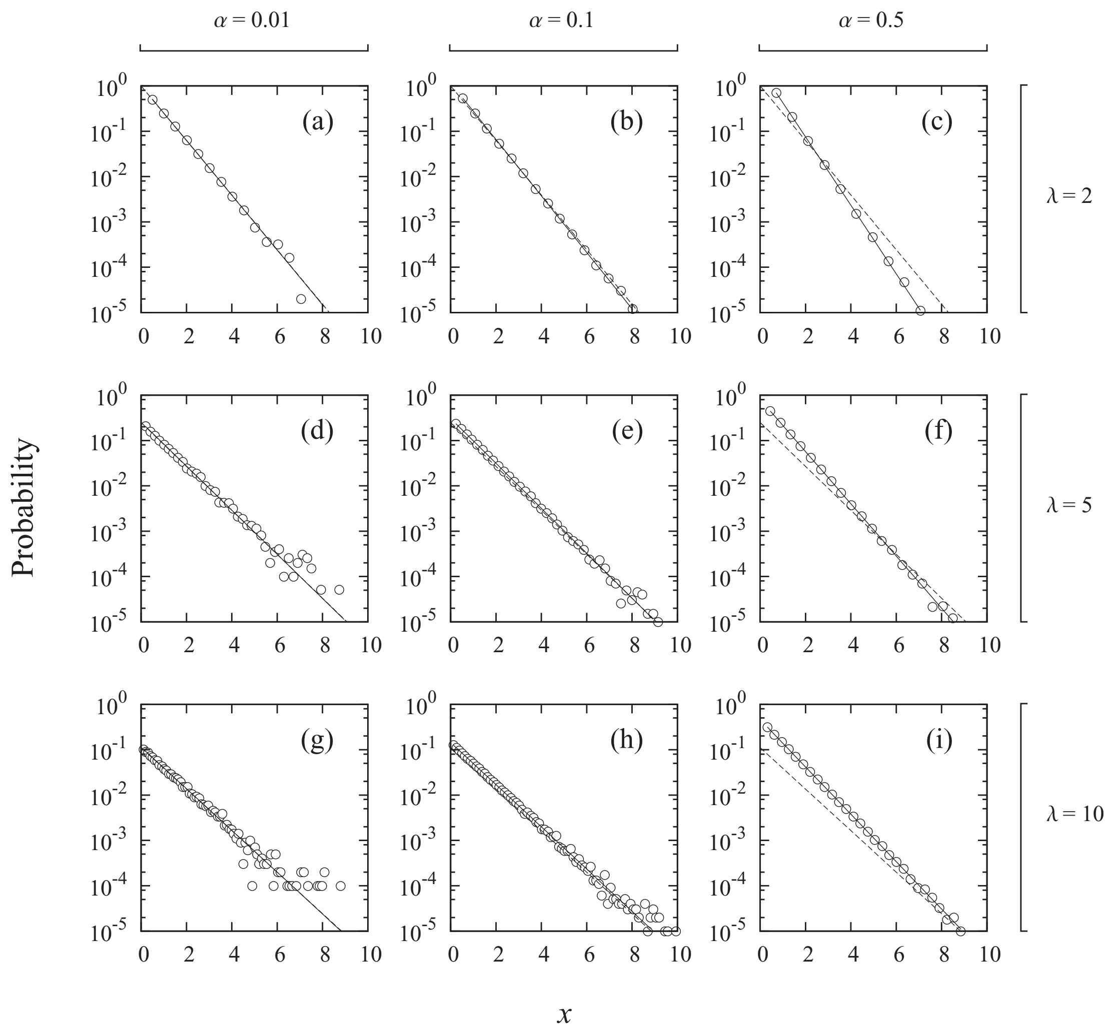


FIG. 5. Comparison between the numerical results of the simulation (white circles), the numerical solution of the general form for $t_i = 10^4$ (solid lines), and analytic curves drawn from the particular solution for a sufficiently small $\alpha$ (dotted lines), for different $\alpha$ and $\lambda$ values. In the case of small values of $\alpha$, as shown in (a), (d), and (g), all results exhibit a good match. At the same time, the mismatch between the particular solution and the others increases for large values of $\alpha$, as shown in (c), (f), and (i).

divergence from the mean-field expectation at the individual level. Figure 3 shows three typical patterns of the growth of the cumulative number of word occurrences in the simulation, especially in the case of $\alpha = 0.1$. These three sampled words (89th, 90th, and 91st) exhibit, respectively, an increase of (a) following, (b) exceeding, and (c) falling behind their expected growth curve given by the mean-field estimation (5). They are created close to each other in terms of time, and yet nevertheless exhibit differing growth courses. This word-by-word individuality observed in the growth fluctuation has not received theoretical treatment before and is what we have focused on in this study.

Now we go back to the expression of the general solution and examine how it behaves by solving Eqs. (12) and (13) numerically. Having introduced the deviation scale $x$ in Eq. (18), $n$ can be replaced with $x\lambda^{\bar{\alpha}}$ in Eqs. (12) and (13), and the probability $P(x)$ is calculated over a certain range of $x$ for some $t_i$ under different $\lambda$ and $\alpha$ values: $\lambda = 2$, 5, 10, and the values of $\alpha$ used in the simulation above. Figure 4 shows the result of $t_i = 10^1$ (black circles), $10^2$ (dark gray), $10^3$ (light gray), and $10^4$ (white); increasing $t_i$ (from black to while circles) causes the distribution to converge to the unique functions in all cases, which are shown as descending straight lines in the figure. Therefore, the resulting probability distribution assembled over all $t_i$ for given $\alpha$ and $\lambda$ is approximately represented by the distribution for sufficiently large $t_i$, and presumably has the form of an exponential function, as we derived in Eq. (20). We compare these numerical results of the general solution with empirically measured $P(x)$ in the simulation (see Fig. 5). The white circles, solid lines, and dotted lines in the figure are, respectively, the simulation result, the general solution for $t_i = 10^4$, and the analytic curve of the particular solution. For all $\alpha$ and $\lambda$ values, the general solution exhibits a good match with the simulation result, and we conclude that our inductive derivation of Eqs. (10) and (12) is valid. For small $\alpha$ values, we also see the particular solution exhibits a good match with the simulation result and the general solution. However, the mismatch between them increases for large $\alpha$ values. This is consistent with our assumption concerning the asymptotic behavior of the particular solution.

In order to obtain a quantitative form of the general solution, we fit the simulation result to the exponential function using the least square method as follows:

$$P(x) = P(x = x_{\min}) \exp\left(-\frac{x - x_{\min}}{x_c}\right)$$
$$= \lambda^{-\bar{\alpha}} \exp\left(-\frac{x - \lambda^{-\bar{\alpha}}}{x_c}\right), \qquad (21)$$

where the only fitted parameter, $x_c$, is the characteristic scale of $x$, and $x_{\min}$ is the value of $x$ when the word occurs only once:

$$x_{\min} = \frac{\min[n_i(s_i)]}{n_i^*(s_i)} = \frac{1}{\lambda^{\bar{\alpha}}}. \qquad (22)$$

$P(x)$ takes its maximum value when $x = x_{\min}$, and the maximum value, $P(x = x_{\min})$, is nothing but the solution of Eq. (6), which is the probability of a word occurring only once. In the asymptotic limit of large $t_i$, the probability Eq. (6) becomes $\lambda^{-\bar{\alpha}}$, thus reducing to Eq. (21). Table I and Fig. 6 show the fitting result. The value of $x_c$ appears to approach 1 for large values of $\lambda$, and to do so more quickly as $\alpha$ is smaller. Thus, we have reached our result in Eq. (20) via the numerically evaluated general solution, which reduces to $\lambda^{-1}\exp(-x + \lambda^{-1})$ in the asymptotic limit of large $\lambda$. This is actually the specific case of Eq. (21) for $x_c = 1$ for sufficiently large $\lambda$ and $\bar{\alpha} = 1$, which we assumed in deriving the particular solution.

TABLE I. Fitted $x_c$ and its standard error of the simulation result in the least-square approximation of Eq. (21) for different values of $\alpha$ and $\lambda$.

| $\alpha$ | $\lambda$ | $x_c$ | Std. Err. of $x_c$ |
|---|---|---|---|
| 0.01 | 2 | 0.720214 | ± 0.384% |
| 0.01 | 5 | 0.887837 | ± 0.5654% |
| 0.01 | 10 | 0.949716 | ± 0.605% |
| 0.1 | 2 | 0.69894 | ± 0.1114% |
| 0.1 | 5 | 0.877731 | ± 0.1502% |
| 0.1 | 10 | 0.935541 | ± 0.1814% |
| 0.5 | 2 | 0.5761 | ± 0.03373% |
| 0.5 | 5 | 0.75459 | ± 0.0219% |
| 0.5 | 10 | 0.832912 | ± 0.08112% |

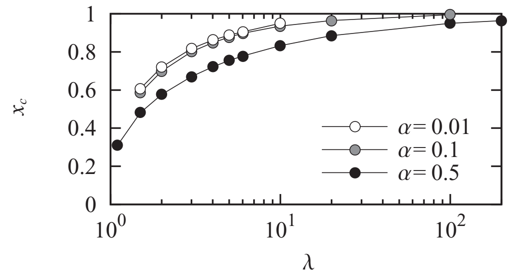


FIG. 6. The relationship between $\lambda$ (the observation timescale), and fitted $x_c$ (the characteristic scale of the deviation). White, light gray, and black circles show the qualified results that have the standard error less than 1%, for $\alpha = 0.01$, 0.1 and 0.5, respectively.

## IV. CONCLUSION

In this article, we have opened a new perspective on preferential attachment dynamics. We have examined the growth fluctuation exhibited by individual elements in a system, and derived the probability distribution of the scale of the deviation from the mean-field estimation, based on the analysis of the Yule-Simon process. The distribution function was represented by the particular form for a sufficiently small $\alpha$, the vocabulary creation rate, that shows exponential decay with an increasing deviation scale. We also obtained the general form of the probability distribution of the number of individual word occurrences depending on the time the word is created, and showed numerically that the solution also follows the exponential decay in the growth fluctuation. We confirmed that the theoretical solutions and the simulation results matched well, concluding that our inductive derivation seems suitable.

The obtained result illustrates that the magnitude of the deviation from the mean-field estimation rapidly drops around the expected value (we saw the characteristic scale of the deviation was approximately 1). This implies that the preferential attachment mechanism indeed makes *the rich get richer*. However, it is strictly bounded by the growth horizon determined by the time the individual element joined the system. Hence, there is no significant difference among individuals if they are temporally scaled. Our work sets the stage for future empirical work to identify any phenomena with individuals exhibiting growth fluctuation that diverges from our theoretical estimation here.

## ACKNOWLEDGMENTS

We sincerely express our gratitude to T. Ikegami, M. Oka, and K. Sato for many fruitful discussions and suggestions, and also to the voluntary reviewers for their valuable contribution to improving the quality of the article.

[1] G. U. Yule, Philos. Trans. R. Soc. London B **213**, 21 (1924).

[2] N. Bacaër, Yule and evolution (1924), in *A Short History of Mathematical Population Dynamics* (Springer-Verlag, London, 2011), pp. 81–88.

[3] M. V. Simkin and V. P. Roychowdhury, Phys. Rep. **502**, 1 (2011).

[4] H. A. Simon, Biometrika **42**, 425 (1955).

[5] G. K. Zipf, *The Psycho-Biology of Language* (Houghton Mifflin Company, Boston, 1935).

[6] Note that Simon himself assumed rather a weaker condition that the probability is not defined for the individual word but for the *class* of words that have the same number of occurrences in the sequence .

[7] H. Mahmoud, *Pólya Urn Models* (Chapman and Hall, CRC, Boca Raton, 2008).

[8] A.-L. Barabási, R. Albert, and H. Jeong, Physica A **272**, 173 (1999).

[9] S. Bornholdt and H. Ebel, Phys. Rev. E **64**, 035104 (2001).

[10] P. L. Krapivsky and S. Redner, Phys. Rev. Lett. **89**, 258703 (2002).

[11] G. Bianconi and A.-L. Barabási, Europhys. Lett. **54**, 436 (2001).

[12] A.-L. Barabási, Nature **435**, 207 (2005).

[13] D. Rybski, S. V. Buldyrev, S. Havlin, F. Liljeros, and H. A. Makse, Sci. Rep. **2**, 560 (2012).

[14] P. L. Krapivsky and S. Redner, J. Phys. A: Math. Gen. **35**, 9517 (2002).

[15] X. Gabaix, Quart. J. Econ. **114**, 739 (1999).

[16] H. D. Rozenfeld, D. Rybski, J. S. Andrade, Jr., M. Batty, H. E. Stanley, and H. A. Makse, Proc. Natl. Acad. Sci. USA **105**, 18702 (2008).

[17] K. Matia, L. A. Nunes Amaral, M. Luwel, H. F. Moed, and H. E. Stanley, J. Am. Soc. Inf. Sci. Technol. **56**, 893 (2005).

[18] D. Rybski, S. V. Buldyrev, S. Havlin, F. Liljeros, and H. A. Makse, Proc. Natl. Acad. Sci. USA **106**, 12640 (2009).