Preferential attachment and growth dynamics in complex systems

Kazuko Yamasaki,[1,2] Kaushik Matia,[1] Sergey V. Buldyrev,[3] Dongfeng Fu,[1] Fabio Pammolli,[4,5] Massimo Riccaboni,[4,5] and H. Eugene Stanley[1]

[1]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA*
[2]*Tokyo University of Information Sciences, Chiba City 265-8501, Japan*
[3]*Department of Physics, Yeshiva University, 500 West 185th Street, New York, New York 10033, USA*
[4]*Faculty of Economics, University of Florence, 50127 Florence, Italy*
[5]*IMT Institute for Advanced Studies, Via S. Micheletto 3, 55100 Lucca, Italy*

(Received 26 June 2004; revised manuscript received 6 February 2006; published 21 September 2006)

Complex systems can be characterized by classes of equivalency of their elements defined according to system specific rules. We propose a generalized preferential attachment model to describe the class size distribution. The model postulates preferential growth of the existing classes and the steady influx of new classes. According to the model, the distribution changes from a pure exponential form for zero influx of new classes to a power law with an exponential cut-off form when the influx of new classes is substantial. Predictions of the model are tested through the analysis of a unique industrial database, which covers both elementary units (products) and classes (markets, firms) in a given industry (pharmaceuticals), covering the entire size distribution. The model's predictions are in good agreement with the data. The paper sheds light on the emergence of the exponent $\tau \approx 2$ observed as a universal feature of many biological, social and economic problems.

DOI: 10.1103/PhysRevE.74.035103 PACS number(s): 89.75.Fb, 05.70.Ln, 89.75.Da, 89.65.Gh

Many complex systems of interest to physicists, biologists, and social scientists [1–6] share two basic similarities in their growth dynamics. (i) The system does not have a steady state but is growing. (ii) Basic units are born, they agglomerate to form classes, and classes grow in size according to a rule of proportional growth [7]. In biological systems, units could be bacteria, and classes would be bacterial colonies. In the context of economic systems, units could be products, and classes would be firms. In social systems, units could be human beings, and classes would be cities.

The probability distribution function $p(k)$ of the class size k of the systems mentioned above has been shown to follow a universal scale-free behavior $p(k) \sim k^{-\tau}$ with $\tau \approx 2$ [1–3,8]. Other possible values of τ are discussed and reported in [9]. Also, for most of the systems $p(k)$ has an exponential cutoff, which is often assumed to be a finite-size effect of the databases analyzed. Several models [4,10–13] explain $\tau \approx 2$ but none explains the exponential cutoff of $p(k)$. Moreover, the models describing $p(k) \sim k^{-\tau}$ are not suitable to describe simultaneously systems for which $p(k) \sim \exp(-\gamma k)$.

In this paper, we present a model with a simple set of rules to describe $p(k)$ for the entire range of k, i.e., a power law with an exponential cutoff. We show that the exponential cutoff of the power law is not due to finite size, but is an effect of the finite-time interval of the evolution. We show how the functional form of $p(k)$ is determined by different scenarios of the model, changing from a pure exponential to a pure power law (with $\tau \approx 2$), via a power law with an exponential cutoff. The predictions of the model are then tested through the analysis of a unique industrial database [14,15], which covers both elementary units (products) and classes (markets, firms) in a given industry (pharmaceuticals).

The model consists of the following rules:

1. At time $t=0$ there exist N classes, each with a single unit [16].

2. At each step (a) with probability b $(0 \leq b \leq 1)$, a new class with a single unit is born; (b) with probability λ $(0 < \lambda \leq 1)$, a randomly selected class grows one unit in size. The selection of the class that grows is made with probability proportional to the number of units it already has ("preferential attachment"); (c) with probability μ $(0 < \mu < \lambda)$, a randomly selected class shrinks one unit in size. The selection of the class that shrinks is done with probability proportional to the number of units it already has ("preferential detachment").

In the continuum limit, the above rules give rise to a master equation for the class size PDF $p(k,t_i,t)$, which is the probability at time t, for a class i introduced at time t_i, to have k units,

$$\frac{\partial p(k,t_i,t)}{\partial t} = \lambda \frac{(k-1)}{n(t)} p(k-1,t_i,t) + \mu \frac{(k+1)}{n(t)} p(k+1,t_i,t) - (\lambda+\mu)\frac{k}{n(t)} p(k,t_i,t), \qquad (1)$$

where $n(t) \equiv N + (\lambda - \mu + b)t$ is the total number of units at simulation step t. Equation (1) is transformed to the master equation of birth and death processes [17] by a new variable s, where $dt/ds = n(t)$ and $\bar{p}(k,s_i,s) \equiv p(k,t_i(s_i),t(s))$. The master equation for $\bar{p}(k,s_i,s)$ has the same form as Eq. (1) after replacing t by s, $p(k,t_i,t)$ by $\bar{p}(k,s_i,s)$, and $n(t)$ by 1, respectively. From the well-known solution of birth and death processes under the initial condition $\bar{p}(1,s_i,s_i)=1$[18], we obtain the solution after transforming back from s to t,

$$p(k,t_i,t) = \begin{cases} \dfrac{\mu}{\lambda} \eta_{t_i,t} & (k=0) \\ (1-\eta_{t_i,t})\left(1-\frac{\mu}{\lambda}\eta_{t_i,t}\right)\eta_{t_i,t}^{k-1} & (k>0) \end{cases} \qquad (2)$$

with

TABLE I. The ATC hierarchical classification. The ATC categorizes drugs at four levels of aggregation according to the organ or system on which they act and their chemical, pharmacological, and therapeutic properties. There are 13 main groups (level A) and 84 pharmacological, subgroups (level B). The levels C and D are pharmacological/therapeutic subgroups. Medicinal products, such as bisphosphonates in the example, are classified according to the main therapeutic use of the main active ingredient. The basic principle is one ATC code for each pharmaceutical formulation. The WHO is responsible to manage the ATC. Over the period of our empirical analysis, the number of classes of levels A and B has remained constant, while the number of classes in levels C and D increased by 3% and 5%, respectively.

Level	Type	N	Code	Content
A	Anatomical main group	13	M	Musculo-skeletal system
B	Therapeutic subgroup	84	M05	Drugs for treatment of bone diseases
C	Pharmacological subgroup	259	M05B	Drugs affecting bone structure and mineralization
D	Chemical subgroup	432	M05BA	Bisphosphonates

$$\eta_{t_i,t} = \frac{1 - R^{\alpha}}{1 - \frac{\mu}{\lambda}R^{\alpha}}, \tag{3}$$

where $R \equiv [t_i + N(\lambda - \mu + b)^{-1}]/[t + N(\lambda - \mu + b)^{-1}]$ and $\alpha \equiv (\lambda - \mu)(\lambda - \mu + b)^{-1}$. $p(k,t_i,t) \propto \eta_{t_i,t}^{k-1}$ is obviously an exponential function of k. Finally, one can obtain $p(k,t)$ by averaging $p(k,t_i,t)$ over units introduced at different t_i as follows:

$$p(k,t) = \frac{N}{N+bt}p(k,0,t) + \frac{b}{N+bt}\int_0^t dt_i p(k,t_i,t). \tag{4}$$

The first term, I_1, is

$$I_1 \propto \exp(-\gamma k) \quad \text{with } \gamma = -\log \eta_{0,t} \sim t^{-\alpha}. \tag{5}$$

TABLE II. The evolution of the number of classes N for different levels of the PHID over 10 years. There are three different cases of N in 6 levels: (i) For levels A and B there is no birth or death of classes (i.e., the number of newly born classes N_b is 0 and the number of dead classes N_d is also 0. (ii) For levels C and D system grows not only with birth and death of units inside classes but also with the birth of classes. The system grows with the birth of new classes to the final N_f classes (259 for level C and 432 for level D). (iii) From the table, the values of $b/(\lambda - \mu)$ estimated to be $N_{b,L}/(N_{b,p} - N_{d,p})$ are 0.0009 (level C), 0.002 (level D), and 0.049 (firms).

Level	A	B	C	D	firms	products
N_f	13	84	259	432	3913	48819
N_b	0	0	8	20	458	12645
N_d	0	0	0	0	252	3361

TABLE III. Correlation coefficient $C(k_b,k_e)$ between the number of born units k_b and existing number of units k_e in classes and $C(k_d,k_e)$ between the number of dead units k_d and k_e in classes for each level in the PHID. The observed correlations justify the assumptions in our model: the preferential birth and death of units (rules 2b and 2c).

Level	A	B	C	D	firms
$C(k_b,k_e)$	0.93	0.87	0.84	0.82	0.70
$C(k_d,k_e)$	0.88	0.86	0.80	0.78	0.75

To obtain the second term I_2 we first substitute $p(k,t_i,t)$ from Eq. (2) in Eq. (4), then change the variable of integration from t_i to η. Hence

$$I_2 = \frac{b\left(t + \frac{N}{\lambda - \mu + b}\right)\left(1 + \frac{b}{\lambda - \mu}\right)\left(1 - \frac{\mu}{\lambda}\right)}{(N+bt)} \times \int_0^{\eta_{0,t}} d\eta \left(\frac{1-\eta}{1 - \frac{\mu}{\lambda}\eta}\right)^{1+[b/(\lambda-\mu)]} \eta^{k-1}. \tag{6}$$

In the limit of $t \to \infty$, we have $\eta_{0,t} \to 1$. Since $(1 - \mu\eta/\lambda)^{-1} \approx 1 + \mu\eta/\lambda$, Eq. (6) can be integrated giving the Yule distribution [13]

$$I_2 = \left(1 + \frac{b}{\lambda - \mu}\right)\left(1 - \frac{\mu}{\lambda}\right)\sum_{m=0}^{\infty} \frac{\left(\frac{\mu}{\lambda}\right)^m}{m!} \frac{\left(\frac{b}{\lambda-\mu} + m\right)!}{\left(\frac{b}{\lambda-\mu}\right)!} \times \int_0^1 d\eta (1-\eta)^{1+[b/(\lambda-\mu)]} \eta^{m+k-1}. \tag{7}$$

In the limit of $t \to \infty$, we obtain from Eq. (7) [19]

$$I_2 \propto k^{-\{2+[b/(\lambda-\mu)]\}}, \tag{8}$$

in which an exponential function has been transformed into a power-law function by integration. This situation is analogous to the one described by the standard preferential attachment model [3], where the power-law distribution also follows from the Yule distribution. In the limit of fixed time t and $k \to \infty$, $I_2 \propto \exp(-\gamma k)/k$, which decays faster than Eq. (5), implying that the distribution of class size k for new classes has an exponential cutoff faster than for the old classes. Thus, the full solution of Eq. (1) is a power law [Eq. (8)] with an exponential cutoff [Eq. (5)]. We observe that these two terms are of the same order in the range $k \gtrsim t^{\alpha}$ for large finite t.

We next present a mean-field interpretation of the result $\tau \approx 2$. At any time t_0, the number of units in the already-existing classes is $n(t_0)$. Suppose a new class consisting of one unit is born at time t_0. According to rules 2b and 2c, its growth rate is proportional to $1/n(t_0)$. Neglecting the effect of the influx of new classes on $n(t_0)$, the average size k of

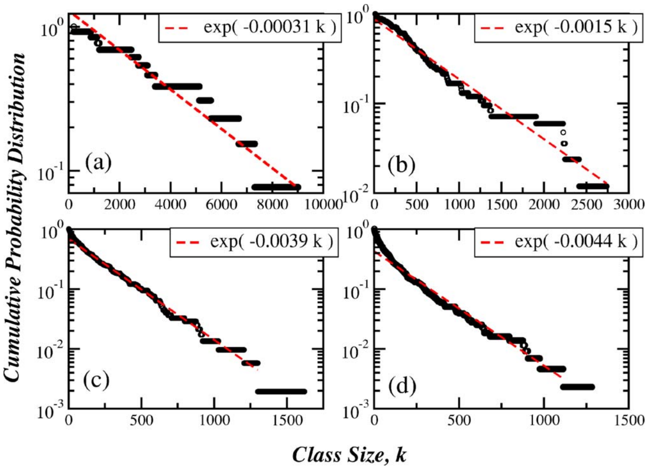

FIG. 1. (Color online) Empirical results for the cumulative probability distribution, $p(k)$, of class size k at different levels. (a)–(d) correspond to levels A–D, respectively. Symbols represent data points in each level (a)–(d), while dashed lines are predictions of the model. The cumulative probability distributions for all levels are reasonably well fit by pure exponentials, as predicted by the model.

this class born at t_0 is proportional to $1/n(t_0)$. So the classes that were born at times $t>t_0$ tend to have an average size measured in terms of k that is smaller than the one of older classes. If we sort the classes according to size, the rank $R(k)$ of a class is proportional to its age $R(k)\propto t_0$. Thus, $k\sim 1/n(t_0)\sim 1/t_0\sim 1/R(t_0)$, coherently with the standard formulation of the Zipf law [1] according to which the size of a class k is inversely proportional to its rank. If we take into account the decrease of the growth rate with the influx of new classes, one can show after some algebra that $k\sim R^{-(\lambda-\mu)/(\lambda-\mu+b)}$, which includes $k\sim R^{-1}$ as a limiting case for $b\to 0$. Since $R(k)$ is the number of classes whose sizes are larger than k, we can write in the continuum limit $R(k)\sim\int_k^\infty p(k)dk$, and hence $p(k)\sim k^{-2-b/(\lambda-\mu)}$.

The full solution of Eq. (1), a power law with an exponential cutoff, can be interpreted as follows. We start with N classes that are colored red, and let the newly born classes be colored blue. Due to the preferential attachment rule, the red classes have on average a number of units that is larger than the blue classes. Thus for large k, $p(k)$ is governed by the exponential distribution of the red classes (*case i*), while for small k, $p(k)$ is governed by the power-law distribution of the blue classes (*case ii*).

We now test the predictions of the model using the pharmaceutical industry database (PHID), a microlevel economic database that allows a fine-grained decomposition of the statistical properties of growth dynamics of business firms in a given industry. PHID records quarterly sales figures of 48 819 pharmaceutical products commercialized in the European Union and North America from September 1991 to June 2001. The products are then classified into different hierarchical levels based on the Anatomic and Therapeutic Classification (ATC) (Table I). Each level has a specific number of classes (Table II).

We observe that, at all different levels, there are positive correlations between the number of units (products) that en-

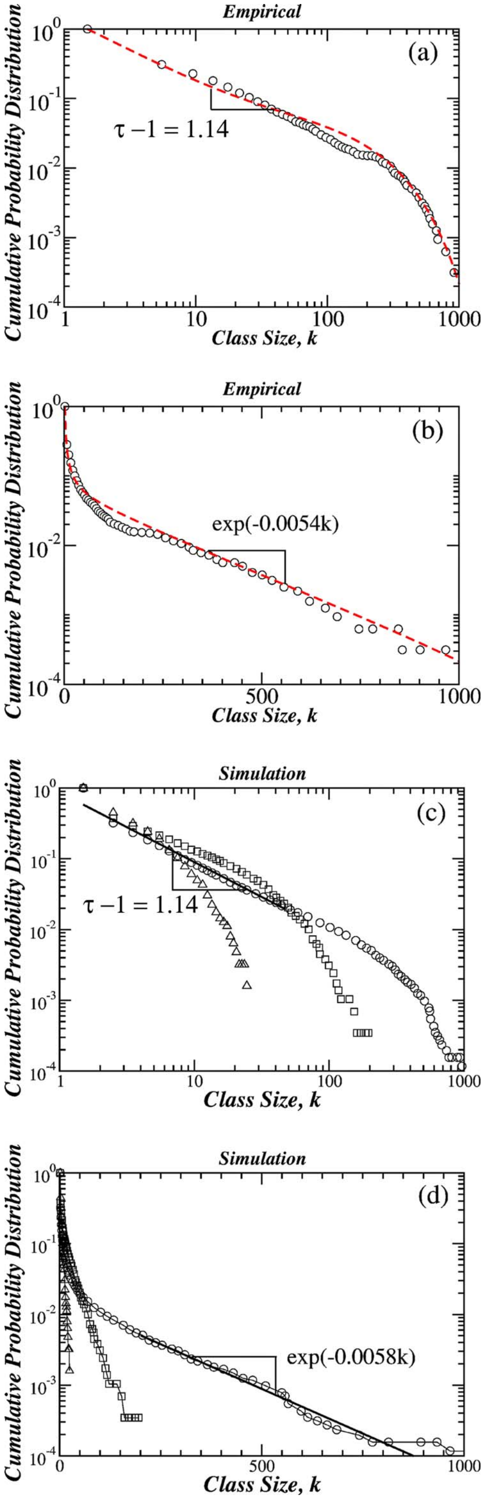

FIG. 2. (Color online) Comparison of empirical results for firms of the PHID and the simulation results. (a) and (c) Log-log plots of the cumulative probability distribution of the class sizes show a power-law decay $k^{-(\tau-1)}$ with $\tau\approx 2$ for $k<200$. The fit (dashed line) in (a) and (b) is obtained by numerical calculation of Eq. (4) with $b=0.1$, $t/N=200$, $\lambda=0.77$, and $\mu=0.33$. (b) and (d) Log-linear plots of the cumulative probability distribution, showing exponential decay for $k>200$. In (c) and (d), ○, □, and △ show the distribution for $t=200\,000$, 20 000, and 2000, respectively. Note that the exponential function gradually changes into a power-law function.

TABLE IV. Comparison of values of the parameters in the model using the data and from the model. γ of the data is estimated by regression in Fig. 1, and γ of the model is estimated using $\gamma = -\log \eta_{0,t}$ [$\eta_{0,t}$ in Eq. (3) is estimated by the value of $b/(\lambda-\mu)$ and N, which is the solution of two equations: $N+(\lambda-\mu+b)t = 48819$ and $N+bt = N_{f,L}$, based on Table II]. τ of the data is estimated by regression in Fig. 2(a) and τ of the model is estimated using $\tau = 2+b/(\lambda-\mu)$ with the numbers of Table II.

(Level)	γ (A)	γ (B)	γ (C)	γ (D)	γ (firms)	τ (firms)
data	0.00031	0.0015	0.0039	0.0044	0.0054	2.14
model	0.00020	0.0013	0.0033	0.0050	0.0173	2.05

ter or exit and the number of units in the classes (Table III). This empirical observation is consistent with a preferential birth and death process (rules 2b and 2c), as described in the model.

For levels A and B, the number of classes did not change during the period of observation and we find an exponential distribution [Figs. 1(a) and 1(b)] as predicted by the limiting *case i* of the model. For levels C and D, a weak departure from the exponential functional form [Figs. 1(c) and 1(d)] can be accounted for within the model, if we consider a slight growth in the number of classes.

The full solution predicted by the model, i.e., a power law followed by the exponential decay of $p(k)$, is observed empirically for firms (Fig. 2), displaying a power law with exponent $\tau = 2.14$ for $k < 200$ and an exponential cutoff for $k > 200$. Coherently with the predictions of the model, the exponential part of $p(k)$ arises from large, diversified, "old" firms, while the power law part of $p(k)$ is produced by young firms.

The estimated parameters are given based on Table II: $b/(\lambda-\mu)$ is estimated to be $N_{b,L}/(N_{b,p}-N_{d,p})$, $\lambda/\mu = N_{b,p}/N_{d,p}$, $N+(\lambda-\mu+b)t = 48\,819$, and $N+bt = N_{f,L}$ (where the subscripts "p," "b," "d," and "f" denote "product," "birth," "death," and "final," respectively, and "L" means either of level A to D or firms). Using $\gamma = -\log \eta_{0,t}$ and $\eta_{0,t}$ in Eq. (3) by eliminating t, γ, and τ can be estimated (Table IV).

The oldest firms within the industry entered it almost 150 years ago, while our data cover only the past decade. Nonetheless, the theoretical estimations of γ and τ based on Eqs. (5) and (8) are surprisingly good, except for γ of firms. This departure can be accounted for if we consider that the real data for firms are shaped not only by *firm entry*, but also by *firm exit*, *mergers and acquisitions*, which are not considered by the model [13,20] because the model does not permit exits of classes consisting of more than one unit. Additional computer simulations show that if we include into our model the possible exit of large classes, the value of γ estimated from the parameters of the model comes to an agreement with the actual one. We show simulation results in Figs. 2(c) and 2(d), and they are in good agreement with the empirical results in Figs. 2(a) and 2(b).

We conclude that our model is in good agreement with the data, for which $p(k)$ is either pure exponential or power law with an exponential cutoff. Our analysis sheds light on the emergence of the exponent $\tau \approx 2$ observed as a universal feature of many biological, social, and economic systems.

We thank A. Bunde and S. Havlin for helpful discussions and suggestions, and NSF and Merck Foundation (EPRIS Program) for financial support.

[1] G. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, 1949).
[2] F. Liljeros *et al.*, Nature (London) **411**, 907 (2001).
[3] H. Jeong *et al.*, Nature (London) **407**, 651 (2000).
[4] S. V. Buldyrev *et al.*, Physica A **330**, 653 (2003).
[5] H. A. Makse *et al.*, Phys. Rev. E **58**, 7054 (1998).
[6] K. Matia *et al.*, J. Am. Soc. Inf. Sci. Technol. **56**, 893 (2005).
[7] R. Gibrat, Bull. Stat. Gén. France **19**, 469 (1930).
[8] R. Kumar *et al.*, Comput. Netw. **31**, 1481 (1999).
[9] M. E. J. Newman, Contemp. Phys. **46**, 323 (2005).
[10] D. Champernowne, Econom. J. **63**, 318 (1953).
[11] J. Fedorowicz, J. Am. Soc. Inf. Sci. **33**, 223 (1982).
[12] W. J. Reed and B. D. Hughes, Phys. Rev. E **66**, 067103 (2002).
[13] Y. Ijiri and H. A. Simon, *Skew Distributions and the Sizes of Business Firms* (North-Holland, Amsterdam, 1977).
[14] G. De Fabritiis *et al.*, Physica A **324**, 38 (2003).
[15] K. Matia *et al.*, Europhys. Lett. **67**, 498 (2004).
[16] The assumption here is necessary to estimate the theoretical value of γ, considering the limited historical information of the PHID.
[17] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes* (Chapman and Hall, London, 1968).
[18] For more general initial conditions, the model predicts the same behavior for the sufficiently large t, but with a different value of the exponential decay constant γ.
[19] The parameter b is related to the parameter b' defined in D. Fu *et al.*, Proc. Natl. Acad. Sci. U.S.A. **102**, 18801 (2005) by the relation $b = b'/(1-b')$.
[20] M. Riccaboni and F. Pammolli, Res. Policy **31**, 1405 (2002).