A simple view of the heavy-tailed sales distributions and application to the box-office grosses of U.S. movies

Ken Yamamoto

Department of Physics, Faculty of Science and Engineering, Chuo University - Kasuga 1, Bunkyo, Tokyo 112–8551, Japan

received 5 August 2014; accepted in final form 8 December 2014
published online 7 January 2015

PACS 89.75.Da – Systems obeying scaling laws
PACS 02.50.Ey – Stochastic processes
PACS 89.65.Gh – Economics; econophysics, financial markets, business and management

Abstract – This letter treats of the power-law distribution of the sales of items. We propose a simple stochastic model which expresses the selling process of an item. This model produces a stationary power-law distribution, whose power-law exponent is analytically derived. Next we compare the model with an actual data set of movie income. We focus on the return on investment (ROI), defined as the gross income divided by the production budget. We confirm that the power-law exponent of the ROI distribution can be estimated from the ratios of income between two adjoining weeks, as predicted by the model analysis. Moreover, an exponential decay of the weekly income is observed both in the model and actual income. Therefore, the proposed model is simple enough, but it can quantitatively describe the power-law sales distribution.

The heavy-tailed distributions, including the power-law and log-normal distributions, have been found in various social phenomena [1,2]. For example, the stock volatility [3], the human mobility [4], and the population of cities [5] follow power-law distributions, and the citation of physics papers [6], the polling score of election [7] and the population of villages [8] follow log-normal distributions. A remarkable implication of the heavy-tailed distribution is that it admits very large elements (statistical outliers) almost inevitably.

The sales of the items typically follows a power-law distribution [9–12], which means that popular items sell far better than niche items. At the same time, a very large number of niche items can bring a non-negligible percentage of sales, and this property is called the long-tail phenomenon [13]. By applying the ideas of the critical exponents and universality class from statistical physics [1,14], the power-law exponent is interpreted as giving information about market and consumer behavior. However, the meaning of the exponent and the process by which it is determined remain unclear.

The aim of this letter is to explain a simple and general mechanism of power-law sales distributions. We propose a simple stochastic model for an item’s selling process. This model produces a stationary power-law distribution, which corresponds to the distribution of the sales amount.

The stochastic model we propose is in the following form:

\[ x_{t+1} = \mu_t x_t, \]  
\[ S_{t+1} = S_t + x_t, \]

where \( t = 1, 2, \ldots \). This set of equations expresses a selling process of a particular item. We regard \( x_t \) as the sales amount in the \( t \)-th period, and set the initial condition \( x_1 = \text{constant} \) for simplicity. We set \( S_1 = 0 \), so that \( S_t = x_1 + \cdots + x_{t-1} \) represents the total sales amount up to \( t - 1 \). Figure 1 depicts the illustration of model (1).
In the graph of $x_t$ as a function of $t$, $S_t$ is given by the area under the $x_t$ curve. The stochastic process $\{x_t\}$ given by eq. (1a) is known as the Gibrat process [15], and $x_t$ approximately follows the log-normal distribution for large $t$ [16]. Note that eq. (1a) implies positive feedback of the sales amount. That is, if an item sells well in a certain time period, it tends to make a large profit also in the following period. The growth rate $\mu_t$ is a random number, and for simplicity, we assume that $\mu_t$ is independent and drawn from the same probability distribution for different $t$. The accumulated sales amount $S_t$ becomes stochastic by the fluctuation of $\mu_t$.

Analysis [17] of eq. (1) reveals that for large $t$, the cumulative distribution of $S_t$ has a stationary power-law tail $P(S_t > s) \propto s^{-\beta}$, which corresponds to the power-law sales distribution. The power-law exponent $\beta$ is given by the positive solution of

$$E(\mu_t^\beta) = 1,$$  \hfill (2)

where $E(\cdot)$ stands for the average. Note that the exponent $\beta$ is determined only from $\mu_t$.

Now, we check whether the proposed model (1) captures the essence of the actual data. We analyzed the movie income, because sufficient information for our study is available free online. We used a week as the unit of time $t$ in eqs. (1) and (2) in order to smooth out the daily fluctuations, e.g., larger audiences on weekends. Intuitively, it is plausible to consider that the variable $S_t$ at large $t$ corresponds to the gross income, but this perspective has a problem. We emphasize here that $x_t$ and $S_t$ in eq. (1) are random variables which describe the selling process of a particular item (movie title). We cannot directly compare the empirical distribution of the gross income made by gathering a number of different movies and the theoretical distribution of $S_t$ (especially the stationary power law $P(S_t > s) \propto s^{-\beta}$), because income of different movies are generally not statistically similar to each other. In fact, the income is associated with the financial scale of the movie, and an expensive movie tends to earn a large income [18].

Here we deduce an appropriate measure for our study, in place of the movie income. For simplicity, we assume that different movies, namely a pair of random variables $(x_i^{(t)}, S_i^{(t)})$ for the movie $i$, share the same dynamics (1), which means that the growth rate $\mu_t$ has the same distribution for different movies. The scale of the movie $i$, which is denoted by $b^{(i)}$, is included only in the initial value $x_i^{(1)} = xb^{(i)}$, where $x$ is a constant. Obviously, $S_i^{(1)}$ and $x_i^{(t)}$ depend on the scale $b^{(i)}$, so they are not statistically similar for different $i$. On the other hand, $\tilde{x}_i^{(t)} = x_i^{(t)}/b^{(i)}$ and $\tilde{S}_i^{(t)} = S_i^{(t)}/b^{(i)}$ again follow eq. (1), whose initial values $\tilde{x}_i^{(1)} = x$ and $\tilde{S}_i^{(1)} = 0$ are constant for all $i$. Thus, $\tilde{s}_i^{(t)}$, instead of income $S_i^{(t)}$, is statistically similar for different $i$. In this study, we chose the production budget for $b^{(i)}$; $\tilde{S}_i^{(b)}$ at large $t$, corresponding to the gross income of the movie $i$ divided by its production budget, is called return on investment (ROI). The ROI is a measure of a movie’s success; a movie is in the black if its ROI is greater than 1 and is in the red otherwise. We expect the empirical distribution of the ROI to be comparable with the theoretical stationary distribution of $S_t$.

Figure 2(a) shows the cumulative distribution of the U.S. gross domestic income for 665 movies released in 2012. This set of data is collected from the free online database Box Office Mojo [19]. This distribution does not possess a clear power-law tail. There are various opinions about the distribution of the gross income; some reports claim that the gross income follows a power law [20,21], whereas others are not [18,22]. In contrast, fig. 2(b) shows the distribution of the ROI of 108 movies that came out in 2012, which exhibits heavy-tailed behavior. We could compute the ROI of only 108 movies because the production budgets of the other movies were not in the Box Office Mojo database. By using the method of maximum likelihood [23], the power-law exponent of the tail, corresponding to an ROI larger than $10^9$, is calculated as $\beta = 1.04$, and this power law is shown as the straight line in fig. 2(b). The standard error on $\beta$ is $\sigma = 0.13$; $\sigma$ becomes large because only 68 movies in fig. 2(b) achieve an ROI greater than unity. To obtain more samples, we used another database, The Numbers [24], which provides a more complete list of production budgets and allowed us to compute the ROI of 3906 movies from 1915 to April 1984. The ROI distributions have power-law tails with exponents of (b) $-1.04$ and (c) $-0.97$. The ROI distributions have power-law tails with exponents of (b) $-1.04$ and (c) $-0.97$. The ROI distributions have power-law tails with exponents of (b) $-1.04$ and (c) $-0.97$.
From these results, the ROI is obtained from the top-100 movies in 2010. A log-normal distribution is suitable in $\mu_t < 1$, and power-law decay appears in $\mu_t > 1$.

Let us check whether eq. (2) gives a satisfactory estimate of $\beta$, assuming that the actual movie income follows model (1). The variable $\mu_t$ is calculated as the ratio of the income on week $t+1$ to that on $t$ (i.e., $\mu_t = x_{t+1}/x_t$). In fig. 3(a), we show the histogram of $\mu_t$ built from the weekly records of the top-100 movies released in 2012 obtained from the Box Office Mojo database. By using the set of $\mu_t$’s, we calculate $E(\mu_t^\beta)$ as a function of $\beta$ (fig. 3(b)), and find that $\beta = 1.10$ is the positive solution of $E(\mu_t^\beta) = 1$. This solution is close to the exponent $\beta = 1.04$ and $0.97$ obtained from the ROI data presented in fig. 2. Similarly, we find $\beta = 1.09$ and $1.08$ from the top-200 movies in 2012 (fig. 3(c)) and the top-100 movies in 2010 (fig. 3(d)), respectively. That is, the estimate of $\beta$ does not depend strongly on the rank or the released year. Based on model (1), we obtain a good estimate of $\beta$, and this result supports the validity of the model.

In fig. 3(a), $\mu_t$ mostly falls between 0 and 1, but there exist a few large values. To see this in more detail, we give further statistical analysis for the distribution of $\mu_t$. The overall cumulative distribution of $\mu_t$ is shown in fig. 4. Clearly, the behavior critically changes at $\mu_t = 1$; the distribution of $\mu_t$ approximately follows a log-normal distribution in $\mu_t < 1$, while the distribution roughly exhibits power-law decay, having exponent $-1.19$, in $\mu_t > 1$. The discontinuous crossover at $\mu_t = 1$ perhaps causes a great disparity between hit movies and poor movies. (The power-law exponent $-1.19$ is not directly related to the exponent $\beta$ of the ROI distribution.)

Next, we discuss the statistics of an individual movie. By taking the logarithm of eq. (1a), we obtain

$$\ln x_{t+1} = \ln x_t + \ln \mu_t.$$ 

This equation means that $x_t$ is a random walk on a logarithmic scale and its average displacement per step is $E(\ln \mu_t)$. Thus, on a linear scale, $x_t$ typically decreases exponentially according to $x_t \propto \exp(-\gamma t)$, where $\gamma = -E(\ln \mu_t)$ is the decay rate. This theoretical outcome is qualitatively correct for actual movies. Figure 5 shows the weekly income of the 1st, 10th, 50th, 100th, and 150th highest grossing movies in the U.S. released in 2012.
In the lower-ranked movies, the weekly records are inclined to be sharply in week 18. The curves for the 100th and 150th movies, which is mainly because theaters showing this movie increased curve for the 1st movie has a jump from week 17 to week 18, Casa de Mi Padre, Madagascar 3, are, respectively, and confirms their exponential decay. Their decay rates for the 100th and 150th highest-grossing movies in 2012 (Marvel’s the Avengers, Chronicle, The Five-Year Engagement, and Casa de Mi Padre, respectively) decreases exponentially. The weekly income per theater of a movie shows a power-law distribution. It has a very simple form. It does not directly consider the effects of consumers’ preference, advertisement, and word of mouse [26]; these factors are condensed into the random variable $\mu_t$. Recall that the aim of this letter is to give a simple mechanism for the power-law behavior. We do not intend a detailed and faithful description of a real phenomenon. Elaborate analysis of a movie market deals with miscellaneous statistical reductions, and tend to be less reliable in comparison to the record of the gross income. We need to study other sales distributions, so as to ascertain that our theoretical model (1) is applicable to actual data.

We consider eq. (1) to be the minimal model for the power-law sales distribution. It has a very simple form. It does not directly consider the effects of consumers’ preference, advertisement, and word of mouse [26]; these factors are condensed into the random variable $\mu_t$. Recall that the aim of this letter is to give a simple mechanism for the power-law behavior. We do not intend a detailed and faithful description of a real phenomenon. Elaborate analysis of a movie market deals with miscellaneous statistical reductions, and tend to be less reliable in comparison to the record of the gross income. We need to study other sales distributions, so as to ascertain that our theoretical model (1) is applicable to actual data.

We conclude that the power-law behavior of the ROI of movies is adequately simulated by eq. (1). Meanwhile, we need to be careful with data bias. The list of the production budgets [24] is incomplete for movies having low production budgets. Moreover, the listed budgets are rough estimates, and tend to be less reliable in comparison to the record of the gross income. We need to study other sales distributions, so as to ascertain that our theoretical model (1) is applicable to actual data.

We consider eq. (1) to be the minimal model for the power-law sales distribution. It has a very simple form. It does not directly consider the effects of consumers’ preference, advertisement, and word of mouse [26]; these factors are condensed into the random variable $\mu_t$. Recall that the aim of this letter is to give a simple mechanism for the power-law behavior. We do not intend a detailed and faithful description of a real phenomenon. Elaborate analysis of a movie market deals with miscellaneous statistical reductions, and tend to be less reliable in comparison to the record of the gross income. We need to study other sales distributions, so as to ascertain that our theoretical model (1) is applicable to actual data.

**Acknowledgments**

The author is grateful to Prof. Yoshihiro Yamazaki for constructive comments. The present work was supported by a Grant-in-Aid for Young Scientists (B).

---

Fig. 5: (Colour on-line) Exponential decay of the weekly gross income. The weekly income of the 1st, 10th, 50th, 100th, and 150th highest-grossing movies in 2012 (Marvel’s the Avengers, Madagascar 3, Chronicle, The Five-Year Engagement, and Casa de Mi Padre, respectively) decreases exponentially. The unit of $\gamma$ is week$^{-1}$. On the other hand, by using the distribution of $\mu_t$ shown in figs. 3(a) and 4, we obtain $-E(\ln \mu_t) = 0.416$ week$^{-1}$. Mathematically, $\gamma > 0$ (equivalently $E(\ln \mu_t) < 0$) is a necessary condition for $S_t$ to have a stationary power-law distribution [17]. A previous study [18] reported the exponential decay of daily movie income, but without proposing a simple mechanism like eq. (1). Furthermore, the weekly income per theater of a movie shows a power-law decay in time [18], but this result cannot be obtained by the model in this letter.

A movie has a running period for which it is shown in theaters. Yet, we have compared the actual ROI distribution and the stationary distribution of $S_t$, without considering the effect of finite running period in the model. We can properly exclude it, based on the discussion as follows. As shown in fig. 5, the weekly income decreases approximately exponentially. Hence, the income in the first several weeks accounts for a large percentage of the gross income. More accurately, if a movie has $T$ running weeks and its weekly income follows the exact exponential decay $x_t = x \exp(-\gamma(t - 1))$, the gross income is given by

$$S_{T+1} = \sum_{t=1}^{T} x_t = \frac{1 - e^{-\gamma T}}{1 - e^{-\gamma}} x.$$  

The gross income of this movie if it has infinite running weeks is $S_\infty = x/(1 - \exp(-\gamma))$, which corresponds to the theoretical stationary value of the gross income. The ratio of $S_{T+1}$ to $S_\infty$ is

$$\frac{S_{T+1}}{S_\infty} = 1 - e^{-\gamma T}.$$  

A typical value of the decay rate is $\gamma = 0.5$ week$^{-1}$, and the average running period of the top-100 movies in 2012 is $T = 15.7$ weeks (see fig. 6 for reference). Using these values, we can estimate $S_{T+1}/S_\infty \approx 0.9996$. Therefore, $S_T$ is essentially the same as $S_\infty$. In other words, the income $S_t$ becomes almost stationary within the typical running weeks. This is why we can ignore the effect of finite running weeks.

We conclude that the power-law behavior of the ROI of movies is adequately simulated by eq. (1). Meanwhile, we need to be careful with data bias. The list of the production budgets [24] is incomplete for movies having low production budgets. Moreover, the listed budgets are rough estimates, and tend to be less reliable in comparison to the record of the gross income. We need to study other sales distributions, so as to ascertain that our theoretical model (1) is applicable to actual data.

We consider eq. (1) to be the minimal model for the power-law sales distribution. It has a very simple form. It does not directly consider the effects of consumers’ preference, advertisement, and word of mouse [26]; these factors are condensed into the random variable $\mu_t$. Recall that the aim of this letter is to give a simple mechanism for the power-law behavior. We do not intend a detailed and faithful description of a real phenomenon. Elaborate analysis of a movie market deals with miscellaneous statistical reductions, and tend to be less reliable in comparison to the record of the gross income. We need to study other sales distributions, so as to ascertain that our theoretical model (1) is applicable to actual data.

We conclude that the power-law behavior of the ROI of movies is adequately simulated by eq. (1). Meanwhile, we need to be careful with data bias. The list of the production budgets [24] is incomplete for movies having low production budgets. Moreover, the listed budgets are rough estimates, and tend to be less reliable in comparison to the record of the gross income. We need to study other sales distributions, so as to ascertain that our theoretical model (1) is applicable to actual data.

We conclude that the power-law behavior of the ROI of movies is adequately simulated by eq. (1). Meanwhile, we need to be careful with data bias. The list of the production budgets [24] is incomplete for movies having low production budgets. Moreover, the listed budgets are rough estimates, and tend to be less reliable in comparison to the record of the gross income. We need to study other sales distributions, so as to ascertain that our theoretical model (1) is applicable to actual data.
Mechanism of heavy-tailed sales distributions and application to movie income (25870743) from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of Japan.

REFERENCES

[1] Buchanan M., Ubiquity: Why Catastrophes Happen (Three Rivers Press, New York) 2000.
[2] Kobayashi N., Kuninaka H., Wakita J., and Matsushita M., J. Phys. Soc. Jpn., 80 (2011) 072001.
[3] Gabaix X., Gopikrishnan P., Plerou V., and Stanley H. E., Nature, 423 (2003) 267.
[4] Rhee I., Shin M., Hong S., Lee K., Kim S. J. and Chong S., IEEE/ACM Trans. Netw., 19 (2011) 630.
[5] Zanette D. H. and Manrubia S. C., Phys. Rev. Lett., 79 (1997) 523.
[6] Redner S., Phys. Today, 58, issue No. 6 (2005) 49.
[7] Fortunato S. and Castellano C., Phys. Rev. Lett., 99 (2007) 138701.
[8] Sasaki Y., Kuninaka H., Kobayashi N. and Matsushita M., J. Phys. Soc. Jpn., 76 (2007) 074801.
[9] Bentley R. A., Madsen M. E. and Ormerod P., Physica A, 388 (2009) 691.
[10] Fenner T., Levene M. and Loizou G., Physica A, 389 (2010) 2416.
[11] Hisano R., Sornette D. and Mizuno T., Phys. Rev. E, 84 (2011) 026117.
[12] Brabazon P. G. and MacCarthy B., Omega, 40 (2012) 302.
[13] Anderson C., The Long Tail (Hyperion, New York) 2006.
[14] Stanley H. E., Introduction to Phase Transitions and Critical Phenomena (Oxford University Press, Oxford) 1987.
[15] Gibrat R., Les Inégalités Économiques (Librairie du Recueil Sirey, Paris) 1931.
[16] Crow E. L. and Shimizu K. (Editors), Lognormal Distributions: Theory and Applications (Marcel Dekker, New York) 1988.
[17] Yamamoto K., Phys. Rev. E, 89 (2014) 042115.
[18] Pan R. J. and Sinha S., New J. Phys., 12 (2010) 115004.
[19] Box Office Mojo, http://www.boxofficemojo.com/.
[20] Sornette D. and Zajdenweber D., Eur. Phys. J. B, 8 (1999) 653.
[21] Sinha S. and Raghavendra S., Eur. Phys. J. B, 42 (2004) 293.
[22] Walls W. D., Appl. Financ. Econ., 15 (2005) 1181.
[23] Clauset A., Shalizi C. R. and Newman M. E. J., SIAM Rev., 51 (2009) 661.
[24] The Numbers, http://www.the-numbers.com/.
[25] Zipf G. K., Human Behavior and the Principle of Least Effort (Addison-Wesley, Reading, Mass.) 1949.
[26] Ishii I., Arakaki H., Matsuda N., Umemura S., Urushidani T., Yamagata N. and Yoshida N., New J. Phys., 14 (2012) 063018.
[27] De Vany A., Hollywood Economics: How Extreme Uncertainty Shapes the Film Industry (Routledge, London) 2004.