Random Division and the Size Distribution of Business Firms

Fiona C. Maclachlan

Department of Economics and Finance
Manhattan College
Riverdale, NY 10471
fiona.maclachlan@manhattan.edu

A computational model of business firm size based on random division is presented. Simulations generate size distributions that are positively skewed with Pareto (power-law) upper tails. Furthermore, the simulated distributions are shown to deviate from the lognormal in ways consistent with some recent empirical findings.

1. Introduction

The positively skewed distribution of firm sizes, with a large number of relatively small firms coexisting alongside fewer larger firms, is an enduring empirical fact in market economies. Typically, when data points conform to a Gaussian distribution, they are supposed to be the result of multiple random factors and not worthy of further investigation. A skewed distribution, on the other hand, calls out for a theory. In 1931, French economist Robert Gibrat applied Jacobus Kapteyn's 1903 work on skew distributions in biology to explain the observed distribution of firm sizes. He speculated that the distribution was lognormal and generated in the steady state by a process of stochastic proportional growth [1]. Gibrat's explanation is the dominant one in the economics literature, having set the direction for subsequent research on the distribution of firm sizes, as well as on the distribution of city sizes. Simon [2], in a variation of the model based on Yule's [3] analysis of the distribution of species across genera, shows that a positively skewed distribution with a Pareto upper tail arises if the stochastic proportional growth process allows for the random addition of new firms. Another variation of the model that shows a Pareto upper tail in the steady state postulates a lower reflective barrier preventing the stochastically growing entity from becoming too small [4].

Stochastic proportional growth models are susceptible, however, to several criticisms. Criticism can be made regarding the realism of the Gibrat model assumption that the growth rate of a firm is independent of its size. And on a more fundamental level, it can be argued that the steady-state construct is not appropriate for economic

https://doi.org/10.25088/ComplexSystems.18.2.227

systems. Such systems are subject to frequent exogenous changes and cannot be safely assumed to ever settle into a steady state. Mandelbrot [5] raises this issue and, in addition, questions the assumption, implicit in diffusion models, that the growth of one economic entity will not affect the growth of the others.

In this paper an alternative simpler explanation for the observed size distribution of firms is presented. Its advantage over existing models is that it does not rely on restrictive assumptions about growth rates, nor does it need to evoke the notion of a steady state. The simpler explanation points to a common cause for apparent power-laws in disparate phenomena, including such things as word frequencies in texts and the sizes of shattered pottery pieces, that are not undergoing a process of growth.

The essential insight of the model is that the random division of a constrained quantity results in a skewed distribution of the part sizes. Consider a line of unit length divided into n intervals by $n-1$ points dropped onto the line at random. If the intervals are ordered from smallest to largest, the expected size of the r^{th} smallest interval is given by

$$\frac{1}{n}\sum_{i=1}^{r}\frac{1}{n+1-i}.$$

Although this expression for the expected value, colloquially known as the broken stick rule, is familiar to students of probability, the size frequency distribution is not so readily characterized. It appears to be an example of a case where a simple rule gives rise to a surprising degree of complexity [6].

Random division has been suggested as an explanation for observed size distributions in a variety of contexts. The correspondence between the broken stick rule and observed market shares has been noted by several authors [7-9]. In zoology, the rule has been used to explain the relative abundance of animals of different species in a given locality. MacArthur [10], for instance, finds that the rule fits census data on birds in tropical forests almost perfectly; and it can be made to fit the more uneven distribution of birds in Quaker Run Valley, New York, if viewed as multiple sticks of different lengths being broken. In linguistics, Perline [11] relates the broken stick rule to Zipf's law for word frequencies. The random division model also shares a resemblance to the so-called "monkeys typing" models used to explain power-law behavior in word frequencies [12]. In these models, a string of letters is imagined to be divided into words by random hits to a notional space bar.

To investigate the extent to which random division provides a satisfactory explanation for empirical firm size distributions, a simple model of market shares is suggested that connects the idea of random division with the division of total revenue between competing firms in a given market. To generate a size distribution for firms economy-wide

https://doi.org/10.25088/ComplexSystems.18.2.227

from the simple model of market shares, it is necessary to make two additional assumptions: the first assumption is about the relative size of revenue in each of the markets, and the second assumption is about the number of firms in each of the markets—that is, the degree of market concentration. Since the size of total spending in an economy can be supposed to be constrained by the quantity of resources, it is reason able to model the share of total spending in each of the markets as being determined by random division. For the degree of market concentration, we will start with the simplest assumption that it is uniform across markets. The assumption is then relaxed, allowing for a varying degree of market concentration, and a greater degree of inequality is found that is more consistent with empirical distributions.

Since random division does not generate a clearly identifiable distribution, the implications of the model are explored through running computer experiments. The aim is to discover the simplest model that will generate distributions that match empirical findings. Empirical distributions of publicly held firms have been found to be approximately lognormal with power-law upper tails. Furthermore, recent work [13, 14] shows the distributions of the log sizes to be skewed and leptokurtic. All these features are replicated in the model presented here.

2. A Model of Market Shares

Assume that markets are monopolistically competitive with each firm producing a unique product, but one that buyers could substitute with the products of other firms in the same market. Let the firm's size be represented by its sales revenue, that is, by its share of the total spending in the market. To model the market shares, consider a theoretical construction introduced by Hotelling in [15]. Firms differentiate their product along a single quantifiable dimension and buyers' preferences are distributed evenly along that dimension. Hotelling uses the example of competing sellers located at different positions on a main street. The buyers are evenly distributed spatially and have a preference for the nearest seller. Hotelling indicates that the construction can be applied more generally citing the example of two competing cider merchants, one selling a sweeter product than the other.

While Hotelling analyzes a situation of two sellers with full information about the competitor's product, the model presented here has the added complexity that the number of sellers is unknown to the competitors, as is the type of product that competing firms offer. Let there be a large number of buyers in each market, and let each buyer have a unique first preference for the quantifiable quality over which firms are differentiating their product. Imagine lining up all the consumers in order of their first preference and assign each a successive integer number. Call these numbers the varieties. Further, assume that

https://doi.org/10.25088/ComplexSystems.18.2.227

preference rankings are such that a consumer will select the variety of product with the integer number closest to that of his or her unique preference. It is not necessary to make any assumption about the distribution of the preferences over the quality, but by assigning each consumer a successive number, we are transforming it into a uniform distribution.

Without knowing what the competition is doing, or even how many competitors are in the field, there is no strategy for a firm that beats random selection when choosing which variety to offer. We can allow that the firms' managers are acting purposively in deciding how to differentiate their products but still assume that, in effect, they are acting as if by random. The market share of a given firm will be determined by the number of consumers whose preference is closer to the firm's variety, than to any other. Figure 1 shows an example of market shares in a market with five firms. The tick marks indicate the firms' varieties and the gray and black shaded regions represent the market shares.

Figure 1. Market shares in a market with five firms. The ticks represent the firms' choices of variety, and the black and gray shaded areas represent the market shares.

The model generates a size distribution that is close, but not identical, to what would be determined by dropping points onto a line at random and considering the intervals. The size frequency histogram of simulated data in Figure 2 shows that the distribution is positively skewed with a few large shares coexisting alongside a larger number of smaller shares. Figure 3 shows the same data organized as a Zipf plot, that is, a rank-size plot using a log scale on both axes. The upper tail is approximately a straight line and could pass a rough test for a power-law. The power-law exponent, given by the slope of the Zipf plot, measures the degree of inequality in the distribution. Alternatively, the data could be presented in the form of a cumulative distribution function. A straight line Zipf plot corresponds to a Pareto distribution whose shape parameter is the reciprocal of the absolute value of the power-law exponent. Zipf's law refers to a Pareto distribution with a shape parameter equal to one. Our random division model of market shares, with a power-law exponent of around -0.3, shows less inequality in the upper tail than would be found under Zipf's law.

https://doi.org/10.25088/ComplexSystems.18.2.227

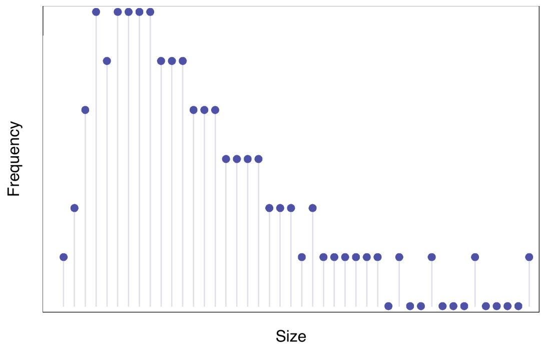

Figure 2. A frequency histogram of share sizes in a market with 100 firms. The share sizes were obtained by running 100 simulations and taking the means of the ranked data.

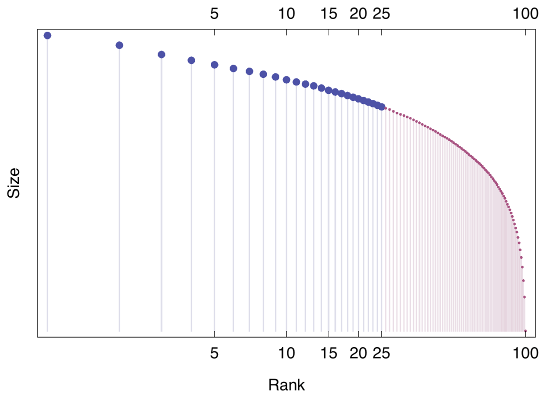

Figure 3. A Zipf plot of market shares in a market with 100 firms. The points were obtained by running 100 simulations and taking the means. Regressing the mean sizes of the top quartile of the firms on their rank gives a coefficient of determination of 0.98 and a slope of -0.31.

Testing the model of market shares against actual observations presents difficulties because real-world market shares do not conform to any neat systematic pattern. As expected with complex phenomena, they take on a variety of configurations, and the most that can be said is that they are generally positively skewed. If we aggregate all the firms in the economy, however, a more distinct pattern emerges. So the next step is to take the model of market shares and embed it into a model for the economy-wide firm size distribution. It is then possible to run computer simulations and compare the resulting firm size distributions with their empirical counterparts.

https://doi.org/10.25088/ComplexSystems.18.2.227

3. A Computational Model of Firm Size

Assume that each firm produces only one type of output and let the firm's size be measured by its annual sales revenue. Randomly divide aggregate spending in the economy over m markets and let S_i, $i = 1, \ldots, m$, represent spending in the i^{th} market. Let there be n_i, $i = 1, \ldots, m$, competing firms comprising the m markets. Let s_{ij}, $j = 1, \ldots, n_i$, be the market share of the j^{th} firm in the i^{th} market and let it be determined according to the model of market shares in Section 2. Thus, the size of the j^{th} firm in the i^{th} market will be given by $S_i s_{ij}$.

Begin with the simplest assumption that the degree of market concentration is uniform across markets, that is, $n_i = n_j$ for all i, j. The resulting distribution is similar in shape to the distribution for market shares illustrated in Figures 1 and 2, but it displays a greater degree of inequality, as evidenced by a steeper slope on the upper tail of the Zipf plot. For example, taking the means from 100 simulations of 100 firms divided evenly between 5, 10, and 25 markets generates upper tail slopes and coefficients of determination of -0.55 (0.98), -0.54 (0.98), and -0.46 (0.97), respectively. While the inequality is greater than in the case of a single market, it is still less than in empirical distributions.

Using a large dataset reporting the number of employees of all tax-paying firms, Axtell [16] finds a power-law exponent of minus one, corresponding to Zipf's law. Unlike other studies that include only the larger publicly-held firms, Axtell finds that a power-law holds for all parts of the distribution, and not just to the upper tail. As would be expected with Zipf's law, the data reveals that the most frequent firm size is the minimum size of zero employees. Interpreting data based on tax records requires care, however, as small firms may not report employees who are working off-the-books. In addition, a number of the small businesses that report zero employees may be artificial economic entities created to reduce tax liabilities by offseting income from other sources. If the number of very small firms is indeed inflated in the data, then the true Zipf plot might dip down at the low end, as predicted by the model.

Taking logs of the simulated sizes and normalizing by subtracting the mean and dividing by the standard deviation, allows us to ascertain if the simulated size distribution is lognormal. If the size distribution is lognormal, then the log of the sizes will be normally distributed with an expected skewness of zero and kurtosis of three. Table 1 provides the mean skewness and Table 2 the mean kurtosis over 100 simulations with different values for the number of firms in each market and for the number of markets. In a recent study using data from Bloomberg, Kaizoji et al. [14] find negative skewness and positive excess kurtosis in the normalized log sizes for U.S. firms.

https://doi.org/10.25088/ComplexSystems.18.2.227

Over the time period 1995-2003, they find a mean negative skewness of 0.34, and a mean kurtosis of 3.66. Interestingly, for Japanese firms over the same period they find a similar kurtosis, but a positive, rather than negative, skew.

Markets \ Firms	20	40	60	80	100
20	−0.628	−0.576	−0.589	−0.627	−0.585
40	−0.704	−0.668	−0.663	−0.694	−0.691
60	−0.723	−0.714	−0.74	−0.739	−0.716
80	−0.765	−0.745	−0.754	−0.701	−0.723
100	−0.681	−0.728	−0.743	−0.77	−0.76
120	−0.808	−0.743	−0.75	−0.757	−0.758
140	−0.805	−0.759	−0.78	−0.739	−0.74
160	−0.767	−0.755	−0.825	−0.745	−0.774
180	−0.812	−0.842	−0.764	−0.772	−0.776
200	−0.784	−0.772	−0.765	−0.82	−0.775

Table 1. Skewness of standardized log sizes for simulations with a uniform number of competitors in each market.

Markets \ Firms	20	40	60	80	100
20	3.4	3.41	3.3	3.6	3.43
40	3.68	3.65	3.58	3.65	3.67
60	3.95	3.89	3.82	3.89	3.89
80	3.82	4.05	4.08	3.93	3.95
100	4.26	4.08	3.97	4.12	3.95
120	4.02	4.28	4.4	4.04	4.26
140	4.08	4.1	3.98	3.96	4.1
160	4.05	4.06	4.23	4.03	4.2
180	4.31	4.11	4.15	4.18	4.23
200	4.08	4.25	3.93	4.29	4.21

Table 2. Kurtosis of standardized log sizes for simulations with a uniform number of competitors in each market.

Under the assumption of a uniform number of competitors across markets, the simulations show features found in empirical distributions: a power-law upper tail, as well as skewness and a positive excess kurtosis of the normalized log sizes. But the degree of inequal-

https://doi.org/10.25088/ComplexSystems.18.2.227

ity in the upper tail in the simulated distributions is low relative to what is found in empirical distributions. To obtain a better match to the empirical distributions, the simplifying assumption that the number of competing firms is uniform across markets can be relaxed. Let the markets be divided equally into three groups: one group contains markets that are highly concentrated with just a few competitors, the next group contains markets that are less concentrated with a moderate number of competitors, and the third group contains markets with many competitors. Varying the number of competitors in the three groups leads to firm size distributions with different power-law exponents. For example, we can match an empirical distribution with a Zipf law upper tail by selecting values of two, 10, and 200 competitors for the three groups of markets. Figure 4 shows the slopes of the upper tail for 1000 simulations using those values for the numbers of competing firms. The mean slope over the 1000 simulations was -1.02 with a mean coefficient of determination of 0.98.

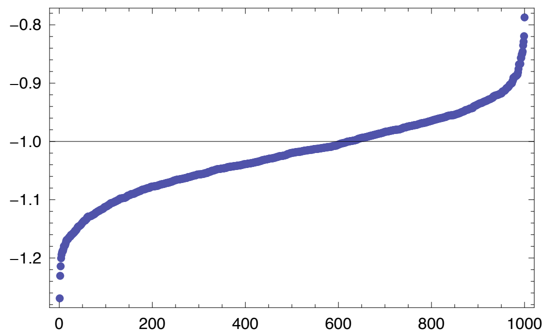

Figure 4. The slopes of the upper quartile tails of Zipf plots for 1000 simulations. Markets in the simulation were assumed equally likely to contain two, 10, or 200 competing firms.

Allowing for varying degrees of market concentration leads to the greater inequality in firm sizes that we find in empirical distributions. The skewness and kurtosis of the normalized log sizes are also affected. Whereas with the uniform number of competing firms the skewness was consistently negative, with varying degrees of market concentration the skewness can be positive. The greater degree of inequality in the sizes is also associated with a greater degree of kurtosis.

4. Conclusion

Power-laws in economic phenomena first became evident with Vilfredo Pareto's observation in the late nineteenth century of a distinctive size frequency distribution for income and wealth. Later in

https://doi.org/10.25088/ComplexSystems.18.2.227

1949, Zipf's discovery of a power-law distribution for city populations and word frequencies reignited interest in the topic. And more recently, the same pattern has shown up in the degree distributions of networks like the World Wide Web, stimulating further research and popular interest [17, 18].

The distributions call out for an explanation beyond randomness because the tendency is to assume that randomness, by itself, would lead to a Gaussian distribution. Gibrat's idea of stochastic proportional growth has provided the dominant explanation. The rich-grow-richer idea of the model is that larger entities are expected on average to grow by larger increments than smaller entities. In the theory of networks, the notion of preferential attachment amounts to much the same thing. The Chinese restaurant process, for instance, compares the stochastically growing entity to the number of patrons sitting at a table in a restaurant. A new patron entering the restaurant will either start a new table with some given probability, or join an already occupied table, with a probability proportional to the number of existing occupants at the table.

While stochastic proportional growth and preferential attachment models make sense in some contexts, it is not immediately clear that they should apply to business firms, as smaller, younger firms are more likely to experience growth spurts than the larger, more mature companies. In addition, it is unrealistic, as Mandelbrot [5] has noted, to apply the concept of the steady state to economic systems that are in a constant state of flux.

The model presented in this paper shows that the essential features of the observed distributions of firm sizes can be derived without invoking any special mechanism beyond random division. While the focus was on the distribution of firm sizes, similar models could probably be worked out for the size distributions of other phenomena exhibiting the same general features.

Acknowledgments

This paper is dedicated to the memory of John E. Reith (1919-2008) whose work on the broken stick rule served as its inspiration. I thank participants at the Wolfram Technology Conference 2006 and the NKS 2007 Wolfram Science Conference, as well as two anonymous referees, for their helpful comments.

https://doi.org/10.25088/ComplexSystems.18.2.227

References

[1] J. Sutton, "Gibrat's Legacy," *Journal of Economic Literature*, **35**(1), 1997 pp. 40-59.

[2] H. A. Simon, "On a Class of Skew Distribution Functions," *Biometrika*, **42**(3-4), 1955 pp. 425-440. doi.10.1093/biomet/42.3-4.425.

[3] G. U. Yule, "A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Wills, F.R.S.," *Philosophical Transactions of the Royal Society of London*, Series B, **213**, 1925 pp. 21-87.

[4] X. Gabaix, "Zipf's Law for Cities: An Explanation," *Quarterly Journal of Economics*, **114**(3), 1999 pp. 739-767. doi.10.1162/003355399556133.

[5] B. B. Mandelbrot, *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*, New York: Springer, 1997.

[6] S. Wolfram, *A New Kind of Science*, Champaign, IL: Wolfram Media, Inc., 2002.

[7] J. E. Cohen, *A Model of Simple Competition (Harvard Universiy Computation Laboratory Annals)*, **41**, Cambridge, MA: Harvard University Press, 1966.

[8] J. E. Reith, "The Generalized Distribution Rule," unpublished paper, 1968.

[9] J. J. Gilman, "Broken Sticks—Why Mergers May Fail to Garner Market Share," *Managerial and Decision Economics*, **13**(5), 1992 pp. 453-456.

[10] R. MacArthur, "On the Relative Abundance of Species," *The American Naturalist*, **94**(874), 1960 pp. 25-36.

[11] R. Perline, "Zipf's Law, the Central Limit Theorem, and the Random Division of the Unit Interval," *Physical Review* E, **54**(1), 1996 pp. 220-223. doi.10.1103/PhysRevE.54.220.

[12] G. A. Miller, "Some Effects of Intermittent Silence," *The American Journal of Psychology*, **70**(2), 1957 pp. 311-314.

[13] M. H. R. Stanley, S. V. Buldyrev, S. Havlin, R. M. Mantegna, M. A. Salinger, and H. E. Stanley, "Zipf Plots and the Size Distribution of Firms," *Economics Letters*, **49**(4), 1995 pp. 453-457.

[14] T. Kaizoji, H. Iyetomi, and Y. Ikeda, "Re-examination of the Size Distribution of Firms," *Evolutionary and Institutional Economics Review*, **2**(2), 2005 pp. 183-198.

[15] H. Hotelling, "Stability in Competition," *The Economic Journal*, **39**(153), 1929 pp. 41-57.

[16] R. L. Axtell, "Zipf Distribution of U. S. Firm Sizes," *Science*, **293**(5536), 2001 pp. 1818-1820. doi .10.1126/science.1062081.

[17] A-L. Barabási, "Linked: The New Science of Networks," *American Journal of Physics*, **71**(4), 2003 pp. 409-410.

[18] D. J. Watts, *Six Degrees: The Science of a Connected Age*, New York: W. W. Norton & Company, Inc., 2003.

https://doi.org/10.25088/ComplexSystems.18.2.227