Market Shares Are Not Zipf-Distributed

Fernando Buendía
School of Economic and Administrative Sciences
Panamericana University Guadalajara
Prolongación Calzada Circunvalación Poniente 49
Zapopan, Jalisco, Mexico, 45010
fernando.buendia@up.edu.mx

Josué Reynoso
School of Engineering
Panamericana University Guadalajara
Prolongación Calzada Circunvalación Poniente 49
Zapopan, Jalisco, Mexico, 45010
josue.reynoso@up.edu.mx

The distribution of market shares within an industry is a relevant performance measure for managers and policy makers, allowing the determination of whether a given industry is competitive or not. Usually, the perfect competition model is the reference to judge the efficiency of markets. However, there are empirical studies that show that one of the most evident features of global firms is that both their size and market are Zipf-distributed. In this paper, it is shown that the distribution of market shares in the world cola-drink market is skewed, but it does not follow Zipf's law; thus it is inadequate to generalize this distribution to any kind of market.

1. Introduction

The best developed and most influential model in the fields of economics, management, marketing, and other social sciences is the competitive-equilibrium paradigm. Two of the most important conditions of this model are price-taking behavior and that the markets are formed by a large number of homogenous firms and consumers. Though many markets are perfectly competitive, most important industries in modern capitalist countries are formed by a small number of firms. Some economists believe that imperfect competition models (oligopolistic and monopolistic competition) are best suited to describe these industries, but they do not consider the fact that in many modern industries a small number of large firms coexist alongside a large number of small firms. The study of company sizes was started by Gibrat [1] and Zipf [2]. Recently, Ramsden and Kiss-Haypál [3] found that Zipf's law does not fit the data for firm sizes of the differ-

https://doi.org/10.25088/ComplexSystems.22.3.221

ent countries they studied. For instance, their analysis of the data on the largest 500 US firms gives a β_2 close to 1.25. However, using the entire population of tax-paying firms in the United States, Axtell [4] shows that Zipf's [2] distribution characterizes firm sizes. Independently of the value β_2 takes, this empirical finding poses a tremendous challenge to conventional wisdom, as well as current economic theories, and reveals the need to build a persuasive account of how performance differences among firms arise.

Another proof of the inadequacy of the perfect and imperfect competition models to describe modern industrial sectors is related to the distribution of market share, the main focus of this paper. In the extant literature it has been shown that market share of the firms participating in these industries are Zipf-distributed. Although these findings are compelling evidence of what any accurate marketing theory has to explain, this paper shows that there is no reason to expect the distribution of market shares to take any particular form for the general run of industries. Specifically, using the data about market shares in the world cola-drink industry analyzed by Buendía [5], it is proven that the distribution of market shares in the world cola-drink industry does not follow Zipf's law. The remainder of this paper is organized as follows. Section 2 reviews literature on market shares distribution. Section 3 provides a brief description of the characteristics of fractal parabolic, stretched exponential, and log-normal distributions. Section 4 shows why power law with cut-off distribution is a better description of the behavior of market shares in the world cola-drink industry than Zipf's distribution. The paper finishes with some conclusions.

2. Distribution of Market Share: A Literature Review

There is enough empirical evidence that shows that, in many industries, market shares follow a highly skewed distribution, where most of the market share is accounted for by a few large companies and the remainder is divided among a large amount of very small competitors. In the 1970s, the Boston Consulting Group (BCG) [6] advanced the hypothesis that market share of the largest competitors of a given industry follows a distribution in which the ratio between the first competitor and the next one, in terms of revenue, is 2:1. A more rigorous analysis was done by Buzzel [7], who applies the BCG hypothesis to the whole set of firms in many industries. He found that market shares fit a semi-logarithmic distribution. Even though he agrees with the fact that market share following a semi-logarithmic distribution is an important finding, he accepts that there are other

https://doi.org/10.25088/ComplexSystems.22.3.221

models that could describe more adequately the same skewed patterns. However, what Buzzell [7] calls semi-logarithmic distribution is in fact an exponential distribution. Specifically, he uses the expression $\log(\mathrm{MS}_r) = k_0 + k_1$ that mathematically is equivalent to an exponential distribution of the form $\mathrm{MS}_r = k_2\, e^{k_1 r}$. Therefore, since it is more general, the latter is more adequate than the former expression.

Riemer et al. [8], using data from 70 industries, found that market shares in these markets are Zipf-distributed. These findings contrast with those of Buzzell [7] in that the empirical data is best described by a power-law distribution rather than an exponential one. Furthermore, while power-law distribution refers to the reproduction of Yule's [9] stochastic process, exponential distribution is related to Gibrat's model of proportional growth. Therefore, power-law distributions cannot be reproduced through the model of proportionated growth that Buzzell suggests. Consequently, it is necessary to find other stochastic models that could take into account the microeconomic variables, and that are capable of reproducing the behavior of a power-law distribution. More recently, Kohli and Sah [10] found that power-law distribution better describes the data of the US sporting goods and food industries. These findings are compatible with those of Riemer et al. [8].

Finally, Buendía [5] found that the data of the market share of the cola-drink industry fits a third-degree polynomial logarithmic function and whose R^2 is 0.9724, a fitness that is higher than the conventional log-log plot found by Riemer et al. [8]: 0.9513. Even though his results have a higher level of fitness, there are other methodologies that instead of using least-squared estimation apply maximum likelihood. The use of these methodologies provides new insights and can help solve the debate about what distribution better describes market share data. But for this, it is necessary to understand the behavior of other kinds of distributions.

3. Fractal Parabolic, Stretched Exponential, Log-Normal, and Other Distributions

In the literature, there are many phenomena whose elements are considered to follow skewed distributions, which implies that they follow a distribution in which the size is proportional to the multiplicative inverse of the rank. This relationship is given by the following equation:

$$S_r = \frac{k}{r}. \tag{1}$$

https://doi.org/10.25088/ComplexSystems.22.3.221

S_r is the size of the element in the r^{th} position, sorted from the biggest to the smallest, and k is a constant. To assess whether or not a determined set of data follows Zipf's distribution, it is common to plot the data using logarithmic scales in both the x and y axes and obtain the equation of this relationship using linear regression. The resulting equation has the form $\log(S_r) = \beta_1 + \beta_2 \log(r + \epsilon_1)$. If β_2 is close to -1, it is said that the data follows Zipf's law. Moreover, if the resulting β_2 value is not close to 1, then the data is considered to follow a generalized Zipf distribution where the exponent q is equal to the additive inverse of β_2 ($q = -\beta_2$):

$$S_r = \frac{k}{r^q}. \tag{2}$$

On the other hand, the generalized Zipf distribution is equivalent to the power-law distribution:

$$p(x) = C\,x^{-\alpha}. \tag{3}$$

Here, $p(x)$ is the probability density function ($p(x) = P(X == x)$), x represents the size, and C is a normalization constant that adjust the y axis to 1. Consequently, equation (2) is the size as a function of the rank; meanwhile, equation (3) is the probabilistic distribution function (PDF) as a function of the size. From this brief analysis, it is possible to conclude that Zipf's distribution is, in fact, a cumulative distribution function (CDF). Therefore, $P(X, \leqq x)$, implicit in Pareto's distribution, is the inverse function of equation (2). This means that the relationship between the exponent q of the generalized Zipf's law and the exponent α of the power law can be obtained by calculating the CDF of equation (3), which in fact is the inverse function of equation (2). This relationship can be obtained as follows. The generalized Zipf distribution in equation (2) entails that there are r points whose value is greater than or equal to k/r^q. In other words, the probability that the size X of a data point is greater than or equal to k/r^q can be expressed as:

$$P\left[X \geqq \frac{k}{r^q}\right] \approx r, \tag{4}$$

which is equivalent to:

$$P[X \geqq y] \approx r. \tag{5}$$

To obtain the PDF from the CDF, it is necessary to calculate the first derivative of equation (5) with respect to y, which yields:

$$P[X == y] \approx y^{-[1+(1/q)]}. \tag{6}$$

https://doi.org/10.25088/ComplexSystems.22.3.221

Since the power-law distribution in equation (3) is equivalent to equation (6), we can then conclude that the relationship of the exponents is given by:

$$q = \frac{1}{\alpha - 1}. \tag{7}$$

According to equation (7), a dataset that follows Zipf's distribution (or a generalized Zipf distribution where $q = 1$) has the same slope, in a log-log plot, as a power-law distribution with exponent $\alpha = 2$. Therefore, there is an equivalence between Pareto and generalized Zipf distributions when the Pareto index is equal to the multiplicative inverse of the generalized Zipf exponent (Pareto index $= 1 / q$).

Recently, Newman [11] has suggested that in many phenomena whose distributions follow (allegedly) power laws, it is common to observe that there is a threshold value below which the power-law behavior does not hold; thus in both Pareto and power-law estimations, there is a cut-off value, commonly referred to as $x_{\min}$. All data below this value is not considered within the distribution.

Another important contribution concerning skewed distributions has to do with other similar mathematical models, such as parabolic fractal, log-normal, or stretched exponentials [12]. Parabolic fractal distribution is similar to a Zipf distribution, but instead of a linear relationship, in a log-log plot of the size as a function of the rank, the relationship is parabolic, as its name indicates (a second-degree polynomial function). Log-normal is a function that, if plotted over a logarithmic scale on the x axis, resembles a normal distribution. The stretched exponential is similar to a power law multiplied by an exponential.

Finally, there is enough evidence that shows that the method of linear regression over log-log plots, to assess whether a set of data follows a power law, is inaccurate. For example, Clauset et al. [13] show that the linear regression over a number of rank-frequency sets of data, to calculate the scaling parameter, is biased almost 3% in a set of 10 000 randomly generated data points. This bias, however small, could get significantly bigger in smaller datasets. Clauset et al. [13] also suggest that maximum likelihood estimation is a better method to assess the goodness of fit between an empirical set of data and a theoretical model.

4. Power Law with Cut-Off Distribution of Market Share in the Cola-Drink Industry

After studying the different kinds of distributions, it is possible to determine which one can better describe the behavior of market shares

https://doi.org/10.25088/ComplexSystems.22.3.221

of the world cola-drink industry. To do so, the models reviewed in Section 3 are applied to the market shares of brands that compete in the cola-drink industry used by Buendía [5]. If the semi-logarithmic (exponential) distribution is utilized, as Buzzel did, the market shares distribution can be explained using Gibrat's law of proportionate growth. In this case, market shares obey the following equation:

$$\log(\mathrm{MS}_r) = k_0 + k_1\, r \tag{8}$$

which is equivalent to the exponential function:

$$\mathrm{MS}_r = k_2\, e^{k_1 r}. \tag{9}$$

MS_r is the market share of the r^{th} competitor, and k_1 and k_2 are constants. If least-squares regression is used, $k_1 = -0.159$, $k_2 = 4782$, and $R^2 = 0.844$, according to equation (9). (If, instead of equation (9), equation (8) is used, the value of k_0 would be 8.473.) As can be observed, in contrast to what Buzzel [7] found, the data does not fit the exponential model. Furthermore, looking at Figure 1, it is easy to observe that the semi-logarithmic (exponential) distribution is not the best description of the data.

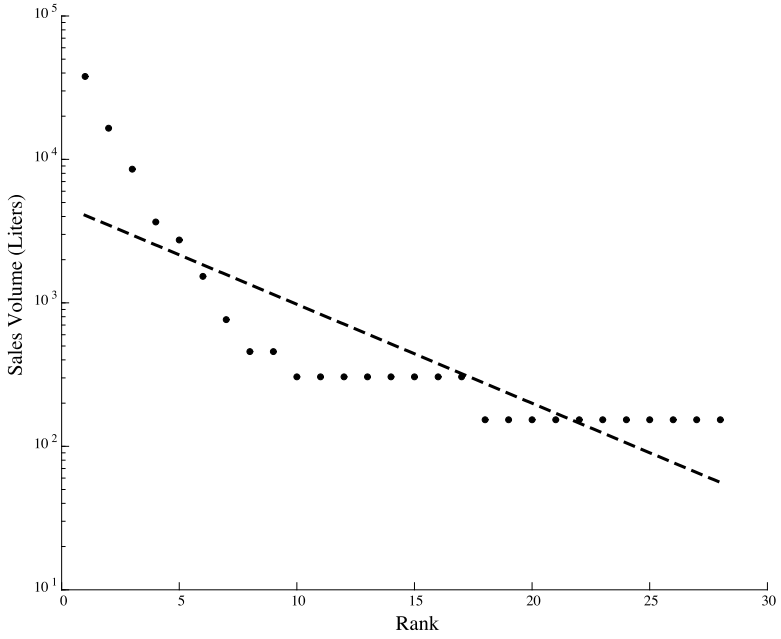

Figure 1. Cola-drink market share. The empirical data was fitted to an exponential (semi-logarithmic) distribution. The y axis is in logarithmic scale.

https://doi.org/10.25088/ComplexSystems.22.3.221

When Zipf's distribution is applied to the dataset, the linear regression over the log-log plot is a straight line. That is to say, by applying equation (2) it is possible to obtain the following values: $k = 36\,369$, $q = -1.79$, and $R^2 = 0.975$. As Figure 2 shows, this regression line better fits the data than the exponential model.

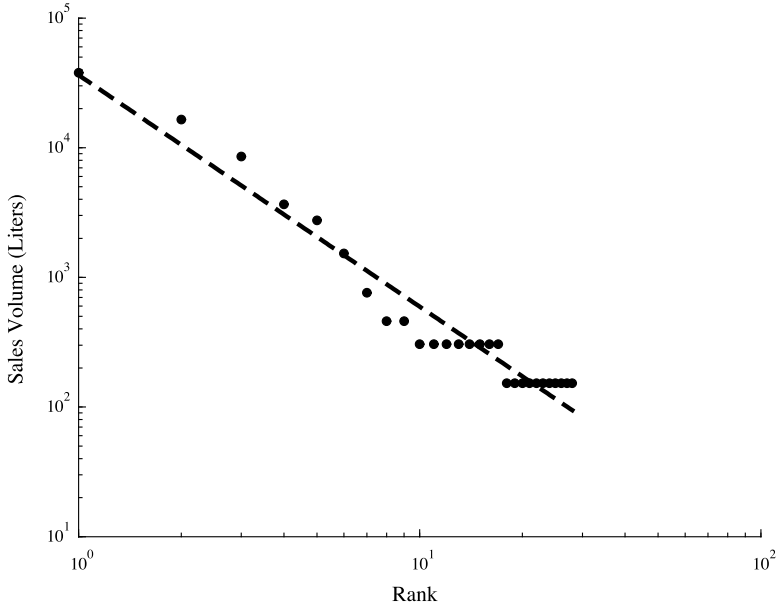

Figure 2. Generalized Zipf distribution. The data was fitted to a generalized Zipf distribution using linear regression. Both rank and volume axes are shown in logarithmic scales.

So far, it is clear that market shares of the world cola-drink industry do not follow a Zipf distribution with a slope of -1. However, according to the generalized Zipf distribution, the slope of the regression line is -1.79 ($q = 1.79$). Since this model is equivalent to a power-law distribution, the application of equation (7) will give an exponent α of 1.56. Given that the R^2 value is higher than the one obtained for the semi-logarithmic model, a power law is a more plausible description of the data. Goldstein et al. [14], Newman [11], and Clauset et al. [13], instead of applying the least-square method over a log-log plot, suggest the use of the maximum likelihood estimation as a more precise approach to obtain the parameters of the distribution. The probability that the values were generated by a specific function

https://doi.org/10.25088/ComplexSystems.22.3.221

is proportional to the likelihood function:

$$l(\alpha \mid x) = \prod_{i=1}^{N} \frac{x_i^{-\alpha}}{\zeta(\alpha)}, \tag{10}$$

where $l(\alpha \mid x)$ is the likelihood function and ζ is the Riemman zeta function. Therefore, by calculating the value of α that maximizes the likelihood function, it is possible to determine the most likely parameters that describe the distribution of empirical data. Since the maximum of the likelihood function is located at the same position as the maximum of the logarithm of the likelihood function, it is easier to maximize the logarithm instead [14]:

$$\begin{aligned} L(\alpha \mid x) &= \log l(\alpha \mid x) \\ L(\alpha \mid x) &= \sum_{i=1}^{N} (-\alpha \log(x_i) - \log(\zeta(\alpha))) \\ L(\alpha \mid x) &= \sum_{i=1}^{N} (\log(x_i) - N \log(\zeta(\alpha))). \end{aligned} \tag{11}$$

If we consider the possibility that there is a cut-off value, we have to modify equation (11) as follows [13]:

$$\begin{aligned} l(\alpha \mid x) &= \prod_{i=1}^{N} \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha} \\ L(\alpha \mid x) &= \sum_{i=1}^{N} \left[\log(\alpha - 1) - \log x_{\min} - \alpha \log \frac{x_i}{x_{\min}} \right] \\ L(\alpha \mid x) &= n(\alpha - 1) - n \log x_{\min} - \alpha \sum_{i=1}^{N} n \frac{x_i}{x_{\min}}. \end{aligned} \tag{12}$$

To find the parameter, first the $x_{\min}$ value must be found, which can be calculated using the Kolmogorov–Smirnov (KS) test as proposed by Clauset et al. [13], and then the derivative of equation (12) is set equal to 0 and solved for α. With this method, it is possible to pin down the companies with the largest market share distributed according to a power law and determine a cut-off value under which this distribution does not hold. In this case $x_{\min} = 1526$, a value that maximizes the likelihood estimation. As Figure 3 shows, this value is the market share of the sixth largest firm, which implies that only the market share of the six largest companies fit the power-law distribution. In this case $\alpha = 1.68$, which is equivalent to $q = 1.47$.

https://doi.org/10.25088/ComplexSystems.22.3.221

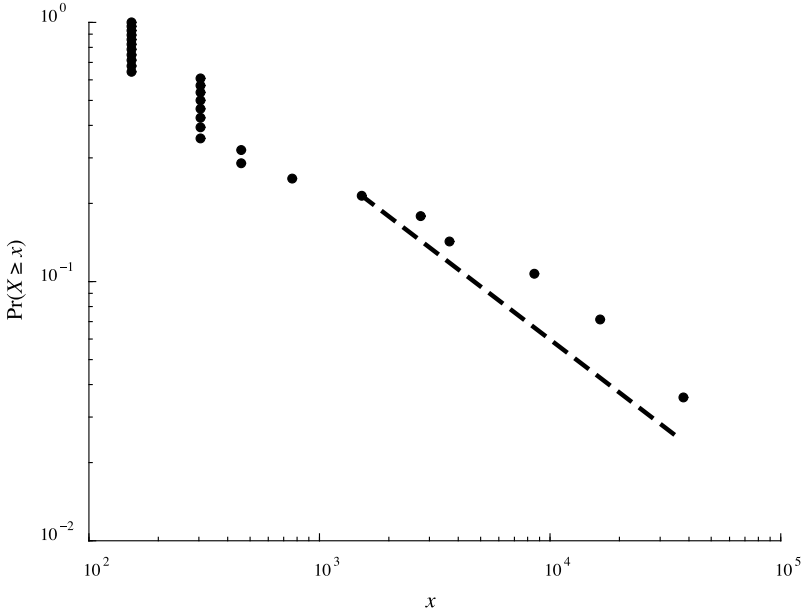

Figure 3. Power law with cut-off. The cola-drink market share data was fitted to a power-law distribution with cut-off.

To determine the goodness of fit of this model, it has been suggested to calculate a goodness-of-fit parameter through the KS statistic between the obtained model and the empirical data. Then it is compared with the KS of a thousand different sets of synthetic data randomly generated with the corresponding model according to the power-law hypothesis [13]. In the case of the world cola-drink industry, the KS statistic between our model and the empirical data is higher than that between the synthetic data and its corresponding model in 82 percent of the cases. (To perform this analysis, we used Aaron Clauset's implementation of the methods provided in http://tuvalu.santafe.edu/~aaronc/powerlaws/.) This result proves that there is not enough evidence to rule out the power law with cut-off hypothesis. Furthermore, if the same analysis is performed using the power law without cut-off ($\alpha = 1.56$), a statistic of 0 percent is obtained. This means that a power law with cut-off is a far better description of the data than a power-law distribution. Unfortunately, given the small size of the dataset and its low resolution, there is not enough evidence to rule out other models with cut-off such as stretched exponential, parabolic fractal, or log-normal distributions.

https://doi.org/10.25088/ComplexSystems.22.3.221

5. Conclusions

This paper shows that power-law distribution with cut-off is the most adequate model to describe the market shares distribution of the world cola-drink industry. Although a power law is a plausible description of the data, there is not enough evidence to rule out competing models such as stretched exponential, log-normal, or parabolic fractal distributions. Furthermore, the low resolution of the data and the small size of datasets, which are a characteristic of market shares in most industries, do not allow for enough confidence in the accuracy of the resulting models. In sum, the literature has suggested that some empirical datasets could be described with specific distributions, but because of the discussion in this paper, it is possible to conclude that such a claim must be taken with reserve.

On the other hand, given that the distribution that better describes the cola-drink industry's market shares is a power law with cut-off, the explanation of the underlying process that produces such a distribution could reside in self-organized criticality, highly optimized tolerance, Yule processes, or Pólya processes. But this is the issue of another paper.

6. References

[1] R. Gibrat, *Les Inégalités Economiques*, Paris: Libraire du Recueil Sirey, 1931.

[2] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Cambridge, MA: Addison-Wesley Press, 1949.

[3] J. J. Ramsden and Gy. Kiss-Haypál, "Company Size Distribution in Different Countries," *Physica A: Statistical Mechanics and Its Applications*, **277**(1–2), 2000 pp. 220–227. doi:10.1016/S0378-4371(99)00572-5.

[4] R. L. Axtell, "Zipf Distribution of US Firm Sizes," *Science*, **293**(5536), 2001 pp. 1818–1820. doi:10.1126/science.1062081.

[5] F. Buendía, "Polynomial Distribution of Market Share," *Business Management Dynamics*, **2**(3), 2012 pp. 22–25.

[6] B. Henderson, "The Rule of Three and Four," *Boston Consulting Group Perspectives*, **187**, 1976.

[7] R. D. Buzzel, "Are there 'Natural' Market Structures?," *Journal of Marketing*, **45**(1), 1981 pp. 42–51.

[8] H. Riemer, S. Mallik, and D. Sudharshan, *Market Shares Follow a Zipf Distribution*, Working Papers Series 02-0125, College of Business, University of Illinois, Urbana-Champaign, IL, 2002. http://www.business.illinois.edu/Working_Papers/papers/02-0125.pdf.

https://doi.org/10.25088/ComplexSystems.22.3.221

[9] G. U. Yule, "A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S.," *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, **213**(402–410),1925 pp. 21–87. doi:10.1098/rstb.1925.0002.

[10] R. Kohli and R. Sah, "Some Empirical Regularities in Market Shares," *Management Science*, **52**(11), 2006 pp. 1792–1798. doi:10.1287/mnsc.1060.0572.

[11] M. E. J. Newman, "Power Laws, Pareto Distributions, and Zipf's Law," *Contemporary Physics*, **46**(5), 2005 pp. 323–351. doi:10.1080/00107510500052444.

[12] J. Laherrère and D. Sornette, "Stretched Exponential Distributions in Nature and Economy: 'Fat Tails' with Characteristic Scales," *The European Physical Journal B*, **2**(4), 1998 pp. 525–539. doi:10.1007/s100510050276.

[13] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," *SIAM Review*, **51**(4), 2009 pp. 661–703. doi:10.1137/070710111.

[14] M. L. Goldstein, S. A. Morris, and G. G. Yen, "Problems with Fitting to the Power-Law Distribution," *The European Physical Journal B*, **41**(2), 2004 pp. 255–258. doi:10.1140/epjb/e2004-00315-6.

https://doi.org/10.25088/ComplexSystems.22.3.221