

Journal Pre-proof

The Zipf-Polylog distribution: Modeling human interactions through social networks

Jordi Valero, Marta Pérez-Casany, Ariel Duarte-López

PII: S0378-4371(22)00454-X

DOI: <https://doi.org/10.1016/j.physa.2022.127680>

Reference: PHYSA 127680

To appear in: *Physica A*

Received date: 24 January 2022

Revised date: 29 April 2022

Please cite this article as: J. Valero, M. Pérez-Casany and A. Duarte-López, The Zipf-Polylog distribution: Modeling human interactions through social networks, *Physica A* (2022), doi: <https://doi.org/10.1016/j.physa.2022.127680>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



The Zipf-Polylog Distribution: Modeling Human Interactions Through Social Networks

Jordi Valero^a, Marta Pérez-Casany^{a,b} and Ariel Duarte-López^{a,b,*}

^aTechnical University of Catalonia, Dpt. Statistics and OR, Avinguda Diagonal, 647, 08028, Barcelona, Spain

^bData Management Group (DAMA-UPC), Barcelona, Spain

ARTICLE INFO

Keywords:

Zipf's Law
Network Analysis
Mixture Distribution
Overdispersion
Degree Sequence

ABSTRACT

The Zipf distribution attracts considerable attention because it helps describe data from natural as well as man-made systems. Nevertheless, in most of the cases the Zipf is only appropriate to fit data in the upper tail. This is why it is important to dispose of Zipf extensions that allow to fit the data in its entire range. In this paper, we introduce the Zipf-Polylog family of distributions as a two-parameter generalization of the Zipf. The extended family contains the Zipf, the geometric, the logarithmic series and the shifted negative binomial with two successes, as particular distributions. We deduce important properties of the new family and demonstrate its suitability by analyzing the degree sequence of two real networks in all its range.


1. Introduction

Zipf's Law is widely used in various fields as an appropriate distribution for modeling data, such as frequencies of frequencies or ranks. Zipf (1949) applied this distribution to linguistics, for which he was interested in modeling the frequency of words appearing in a text a fixed number of times. It has been used in many different phenomena such that the underlying process makes the majority of the objects to be small, and very few objects to be very large. One can find many references of application of the Zipf model, also known as power law model, in ecology, financial market, internet topology, web visits, and demography, among others. Just to mention two of the most recent ones, the reader can look at the paper by Chacoma and Zanette (2021) for a prove that in linguistics, the frequency-rank relationship depends on the type of token considered and consequently, it reflects linguistic features related to the grammatical function of the token. In Asif, Hussain, Aşghar, Hussain, Raftab, Shah and Khan (2021) it is proved that the upper tail of the wealth distribution in the period 2010-2020 also follows a Zipf model, and that the Covid19 pandemic increased the disparity in wealth allowing the rich be richer.

The role of the Zipf distribution for modeling the degree sequence of a network has an special interest, because it is a particular case of power law distribution and it is known that many real networks have a degree sequence power law distributed. Based on that, there are several methodologies that allow to generate random networks that mimic the characteristics of the real ones, for example Barabási and Albert (1999); Barigozzi, Brownlees, Lugosi et al. (2018); Chung, Chung, Graham, Lu, Chung et al. (2006).

Hill and Woodroffe (1975) prove that the Zipf distribution is the limit distribution of the proportion of classes with exactly x units in a classification problem of N units in M categories, when the number of units to be classified tends to infinity. This result is important, because it explains a large number of situations in which Zipf's law appears. Nevertheless, as many research papers such as Newman (2005) and Dyer and Owen (2012) have pointed out, real data usually follow a Zipf pattern only in the tail of the distribution, although in some cases it may occur in only the central part of the distribution. In Broido and Clauset (2019), a diverse corpus of degree sequences in real-world networks from different domains is fitted by means of a power law distribution, which is the Zipf distribution defined only for values above a given threshold. In most of the degree sequences analyzed, no evidence of clear Zipf behavior has been observed. The ubiquity of the power law is a controversial topic in the network science community. The controversy was re-opened after the publication of the aforementioned paper by Broido and Clauset. The work by Holme (2019) intends to find a consensus between the group of researchers assuming scale-free networks as ideal objects and the

*Corresponding author

 jordi.valero@upc.edu (J. Valero); marta.perez@upc.edu (M. Pérez-Casany); ariel.duarte.lopez@upc.edu (A. Duarte-López)

ORCID(s): 0000-0002-7827-0225 (J. Valero); 0000-0003-3675-6902 (M. Pérez-Casany); 0000-0002-7432-0344 (A. Duarte-López)

The Zipf-Polylog Model

group interpreting these systems as particular objects part of the real world that are fine and fitted. For the author, both points of view are correct if one distinguishes between finite real networks and their projection to infinity. The author points out that applying concepts related to large-size systems to finite networks may be inexact. The family of distributions introduced in this paper allows for a more accurate fitting of the degree sequence of finite networks.

The objective of this paper is to propose the Zipf-Polylog family of distributions as a Zipf extension that can properly fit a large amount of real data sets not only in the tail, but across the whole range. The proposed model is a verified two-parameter exponential family containing the Zipf, the logarithmic series, the geometric distribution and the shifted negative binomial distribution with two successes as particular cases. In addition, any distribution in the Zipf-Polylog family that is not in the Zipf family has moments of any order, and verifies the Gauss principle. Moreover, it has maximum Shannon entropy under certain conditions. We also establish the hypothesis under which the extended family is a zero-truncated mixed Poisson (ZTMP) and/or a mixture of zero-truncated Poisson distributions (MZTP).

To illustrate the behavior of the presented model, we fit the degree sequences of two real networks: the Rovira i Virgili University email network that was analyzed in Guimera, Danon, Diaz-Guilera, Giralt and Arenas (2003); and the Facebook network considered in the paper by Traud, Mucha and Porter (2012a). We compare the results with the following: the results obtained by the Marshall-Olkin Extended Zipf distribution (MOEZipf) in Pérez-Casany and Casellas (2013) and Duarte-López, Pérez-Casany and Valero (2021), which was proven by Duarte-López, Prat-Pérez and Pérez-Casany (2015) to be appropriate for fitting this type of data; the Discrete Gaussian Exponential distribution (DGX) defined in Bi, Faloutsos and Korn (2001); and the zero-truncated Zipf-PSS.

2. The Zipf Distribution

The Zipf distribution (Zipf, 1949), also known as either a discrete Pareto or a zeta distribution, is a uni-parametric distribution defined in strictly positive integer numbers, such that its probabilities change inversely to a power of the values. Since it is a markedly skewed distribution, one may observe in a sample from this model values that sometimes differ by orders of magnitude. The Zipf distribution is highly recommended for modeling ranks and frequencies of frequency data. For example, Malone and Maher (2012) analyze its suitability for describing the frequency of chosen passwords. The use of this distribution for predicting consumer visitation patterns is shown in Krumme, Llorente, Cebrian, Moro et al. (2013). Their research evidences that, independently of shopper preferences, the Zipf distribution can be used to describe how frequently a client visits a store. Recently, in Chen (2021) it is proved that the level of urbanization of a country with a large population is related to the Zipf's parameter. Small values of the parameter are related to highly urbanized countries.

Although highly used in practice, the Zipf distribution has important limitations to fitting real data, as mentioned before. This is because plotting probabilities in log-log scale usually leads to pattern of real data being top-concave (Newman, 2005; Clauset, Shalizi and Newman, 2009; McKelvey et al., 2018), which is contrary to the expected linearity of a Zipf distribution. As a consequence, as already said, in many instances the Zipf is fitted only in the tail, with the corresponding loss of information.

It is said that a random variable (r.v.) X follows a Zipf distribution with parameter α if, and only if, its probability mass function (PMF) is equal to:

$$P(X = x) = \frac{x^{-\alpha}}{\zeta(\alpha)}, \quad x \in \{1, 2, 3, \dots\} \text{ and } \alpha > 1,$$

where $\zeta(\alpha) = \sum_{x=1}^{+\infty} x^{-\alpha}$ is the Riemann zeta function. The probability generating function (PGF) of a Zipf distributed r.v. is equal to:

$$h_X(z) = E(z^X) = \sum_{x=1}^{+\infty} \frac{z^x x^{-\alpha}}{\zeta(\alpha)} = \frac{Li_\alpha(z)}{Li_\alpha(1)}, \quad |z| < 1 \text{ and } \alpha > 1, \quad (1)$$

where the $Li_\alpha(z)$ is known as the *polylogarithm function* or *Li function of order α* , and it is equal to:

$$Li_\alpha(z) = \sum_{x=1}^{+\infty} \frac{z^x}{x^\alpha}. \quad (2)$$

The Li function of order α is defined for any arbitrary complex number α and any complex number z , such that $|z| < 1$. Nevertheless, using analytic prolongation, the Li function is defined throughout the complex plane. For $Re(\alpha) > 0$,

The Zipf-Polylog Model

and all z except for z that are real and larger or equal to one, the polylogarithm function may be expressed in terms of the integral of Bose-Einstein distribution as follows:

$$Li_{\alpha}(z) = \frac{1}{\Gamma(\alpha)} \int_0^{\infty} \frac{t^{\alpha-1}}{\frac{\exp(t)}{z} - 1} dt, \quad (3)$$

which can be checked by computing the Taylor expansion of the integrand and integrating termwise. Concerning to this paper, both α and z will only take values in the real line and $\alpha > 0$.

Important to observe that $Li_{\alpha}(1) = \zeta(\alpha)$ and thus, the Li function may be seen as an extension of the Riemann zeta function. Also if $\alpha = 1$, one has that $Li_1(z) = -\log(1 - z)$ which justifies the name of polylogarithm.

With respect to the moments of the Zipf distribution, it is known that the k -th moment, $k \in \mathbb{Z}^+$, is finite if, and only if, $\alpha > k + 1$. In this case, it is equal to:

$$E[X^k] = \frac{\zeta(\alpha - k)}{\zeta(\alpha)}, \quad \alpha > k + 1. \quad (4)$$

Based on (4), $k = 1$ allows one to directly obtain the first moment of the distribution that appears in (5). Computing (4) for $k = 1$ and $k = 2$ provides the variance of the distribution that appears in (6).

$$E[X] = \frac{\zeta(\alpha - 1)}{\zeta(\alpha)}, \quad \alpha > 2, \quad (5)$$

$$Var[X] = \frac{\zeta(\alpha - 2)\zeta(\alpha) - \zeta(\alpha - 1)^2}{\zeta(\alpha)^2}, \quad \alpha > 3. \quad (6)$$

Applying the logarithm to a Zipf distributed r.v., it is guaranteed that the transformed variable has moments of any order. This is a consequence of the fact that the logarithm reduces the data variability. Moreover, if x_1, x_2, \dots, x_n is a sample from an r.v. X with a Zipf(α) distribution, the maximum likelihood estimation (MLE) of α is equal to the solution of the equation:

$$E[\log(X)] = \frac{1}{n} \sum_{i=1}^n \log(x_i) = \overline{\log(x)}, \quad (7)$$

which is equivalent to applying the moment-method estimation to the logarithm of the variable. In Visser (2013), it is proved that the Zipf distribution is a discrete uni-parametric distribution with support on the strictly positive integer values with maximum Shanon entropy, for a fixed value of $\overline{\log(x)}$.

3. The Zipf-Polylog Generalization

In the first part of this section, a two-parameter generalization of the Zipf distribution is defined by means of adding an additional parameter to its PGF. The second part is devoted to proving the main properties of the presented model. As mentioned previously, the Zipf distribution is not flexible enough to model many real data sets, due to the fact that they usually show a top-concave pattern at the low values when plotted in log-log scale. This is one of the reasons why it is interesting to find Zipf generalizations that can adapt real observations to its entire range.

3.1. Definition

An r.v. Y is said to follow a Zipf-Polylog distribution with parameters $(\alpha, \beta) \in (-\infty, +\infty) \times (0, 1) \cup (1, +\infty) \times \{1\}$ (which from now on will be denoted by Zipf-Polylog(α, β)) if and only if, for any $z \in (-\infty, 1)$, its PGF is equal to:

$$h_Y(z) = \begin{cases} \frac{Li_{\alpha}(\beta z)}{Li_{\alpha}(\beta)} & \text{if } \beta \neq 1 \text{ and } \alpha \in (-\infty, +\infty) \\ \frac{Li_{\alpha}(z)}{Li_{\alpha}(1)} & \text{if } \beta = 1 \text{ and } \alpha > 1. \end{cases} \quad (8)$$

The Zipf-Polylog Model

To see that (8) defines a real PGF, it is only necessary to prove that: it takes the value one at one; it is analytical in an interval that contains the zero value; and the coefficients of the series expansion at zero are all positive. The first condition is true because

$$h_Y(1) = \frac{Li_\alpha(\beta)}{Li_\alpha(\beta)} = 1.$$

The second condition is also true because, by (2), $h_Y(z)$ is defined by means of a series expansion centered at zero. Finally, the third condition is true because the n -th coefficient is equal to $(Li_\alpha(\beta))^{-1} \beta^n n^{-\alpha} \geq 0$.

The support of a Zipf-Polylog distributed r.v. is the same as that of the Zipf distribution, i.e., strictly positive integer numbers. This is because

$$P(Y = 0) = h_Y(0) = Li_\alpha(0)/Li_\alpha(\beta) = 0.$$

To obtain the PMF of a Zipf-Polylog distribution, it is enough to see that

$$h_Y(z) = \sum_{x=1}^{+\infty} \frac{\beta^x x^{-\alpha}}{Li_\alpha(\beta)} z^x,$$

and, thus, according to the definition of the PGF, the probabilities are equal to:

$$P(Y = x) = \frac{\beta^x x^{-\alpha}}{Li_\alpha(\beta)}, \quad x = 1, 2, \dots \quad (9)$$

Observe that by defining $\gamma = -\log(\beta)$, the Zipf-Polylog distribution turns out to be the discrete version of the *power law distribution with exponential cut-off*, which appears in the paper by Clauset et al. (2009). Their paper states that the power law distribution is on the boundary of the parameter space. In what follows, this result is extended proving that it also contains other known families of distributions in the interior of its parameter space.

When $\alpha = 1$, (9) is plainly the PMF of the logarithmic-series distribution because $Li_1(\beta) = -\log(1 - \beta)$, as observed before. Moreover, if $\alpha = 0$ we obtain the geometric distribution with support $\{1, 2, 3, \dots\}$ and probability of success $p = 1 - \beta$. Finally, if $\alpha = -1$, given that $Li_{-1}(\beta) = \beta(1 - \beta)^{-2}$, (9) is equal to the PMF of a shifted negative binomial distribution with $r = 2$ successes and probability of success $p = 1 - \beta$. Figure 1 contains the parameter space of the Zipf-Polylog family, with the Zipf on the boundary, and the logarithmic-series; the geometric; and the shifted negative binomial distributions in the interior of the parameter space. Note that the three-parametric family of distributions known as *Lerch distribution* also contains the Zipf, the logarithmic series and the geometric distributions as particular cases (see Zörnig and Altmann, 1995).

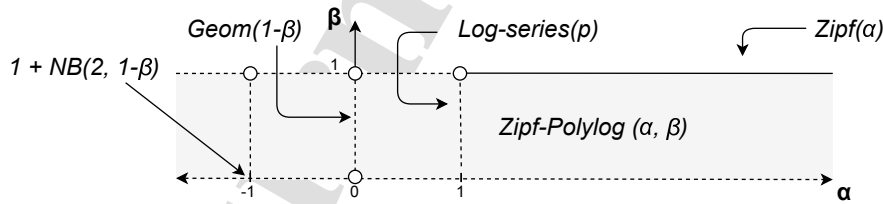


Figure 1: Parameter space of the Zipf-Polylog family of distributions with the geometric, the log-series, the Zipf and the shifted negative binomial with $r = 2$ families as particular cases.

As a consequence of (9), any distribution in the Zipf-Polylog family may be seen as a weighted version of a distribution in the Zipf family. If $\alpha > 1$, the Zipf-Poly(α, β) is the weighted version of the Zipf(α) distribution with weight function $w(x; \beta) = \beta^x > 0$. For $\alpha \in (0, 1)$, it is the weighted version of a Zipf($\alpha + 1$) with weight $w(x; \beta) = \beta^x x$. For $\alpha < -1$, it may be seen as a weighted version of a Zipf($-\alpha$) distribution with weight function $w(x; \beta, \alpha) = \beta^x x^{-2\alpha}$. Finally, for $\alpha \in (-1, 0)$ it is a weighted version of a Zipf($\alpha + 2$) with weight function $w(x; \beta, \alpha) = \beta^x x^2$.

The concept of *weighted distribution* first originates in Fisher (1934) and later became well established in Patil and Rao (1978). According to these authors, a weighted distribution is needed when the probability of observing a value x depends on of the size of the value.

The Zipf-Polylog Model

In our case, if we assume that the data come from an r.v. X with a Zipf(α) distribution, with a given $\alpha > 1$, and that

$$P(\text{Recording } x|X = x) = \beta^x,$$

then, the sample comes from a Zipf-Polylog(α, β). As a consequence, the β parameter may be interpreted as the probability of observing the value 1 when this is the true value. See Saghir, Hamedani, Tazeem and Khadim (2017) for a recent review on weighted distributions.

Figure 2 shows the probabilities of the Zipf-Polylog(α, β) for a fixed α and different values of β . More exactly, α has been taken to be equal to $-0.8, -0.5$ and 2.3 and β equal to $0.05, 0.1, 0.3$ and 0.7 . When $\alpha > 1$ (bottom part of the plot), the probabilities for $\beta = 1$ (Zipf) are also included. On the left-hand side, the probabilities are shown in the natural scale and, on the right-hand side, in the log-log scale. In the plot we can observe that, independently of the α value, the largest probability at one is attained for the smaller value of β . In fact, the probability at one as a function of β is equal to: $f(\beta) = \beta/Li_\alpha(\beta)$, and its derivative is equal to $f'(\beta) = Li_\alpha(\beta) - Li_{\alpha-1}(\beta) < 0$, which proves that this probability decreases by increasing β . For the remaining values, the probabilities increase by increasing the β value. We also observe a mode on the interior of the distribution for negative α and sufficiently large β . Comparing the three parts of the plot reveal that, independently of the value β , the probabilities tend to concentrate in the first values when α increases.

Figure 3 contains the probabilities for $\beta = 0.5$ and $\alpha = -3, -0.6, 0.5, 1.5$ and 2 . Observe that, with the exception of the initial integer values, the probabilities decrease by increasing α . One can also see a mode in the interior of the distribution for the lowest value of α .

3.2. Properties

This section is devoted to proving the main properties of the presented model. We first prove that the Zipf-Polylog is a two-parameter exponential family. Then, we show that the distributions not on the boundary of the parameter space can have moments of any order; and we describe the ratio of two consecutive probabilities. We end the section by proving that the Zipf-Polylog may be interpreted, under certain conditions, as a mixture distribution.

Theorem 1. *The Zipf-Polylog is a bi-parametrical exponential family with canonical parameter $\theta = (\alpha, -\log(\beta))$ and canonical statistic $T(x) = (-\log(x), -x)$.*

Proof. The Zipf-Polylog distribution may be parametrized in terms of (α, γ) , with $\gamma = -\log \beta$. With the new parametrization, the PGF and PMF are, respectively, equal to:

$$h_Y(z) = \frac{Li_\alpha(z \exp(-\gamma))}{Li_\alpha(\exp(-\gamma))}, \text{ and}$$

$$P(Y = x) = \frac{x^{-\alpha} \exp(-\gamma x)}{Li_\alpha(\exp(-\gamma))} = \frac{\exp(-\alpha \log(x) - \gamma x)}{Li_\alpha(\exp(-\gamma))}. \quad (10)$$

At the right-hand side of (10), one has that the Zipf-Polylog is an exponential family of order two, with canonical parameter $\theta = (\alpha, \gamma)$, parameter space $\Theta = (-\infty, +\infty) \times (0, +\infty) \cup (1, +\infty) \times \{0\}$, and canonical statistic $T(x) = (-\log(x), -x)$. \square

Observe that the Zipf-Polylog is not a regular exponential family in its entire space, since it has the Zipf model at the boundary ($\gamma = 0$); but it is regular if one considers the family defined in the interior of its parameter space. From the general theory of exponential families (Barndorff-Nielsen, 2014), one has that if x_1, x_2, \dots, x_n is a sample from an r.v. Y with a Zipf-Polylog distribution, and one defines $\overline{\log(x)} = 1/n \sum_{i=1}^n \log(x_i)$, then $t(x) = (\bar{x}, \overline{\log(x)})$ is a minimal and sufficient statistic. Also, the MLE of the parameter vector is the solution of the following system of equations:

$$\left. \begin{aligned} E[Y] &= \bar{x} \\ E[\log(Y)] &= \overline{\log(x)} \end{aligned} \right\},$$

which has a unique solution if $t(x)$ belongs to the interior of the convex hull of $t(\mathcal{N})$, being \mathcal{N} the space where takes values $t(x)$. Note that from (7), the second equation to be solved is the same as the one for finding the MLE for the Zipf distribution. The first equation corresponds to the Gauss Principle (see Teicher, 1961).

The Zipf-Polylog Model

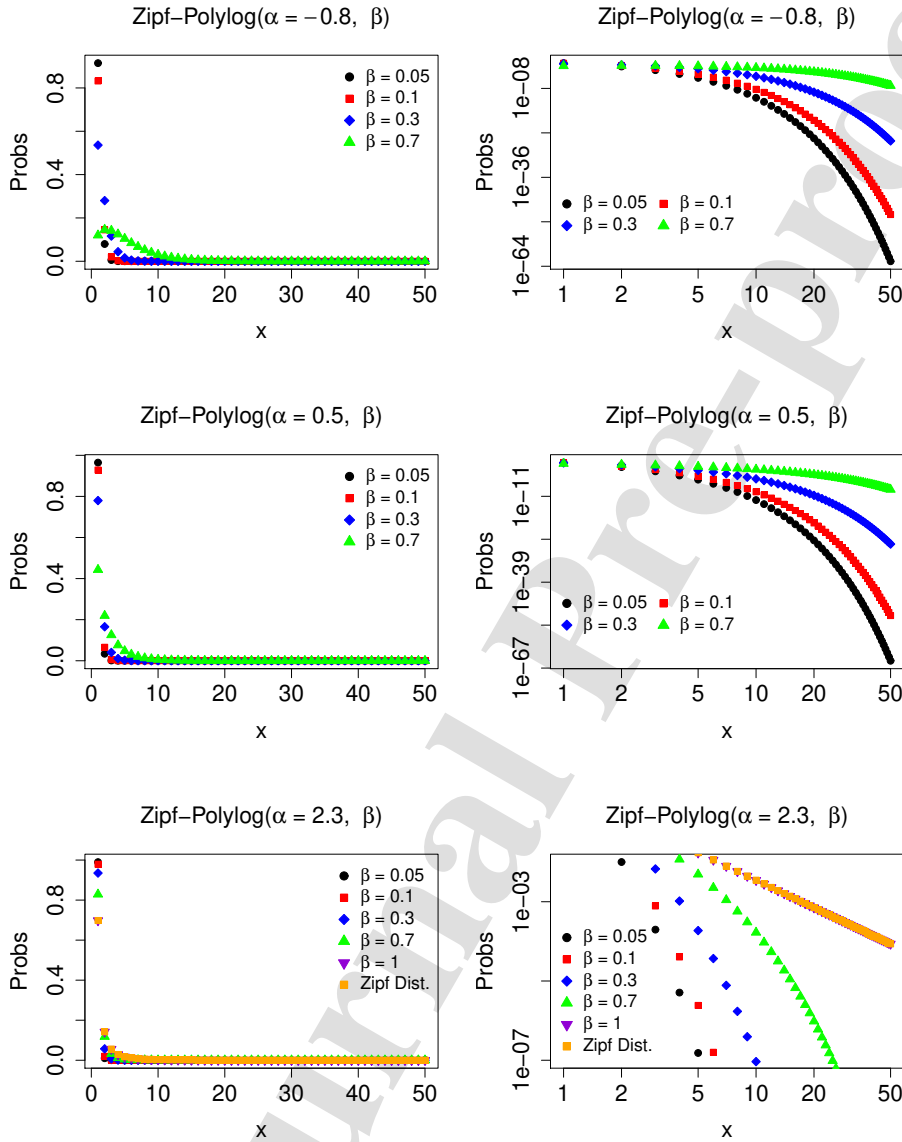


Figure 2: PMF of the Zipf-Polylog(α, β). At the top are negative values of the α parameter and different values of β . At the bottom are positive α values and different values of β . Both cases include, respectively, the plots in normal scale and in log-log scale. The probabilities of the Zipf distribution are included when $\alpha > 1$ and $\beta = 1$.

The Zipf-Polylog distribution is mentioned in Visser (2013) as a *hybrid geometric/power model* that is proven to be a bi-parametric model with support $\{1, 2, \dots\}$ and maximum Shannon entropy once \bar{x} and $\overline{\log(x)}$ are fixed. Note that the author lacks precision when saying that the model is defined for any $\beta < 1$. This is because the odd negative integer values for β have negative probabilities; thus, negative integer values for β are not possible.

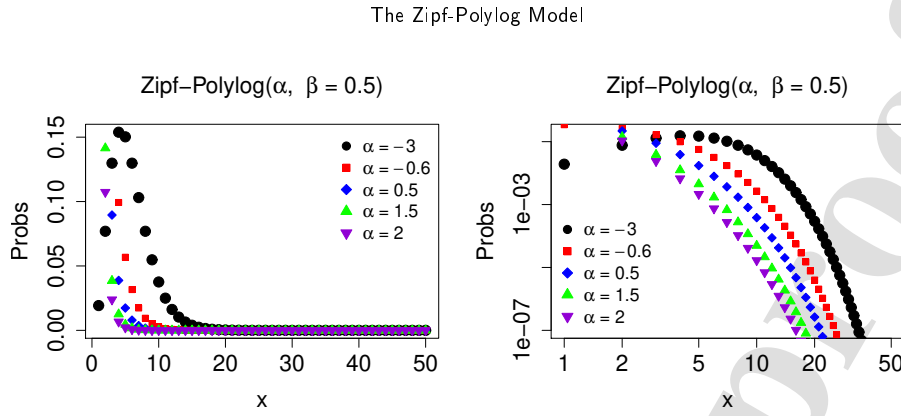


Figure 3: PMF of the Zipf-Polylog(α, β) for $\alpha = -3, -0.6, 0.5, 1.5, 2$ and $\beta = 0.5$. The left-hand side shows the probabilities in normal scale; while the right-hand side shows the same probabilities in log-log scale.

Proposition 1. If $Y \sim \text{Zipf-Polylog}(\alpha, \beta)$ with fixed $\alpha \in \mathbb{R}$ and $\beta \in (0, 1)$, then $E(Y^k) < +\infty$ for any $k \geq 1$.

Proof. Given that the moments of a distribution may be obtained by means of the factorial moments (Johnson, Kemp and Kotz, 2005), it is enough to prove that the factorial moments are finite. Denoting by μ'_k the factorial moment of order k of Y , we have:

$$\mu'_k = \left. \frac{\partial^k}{\partial t^k} h_Y(t) \right|_{t=1} < +\infty, \quad (11)$$

with $h_Y(z)$ being analytical in $(-\infty, 1]$. □

The next proposition explains the relationship between the ratio of two consecutive probabilities of an r.v. with a Zipf-Polylog distribution and the same ratio for a Zipf distribution with the same α parameter.

Proposition 2. If $Y \sim \text{Zipf-Polylog}(\alpha, \beta)$ with $\alpha > 1$, and $X \sim \text{Zipf}(\alpha)$, then the ratio of two consecutive probabilities of Y is proportional to the ratio of two consecutive probabilities of X , with β being the constant of proportionality.

Proof. By (9) we have:

$$\frac{P(Y = x + 1)}{P(Y = x)} = \frac{(x + 1)^{-\alpha} \beta^{x+1}}{(x)^{-\alpha} \beta^x} = \beta \left(\frac{x + 1}{x} \right)^{-\alpha} = \beta \frac{P(X = x + 1)}{P(X = x)}.$$

□

The following result states that any Zipf-Polylog with a positive value of α is a mixture of geometric distributions (see Johnson et al. (2005) for more details about the concept of Mixture Distributions).

Theorem 2. The Zipf-Polylog(α, β) distribution with $\alpha > 0$ and $\beta \in (0, 1)$ is a mixture of geometric distributions parametrized by means of $s = \log(\beta) - \log(1 - p) \in (\log(\beta), +\infty)$, with mixing distribution equal to

$$f(s; \alpha, \beta) = \frac{\frac{s^{\alpha-1}}{\exp(s)-\beta}}{\int_0^{+\infty} \frac{t^{\alpha-1}}{\exp(t)-\beta} dt} = \frac{\beta}{\Gamma(\alpha) Li_\alpha(\beta)} \frac{s^{\alpha-1}}{\exp(s) - \beta}. \quad (12)$$

Proof. The PGF of the geometric distribution parametrized with $s = \log(\beta) - \log(1 - p)$ is equal to:

$$\frac{pz}{1 - (1 - p)z} = \frac{(1 - \beta \exp(-s))z}{1 - \beta \exp(-s)z} = \frac{(\exp(s) - \beta)z}{\exp(s) - \beta z}. \quad (13)$$

The Zipf-Polylog Model

Thus, to prove the theorem, it is necessary to check that:

$$\frac{Li_{\alpha}(\beta z)}{Li_{\alpha}(\beta)} = \int_0^{+\infty} \frac{(\exp(s) - \beta)z}{\exp(s) - \beta z} f(s; \alpha, \beta) ds, \quad (14)$$

with $f(s; \alpha, \beta)$ defined as in (12). By substituting $f(s; \alpha, \beta)$ for its expression and taking into account (3), we have:

$$\int_0^{+\infty} \frac{(\exp(s) - \beta)z}{\exp(s) - \beta z} \frac{s^{\alpha-1}}{\exp(s) - \beta} ds = \frac{\beta z}{\beta \int_0^{+\infty} \frac{t^{\alpha-1}}{\exp(t) - \beta} dt} \int_0^{+\infty} \frac{s^{\alpha-1}}{\exp(s) - \beta z} ds = \frac{Li_{\alpha}(\beta z)}{Li_{\alpha}(\beta)}, \quad (15)$$

which proves the theorem. \square

Theorem 1 of Valero, Pérez-Casany and Ginebra (2010) characterizes the families of distributions with finite mean that are ZTMP, based on their PGF. The theorem states that a PGF $h(z)$ is the PGF of a ZTMP distribution if and only if it verifies that:

- (a) $h(0)=0$, $h(1) = 1$ and $h'(1) < +\infty$;
- (b) it is analytical in $(-\infty, 1)$;
- (c) all the coefficients of the series expansion of $h(z)$ around any point $z_0 \in (-\infty, 1)$ are strictly positive, except for the constant term that may be negative or zero; and
- (d) $\lim_{z \rightarrow -\infty} h(z) = -L$, with L being a finite strictly positive number.

Theorem 2 of the same paper establishes that the PGFs of MZTP distributions need to verify the first three conditions of Theorem 1, but not the last one. As a consequence, any ZTMP distribution is an MZTP distribution, but not the other way around. The characterizations are also true if the distribution has no finite mean. Theorem 4 establishes when the Zipf-Polylog is an MZTP, a ZTMP or none of them. Previously, we show that the geometric distribution is an MZTP, because this is necessary to prove Theorem 4. This results is stated in the next theorem.

Theorem 3. *The geometric distribution with parameter $p \in (0, 1)$ and domain $\{1, 2, \dots\}$ is an MZTP distribution with mixing distribution:*

$$f(\lambda; p) = \frac{p}{(1-p)^2} \exp(-\lambda/(1-p))(\exp(\lambda) - 1), \quad \lambda \in (0, +\infty). \quad (16)$$

Proof. The PGF of the geometric(p) distribution (which support the positive integers that are equal to or larger than one) is equal to $pz/(1 - qz)$, where $q = 1 - p$. Given that the PGF of the zero-truncated Poisson distribution is equal to $(\exp(\lambda z) - 1)/(\exp(\lambda) - 1)$ and that the PGF of an MZTP distribution is the integral, with respect to λ , of the PGF of the zero-truncated Poisson distribution multiplied by the density function of the mixing distribution, proving the proposition is equivalent to see that:

$$\frac{pz}{1 - qz} = \int_0^{+\infty} \frac{\exp(\lambda z) - 1}{\exp(\lambda) - 1} f(\lambda; p) d\lambda,$$

with $f(\lambda; p)$ defined as in (16). Substituting $f(\lambda; p)$ for its corresponding expression and taking into account that $z - 1/(1 - p) < 0$ because $z < 1$ and $p \in (0, 1)$, we have:

$$\begin{aligned} & \int_0^{+\infty} \frac{\exp(\lambda z) - 1}{\exp(\lambda) - 1} f(\lambda; p) d\lambda = \\ & = \frac{p}{(1-p)^2} \int_0^{+\infty} \left[\exp(\lambda(z - 1/(1-p))) - \exp(-\lambda/(1-p)) \right] d\lambda \\ & = \frac{p}{(1-p)^2} \left[\frac{\exp(\lambda(z - 1/(1-p)))}{z - 1/(1-p)} \Big|_0^{+\infty} + \frac{\exp(-\lambda/(1-p))}{1/(1-p)} \Big|_0^{+\infty} \right] \end{aligned}$$

The Zipf-Polylog Model

$$= \frac{-p(1-p)}{(1-p)^2} \left[\frac{1}{z(1-p)-1} + 1 \right] = \frac{pz}{1-(1-p)z}. \quad (17)$$

□

Theorem 4. The Zipf-Polylog(α, β) distribution verifies that:

a) if $\alpha > 0$, it is an MZTP distributions with mixing distribution defined for $\lambda > 0$ and equal to:

$$f(\lambda; \alpha, \beta) = \frac{\exp(\lambda) - 1}{\beta \Gamma(\alpha) Li_\alpha(\beta)} \int_0^{+\infty} s^{\alpha-1} \exp(s - \frac{\lambda}{\beta} \exp(s)) ds, \quad (18)$$

and it is not a ZTMP.

b) if $\alpha = 0$, it is an MZTP distribution and also a ZTMP distribution.

c) if $\alpha < 0$, it is neither an MZTP nor a ZTMP.

Proof. To prove the first statement of a), it is necessary to see that

$$\frac{Li_\alpha(\beta z)}{Li_\alpha(\beta)} = \int_0^{+\infty} \frac{\exp(\lambda z) - 1}{\exp(z) - 1} f(\lambda; \alpha, \beta) d\lambda,$$

where $f(\lambda; \alpha, \gamma)$ is defined as in (18). From Theorem 2 we have

$$\frac{Li_\alpha(\beta z)}{Li_\alpha(z)} = \int_0^{+\infty} \frac{(\exp(s) - \beta)z}{\exp(s) - \beta z} \frac{\beta}{\Gamma(\alpha) Li_\alpha(\beta)} \frac{s^{\alpha-1}}{\exp(s) - \beta} ds. \quad (19)$$

Moreover, taking into account (13), by Theorem 3 we have:

$$\frac{(\exp(s) - \beta)z}{\exp(s) - \beta z} = \int_0^{+\infty} \frac{\exp(\lambda z) - 1}{\exp(\lambda) - 1} f^*(\lambda; s) d\lambda,$$

where

$$f^*(\lambda; s) = f(\lambda; 1 - \beta \exp(-s)) = \frac{\exp(s) - \beta}{\beta^2} \exp(-\frac{\lambda}{\beta} \exp(s)) (\exp(\lambda) - 1).$$

Thus, we have:

$$\frac{(\exp(s) - \beta)z}{\exp(s) - \beta z} = \int_0^{+\infty} \frac{\exp(\lambda z) - 1}{\exp(\lambda) - 1} \frac{\exp(s) - \beta}{\beta^2} \exp(-\frac{\lambda}{\beta} \exp(s)) (\exp(\lambda) - 1) d\lambda.$$

Now, substituting the last equality in (19) gives:

$$\begin{aligned} \frac{Li_\alpha(\beta z)}{Li_\alpha(z)} &= \int_0^{+\infty} \left[\int_0^{+\infty} \frac{\exp(\lambda z) - 1}{\exp(\lambda) - 1} \frac{\exp(s) - \beta}{\beta^2} \exp(-\frac{\lambda}{\beta} \exp(s)) (\exp(\lambda) - 1) d\lambda \right] \\ &\quad \frac{\beta}{\Gamma(\alpha) Li_\alpha(\beta)} \frac{s^{\alpha-1}}{\exp(s) - \beta} ds = \\ &= \int_0^{+\infty} \frac{\exp(\lambda z) - 1}{\exp(\lambda) - 1} \left[\frac{\exp(\lambda) - 1}{\beta \Gamma(\alpha) Li_\alpha(\beta)} \int_0^{+\infty} \exp(s - \frac{\lambda}{\beta} \exp(s)) s^{\alpha-1} ds \right] d\lambda \\ &= \int_0^{+\infty} \frac{\exp(\lambda z) - 1}{\exp(z) - 1} f(\lambda; \alpha, \beta) d\lambda, \end{aligned}$$

and thus, any Zipf-Polylog with a positive value of α is an MZTP distribution. To see that it is not a ZTMP distribution it is enough to see that:

$$\lim_{t \rightarrow -\infty} h_Y(t) = \lim_{t \rightarrow -\infty} \frac{Li_\alpha(\beta t)}{Li_\alpha(\alpha)} = -\infty.$$

The Zipf-Polylog Model

To prove *b*) it is necessary to remember that when $\alpha = 0$, the Zipf-Polylog reduces to the geometric distribution with parameter $p = 1 - \beta$ and, in this case, it is an MZTP distribution, as proved in Theorem 3. Moreover, given that

$$\lim_{z \rightarrow -\infty} \frac{pz}{1 - (1-p)z} = -\frac{p}{1-p},$$

the Zipf-Polylog is also a ZTMP distribution.

Let us now prove *c*). To that end, we see that when $\alpha < 0$, $h'_Y(z) < 0$ at some interval on the negative real line, this means that condition (c) of Theorem 1 of Valero et al. (2010) is not verified. To prove that the first derivative of $h_Y(z)$ is negative, it is enough to see that the first derivative of the Li_α function is also negative. This is proved by distinguishing whether or not α is a negative integer.

- 1) Assume that α is a negative integer value. Then, taking into account that the Li_α function for integer values of α verifies:

$$z \frac{\partial Li_\alpha(z)}{\partial z} = Li_{\alpha-1}(z),$$

we have that when $\alpha = -1$,

$$\frac{\partial Li_{-1}(z)}{\partial z} = \frac{1}{z} Li_0(z) = \frac{1}{z} \frac{z}{1-z} = \frac{1}{1-z}, \quad (20)$$

and it is negative when $z < -1$. For $\alpha = -n$, by applying (20) recursively n times, we have that for certain real values a_1, a_2, \dots, a_{n-2} ,

$$\frac{\partial Li_{-n}(z)}{\partial z} = \frac{z(z^{n-1} + a_{n-2}z^{n-2} + \dots + a_1z + 1)}{(1-z)^n},$$

from which we have $\forall n \ Li'_{-n}(0) = 0$. Moreover, given that $\lim_{z \rightarrow -\infty} Li_{-n}(z) = 0$, it must be negative at a certain interval on the negative real line.

- 2) If α is negative but not an integer number, we also have $Li_\alpha(0) = 0$, and given that

$$Li'_\alpha(z) = 1 + \frac{z}{2^{\alpha-1}} + \frac{z^2}{3^{\alpha-1}} + \frac{z^3}{4^{\alpha-1}} + \dots,$$

we have $Li'_\alpha(0) = 1$. Moreover,

$$\lim_{z \rightarrow 0^-} Li_\alpha(z) = \lim_{u \rightarrow +\infty} Li_\alpha(-\exp(-u)) = \frac{-u^\alpha}{\Gamma(\alpha + 1)} = 0,$$

which proves that, at some interval on the negative real line, $Li'_\alpha(z) < 0$. Consequently, for negative values of α , the Zipf-Polylog is neither a ZTMP nor an MZTP distribution. See Figure 4 for representations of the $Li_\alpha(z)$ function for $\alpha = -0.8, -1, -1.7$ and -2.55 , and $z \in (-\infty, 0)$. □

In economics, informetrics and information sciences it is quite usual to consider a Lorenz curve (LC) as an instrument that allows to see, for instance, the proportion of income earned by a given percentage of population. In epidemiology it is also used to illustrate the exposure-disease association. In the paper by Sarabia, Gómez-Déniz, Sarabia and Prieto (2010) the authors explain how from an initial LC $L_0(\cdot)$, it is possible to generate a parametric family of LCs that contains the initial one in the limit, and that is more flexible as a consequence of having an additional parameter. The generalization arises from compounding the PGF of a discrete and strictly positive r.v. X with $L_0(\cdot)$ (see p. 527 of the aforementioned paper). The authors consider different distributions for X and, in particular, analyze the case of the Zipf distribution. If instead of the Zipf family we use the Zipf-Polylog as a distribution family for X , one obtains that the resulting LCs are equal to:

$$L_{ZP}(p) = \begin{cases} \frac{Li_\alpha(L_0(p))}{Li_\alpha(1)} & \text{if } \beta = 1 \\ \frac{Li_\alpha(\beta L_0(p))}{Li_\alpha(\beta)} & \text{if } \beta \neq 1 \end{cases}$$

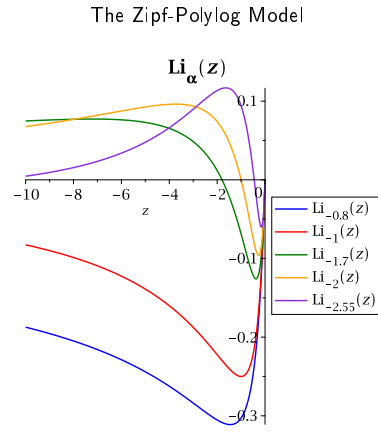


Figure 4: Plots of the $Li_{\alpha}(z)$ function for $\alpha = -0.8, -1, -1.7$ and -2.55 , and $z \in (-\infty, 0)$

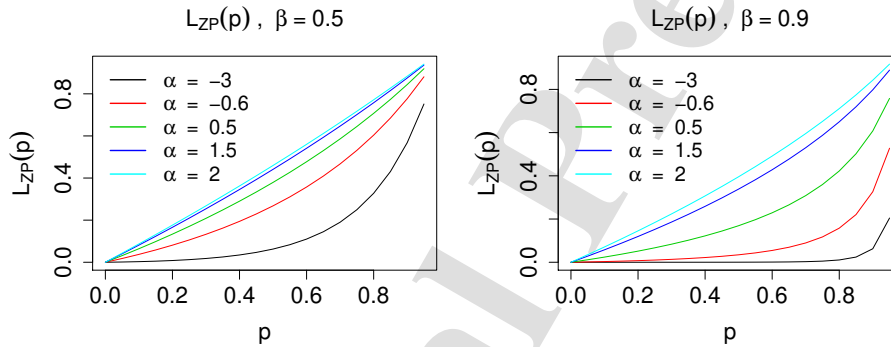


Figure 5: LCs associated with the Zipf-Polylog distribution taking as initial LC $L_0(p) = p$. On the left-hand side for $\alpha = -3, -0.6, 0.5, 1.5, 2$ and $\beta = 0.5$ and, on the right-hand side, for $\beta = 0.9$ and the same values of α

Figure 5 contains the plots of the LCs associated to the Zipf-Polylog distribution taking as initial LC $L_0(p) = p$. On the left-hand side for $\alpha = -3, -0.6, 0.5, 1.5, 2$ and $\beta = 0.5$ and, on the right-hand side, for $\beta = 0.9$ and the same values of α . Figure 6 contains some of the LCs corresponding to the Zipf distribution, more exactly the ones associated to $\beta = 1$ and $\alpha = 1.5, 2, 3.5, 5, 6.5$. We leave as future work to do more research around the role of the proposed family of distributions in the econometric environment.

3.3. The Zipf-Polylog in regression models

It would be of a great interest to consider a response variable with a Zipf-Polylog distribution in the presence of covariates. This not part of the objectives pursued in this paper but in this section, we introduce how this can be done.

We consider the Zipf-Polylog distributions with parameters in the interior of the parameter space, that is $\beta \in (0, 1)$ and $\alpha \in (-\infty, +\infty)$. First, we need to take into account that, if $Y \sim \text{Zipf-Polylog}(\alpha, \beta)$ with $\beta \in (0, 1)$, then

$$\mu = E(Y) = Li_{\alpha-1}(\beta) / Li_{\alpha}(\beta). \quad (21)$$

Observe that parameter β is the argument of both Li functions and that α is related to the particular Li functions considered. If we assume that parameter α changes with the covariates and that β remains constant, for a given realization vector $y^t = (y_1, y_2, \dots, y_n)$ of the response vector Y^t , and for a given monotone and invertible function $\eta(\cdot)$, the model may be written as:

$$\eta(\mu_i) = X_i^t \gamma \quad ,$$

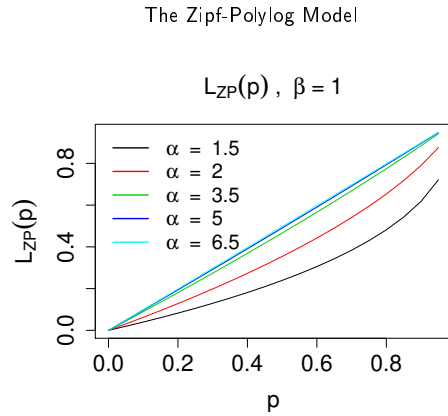


Figure 6: LCs associated with the Zipf distribution for $\alpha = 1.5, 2, 3.5, 5, 6.5$, and $\beta = 1$.

being $X_i = (1, x_{i1}, x_{i2}, \dots, x_{ik})$ the experimental conditions under which y_i has been observed, and γ the parameter vector. The γ parameter vector is proposed to be estimated by maximum likelihood, that is maximizing the log-likelihood function that is equal to:

$$l(\alpha_i, \beta; y) = \log(\beta) \sum_{i=1}^n y_i - \sum_{i=1}^n \alpha_i \log(y_i) - \sum_{i=1}^n \log(Li_{\alpha_i}(\beta)).$$

The maximum needs to be found numerically, and given that we do not have an explicit expression for α_i as a function of the mean, at each step it is necessary to solve n times equation (21). Also it will be necessary to invert n times the function $\eta(\cdot)$ which can require a lot of computational time. All the process can be done by using the functions *optim* or *nlm* from R as it is suggested in the paper by Sáez-Castillo and Conde-Sánchez (2013). To increase the model flexibility it could also be assumed that parameter β evolves as a function of some covariates.

4. Applications

The aim of this section is to illustrate the performance of the Zipf-Polylog family of distributions when it is used to fit real data. In particular, the two case studies presented belong to the field of Network Analysis and the distribution is used to fit the degree sequence of real networks. In order to ensure the suitability of the presented model, the results are compared with those obtained by other bi-parametric families of distributions: the DGX (Bi et al., 2001) and the MOEZipf (Pérez-Casany and Casellas, 2013; Duarte-López et al., 2021), both of which with support in the strictly positive integers; and the zero truncation of the Zipf-PSS (Duarte-López, Pérez-Casany and Valero, 2020). The last one requires to be zero-truncated because its support is the positive integer values including the zero. The work by Duarte-López et al. (2015) demonstrates the suitability of the MOEZipf model for fitting this type of data and, thus, that it is an appropriate model for comparison. For all the families, the parameter estimates are the m.l.e. For the particular cases of MOEZipf and Zipf-PSS, the m.l.e. were obtained with the R-package *zipfextR* (Duarte-López and Pérez-Casany, 2020). The models are compared by means of the log-likelihood and the Akaike Information Criterion (AIC). At the end of each example, the Likelihood Ratio Test (LRT) is performed to compare the Zipf model with its Zipf-Polylog extension. Since the Zipf distribution belongs to the boundary of the parameter space, the likelihood ratio statistic follows a 50:50 mixture of χ_0^2 and χ_1^2 (Self and Liang, 1987).

4.1. Case Study 1: University Rovira i Virgili, E-mail Network

The first case study analyzes the degree sequence of the undirected e-mail network at the University Rovira i Virgili (URV) in the year 2003. This data set was created by researchers in this institution and it is analyzed in the paper by Guimera et al. (2003), in which the authors inspect the self-similarity structure of the network. In their words, this is the structure replication at different levels of the communication network. The network comprises a total of 1133 nodes, all of them belonging to the giant component; and there are neither loops nor multi-edges. In this particular

The Zipf-Polylog Model

Table 1

Fitted distributions jointly with their parameter estimates, confidence intervals, log-likelihood and the AIC goodness-of-fit measure, for the degree sequence of the e-mail network of the URV.

| Distribution | $\hat{\alpha}$ | StdError | CI | $\hat{\beta}$ | StdError | CI | Log-likelihood | AIC |
|--------------|-----------------------|----------|------------------|------------------------|----------|-------------------|----------------|------------------|
| Zipf-Polylog | $\hat{\alpha}=0.1774$ | 0.0553 | (0.0689, 0.2859) | $\hat{\beta}=0.9108$ | 0.0059 | (0.8994, 0.9223) | -3632.2648 | 7268.5295 |
| DGX | $\hat{\mu}=1.7524$ | 0.0345 | (1.6848, 1.82) | $\hat{\sigma}=1.0924$ | 0.0275 | (1.0385, 1.1463) | -3673.7135 | 7351.4271 |
| MOEZipf | $\hat{\alpha}=2.4980$ | 0.0446 | (2.4105, 2.5855) | $\hat{\beta}=28.8413$ | 3.4213 | (22.1355, 35.547) | -3698.9643 | 7401.9286 |
| zt-Zipf-PSS | $\hat{\alpha}=2.0056$ | 0.0271 | (1.9524, 2.0588) | $\hat{\lambda}=3.1180$ | 0.1086 | (2.9051, 3.3309) | -3722.6859 | 7449.3719 |
| Zipf-PE | $\hat{\alpha}=2.0102$ | 0.0261 | (1.959, 2.0613) | $\beta=5.9809$ | 0.3121 | (5.3691, 6.5927) | -3770.9683 | 7545.9366 |
| Zipf | $\hat{\alpha}=1.4374$ | 0.0132 | (1.4116, 1.4632) | - | - | - | -4106.6291 | 8215.2582 |

network, an edge is created between two nodes if user A sends an email to user B and user B sends an email to user A. The number of edges in the network is 5451. This data set can be downloaded from the network repository KONECT (Kunegis, 2013).

Table 1 contains the results obtained after fitting the Zipf model and the four models mentioned in this section's introduction. It can be observed that the best fit is obtained with the Zipf-Polylog distribution, since this is the one that gives the minimum value of the AIC and the maximum log-likelihood. Figure 7 shows the fits obtained by the four models using the real data. Observe not only that the MOEZipf and the zero-truncated Zipf-PSS distributions behave very similarly, but also that the DGX gives a slightly better fit than the previous two models, because it shows a larger curvature at the beginning. Nevertheless, the Zipf-Polylog is the only one that is able to give a probability at one that is pretty close to the real one, and unlike the other distributions, it does not show a linear pattern for values greater than 10.

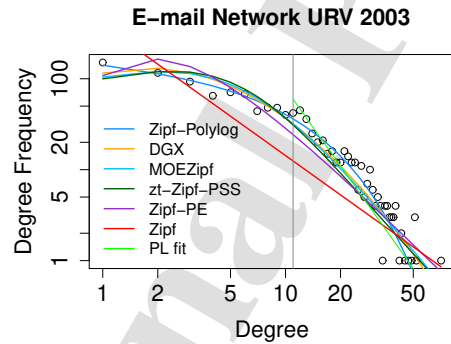


Figure 7: Degree sequence of the URV e-mail network for the year 2003, and the fit obtained by each of the considered models.

LRT is performed to compare the Zipf-Polylog distribution with that of the Zipf, that is, to test if $\beta = 1$ vs. $\beta < 1$. Given that for the Zipf distribution the log-likelihood is equal to -4106.629 and for the Zipf-Polylog it is equal to -3632.26 , the likelihood ratio statistics is equal to $-2[-4106.629 - (-3632.26)] = 948.738$. Under the null hypothesis, the likelihood ratio statistic follows a 50:50 mixture of χ_0^2 and χ_1^2 . Thus, the critical value for $\alpha = 0.05$ is equal to $0.5 \chi_{0.95,1}^2 = 0.5 \cdot 3.84 = 1.92$. Given that $948.738 \geq 1.92$, we clearly reject the null hypothesis and conclude that the Zipf-Polylog gives a better fit than the Zipf model.

Observe that the parameter estimates of the Zipf-Polylog do not allow for their direct interpretation as a weighted version, because α is smaller than one. Nevertheless, by transforming the model as suggested in Section 3, defining $\alpha^* = \hat{\alpha} + 1$ and considering the weight function $w(x; \beta) = x \cdot 0.91^x$, one can assume that the data follow a weighted version of a Zipf(1.18) distribution. Parameter $\hat{\beta} = 0.91$ is interpreted as the probability of observing that the degree of a node is one when it is actually equal to one. Hence, values of $\hat{\beta}$ close to one ensure that almost all the nodes with degree one are observed to be like they are in reality. On the other hand, as a consequence of Theorem 4, the

The Zipf-Polylog Model

Table 2

Fitted distributions jointly with their parameter estimates, confidence intervals, log-likelihood and the AIC goodness-of-fit measure, for the Facebook degree sequence at the University of California.

| Distribution | $\hat{\alpha}$ | Std Error | CI | $\hat{\beta}$ | Std Error | CI | Log-likelihood | AIC |
|--------------|--------------------------|-----------|--------------------|--------------------------|-----------|----------------------|----------------|-------------|
| Zipf-Polylog | $\hat{\alpha} = -0.0566$ | 0.0160 | (-0.0879, -0.0252) | $\hat{\beta} = 0.9789$ | 0.0004 | (0.9782, 0.9797) | -44059.6760 | 88123.3521 |
| MOEZipf | $\hat{\alpha} = 2.5038$ | 0.0138 | (2.4767, 2.5309) | $\hat{\beta} = 401.8319$ | 23.5524 | (355.6692, 447.9946) | -44847.7235 | 89699.4471 |
| DGX | $\hat{\mu} = 3.4084$ | 0.0129 | (3.383, 3.4335) | $\hat{\sigma} = 1.2084$ | 0.0095 | (1.19, 1.2274) | -44936.4704 | 89876.9409 |
| zt-Zipf-PSS | $\hat{\alpha} = 1.7104$ | 0.0049 | (1.7008, 1.7201) | $\hat{\lambda} = 6.6136$ | 0.0701 | (6.4763, 6.7509) | -46066.8610 | 92137.7219 |
| Zipf-PE | $\hat{\alpha} = 1.7408$ | 0.0056 | (1.7298, 1.7518) | $\hat{\beta} = 11.4596$ | 0.2095 | (11.0489, 11.8702) | -46666.5545 | 93337.1091 |
| Zipf | $\hat{\alpha} = 1.2542$ | 0.0027 | (1.2489, 1.2595) | - | - | - | -51935.0908 | 103872.1816 |

Zipf-Polylog($\hat{\alpha}, \hat{\beta}$) is an MZTP distribution. Thus, it is possible to say that the number of connections of the nodes come from a zero-truncated Poisson distribution, although each node has a different Poisson parameter.

4.2. Case Study 2: Facebook 100, the University of California, Santa Cruz network

The second analyzed data set appears in Traud et al. (2012a). In their work, the authors studied the complete Facebook network of 100 universities and colleges in the United States on a non-specified day in September 2005, with the aim of comparing homophily and determining its community structure. The comparison was made using partitions of data that was based on categorical information collected for each user, such as, gender, major, class year, etc. The authors remark that at the time the data were collected, it was necessary to have an .edu e-mail address for being able to create a Facebook profile. A peculiarity of this dataset is that the links between different institutions are ignored, which allows for unconnected networks, one for each of the different institutions considered.

In this particular example, the degree sequence associated with the University of California, Santa Cruz (UCSC) is analyzed. The network comprises a total of 8979 nodes and 224578 edges. The degree sequence is available through the git-hub repository: <https://github.com/adbroido/SFAnalysis>, mentioned in Broido and Clauset (2019).

Table 2 contains the results obtained after fitting the same models as before. Similarly to the previous example, the Zipf-Polylog family of distributions provides the best fit because it is the one that gives not only the maximum value of the log-likelihood, but also the minimum value of the AIC. As can be appreciated in the table, the goodness-of-fit obtained by the DGX and the MOEZipf models are quite similar, but not as good as that of the Zipf-Polylog. The worst two-parametric models are clearly the zero-truncated Zipf-PSS and the Zipf-PE. Figure 8 illustrates the performance of each one of the models jointly with the real observations. Observe that the Zipf-Polylog is the only one able to adjust the frequency of the smallest degrees. The DGX, the MOEZipf, the zero truncated Zipf-PSS and the Zipf-PE do not fit the real observations properly, since on the one hand they underestimate the first integer values and, on the other, they overestimate the middle values. In addition, these distributions also show a heavier right-hand tail than the Zipf-Polylog, which decays similarly to the real data.

By means of the log-likelihood values of the Zipf and Zipf-Polylog models (see Table 2), the likelihood ratio statistic is computed and comes out equal to $-2[-51935.09 - (-44059.68)] = 15750.82$, which is clearly larger than the critical value 1.92. Hence, the null hypothesis is rejected with a significance level of 0.05, and β is significantly different from one.

Given that $\hat{\alpha} \in (-1, 0)$, one can assume that the data follow a weighted Zipf($\hat{\alpha} + 2 = 1.94$) distribution with weight function $w(x; \beta) = x^2 \cdot 0.98^x$. Since $\hat{\beta} = 0.98$ is close to one, the same interpretation made in the first example continues to be valid in this case. Based on Theorem 4, this dataset is not fitted by an MZTP because $\hat{\alpha}$ is negative.

4.3. Large scale analysis

The goal of this section is to provide an overview of the performance of the Zipf-Polylog distribution when applied to a large-scale study. We show the results obtained by fitting all the degree sequences of the networks in the ICON (Clauset, Tucker and Sainz, 2020) repository of the University of Colorado. These networks are also analyzed in the work by Broido and Clauset (2019) already mentioned in the paper. They are interesting because they describe phenomena of multiple domains as: biology, technology, social networks, or transportation among others.

The whole collection of degree sequences appears in the github repository mentioned in Subsection 4.2, and it contains 3649 degree sequences. In the analysis we have excluded the sequences that verify one of the three following conditions: i) it contains information about isolated nodes, ii) it has less than 30 value on its support, and iii) it takes more than two minutes to fit at least one of the six distribution families considered. The final number of sequences fitted is 2759.

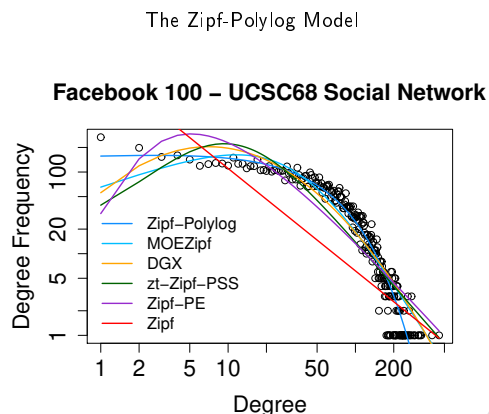


Figure 8: Degree sequence of the Facebook network at UCSC, and the fit obtained by each of the considered models.

Table 3

For each probability distribution, the table contains the number of the degree sequences (%) for which the model provides the best fit, based on the AIC criterion.

| Distribution | Total(%) |
|---------------------|-------------|
| Zipf-PE | 1740 (63%) |
| zt-Zipf-PSS | 522 (19%) |
| Zipf-Polylog | 316 (11.5%) |
| MOEZipf | 89 (3.2%) |
| DGX | 86 (3.1%) |
| Zipf | 6 (0.2%) |
| Total | 2759 |

Table 3 contains the ranking of the models based on the number of sequences in which they provide the best fit based on the AIC criterion. When we say that a probability distribution provides the best fit, it means that it is the best among the ones considered in this research paper. Moreover, this does not necessary imply that the data set is properly fitted by the proposed model. It is also important to mention that, in many cases, the models Zipf-PE, MOEZipf, and zt-Zipf-PSS provides similar results.

The Zipf-Polylog is the third distribution in the ranking and provides the best fits for 316 sequences. One can assume that this distribution is appropriate for fitting sequences that show a moderate decrease in the probabilities at the first values, and an abruptly decreasing of the probabilities for large values. This can be seen in the four examples presented in Figure 9. On the top left- and right- hand side of that figure, we show the degree sequences of the within-college social network (thefacebook.com) for Dartmouth and Harvard respectively. In these networks the nodes represent users, and the edges are created if two users are friends. This data set appears in the work by (Traud, Mucha and Porter, 2012b). On the bottom-left hand side, we show the degree sequence obtained using Roget's Thesaurus book (Knuth, 1993). This network belongs to the domain of Informational Language. The nodes represent the categories in the 1879 edition of Roget's Thesaurus of English Words and Phrases, and edges represent references among categories. On the bottom-right hand side, we show the fits obtained for the degree sequence of the scientific collaboration network, of the community of Astrophysics (Newman, 2001) from 1995 to 1999. The nodes represent scientists, and an edge is placed between a pair of nodes if the scientists co-author a paper.

5. Conclusions

In this paper a new family of discrete probability distributions that generalizes the Zipf distribution is proposed. Some advantages of the introduced family are:

- It has finite moments of any order out of the boundary of the parameter space.

The Zipf-Polylog Model

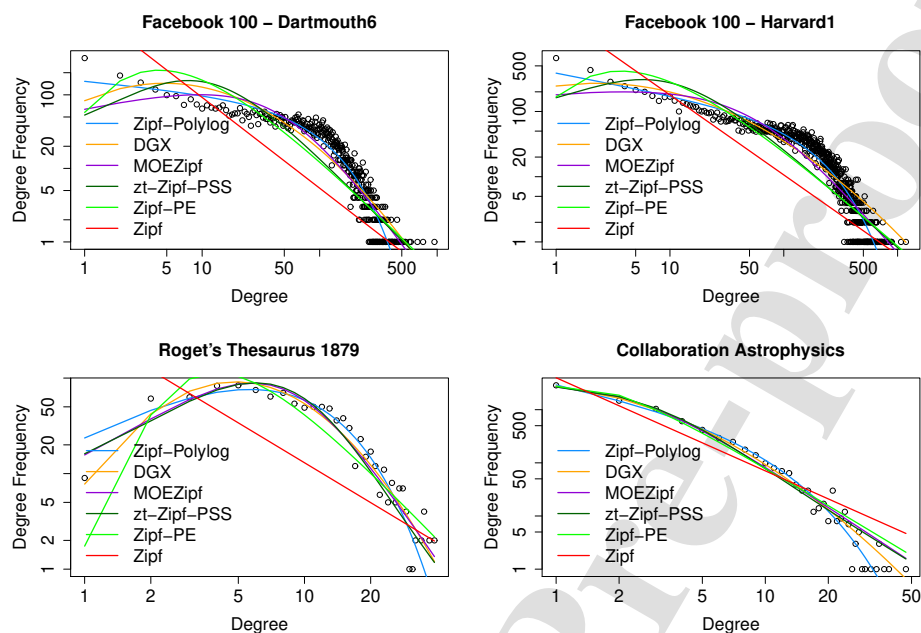


Figure 9: On the top, the fits obtained with the degree sequences of the Facebook 100 network. On the left-hand side for Dartmouth, and on the right-hand side for Harvard. On the bottom-left hand side, the fits obtained for the Roget's Thesaurus book and, on the bottom-right hand side, for the collaboration network of astrophysics on 1995-1999.

- It contains several classical distributions as particular cases.
- It is able to properly fit frequencies of frequency data from many different areas, in particular when in log-log scale they show a concave pattern at the beginning, and then the probabilities decrease sharply.

Acknowledgements

This research was supported in part by the grants TIN2017-89244-R from MINECO (Ministerio de Economía, Industria y Competitividad), Spain; the recognition 2017SGR-856 (MACDA) and the grant 2018_FI_B2_00064 from AGAUR (Generalitat de Catalunya), Spain.

References

- Asif, M., Hussain, Z., Asghar, Z., Hussain, M.I., Raftab, M., Shah, S.F., Khan, A.A., 2021. A statistical evidence of power law distribution in the upper tail of world billionaires's data 2010–20. *Physica A: Statistical Mechanics and its Applications*, 126198.
- Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. *science* 286, 509–512.
- Barigozzi, M., Brownlee, C., Lugosi, G., et al., 2018. Power-law partial correlation network models. *Electronic Journal of Statistics* 12, 2905–2929.
- Barndorff-Nielsen, O., 2014. Information and exponential families in statistical theory. *Wiley Series in Probability and Statistics*, John Wiley & Sons, Ltd., Chichester. URL: <https://doi.org/10.1002/9781118857281>, doi:10.1002/9781118857281. reprint of the 1978 original [MR0489333].
- Bi, Z., Faloutsos, C., Korn, F., 2001. The DGX distribution for mining massive, skewed data, in: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 17–26.
- Broido, A.D., Clauset, A., 2019. Scale-free networks are rare. *Nature communications* 10, 1–10.
- Chacoma, A., Zanette, D.H., 2021. Word frequency–rank relationship in tagged texts. *Physica A: Statistical Mechanics and its Applications* 574, 126020.
- Chen, Y., 2021. Exploring the level of urbanization based on Zipf's scaling exponent. *Physica A: Statistical Mechanics and its Applications* 566, 125620.

The Zipf-Polylog Model

- Chung, F., Chung, F.R., Graham, F.C., Lu, L., Chung, K.F., et al., 2006. Complex graphs and networks. 107, American Mathematical Soc.
- Clauset, A., Shalizi, C.R., Newman, M.E.J., 2009. Power-law distributions in empirical data. *SIAM Rev.* 51, 661–703. URL: <https://doi.org/10.1137/070710111>, doi:10.1137/070710111.
- Clauset, A., Tucker, E., Sainz, M., 2020. The colorado index of complex networks (2016). URL <https://icon.colorado.edu>.
- Duarte-López, A., Pérez-Casany, M., 2020. zipfextR: Zipf Extended Distributions. URL: <https://CRAN.R-project.org/package=zipfextR>. R package version 1.0.2.
- Duarte-López, A., Pérez-Casany, M., Valero, J., 2020. The Zipf-Poisson-stopped-sum distribution with an application for modeling the degree sequence of social networks. *Comput. Statist. Data Anal.* 143, 106838, 16. URL: <https://doi.org/10.1016/j.csda.2019.106838>, doi:10.1016/j.csda.2019.106838.
- Duarte-López, A., Pérez-Casany, M., Valero, J., 2021. Randomly stopped extreme Zipf extensions. *Extremes*, 1–34.
- Duarte-López, A., Prat-Pérez, A., Pérez-Casany, M., 2015. Using the Marshall-Olkin extended Zipf distribution in graph generation, in: *European Conference on Parallel Processing*, Springer. pp. 493–502.
- Dyer, J.S., Owen, A.B., 2012. Correct ordering in the Zipf-Poisson ensemble. *J. Amer. Statist. Assoc.* 107, 1510–1517. URL: <https://doi.org/10.1080/01621459.2012.734177>, doi:10.1080/01621459.2012.734177.
- Fisher, R.A., 1934. The effect of methods of ascertainment upon the estimation of frequencies. *Annals of eugenics* 6, 13–25.
- Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., Arenas, A., 2003. Self-similar community structure in a network of human interactions. *Physical review E* 68, 065103.
- Hill, B.M., Woodroffe, M., 1975. Stronger forms of Zipf's law. *J. Amer. Statist. Assoc.* 70, 212–219. URL: [http://links.jstor.org/sici?sici=0162-1459\(197503\)70:349<212:SF0ZL>2.0.CO;2-X&origin=MSN](http://links.jstor.org/sici?sici=0162-1459(197503)70:349<212:SF0ZL>2.0.CO;2-X&origin=MSN).
- Holme, P., 2019. Rare and everywhere: Perspectives on scale-free networks. *Nature communications* 10, 1–3.
- Johnson, N.L., Kemp, A.W., Kotz, S., 2005. *Univariate discrete distributions*. Wiley Series in Probability and Statistics. third ed., Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. URL: <https://doi.org/10.1002/0471715816>, doi:10.1002/0471715816.
- Knuth, D.E., 1993. *The Stanford GraphBase: a platform for combinatorial computing*. volume 1. AcM Press New York.
- Krumme, C., Lorente, A., Cebrían, M., Moro, E., et al., 2013. The predictability of consumer visitation patterns. *Scientific reports* 3, 1645.
- Kunegis, J., 2013. Konect: the koblenz network collection, in: *Proceedings of the 22nd International Conference on World Wide Web*, ACM. pp. 1343–1350.
- Malone, D., Maher, K., 2012. Investigating the distribution of password choices, in: *Proceedings of the 21st international conference on World Wide Web*, ACM. pp. 301–310.
- McKelvey, B., et al., 2018. Using maximum likelihood estimation methods and complexity science concepts to research power law-distributed phenomena, in: *Handbook of Research Methods in Complexity Science*. Edward Elgar Publishing.
- Newman, M.E., 2001. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences* 98, 404–409.
- Newman, M.E., 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary physics* 46, 323–351.
- Patil, G.P., Rao, C.R., 1978. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics* 34, 179–189. URL: <https://doi.org/10.2307/2530008>, doi:10.2307/2530008.
- Pérez-Casany, M., Casellas, A., 2013. Marshall-Olkin Extended Zipf Distribution. arXiv preprint arXiv:1304.4540.
- Sáez-Castillo, A., Conde-Sánchez, A., 2013. A hyper-poisson regression model for overdispersed and underdispersed count data. *Computational Statistics & Data Analysis* 61, 148–157.
- Saghir, A., Hamedani, G., Tazeem, S., Khadim, A., 2017. Weighted Distributions: A Brief Review, Perspective and Characterizations. *International Journal of Statistics and Probability* 6, 109.
- Sarabia, J.M., Gómez-Déniz, E., Sarabia, M., Prieto, F., 2010. A general method for generating parametric Lorenz and Leimkuhler curves. *Journal of Informetrics* 4, 524–539.
- Self, S.G., Liang, K.Y., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* 82, 605–610. URL: [http://links.jstor.org/sici?sici=0162-1459\(198706\)82:398<605:APOMLE>2.0.CO;2-2&origin=MSN](http://links.jstor.org/sici?sici=0162-1459(198706)82:398<605:APOMLE>2.0.CO;2-2&origin=MSN).
- Teicher, H., 1961. Maximum likelihood characterization of distributions. *The Annals of Mathematical Statistics* 32, 1214–1222.
- Traud, A.L., Mucha, P.J., Porter, M.A., 2012a. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications* 391, 4165–4180. URL: <http://www.sciencedirect.com/science/article/pii/S0378437111009186>, doi:<https://doi.org/10.1016/j.physa.2011.12.021>.
- Traud, A.L., Mucha, P.J., Porter, M.A., 2012b. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications* 391, 4165–4180.
- Valero, J., Pérez-Casany, M., Ginebra, J., 2010. On zero-truncating and mixing Poisson distributions. *Adv. in Appl. Probab.* 42, 1013–1027. URL: <https://doi.org/10.1239/aap/1293113149>, doi:10.1239/aap/1293113149.
- Visser, M., 2013. Zipf's law, power laws and maximum entropy. *New Journal of Physics* 15, 043021.
- Zipf, G.K., 1949. *Human Behaviour and the Principle of Least-Effort*. Cambridge MA edn.
- Zörnig, P., Altmann, G., 1995. Unified representation of Zipf distributions. *Comput. Statist. Data Anal.* 19, 461–473. URL: [https://doi.org/10.1016/0167-9473\(94\)00009-8](https://doi.org/10.1016/0167-9473(94)00009-8), doi:10.1016/0167-9473(94)00009-8.

Jordi Valero: Conceptualization, Methodology. **Marta Pérez-Casany:** Conceptualization, Methodology, Writing- Reviewing and Editing. **Ariel Duarte-López:** Data curation, Conceptualization, Software, Writing- Reviewing and Editing.

Journal Pre-proof

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: