**Extremes manuscript No.**
(will be inserted by the editor)

# Randomly Stopped Extreme Zipf Extensions

**Ariel Duarte-López** · **Marta Pérez-Casany** ·
**Jordi Valero**

Received: date / Accepted: date

**Abstract** In this paper, we extend the Zipf distribution by means of the Randomly Stopped Extreme mechanism; we establish the conditions under which the maximum and minimum families of distributions intersect in the original family; and we demonstrate how to generate data from the extended family using any Zipf random number generator. We study in detail the particular cases of geometric and positive Poisson stopping distributions, showing that, in log-log scale, the extended models allow for top-concavity (top-convexity) while maintaining linearity in the tail. We prove the suitability of the models presented, by fitting the degree sequences in a collaboration and a protein-protein interaction networks. The proposed models not only give a good fit, but they also allow for extracting interesting insights related to the data generation mechanism.

**Keywords** Zipf distribution · Randomly Stopped Extreme distributions · Heavy-tail distributions · Degree sequence · Power law

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.

A. Duarte-López
Data Management Group (DAMA-UPC); Dept. of Statistics and OR;
Technical University of Catalonia
E-mail: ariel.duarte.lopez@upc.edu

M. Pérez-Casany
Data Management Group (DAMA-UPC); Dept. of Statistics and OR;
Technical University of Catalonia
Avinguda Diagonal, 647
08028 Barcelona, Spain
Tel.: +34-934011726
E-mail: marta.perez@upc.edu

J. Valero
Dept. of Statistics and OR;
Technical University of Catalonia
E-mail: jordi.valero@upc.edu

**Mathematics Subject Classification (2010)** 60E05 · 62-07 · 62E99

## 1 Introduction

Discrete Power Law (DPL) distributions are those families of distributions such that the probabilities are inversely proportional to a positive power of the value itself. When the support of the DPL is the strictly positive integer values equal or larger than 1, we obtain the Zipf distribution (Zipf, 1949). The popularity of the Zipf distribution has increased over the years because it provides a reasonable fit to data that originates from dissimilar areas. Some examples of its applications can be found in assessing the quality of the peer review process (Ausloos et al., 2016), mobility patterns (Ectors et al., 2018), and the arts (Manaris et al., 2005). In the last years several authors have pointed out that, in practice, few empirical phenomena obey DPLs for all values of $x$, and more often the DPL applies only to values greater than a given threshold ($x_{min}$), see Clauset et al. (2009) and McKelvey et al. (2018). In the network analysis environment, networks with a DPL degree distribution are also called *scale free* networks see Barabási and Pósfai (2016). Recently the work by Broido and Clauset (2019) has analyzed a large corpus of degree sequences of graphs coming from many different research areas, and has confirmed that only a small percentage of those are what they denote as ”pure scale free”. The goal of this paper is to define two-parameter Zipf extensions that, on one side, perform similarly to the Zipf when modeling the tail of the data and, on the other side, allow for fitting the data in all its range without requiring the selection of an $x_{min}$ value.

The main reason for the real degree sequences to deviate from DPL behavior is that, when the probabilities of a DPL are plotted in log-log scale, one obtains a straight line, and degree sequences of real networks tend to show a top-concave (less frequent top-convex) pattern that is not adapted by DPL distributions. Thus, the additional parameter has to guarantee this flexibility for the initial values.

The concept of Randomly Stopped Extreme distribution (RSED) is used to extend the Zipf distribution. The RSEDs are the distribution families defined as the minimum or the maximum of a random number of independent and identically distributed (i.i.d.) random variables (r.v.). The name RSED was introduced by Pérez-Casany et al. (2016) at the ICOSDA 2016 conference. However, these kinds of distributions have been widely studied in the literature. See Louzada et al. (2012) for a formal definition of the RSED. These distribution families are applied mainly to lifetime scenarios in which the information of a particular event is not observed and one instead has the information about the minimum or maximum of a random number of events. The survey by Tahir and Cordeiro (2016) introduces several distributions that belong to this class. However, in most of the cases that appear in the literature, the extended distribution is a continuous distribution and the extension of a discrete family is less frequent. See Gómez-Déniz (2010) for an extension of the geometric distribution.

The paper is organized as follow: Section 2 introduces the Zipf distribution and its main characteristics. Section 3 focuses on the concept of RSED and presents two new results related to this concept. Section 4 particularizes on RSE Zipf generalizations, and Section 5 is devoted to two particular extensions obtained by assuming geometric and positive Poisson stopping distributions. Section 6 shows the suitability of the models proposed through the analysis of the degree sequence of two real networks. The fits obtained by the presented models are also compared with those from other bi-parametric models, such as: the discrete Gaussian Exponential (DGX) (Bi et al., 2001), the Zipf-Polylog (Valero et al., 2020) and the positive version of the Zipf-PSS (Duarte-López et al., 2020) distribution. In Section 7, we explain how to synthetically generate networks with a degree sequence that follows one of our extensions. This is a research that is in its early stage, but with encouraging results. Finally, the main conclusions are stated in Section 8.

## 2 The Zipf distributions and its limitations

A random variable (r.v.) is said to follow a DPL when its probability mass function (PMF) is equal to:

$$P(X=x)=\begin{cases}\frac{x^{-\alpha}}{\zeta(\alpha,x_{min})} & \forall x\geq x_{min}\\ 0 & \text{otherwise,}\end{cases} \tag{1}$$

where $\alpha>1, x_{min}>0$ and $\zeta(\alpha,x_{min})=\sum_{i=x_{min}}^{+\infty} i^{-\alpha}=\sum_{i=0}^{+\infty}(i+x_{min})^{-\alpha}$ is the Hurwitz zeta function.

For the particular case when $x_{min}=1$ in (1) the Zipf distribution is obtained. Thus, it is said that a r.v. $X$ follows a Zipf distribution with parameter $\alpha>1$ if, and only if, its PMF is equal to:

$$P(X=x)=\frac{x^{-\alpha}}{\zeta(\alpha)}\ ,x=1,2,...,\ \alpha>1, \tag{2}$$

where $\zeta(\alpha)=\sum_{i=1}^{+\infty} i^{-\alpha}$ is the Riemann Zeta function. Observe that the parameter space of the Zipf distribution is the set of values where the Riemann zeta function converges, which is $(1,+\infty)$.

The Zipf distribution is a one-parametric distribution defined on the strictly positive integer numbers, where the 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.

As any DPL distribution, it is highly recommended for modeling two types of data: rank and frequencies of frequency. An example of rank data is, for instance, the list of the world's billionaires [1] provided by Forbes. There the richest people in the world are ranked based on the fortune that they own. For frequencies of frequency data, one understands data that are frequency tables of counts. For instance, let us assume that we known the number of followers that each Instagram account has, if we group them by the number of followers, and then we count how many accounts each group has, it gives place to the frequencies of frequency table having in the first column the category and, in the second column, the amount of accounts of that

[1] https://www.forbes.com/billionaires/list/;

category. The data sets considered in Section 6 are frequencies of frequency data. Notwithstanding, they could also be analyzed in terms of ranks.

In what follows we point out the main characteristics of the Zipf distribution. By taking logarithm in both sides of (2) one has that when the probabilities are plotted in log-log scale they show a straight line with a slope equal to $-\alpha$ and an intercept equal to $\log(\zeta(\alpha))$. Figure 1 shows the probabilities of the Zipf for different values of the $\alpha$ parameter in log-log scale. Observe that when the $\alpha$ parameter increases, the probabilities concentrates at the low values.

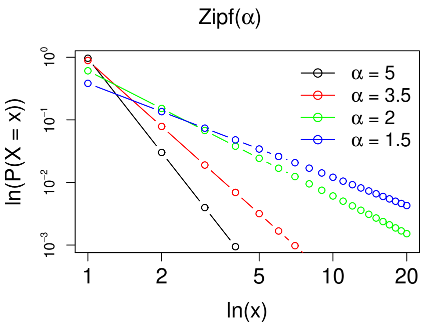

Fig. 1: PMFs of the Zipf distribution for $\alpha = 1.5, 2, 3.5$ and 5 in log-log scale.

The survival function (SF) and the cumulative density function (CDF) of the Zipf distribution with parameter $\alpha$ are respectively equal to:

$$\overline{F}_\alpha(x) = P(X > x) = \frac{1}{\zeta(\alpha)} \sum_{i=x+1}^{+\infty} i^{-\alpha} = \frac{\zeta(\alpha, x+1)}{\zeta(\alpha)},\ \alpha > 1, \tag{3}$$

$$F_\alpha(x) = 1 - \frac{\zeta(\alpha, x+1)}{\zeta(\alpha)} = \frac{\zeta(\alpha) - \zeta(\alpha, x+1)}{\zeta(\alpha)},\ \alpha > 1. \tag{4}$$

The *k-th* moment of the Zipf, $k \in \mathbb{Z}^+$ is equal to:

$$E[X^k] = \sum_{x=1}^{+\infty} \frac{x^k x^{-\alpha}}{\zeta(\alpha)} = \frac{\zeta(\alpha - k)}{\zeta(\alpha)},$$

and thus, it is finite if, and only if, $\alpha > k+1$ because $\zeta(\alpha - k)$ needs to be finite. In particular, the first moment only exists if $\alpha > 2$ and in that case, it is equal to:

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

Moreover, if $x_1, x_2, \ldots, x_n$ is a sample from an r.v. X with a Zipf($\alpha$) distribution, the log-likelihood function is equal to:

$$\ell(\alpha; x_1, x_2, \ldots, x_n) = -\alpha \sum_{i=1}^{n} \log(x_i) - n \log(\zeta(\alpha)).$$

Thus, the maximum likelihood estimation (MLE) of $\alpha$ is obtained by solving the equation:

$$-\sum_{i=1}^{n}\log(x_i) - n\frac{\zeta'(\alpha)}{\zeta(\alpha)} = 0,$$

and given that $\zeta'(\alpha) = \sum_{i=1}^{+\infty} i^{-\alpha}\log(i)$, it is equivalent to solve:

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

Observe that this equation is equivalent to applying the moment-method estimation to the logarithm of the variable. Applying the logarithm to a Zipf distributed r.v., i.e. considering the r.v. $\log(X)$, 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. The MLE of the Zipf distribution when necessary, can be computed numerically.

The suitability of the Zipf distribution has been widely demonstrated in dissimilar areas. For example, the classical book by Zipf (1949) shows, from among other examples, that this distribution provides accurate results when it is used for fitting the frequency of the words in a given text. More recently, it has been used in the work by Ausloos et al. (2016) to assess the quality of the peer review process. In particular, the analyzed data came from peer review reports of the Journal of the Serbian Chemical Society.

In addition, Ectors et al. (2018) have shown that the Zipf distribution also emerges in the frequency at which people conduct their daily activities; this contribution can be directly used for validating travel demand models. Other examples from a completely unrelated area appear in the paper by Manaris et al. (2005), where the Zipf distribution has been used in music classification for measuring the proportion of various parameters, such as harmonic consonance and duration, among others. Moreover, it has been used for automatic detection of regions of interest in digital images (Caron et al., 2007).

The same occurs in network analysis, where the distribution is considered reasonable for fitting the degree sequence of a real network. For example, the work by Adamic and Huberman (2002) has shown that it provides the best fit for the connections of the Internet backbone as well as for the connections included in the World Wide Web.

However, even though Zipf's law seems to govern multiple natural and man-made systems, it has an intrinsic limitation: it lacks flexibility, which is a consequence of being a one-parameter distribution. When the probabilities are plotted in double logarithmic scale, the distribution always exhibits a straight line (see Figure 1). However, real data usually deviate from this type of pattern and generally show linearity only in the tail. Moreover, for small values, a top-concave pattern is often observed while a top-convex one is seen less often. Figure 2 shows several plots in log-log scale of the degree distributions of real networks. These plots illustrate a clear deviation from pure DPL behavior. On the upper left-hand side is the in-degree sequence of a communication network representing emails exchanged in a European institution. The upper right-hand side shows the degree sequence of the *Arabidopsis thaliana* comprehensive knowledge network. Finally, the example at the bottom corresponds to the Facebook network of the University of California, Santa Cruz in 2005. In Section 6 the second data set is fitted by means of the Zipf extensions proposed in this paper.

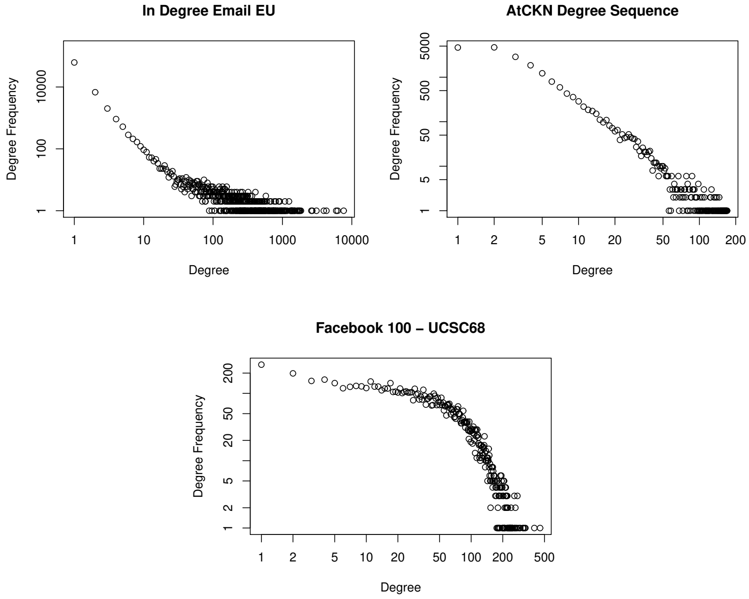

Fig. 2: Examples of degree sequences of real networks plotted in log-log scale. On the upper left-hand side is the in-degree sequence of a communication network representing emails exchanged in a European institution. The upper right-hand side shows the degree sequence of the *Arabidopsis thaliana* comprehensive knowledge network. Finally, the example at the bottom corresponds to the Facebook network of the University of California, Santa Cruz in 2005.

According to McKelvey et al. (2018), in just a few scenarios the power law pattern appears in the entire range of values. In most of the cases, this pattern is observed only for values over a given threshold. This threshold separates two behaviors: the first one tends to be Gaussian; and the second one, which corresponds to the tail, follows a DPL. This implies that fitting a DPL to many data sets requires the selection of a plausible cut-off point, $x_{min}$. A widely used practice is the methodology proposed by Clauset et al. (2009), which seeks a cut-off point and fits the distribution of the data at values that are larger or equal than the selected cut-off. In Drees et al. (2020) it is proved that Clauset's methodology tends to select values of the cut-off that are too high. A consequence of this is that the Hill estimator for parameter $\alpha$ ( see, Hill (1975)) has large variances. The determination of a cut-off point implies a loss of information and creates the need to generalize the Zipf in such a way that the extension is able to fit the data in their entire range, while maintaining the linearity in the tail. In Section 6 we show that by using the models proposed in this paper it is not necessary to choose an $x_{min}$ value because they allows for fitting the data in all its range while performing similarly to the Zipf in the tail.

The next section introduces the concept of RSED as a way of generalizing a family of probability distributions. In addition, it states the conditions under which the original family is included in the extended family. At the end we present a methodology for generating random data from an extended family based on a random number generator of the original family.

## 3 Randomly Stopped Extreme distributions

In practice, maximums (less often minimums) of i.i.d. copies of an r.v. $X$ are used in the lifetime and reliability studies of many research areas, such as physics, computer science, industry, public health, and communications, among others. See for instance Kuş (2007) and Cancho et al. (2011) for the definition of the Poisson-exponential lifetime distribution in terms of, respectively, minimums and maximums. The work by Tahir and Cordeiro (2016) reviews the different classes of compound distributions and introduces several examples of distributions that can be described as Randomly Stopped Extreme distributions.

In this section, we first introduce the concept of *Randomly Stopped Extreme distribution* (RSED) and then point out two important results related to the RSEDs.

### 3.1 Definition

Let $X$ be an r.v. with parameter vector $\alpha$ and cumulative density function (CDF) $F_X(x;\alpha)$; and let $N$ be a discrete r.v. defined in the strictly positive integer numbers, independent of $X$, and with probability generating function (PGF) $h_N(t;\theta)$, with $\theta$ being the parameter vector. The r.v.'s defined as:

$$Y_{X;N}^{max} = max(X_1, X_2, \cdots, X_N) \quad \text{and} \quad Y_{X;N}^{min} = min(X_1, X_2, \cdots, X_N),$$

where $X_i$ are i.i.d. copies of $X$, have their CDF and survival function (SF), respectively, equal to:

$$F_{Y_{X;N}^{max}}(x;\alpha,\theta) = h_N(F_X(x;\alpha),\theta) \quad \text{and} \quad S_{Y_{X;N}^{min}} = h_N(S_X(x;\alpha),\theta), \tag{6}$$

with $S_X(x;\theta)$ being the SF of $X$ (see, Louzada et al., 2012). The distribution of $Y_{X;N}^{max}$ and $Y_{X;N}^{min}$ are called, by definition, RSED, since they are the distribution of a maximum or minimum (extreme) of a random number of independent copies of $X$ (see, Pérez-Casany et al., 2016).

It is important to observe that $X$ is associated with the phenomena under study while $N$ is related to the number of observations of $X$ that one has in a given period of time or in a given space. RSEDs appear in real situations when one observes only the variable of interest when it is larger (smaller) than a given threshold. For instance, one may be interested in buying foreign currency only when the price is lower than a given value. In such a case, $X_i$ will be the price of the currencies that are smaller than the threshold in a given period of time, and $N$ will be the number of currencies that have a price smaller than the threshold.

The distributions of the r.v.'s $X$ and $N$ are, respectively, denoted by *stopped* and *stopping* distributions. This allows for a parallelism between RSED and Stopped Sum distributions, i.e., the distributions that appear as a random sum of i.i.d. copies of a given r.v. $X$. The stopped distribution of an RSED serves as the secondary distribution of a stopped sum, while the stopping distribution represents the primary distribution. In the case of stopped sum distributions, the PGF of the final variable is equal to the composition of the PGF of the primary distribution with the PGF of the secondary distribution. This is the reason why they are also known as *compound distributions*. Based on (6), for RSEDs one compose the PGF of the stopping distribution with the CDF (maximums) or the SF (minimums) of the stopping distribution to obtain, respectively, the CDF of the maximum or the SF of the minimum.

Randomly Stopped Extreme and Stopped-Sum distributions are two mechanisms that allow us to generalize the distribution of $X$. Both transformations help us better understand the mechanism that generates the data. The next section is devoted to the extensions of the Zipf distribution obtained by applying RSED. See Duarte-López et al. (2020) for the generalization of the Zipf distribution based on Poisson-stopped-sums.

By restricting $N$ to being a strictly positive integer r.v., one avoids computing the maximum (minimum) of the empty set. Thus, one may assume for instance, that $N$ follows either a strictly positive geometric distribution or a logarithmic series distribution. One may also consider as a distribution for $N$, any zero truncation of a positive integer distribution. In this latter case, one has to take into account that if $N^{zt}$ denotes the zero-truncated version of $N$, then its PGF is equal to:

$$h_{N^{zt}}(t;\theta) = \frac{h_N(t;\theta) - h_N(0;\theta)}{1 - h_N(0;\theta)}. \tag{7}$$

For example, if $N$ is Poisson distributed with $\lambda > 0$, given that $h_N(t) = e^{\lambda(t-1)}$, one has that,

$$h_{N^{zt}}(t;\lambda) = \frac{e^{\lambda t} - 1}{e^{\lambda} - 1}. \tag{8}$$

As a consequence of the fact that $\lim_{\lambda \to 0} h_{N^{zt}}(t;\lambda) = t$, it is possible to consider $[0,+\infty)$ as the parameter space of the zero-truncated Poisson distribution, where $\lambda = 0$ corresponds to the degenerate distribution at one. The PGFs of the positive negative binomial and the positive Hermite distributions appear in Table 2, and they have been obtained in a similar way.

Taking into account (6) in the case where the zero truncation of $N$ is required, the CDF of the maximum and the SF of the minimum are equal to:

$$F_{Y^{max}_{X;N}}(x;\alpha,\theta) = h_{N^{zt}}(F_X(x;\alpha),\theta) = \frac{h_N(F_X(t;\alpha);\theta) - h_N(0;\theta)}{1 - h_N(0;\theta)}, \tag{9}$$

| $N$ Dist. | $E[N]$ | $h_{N^{zt}}(t;\theta)$ | Param. space | $X$ Dist. |
|---|---|---|---|---|
| geometric | $\frac{1}{p}$ | $\frac{pt}{1-(1-p)t}$ | $[0,1]$ | $p=1$ |
| zt. Poisson | $\frac{\lambda}{1-e^{-\lambda}}$ | $\begin{cases} \frac{e^{\lambda t}-1}{e^{\lambda}-1} & \text{if } \lambda > 0 \\ t & \text{if } \lambda = 0 \end{cases}$ | $[0,+\infty)$ | $\lambda = 0$ |
| zt. Hermite | $\frac{\theta-2\beta}{1-e^{-(\theta+\beta)}}$ | $\frac{e^{\theta t+\beta t^2}-1}{e^{\theta+\beta}-1}$ | $[0,+\infty)\times[0,+\infty)$ | $\theta=\beta=0$ |
| log-series | $-\frac{\theta}{\log(1-\theta)(1-\theta)}$ | $\frac{\ln(1-\theta t)}{\ln(1-\theta)}$ | $(0,1)$ | $\theta=0$ |
| zt. neg.bin | $-\frac{\theta\beta}{(1-\theta)\theta^{\beta}}$ | $\frac{(\frac{1-\theta}{1-\theta t})^{\beta}-(1-\theta)^{\beta}}{1-(1-\theta)^{\beta}},$ | $(0,1)\times(0,+\infty)$ | $\theta=0$ |

Table 1: Some possible stopping distributions together with their PGFs, parameter spaces and the parameter values that gives the family of distributions of $X$.

and

$$S_{Y_{X;N}^{min}}(x;\alpha,\theta) = h_{N^{zt}}(S_X(x;\alpha),\theta) = \frac{h_N(S_X(t;\alpha);\theta) - h_N(0;\theta)}{1-h_N(0;\theta)}. \qquad (10)$$

## 3.2 Two interesting results

In this section, we prove two theorems. The first one establishes a condition under which the random stopped extensions contain the family of distributions of $X$ as a particular case. The second theorem explains how to generate data in the extended family based on a random data generator of the family of distributions of $X$.

**Theorem 1** *If $N$ is defined in the strictly positive integer values and a value $\theta_0$ exists in the parameter space, such that $h_N(t;\theta_0) = t$, then the distribution of $X$ belongs to both sets of maximum and minimum stopped extreme distributions.*

*Proof.* Given that $h_N(t;\theta_0) = t$, from (6) one has that:

$$F_{Y_{X;N}^{max}}(x;\alpha,\theta) = h_N(F_X(x;\alpha),\theta_0) = F_X(x;\alpha),$$

and that:

$$S_{Y_{X;N}^{min}}(x;\alpha,\theta) = h_N(S_X(x;\alpha),\theta_0) = S_X(x;\alpha).$$

□

Observe that saying $h_N(t;\theta_0) = t$ is equivalent to saying that the family contains the degenerate distribution at one, as a particular case. This is the case for the stopping distributions considered in Table 1. Their corresponding RSEDs contain the family of the distribution of $X$ as a particular case, for the parameter values that appear in the last column.

The next theorem shows how to generate random numbers from the RSED families, based on knowing how to generate random data from the baseline family. This is important, because one may use any random number generator implemented in any statistical software for the baseline family, and then easily generate data from the extended families. Thus, even if the CDF of the extended distribution is rather complicated, simulating data from it is computationally simple.

**Theorem 2** *Let $Y$ be an r.v. with an RSED. To generate a random value from $Y$ is enough to follow the next steps and to:*

1) *uniformly generate a value $u$ in $(0,1)$;*
2) *compute the value $u'$ in the following way:*
   a) *if $Y$ is a maximum, then $u' = h_N^{-1}(u(1-h_N(0))) + h_N(0)$, and*
   b) *if $Y$ is a minimum, then $u' = 1 - h_N^{-1}(1-u(1-h_N(0)))$;*
3) *apply the inversion method to $u'$ using the distribution of $X$.*

*Proof.* We first prove the theorem for maximums. Given a value $u \in (0,1)$, to apply the inversion method to the distribution of $Y$ is equivalent to finding the smaller value of $x$, such that $u \leq F_Y(x;\alpha,\theta)$. Taking into account (9), this is equivalent to finding the minimum value of $x$, such that:

$$u \leq \frac{h_N(F_X(x;\alpha);\theta) - h_N(0;\theta)}{1-h_N(0;\theta)},$$

which is equivalent to saying that:

$$h_N^{-1}(u(1-h_N(0;\theta))) + h_N(0;\theta) \leq F_X(x;\alpha) \Leftrightarrow u' \leq F_X(x;\alpha),$$

with $u' = h_N^{-1}(u(1-h_N(0))) + h_N(0)$.

To prove the theorem for minimums, one has that $u \leq F_Y(x;\alpha,\theta) \Leftrightarrow u \leq 1 - S_Y(x;\alpha,\theta)$, and by (10), this is equivalent to saying that:

$$\begin{aligned}
u \leq \frac{1-h_N(S_X(x;\alpha);\theta)}{1-h_N(0;\theta)} &\Leftrightarrow \\
u(1-h_N(0;\theta)) \leq 1 - h_N(S_X(x;\alpha);\theta) &\Leftrightarrow \\
h_N(S_X(x;\alpha);\theta) \leq 1 - u(1-h_N(0;\theta)) &\Leftrightarrow \\
S_X(x;\alpha) \leq h_N^{-1}(1-u(1-h_N(0;\theta));\theta) &\Leftrightarrow \\
1 - F_X(x;\alpha) \leq h_N^{-1}(1-u(1-h_N(0;\theta));\theta) &\Leftrightarrow \\
1 - h_N^{-1}(1-u(1-h_N(0,\theta));\theta) \leq F_X(x;\alpha) &\Leftrightarrow \\
u' \leq F_X(x;\alpha), &
\end{aligned}$$

with $u' = 1 - h_N^{-1}(1-u(1-h_N(0)))$. □

## 4 Randomly Stopped Extreme Zipf distributions

By RSEZipf distribution, we denote the RSED that assumes that $X$ follows a Zipf distribution. Table 2 contains the CDFs of the maximums as well as the SFs of the minimums of the RSEZipf obtained by considering the stopping distributions that appear in Table 1. They were obtained by compounding the PGF of the stopping

| Stopping distrib. | $F_{Y_N^{max}}$ | $S_{Y_N^{min}}$ |
|---|---|---|
| geometric | $\frac{1-\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}}{1+(\frac{1}{\theta}-1)\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}}$ | $\frac{\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}}{\frac{1}{\theta}+(1-\frac{1}{\theta})\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}}$ |
| log.series | $\frac{\ln(1-\theta(1-\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}))}{\ln(1-\theta)}$ | $\frac{\ln(1-\theta\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)})}{\ln(1-\theta)}$ |
| zero-trunc. Poisson | $\frac{e^{\theta(1-\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)})}-1}{e^{\theta}-1}$ | $\frac{e^{\theta\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}}-1}{e^{\theta}-1}$ |
| zero-trunc. neg. bin. | $\frac{(\frac{1-\theta}{1-\theta(1-\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)})})^{\beta}-(1-\theta)^{\beta}}{1-(1-\theta)^{\beta}}$ | $\frac{(\frac{1-\theta}{1-\theta\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}})^{\beta}-(1-\theta)^{\beta}}{1-(1-\theta)^{\beta}}$ |
| zero-trunc. Hermite | $\frac{e^{\theta(1-\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)})+\beta((1-\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}))^2}-1}{e^{\theta+\beta}-1}$ | $\frac{e^{\theta\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}+\beta(\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)})^2}-1}{e^{\theta+\beta}-1}$ |

Table 2: CDF for the maximum and SF for the minimum of the random extreme Zipf generalizations, considering the following types of stopping distributions: geometric, logarithmic series, positive Poisson, positive negative binomial, and positive Hermite.

distribution (that appears in the second column of Table 1) with: the CDF of the Zipf (4) (in the case of maximums); and the SF of the Zipf (3) (in the case of minimums).

The next theorem states the asymptotically relation of the tails of the Zipf and any RSEZipf distribution. The results obtained are a consequence of the work by Jessen and Mikosch (2006, p. 18–20). For these particular result, bear in mind that $f(x) \sim g(x)$, as $x \to +\infty$ is equivalent to saying that $f(x)/g(x) \xrightarrow[x\to+\infty]{} 1$ if $g(x) \neq 0$, and it is equivalent to $f(x) = o(1)$ if $g(x) = 0$.

**Theorem 3** *The tail of an r.v $Y \sim$ RSEZipf$(\alpha,\beta)$ is asymptomatically related to the tail of an r.v. $X \sim$ Zipf$(\alpha)$. More precisely:*

*a) if $Y$ is a minimum, then $P(Y > x) \sim P(N = n_0)[P(X > x)]^{n_0}$, where $n_0$ is the smallest positive integer, such that $P(N = n_0) > 0$,*

*b) if $Y$ is a maximum, then $P(Y > x) \sim E[N]P(X > x)$, where $N$ is the stopping distribution and $E[N] < +\infty$.*

*Proof.* These results hold when $X$ is assumed to be a regularly varying function. According to Gulisashvili (2012, p. 220), the Zipf distribution is a regularly varying function, since it is a Pareto-type distribution. In the author's words, a function $f$ belongs to the kind of Pareto-type distributions if it is asymptotically equivalent to a regularly varying function, which implies that $f$ is also a regularly varying function. For an extended proof of these results, the reader is encouraged to review the work of Jessen and Mikosch (2006, p. 18–19). Consequently, any extension of the Zipf distribution obtained by RSED will have a linear tail independently of the stopping distribution. □

For a generalization of the Zipf distribution with a non-linear tail, see Valero et al. (2020).

## 5 The MOEZipf and the Zipf-PE generalizations

In this section, the two RSEZipf distributions corresponding to the geometric and the positive Poisson stopping distributions are analyzed in detail. The first one is denoted by MOEZipf, because it also the result of applying the Marshall-Olkin transformation (MO) (Marshall and Olkin, 1997) to the Zipf distribution. The second one is denoted by Zipf-PE (Zipf-Poisson Extreme), and it is aligned with the framework proposed by Ramos et al. (2018).

### 5.1 The Marshall-Olkin Extended Zipf distribution

The MO transformation allows us to extend a family of probability distributions by adding an extra parameter. The authors in their work also demonstrate that this transformation concurs with the RSED definition, since the extended family of distributions can be interpreted as the minimum (maximum) of a geometric number of independent r.v.'s (Marshall and Olkin, 1997, p. 646). Consequently, applying the MO transformation to (3) results in the SF (11) of the MOEZipf distribution, that is equal to.

$$\overline{F}_{\alpha,\beta}(x) = \frac{\beta\,\overline{F}_{\alpha}(x)}{1-\overline{\beta}\,\overline{F}_{\alpha}(x)} = \frac{\beta\,\zeta(\alpha,x+1)}{\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)}, \alpha > 1, \beta > 0. \tag{11}$$

The MOEZipf distribution was originally defined and analyzed by Casany and Casellas (Pérez-Casany and Casellas, 2013; Casellas, 2013). The main results presented by these authors are:

a) for sufficiently large $x$, $\beta$ is obtained as the limit of the ratio of the MOEZipf and Zipf probabilities;
b) for large values of $x$, the $\log(P(Y=x))$ is a linear function of the $\log(x)$;
c) the *k-th* moment of the MOEZipf distribution exists only if $\alpha > k+1$;
d) the ratio of two consecutive MOEZipf probabilities is greater than that of probabilities coming from a Zipf distribution with the same $\alpha$ if $\beta > 1$, otherwise it is smaller than the Zipf one. They also define the MLE and the moment-method estimation for the MOEZipf distribution.

In what follows, we complement the results mentioned in the papers above by including new properties of the MOEZipf and by extending one of the existing ones.

Based on the definition, the MOEZipf distribution has support on the strictly positive integer values, and its parameters are the $\alpha$ parameter of the Zipf distribution and the $\beta$ parameter of the geometric distribution. Thus, the parameter space is: $(1,+\infty) \times (0,+\infty)$. The PMF of the MOEZipf distribution can be derived from (11) by computing $\overline{F}_{\alpha,\beta}(x-1) - \overline{F}_{\alpha,\beta}(x)$ and it is equal to:

$$P(Y=x) = \frac{x^{-\alpha}\,\beta\,\zeta(\alpha)}{[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)]\,[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)]}, x = 2,3,4,\ldots, \tag{12}$$

and,

$$P(Y=1) = 1 - \overline{F}_{\alpha,\beta}(1) = \frac{1}{\zeta(\alpha) - \overline{\beta}\zeta(\alpha,2)}.$$

Observe that $P(Y=1)$ is equal to (12) at $x=1$; and thus (12) is the PMF in the entire support.

Figure 3 shows the PMFs of the MOEZipf distribution in log-log scale for $\alpha = 2.1$ and different values of the $\beta$ parameter. Observe how the value of the $\beta$ parameter influences the top-concavity (top-convexity) of the distribution in log-log scale. For $\beta$ values smaller than one, the distribution is top-convex; while for $\beta$ values larger than one it is top-concave. When $\beta = 1$, the probabilities are equal to those of a Zipf distribution with the same $\alpha$ parameter.

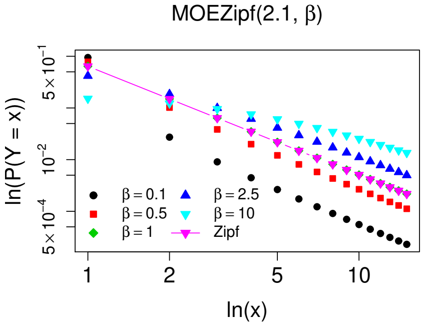

Fig. 3: PMFs of the MOEZipf distribution in log-log scale for $\alpha = 2.1$ and $\beta = 0.1, 0.5, 1, 2.5$ and 10.

The next proposition establishes the conditions under which a MOEZipf distribution can be interpreted in terms of maximums or minimums. It also proves that each distribution in the maximum family has a dual distribution in the minimum family.

**Proposition 1** *Let $Y$ be a MOEZipf distributed r.v. with parameters $(\alpha,\beta)$. Then:*

*i) If $\beta > 1$, $Y$ corresponds to a maximum of i.i.d. Zipf($\alpha$) r.v.'s, where the r.v. $N$ follows a geometric distribution with parameter $\theta = 1/\beta$.*

*ii) If $\beta < 1$, $Y$ corresponds to a minimum of i.i.d. Zipf($\alpha$) r.v.'s, where the r.v. $N$ follows a geometric distribution with parameter $\theta = \beta$.*

*iii) If $\beta = 1$, $Y$ follows a Zipf($\alpha$) distribution and may be seen as a maximum as well as a minimum of i.i.d. Zipf($\alpha$) r.v.'s, where the r.v. $N$ follows a geometric distribution with probability at one equal to one, i.e., a degenerate distribution at one.*

*Proof.* From (11) the CDF of $Y$ is equal to:

$$F_{\alpha,\beta}(x) = 1 - \overline{F}_{\alpha,\beta}(x) = \frac{\zeta(\alpha) - \zeta(\alpha,x+1)}{\zeta(\alpha) - (1-\beta)\zeta(\alpha,x+1)} = \frac{1 - \frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}}{1 + (\beta-1)\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}}.$$

Assuming that $\beta > 1$, the middle part of the first row of Table 2 shows that this corresponds to a maximum of i.i.d. Zipf($\alpha$) r.v.'s, with a geometric stopping distribution with parameter $\theta = 1/\beta$, which proves (i).

By dividing the SF of the MOEZipf($\alpha,\beta$) that appears in (11) by $\zeta(\alpha)$, one has that:

$$\overline{F}_{\alpha,\beta}(x) = \frac{\beta\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}}{1-(1-\beta)\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}} = \frac{\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}}{\frac{1}{\beta}+(1-\frac{1}{\beta})\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}},$$

which, as we can see at right hand side of the first row of Table 2, corresponds to the SF of an RSED, with a Zipf($\alpha$) distribution as the secondary distribution and a geometric distribution with parameter $\theta = \beta$ as the primary distribution, which proves (ii).

Using Theorem 1 when $\beta = 1$, (11) it is equal to the SF of a Zipf($\alpha$) distribution, which can be interpreted as a maximum as well as a minimum RSED with a geometric distribution degenerated at one, which proves (iii). □

To better understand the SF of the MOEZipf distribution and, from there, be able to deduce further properties of the distribution, the next lemma analyzes the sign and monotonicity of the function that appears in its denominator. Its results are illustrated in Figure 4. For any $\alpha > 1$ and $\beta > 0$, let us define the function

$$h_{(\alpha,\beta)}(x) = \zeta(\alpha) - (1-\beta)\zeta(\alpha,x+1),\ x \geq 1. \tag{13}$$

**Lemma 1** *The function $h_{(\alpha,\beta)}(x)$ defined for $x \geq 1$ verifies that:*

*a) If $\beta \in (0,1)$, it is an increasing concave function in $[1,+\infty)$ that takes values in the interval $[\beta\zeta(\alpha),\zeta(\alpha))$. Consequently, $\forall x \geq 1$, $\beta\zeta(\alpha) \leq h_{(\alpha,\beta)}(x) \leq \zeta(\alpha)$.*
*b) If $\beta > 1$, it is a decreasing convex function in $[1,+\infty)$ that takes values in the interval $(\zeta(\alpha),\beta\zeta(\alpha)]$. Consequently, $\forall x \geq 1$, $\zeta(\alpha) \leq h_{(\alpha,\beta)}(x) \leq \beta\zeta(\alpha)$.*
*c) If $\beta = 1$, it is a constant function equal to $\zeta(\alpha)$.*

*Proof.* The first two derivatives of the function $h_{\alpha,\beta}(x)$ are equal to:

$$\begin{aligned} h'_{(\alpha,\beta)}(x) &= \alpha(1-\beta)\zeta(\alpha+1,x+1), \text{and} \\ h''_{(\alpha,\beta)}(x) &= -\alpha(\alpha+1)(1-\beta)\zeta(\alpha+2,x+1). \end{aligned} \tag{14}$$

Taking into account that $\zeta(\alpha,x) \geq 0\,\forall\,\alpha > 0$ and $x \geq 1$, proving a) merely requires observing, first, that for $\beta \in (0,1)$, $h'_{(\alpha,\beta)}(x) \geq 0$ and $h''_{(\alpha,\beta)}(x) \leq 0$ and, second, that $h_{(\alpha,\beta)}(1) = \beta\zeta(\alpha)$ and $\lim_{x\to+\infty} h_{(\alpha,\beta)}(x) = \zeta(\alpha)$. As an increasing function, the interval where it takes values es equal to $[\beta\zeta(\alpha),\zeta(\alpha))$. Proving b), is a matter of observing that for $\beta > 1$, $h'_{(\alpha,\beta)}(x) \leq 0$ and $h''_{(\alpha,\beta)}(x) \geq 0$. As a decreasing function, the interval where it takes values is now equal to: $(\zeta(\alpha),\beta\zeta(\alpha)]$. The proof of c) is straightforward. □

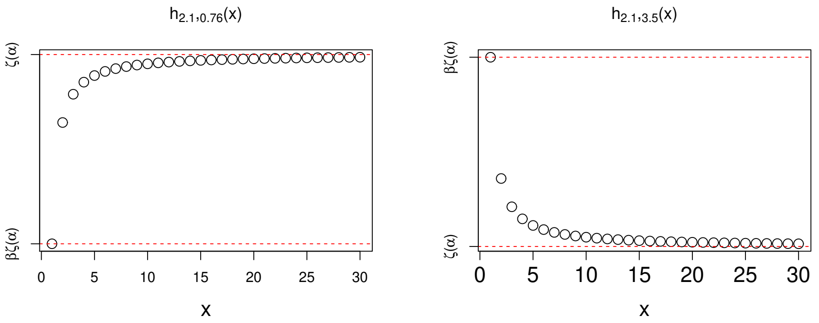

Fig. 4: Function $h_{(\alpha,\beta)}(x)$ for $\alpha = 2.1$. On the left hand side, for $\beta = 0.76$ (minimum) and, on the right hand side, for $\beta = 3.5$ (maximum). The function limits are represented by a dash line.

The next proposition establishes a condition under which the MOEZipf distribution is log-convex. Note that the log-convexity is *sufficient* criteria for stating that the distribution is *infinitely divisible* (Johnson et al., 2005).

**Proposition 2** *Let Y be an r.v., such that Y* $\sim$ *MOEZipf*$(\alpha,\beta)$*, with* $\beta \in (0,1]$*. Then, Y has a log-convex distribution.*

*Proof.* As stated in Johnson et al. (2005), a discrete distribution is said to be log-convex if and only if,

$$\frac{P(Y=x)P(Y=x+2)}{(P(Y=x+1))^2} \geq 1. \tag{15}$$

Thus, it is necessary to prove that (15) holds for $\beta \in (0,1]$. From (12), one has that (15) is equivalent to:

$$\frac{P(Y=x)P(Y=x+2)}{(P(Y=x+1))^2} = \left(\frac{x(x+2)}{(x+1)^2}\right)^{-\alpha} \left(\frac{\frac{h_{\alpha,\beta}(x+1)}{h_{\alpha,\beta}(x)}}{\frac{h_{\alpha,\beta}(x+3)}{h_{\alpha,\beta}(x+2)}}\right) \geq 1. \tag{16}$$

Given that for $x \geq 1$ $x(x+2)/(x+1)^2 < 1$, the first term of the product on the right hand side of the equality that appear in (16) is larger than one. Thus, it is enough to prove that the second term is also larger than one. Defining

$$g(x) = \frac{h_{\alpha,\beta}(x+1)}{h_{\alpha,\beta}(x)},$$

the second term is equal to $g(x)/g(x+2)$. Observe that

$$\begin{aligned} g(x) =& \frac{\zeta(\alpha)-\overline{\beta}\zeta(\alpha,x+2)}{\zeta(\alpha)-\overline{\beta}[(x+1)^{-\alpha}+\zeta(\alpha,x+2)]} = \\ =& \frac{\zeta(\alpha)-\overline{\beta}\zeta(\alpha,x+2)}{\zeta(\alpha)-\overline{\beta}(x+1)^{-\alpha}-\overline{\beta}\zeta(\alpha,x+2)]} = \left[1-\frac{\overline{\beta}(x+1)^{-\alpha}}{h_{(\alpha,\beta)}(x+2)}\right]^{-1}. \end{aligned}$$

If $\beta \in (0,1)$, by Lemma 1, one has that $(x+1)^{\alpha}h_{(\alpha,\beta)}(x+2)$ is an increasing function of $x$, and consequently,

$$\left(1-\frac{\overline{\beta}(x+1)^{-\alpha}}{h_{(\alpha,\beta)}(x+2)}\right)^{-1}$$

decreases by increasing the value of $x$. As $g(x)$ is a decreasing function of $x$, one has that $g(x)/g(x+2) \geq 1$, which is what we wanted to see. □

Figure 5 shows the behavior of the ratio that appears on the left hand side of equation (16) for $\alpha = 2.34$ and $\alpha = 5$. In both cases $\beta = 0.1, 0.6, 2, 10$ and 22. Observe that, for $\beta < 1$, the distribution is log-convex independently of the value of $\beta$. However, for $\beta > 1$, the function can be log-convex or log-concave.

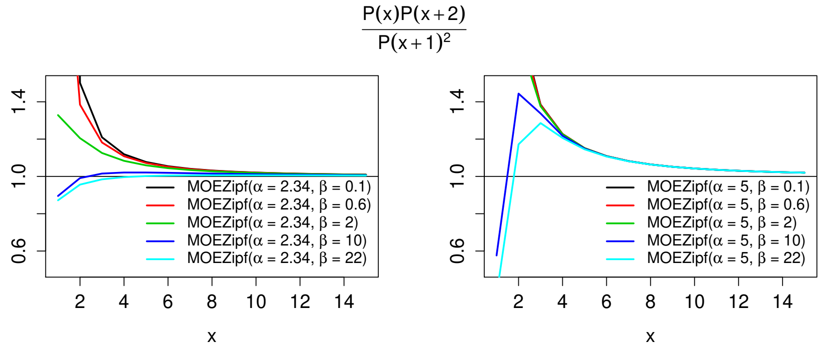

Fig. 5: Behavior of the ratio that appears in equation (16). On the left hand side, for $\alpha = 2.34$ and, on the right hand side, for $\alpha = 5$. In both cases, $\beta = 0.1, 0.6, 2, 10$ and 22.

The next proposition establishes the relationship between the probability values of a MOEZipf and a Zipf distribution with the same $\alpha$ value. This proposition extends Proposition 3.3 by Pérez-Casany and Casellas (2013, p. 6), where only the lower bounds are stated.

**Proposition 3** *Let $Y$ and $X$ be two r.v.'s, such that $Y \sim$ MOEZipf$(\alpha,\beta)$ and $X \sim$ Zipf$(\alpha)$. Then, $\forall x \geq 1$,*

*a) if $\beta \in (0,1)$, then $\beta P(X=x) \leq P(Y=x) \leq \frac{1}{\beta}P(X=x)$,*

*b) if $\beta > 1$, then $\frac{1}{\beta} P(X=x) \leq P(Y=x) \leq \beta P(X=x)$,*
*c) if $\beta = 1$, then $P(Y=x) = P(X=x)$.*

*Proof.* Considering $\beta > 1$ according to Lemma 1, one has that $h_{(\alpha,\beta)}(x)$ is a decreasing function of $x$, and that $h_{(\alpha,\beta)}(x) \leq h_{(\alpha,\beta)}(1) = \beta\,\zeta(\alpha)$, $\forall x \geq 1$. Thus,

$$P(Y=x) = \frac{\beta\,\zeta(\alpha)\,x^{-\alpha}}{h_{(\alpha,\beta)}(x)\,h_{(\alpha,\beta)}(x+1)} \geq \frac{\beta\,\zeta(\alpha)\,x^{-\alpha}}{\beta^2\,\zeta^2(\alpha)} = \frac{1}{\beta}P(X=x),$$

which proves the left hand side of $b$); to see the inequality on the right hand side, it is necessary to take into account that $h_{(\alpha,\beta)}(x) \geq \zeta(\alpha)$, $\forall x \geq 1$, and that

$$P(Y=x) = \frac{\beta\,x^{-\alpha}\,\zeta(\alpha)}{h_{(\alpha,\beta)}(x)\,h_{(\alpha,\beta)}(x+1)} \leq \frac{\beta\,x^{-\alpha}\,\zeta(\alpha)}{\zeta(\alpha)^2} = \beta\,P(X=x).$$

Point $a$) is proved in a similar way using the results of Lemma 1 $a$). Finally, c) is a direct consequence of the definition of the MOEZipf distribution. □

The next theorem relates the tail of the MOEZipf distribution to the tail of the Zipf distribution with the same parameter $\alpha$.

**Theorem 4** *Let $Y$ and $X$ be two r.v.'s, such that $Y \sim$ MOEZipf$(\alpha,\beta)$ and $X \sim$ Zipf$(\alpha)$. The tail of $Y$ is asymptotically equivalent to $\beta$ times the tail of $X$, $\forall\,\beta > 0$.*

*Proof.* We will distinguish if $Y$ is a maximum ($\beta > 1$) or if it is a minimum ($\beta \in (0,1)$). If Y is a minimum, given that $n_0 = 1$ and that $P(N=1) = (1-\beta)^{n_0-1}\beta = \beta$, then, from Theorem 3 a), one has that:

$$P(Y > x) \sim P(N=1)P(Y > x),$$

which implies that

$$P(Y > x-1) - P(Y > x) \sim P(N=1)[P(X > x-1) - P(X > x)] \Leftrightarrow$$
$$P(Y=x) \sim \beta\,P(X=x).$$

If $Y$ is a maximum, from Proposition 1, one has that $E[N] = 1/(1/\beta) = \beta$. And thus, from Theorem 3 b), one has that:

$$P(Y > x) \sim E[N]P(X > x),$$

which implies that

$$P(Y > x-1) - P(Y > x) \sim E[N][P(X > x-1) - P(X > x)] \Leftrightarrow$$
$$P(Y=x) \sim \beta\,P(X=x).$$

□

Figure 6 illustrates the results stated in Theorem 4 for $\alpha = 2.8$ and $\beta = 0.3$ (left hand side) and $\alpha = 2.8$ and $\beta = 4.86$ (right hand side). Observe that, for the parameter values considered, the convergence of the probabilities is faster when the distribution is defined in terms of minimums.

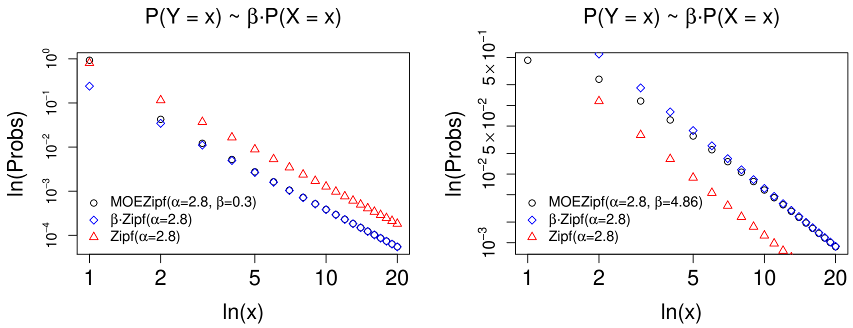

Fig. 6: Probabilities of the Zipf and the MOEZipf distributions with the same $\alpha$ parameter in log-log scale, jointly with $\beta$ times the probability of the Zipf. The MOEZipf on the left-hand side is defined in terms of minimums and, on the right-hand side, it is defined in terms of maximums.

### 5.2 The Zipf-Poisson Extreme distribution

The paper by Ramos et al. (2018) proposes a unified framework for generalizing a family of distributions, which corresponds to an RSED with a positive Poisson stopping distribution. The results of this paper intersect with those presented by Pérez-Casany et al. (2016). In their applications, the authors focus on extending the continuous distributions: exponential, Weibull and Generalized Extreme Value. In this section we focus on extending the Zipf distribution, although others discrete distributions may similarly be considered.

The Zipf-PE family of distributions is obtained when the r.v. $N$ is assumed to be a positive Poisson distribution. The resulting distribution has support on the strictly positive integer numbers, and its parameters are the $\alpha$ parameter of the Zipf and the $\beta$ parameter of the positive Poisson.

Considering $Y$ as an r.v. with a Zipf-PE distribution, the third row of Table 2 shows the CDF of $Y$ if it corresponds to a maximum, and the SF of $Y$ if it corresponds to a minimum. Nevertheless, a little bit of algebra reveals that for any $\alpha > 1$, the CDF of $Y$ when it corresponds to a minimum has the same expression as the CDF of a maximum, but for negative values of $\beta$. Thus, the CDF of any Zipf-PE is equal to:

$$F_{(\alpha,\beta)}(x) = \begin{cases} \frac{e^{\beta\left(\frac{\zeta(\alpha)-\zeta(\alpha,x+1)}{\zeta(\alpha)}\right)}-1}{e^{\beta}-1}, \beta \in \mathbb{R}\backslash\{0\}, \\ 1 - \frac{\zeta(\alpha,x+1)}{\zeta(\alpha,x)}, \beta = 0, \end{cases} \tag{17}$$

where positive values of $\beta$ correspond to maximums of a Po$(\beta)$ number of copies, and negative values correspond to minimums of a Po$(-\beta)$ number of copies. Moreover, as mentioned in Section 3, the parameter space of the zero-truncated Poisson distribution includes the zero value that corresponds to the degenerate distribution at one. That is the reason why the value $\beta = 0$ is also included in (17) and, in this case, $Y$ follows the baseline distribution, that is, the Zipf$(\alpha)$ distribution.

From (17) $\forall\alpha > 1$ and $x \geq 2$, one can obtain the PMF of Y as follows:

$$\begin{aligned} P(Y=x) &= F_{(\alpha,\beta)}(x) - F_{(\alpha,\beta)}(x-1) = \\ &= \frac{e^{\beta\left(\frac{\zeta(\alpha)-\zeta(\alpha,x+1)}{\zeta(\alpha)}\right)} - e^{\beta\left(\frac{\zeta(\alpha)-\zeta(\alpha,x)}{\zeta(\alpha)}\right)}}{e^{\beta}-1} \\ &= \begin{cases} \frac{e^{\beta}\, e^{\frac{-\beta\,\zeta(\alpha,x)}{\zeta(\alpha)}}\left(e^{\frac{\beta x^{-\alpha}}{\zeta(\alpha)}}-1\right)}{e^{\beta}-1}, & \beta \in \mathbb{R}\backslash\{0\}, \\ \frac{x^{-\alpha}}{\zeta(\alpha)}, & \beta = 0. \end{cases} \end{aligned} \tag{18}$$

For $x = 1$,

$$P(Y=1) = F_{(\alpha,\beta)}(1) = \begin{cases} \frac{e^{\frac{\beta}{\zeta(\alpha)}}-1}{e^{\beta}-1}, & \beta \in \mathbb{R}\backslash\{0\}, \\ \frac{1}{\zeta(\alpha)}, & \beta = 0, \end{cases} \tag{19}$$

which is equal to (18) at $x = 1$. Thus (19) is the PMF in the entire support. Figure 7 shows the PMFs of the Zipf-PE distribution in log-log scale for $\alpha = 2.1$ and different values of the $\beta$ parameter. Note that the $\beta$ parameter influences the top-concavity (top-convexity) at the low values of the distribution. For $\beta > 0$, the distribution is top-concave while, for $\beta < 0$, the distribution is top-convex.

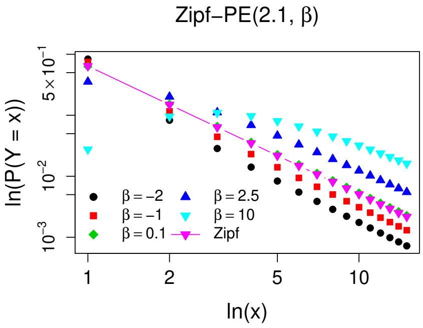

Fig. 7: PMFs of the Zipf-PE distribution in log-log scale for $\alpha = 2.1$ and $\beta = -2, -1, 0.1, 2.5$ and 10.

The next proposition states that the probability at one of an r.v. with a Zipf-PE$(\alpha,\beta)$ distribution is always smaller (larger) than the probability at one of a Zipf distribution with the same parameter $\alpha$, depending on the sign of $\beta$. Negative values of $\beta$ inflate the probability at one while positive values deflate it. This is reasonable because $\beta < 0\,(\beta > 0)$ corresponds to minimums (maximums) and, thus, inflates (deflates) the probabilities of the first values.

**Proposition 4** *Let $Y$ and $X$ be two r.v.'s, such that $Y \sim$ Zipf-PE$(\alpha,\beta)$ and $X \sim Zipf(\alpha)$. Then, $P(Y=1) \leq (\geq) P(X=1)$ for all $\beta > 0 (\beta < 0)$, and the equality holds only when $\beta = 0$.*

*Proof.* If $\beta \neq 0$, taking into account (19) it is necessary to prove that:

$$P(Y=1) \leq (\geq) P(X=1) \Leftrightarrow \frac{e^{\frac{\beta}{\zeta(\alpha)}}-1}{e^{\beta}-1} - \frac{1}{\zeta(\alpha)} \leq (\geq) 0.$$

Let us define the function

$$g(x) = \frac{e^{\beta x}-1}{e^{\beta}-1} - x, \forall x \in [0,1].$$

Observe that $g(x)$ is a continuous and differentiable function in $(0,1)$, which verifies that $g(0) = g(1) = 0$. Applying Bolzano's theorem (Apostol, 1974, p. 84), we have that exists a value $x_0 \in (0,1)$, such that $g'(x_0) = 0$. Differentiating, one has:

$$g'(x_0) = 0 \Leftrightarrow e^{\beta x_0} = \frac{e^{\beta}-1}{\beta} \Rightarrow x_0 = \frac{1}{\beta}\log(\frac{e^{\beta}-1}{\beta}).$$

By computing the second derivative of $g(x)$, one has:

$$g''(x) = \frac{e^{\beta x}\beta^2}{e^{\beta}-1},$$

which is positive if $\beta > 0$, and negative otherwise. Thus, if $\beta > 0$, $x_0$ is a minimum, $g(x) \leq 0 \forall a \in [0,1]$, and, in particular, $g(\frac{1}{\zeta(\alpha)}) \leq 0$. In contrast, if $\beta < 0$, $x_0$ is a maximum, $g(x) \geq 0 \forall x \in [0,1]$, and, in particular, $g(\frac{1}{\zeta(\alpha)}) \geq 0$. □

### 5.2.1 Moments

The following proposition assesses the condition under which the *k-th* moment of a Zipf-PE distribution is finite, which is the same as for the Zipf and the MOEZipf family of distributions.

**Proposition 5** *The k-th moment of a Zipf-PE distribution exists and is finite if, and only if, $\alpha > k+1$.*

*Proof.* Let $Y$ and $X$ be two r.v.'s, such that $Y \sim$ Zipf-PE$(\alpha,\beta)$ and $X \sim$ Zipf$(\alpha)$. As mentioned in Section 4, the *k-th* moment of the Zipf distribution converges if, and only if, $\alpha > k+1$. Applying the comparison criteria of convergence of series of positive terms, one has:

$$\lim_{x\to+\infty} \frac{P(Y=x)x^k}{P(X=x)x^k} = \lim_{x\to\infty} \frac{\frac{e^{\beta} e^{\frac{-\beta\zeta(\alpha,x)}{\zeta(\alpha)}}\left(e^{\frac{\beta x^{-\alpha}}{\zeta(\alpha)}}-1\right)}{e^{\beta}-1}}{\frac{x^{-\alpha}}{\zeta(\alpha)}} =$$

$$= \frac{e^{\beta}\zeta(\alpha)}{e^{\beta}-1} \cdot \lim_{x\to+\infty} e^{\frac{-\beta\zeta(\alpha,x)}{\zeta(\alpha)}} \cdot \lim_{x\to+\infty} \frac{e^{\frac{\beta x^{-\alpha}}{\zeta(\alpha)}}-1}{x^{-\alpha}}.$$

Given that $\zeta(\alpha,x)$ tends to zero when $x$ tends to $+\infty$, $\lim_{x\to+\infty} e^{\frac{-\beta\zeta(\alpha,x)}{\zeta(\alpha)}} = 1$. Moreover, applying L'Hôpital rule, one has:

$$\lim_{x\to+\infty} \frac{e^{\frac{\beta x^{-\alpha}}{\zeta(\alpha)}}-1}{x^{-\alpha}} = \lim_{x\to+\infty} e^{\frac{\beta x^{-\alpha}}{\zeta(\alpha)}} \frac{\beta}{\zeta(\alpha)} = \frac{\beta}{\zeta(\alpha)}.$$

Thus,

$$\lim_{x\to+\infty} \frac{P(Y=x)x^k}{P(X=x)x^k} = \frac{e^{\beta}\,\zeta(\alpha)}{e^{\beta}-1}\,\frac{\beta}{\zeta(\alpha)} = \frac{\beta}{1-e^{-\beta}}, \neq 0, +\infty.$$

Since the limit $\beta/(1-e^{-\beta})$ is a constant value different from zero, the *k-th* moment of the Zipf-PE$(\alpha,\beta)$ distribution converges if, and only if, the *k-th* moment of the Zipf$(\alpha)$ converges, that is, when $\alpha > k+1$. □

Figures 8 shows the behavior of the mean as: a function of $\alpha$ for $\beta = -1, -0.5, 1.5$ and 3 (on the left hand side); and as a function of $\beta$ for $\alpha = 2.5, 4.8, 7.5, 20$ (on the right hand side). A similar plot for the variance appears in Figure 9: on left hand side as a function of $\alpha$ for $\beta = -1, -0.5, 1.5$ and 3; and on the right hand side as a function of $\beta$ for $\alpha = 3.5, 4.8, 7.5$ and 20. Note that, on the left hand side of both figures, the $E[Y]$ and the $Var[Y]$ are not only decreasing functions of $\alpha$, but they decrease faster as $\beta$ becomes smaller. On the right-hand side of both figures can be observed that the $E[Y]$ and $Var[Y]$ are increasing functions of $\beta$, with a slope that decreases when $\alpha$ increases.

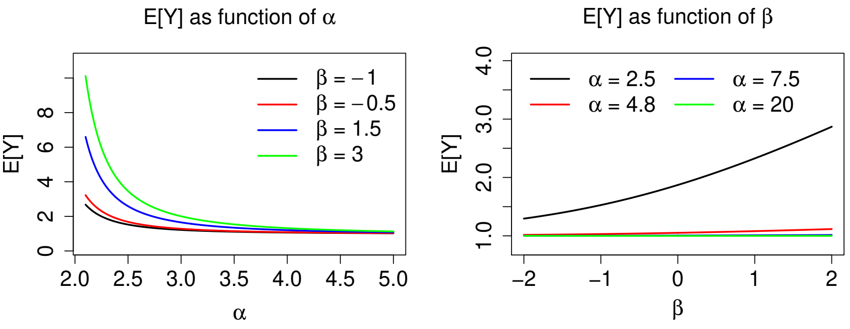

Fig. 8: Mean values of a Zipf-PE$(\alpha,\beta)$ distribution. On the left hand side: as a function of $\alpha$ for $\beta = -1, -0.5, 1.5$ and 3. On the right hand side: as a function of $\beta$ for $\alpha = 2.5, 4.8, 7.5$ and 20.

**Proposition 6** *Let $Y$ and $X$ be two r.v.'s, such that $Y \sim$ Zipf-PE$(\alpha,\beta)$ and $X \sim$ Zipf$(\alpha)$. Then, the ratio of two consecutive probabilities of $Y$ is equal to:*

$$\frac{P(Y=x+1)}{P(Y=x)} = \frac{e^{\beta\,P(X=x+1)}-1}{1-e^{-\beta\,P(X=x)}}.$$

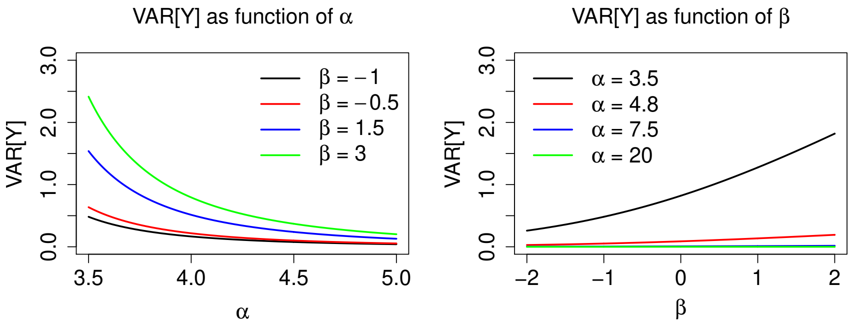

Fig. 9: Variance values of a Zipf-PE$(\alpha,\beta)$ distribution. On the left hand side: as a function of $\alpha$ for $\beta = -1, -0.5, 1.5$ and 3. On the right hand side: as a function of $\beta$ for $\alpha = 3.5, 4.8, 7.5$ and 20.

*Proof.* From (18) one has:

$$\frac{P(Y=x+1)}{P(Y=x)} = \frac{e^{\beta}\, e^{\frac{-\beta\zeta(\alpha,x+1)}{\zeta(\alpha)}} \left(e^{\frac{\beta\,(x+1)^{-\alpha}}{\zeta(\alpha)}} - 1\right)}{e^{\beta}\, e^{\frac{-\beta\zeta(\alpha,x)}{\zeta(\alpha)}} \left(e^{\frac{\beta\,x^{-\alpha}}{\zeta(\alpha)}} - 1\right)} = e^{\frac{\beta\,x^{-\alpha}}{\zeta(\alpha)}} \frac{\left(e^{\frac{\beta\,(x+1)^{-\alpha}}{\zeta(\alpha)}} - 1\right)}{\left(e^{\frac{\beta\,(x)^{-\alpha}}{\zeta(\alpha)}} - 1\right)}$$

$$= \frac{e^{\frac{\beta\,(x+1)^{-\alpha}}{\zeta(\alpha)}} - 1}{1 - e^{\frac{-\beta\,x^{-\alpha}}{\zeta(\alpha)}}} = \frac{e^{\beta\,P(X=x+1)} - 1}{1 - e^{-\beta\,P(X=x)}}.$$

□

Figure 10 shows the behavior of this ratio for $\alpha = 2.1$ and $\beta = -3$ and 3. The ratio of the Zipf$(\alpha)$ is also included in order to facilitate comparison between both distributions. Note that, when $\beta > 0$, the ratio associated with the Zipf-PE$(\alpha,\beta)$ converges faster to that of the Zipf distribution. In contrast, when $\beta < 0$, the convergence is not that fast, even though it also converges to that of the Zipf. In general, the most significant difference occurs at the initial values of $x$, which is another manner of observing the flexibility of the Zipf-PE distribution at the first integer values. In addition, by increasing the value of $x$, the ratio of all the distributions tends to one. Moreover, independently of the $\beta$ value, those values in the tail of the distribution behave similarly to those of the Zipf distribution, which is proven in the Theorem 5.

The next theorem establishes the relationship between the tail of the Zipf-PE and the tail of the Zipf distributions.

**Theorem 5** *The tail of an r.v. $Y \sim$ Zipf-PE$(\alpha,\beta)$ is asymptotically related to the tail of an r.v. $X \sim$ Zipf$(\alpha)$, in such a way that:*

*a) if $\beta < 0$, then $Y$ is a minimum and,*

$$P(Y=x) \sim \frac{-\beta\, e^{\beta}}{1-e^{\beta}} P(X=x),$$

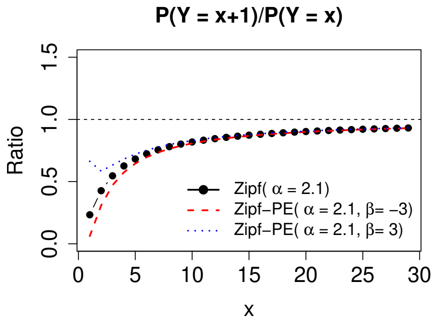

Fig. 10: Ratio of two consecutive Zipf-PE probabilities for $\alpha = 2.1$, with $\beta = -3$ and 3, respectively.

*b) if* $\beta > 0$*, then* $Y$ *is a maximum and,*

$$P(Y = x) \sim \frac{\beta}{(1 - e^{\beta})} P(X = x).$$

*Proof.* From Theorem 3 one has that, if $\beta < 0$, then $n_0 = 1$ and $P(N = 1) = -\beta\, e^{\beta}/1 - e^{\beta}$. Consequently,

$$P(Y > x) \sim P(N = n_0)[P(Y > x)]^{n_0},$$

is equivalent to:

$$P(Y > x-1) - P(Y > x) \sim P(N = n_0)[P(X > x-1) - P(X > x)]^{n_0} \Leftrightarrow$$

$$P(Y = x) \sim \frac{-\beta\, e^{\beta}}{1 - e^{\beta}} P(X = x),$$

which proves a). If $\beta > 0$, then $E[N] = \beta/(1 - e^{-\beta})$ and, consequently,

$$P(Y > x) \sim E[N] P(X > x)$$

is equivalent to:

$$P(Y > x-1) - P(Y > x) \sim E[N][P(X > x-1) - P(X > x)] \Leftrightarrow$$

$$P(Y = x) \sim \frac{\beta}{1 - e^{-\beta}} P(X = x),$$

which proves b). □

Figure 11 shows the results achieved in the previous Theorem. Observe that for the parameter values used, the equivalence between the tails of both distributions emerges for $x \geq 10$.

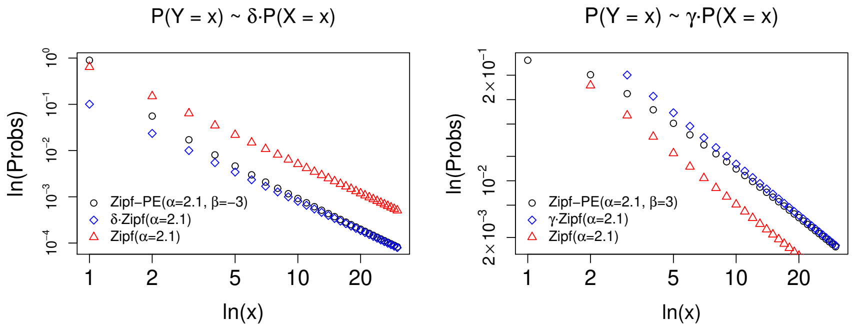

Fig. 11: The probabilities of the Zipf and Zipf-PE distributions with the same $\alpha$ parameter. On the left-hand side, jointly with $\delta = -\beta\, e^{\beta}/(1-e^{\beta})$ times the probability of the Zipf. On the right-hand side, jointly with $\gamma = \beta/(1-e^{\beta})$ times the probability of the Zipf. In both plots, the probabilities are shown in log-log scale. On the left hand side: defined in terms of the minimum family. On the right hand side: in terms of the maximum family.

## 6 Applications

The aim of this section is to illustrate the performance of both the MOEZipf and Zipf-PE families of distributions when they are used to fit real data. We analyze two data sets that contain the degree sequences of real networks. Each degree sequence contains information on nodes that have at least one connection in the network. Thus, isolated nodes are not taken into account. In both examples we assume independent observations. The work by Duarte-López et al. (2015) shows the suitability of the MOEZipf distribution when it is used to fit this type of data. Moreover, the work by Güney et al. (2017) has also used the MOEZipf distribution for modeling the frequency of cancer types in Turkey during the period 2007-2011.

Other bi-parametric models have also been considered, such as the Discrete Gaussian Exponential (DGX) (Bi et al., 2001), the Zipf-Polylog (Valero et al., 2020) and the positive version of the Zipf-PSS (Duarte-López et al., 2020); and their fits have been compared to those associated with the proposed distributions. An implementation of the MOEZipf, the Zipf-PE, the Zipf-Polylog and the Zipf-PSS distributions can be found in the R package *zipfextR* (Duarte-López and Pérez-Casany, 2018), which is available through the CRAN repository.

The Akaike Information Criterion (AIC) and the log-likelihood, at the maximum likelihood parameter estimates, were chosen as goodness-of-fit criteria. The Likelihood Ratio Test (LRT) was performed to compare the Zipf model with the Zipf extension that provided the best fit for each particular data set.

For illustrative purposes, we also include the results obtained from the methodology presented by Clauset et al. (2009). By means of this procedure, the power law is fitted for values above a certain threshold at which the distribution holds. To that aim, the authors estimate a cut-off value, $x_{min}$, by minimizing the value of the

Kolmogorov-Smirnov statistic $(D)$; then, the $\alpha$ parameter of the power law distribution is obtained using MLE on the truncated data. It is important to say that our models are not compared with the fits obtained by means of Clauset's methodology, basically because the domain of the two approaches are different.

### 6.1 Application 1: Collaboration Network

Collaboration networks are important because they play an important role in measuring how knowledge spreads. Furthermore, they allow detecting strategical research collaborations. The co-authorship network studied in this section was created and analyzed in the paper by Molontay and Nagy (2019). Their work is a tribute to the work developed by the network science community in the last 20 years. During the network construction phase, the authors used the Web of Science bibliographic database to collect all the network science papers published in the period 1998-2019. The authors classify a publication as a *network science paper* if it cites at least one of the following important papers: Barabási and Albert (1999), Watts and Strogatz (1998) or Girvan and Newman (2002). After conducting an accurate pre-processing step, they obtained a dataset of 29528 different papers, leading to 52406 authors representing nodes in the network. An edge is created between two authors if they co-authored at least one network science paper. The data set containing this undirected network is accessible through the git-hub repository: `https://github.com/marcessz/Two-Decades-of-Network-Science`.

Table 3 summarizes the main statistical properties of the network and its degree sequence. From the total number of authors, 851 are reported as isolated nodes. This means that these authors have not shared any publications with the other members of the network.

In this analysis we have considered only those authors that have at least one connection in the network. Consequently, the isolated nodes are not included as part of this analysis. For those interested in the analysis of this kind of authors we strongly recommend the use of the Zipf-PSS distribution Duarte-López et al. (2020).

Table 3: Characteristics of the degree sequence: number of nodes (**N**); number of edges (**E**); (**Range**); (**Mean**); variance (**Var**); skewness (**Skew**).

| **N** | **E** | **Range** | **Mean** | **Var** | **Skew** |
|---|---|---|---|---|---|
| 52406 | 329181 | 443 | 12.7701 | 2310.7120 | 6.8616 |

Table 4 contains the parameter estimates and their confidence intervals, as well as the log-likelihood and AIC values for all the considered distributions used in the first part of the study. Without including the isolated nodes, the Zipf-PE distribution is the one that provides the best fit to the data, closely followed by the positive Zipf-PSS. Figure 12 shows the fits obtained by each considered distribution. In general, the distribution families with a clear long right tail are the ones providing the best fit to the real observations. On the other hand, by applying Clauset's methodology just for illustrating, the cut-off point is fixed to be equal to 4, which implies that 43.3% of the authors in the network are not considered.

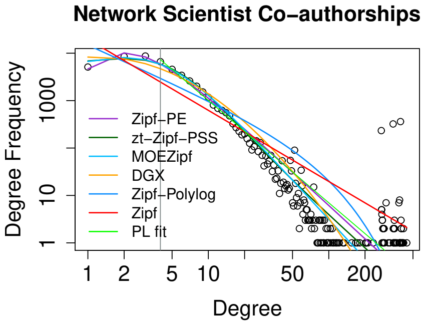

Fig. 12: Degree sequence of the co-authorship network and the fit obtained by each one of the considered models. In addition, the fit obtained using the methodology proposed by Clauset et al. (2009) is also included.

For this data set and the Zipf-PE distribution, we interpret $N$ as the number of papers published by an author in this period of time. Variable $X$ is interpreted as the number of co-authors in a given publication, which has sense to be Zipf distributed since there are papers with very few authors and others with a large number of authors. Finally, the maximum of the number of co-authors for all the publications published by a person has sense to be a good approximation of the total number of co-authors, which corresponds to variable $Y$. Thus, based on the parameter interpretation of the Zipf-PE distribution, we can say that an author published an average of $\widehat{E[N]} = 8.2382$ papers in the period 1998-2019. For each paper published, the number of co-authors in a given publication is estimated by $\widehat{E[X]} = 2.48779$. Finally, $\widehat{E[Y]} = 9.52$ is an estimation of the total number of co-authors of an author in this period of time.

Performing the LRT, we can confirm that the Zipf-PE obtains a better fit than that of the Zipf, thus ensuring the importance of the extra parameter included in the new model when fitting the data in its entire range. As in the previous examples, the significance level considered is $\alpha = 0.05$, which leads to a critical point equal to $\chi^2_{0.95,1} = 3.84$. The LR statistic for this degree sequence is equal to $-2[-165879.1326 - (-146709.7874)] = 38338.69$, which means that the null hypothesis of the Zipf distribution is clearly rejected, thus ensuring the superiority of the Zipf-PE in providing a better fit to the data.

Table 7 in Appendix A contains, for the first 15 integer values, their relative frequencies jointly with the theoretically probabilities of the Zipf($\hat{\alpha}$) and Zipf-PE($\hat{\alpha}, \hat{\beta}$) distributions. It also shows the ratio of two consecutive frequencies as well as the ratio of two consecutive probabilities of the already mentioned distributions. This table allows us to observe that the probabilities of the Zipf-PE are closer to the observed

Table 4: The parameter estimations for each analyzed distribution, their confidence intervals, log-likelihood, and AIC goodness-of-fit measures.

| Distribution | $param_1$ | $CI_{param_1}$ | $param_2$ | $CI_{param_2}$ | Log-like | AIC |
|---|---|---|---|---|---|---|
| **Zipf-PE** | $\hat{\alpha} = 2.3442$ | (2.3339, 2.3544) | $\hat{\beta} = 8.2636$ | (8.1259, 8.4012) | **-146709.7874** | **293423.5748** |
| zt-Zipf-PSS | $\hat{\alpha} = 2.2364$ | (2.2259, 2.247) | $\hat{\lambda} = 2.6946$ | (2.6691, 2.7201) | -146935.1020 | 293874.2040 |
| MOEZipf | $\hat{\alpha} = 2.7668$ | (2.7509, 2.7827) | $\hat{\beta} = 26.1943$ | (25.3232, 27.0654) | -147547.3002 | 295098.6003 |
| DGX | $\hat{\mu} = 1.4308$ | (1.4195, 1.442) | $\hat{\sigma} = 1.1161$ | (1.107, 1.1253) | -151357.8829 | 302719.7658 |
| Zipf-Polylog | $\hat{\alpha} = 1.1366$ | (1.1271, 1.1461) | $\hat{\beta} = 0.9863$ | (0.9858, 0.9868) | -160820.0267 | 321644.0533 |
| Zipf | $\hat{\alpha} = 1.5001$ | (1.4957, 1.5045) | - | - | -165879.1326 | 331760.2651 |

frequencies than those of the Zipf. Also, it shows that while the ratios of the Zipf distribution are always decreasing, those of the Zipf-PE distribution can increase and later decrease when the data show this pattern.

### 6.2 Application 2: Protein-Protein Interaction Network

Network analysis is also a profitable tool in the field of biology, as it helps model the interactions of organisms and proteins, among other objects of study. Therefore, the second application example focuses on analyzing the degree distribution of the *Arabidopsis thaliana* comprehensive knowledge network (AtCKN), see Ramšak et al. (2018). This network is the result of combining a plant immune signaling model with three extra layers of information: protein-protein interactions (PPI); transcriptional regulation; and regulation through microRNA. The resulting network is composed of 20011 nodes and 58901 edges (see Table 5).

Most of the nodes in the network have less than or equal to 30 interacting partners (19462 proteins; 97.26%), which is double the number of pure protein-protein interaction networks (Lee et al., 2010). This can probably be attributed to the fact that AtCKN not only includes protein-protein type reactions, but also transcriptional regulation (protein to gene) and regulations through microRNA (miRNA to gene). Proteins with a very large number of interactions in AtCKN belong to various transcription factor families, which in turn increases the number of interacting partners.

Table 5 contains the main statistics for the AtCKN degree sequence. Again, the data show large variability as well as high skewness value, which allows us to hypothesize the suitability of the proposed models.

Table 5: Characteristics of the degree sequence: number of nodes (**N**); number of edges (**E**); (**Range**); (**Mean**); variance (**Var**); skewness (**Skew**).

| N | E | Range | Mean | Var | Skew |
|---|---|---|---|---|---|
| 20011 | 58901 | 4688 | 5.89 | 143.75 | 7.16 |

Table 6 contains the maximum likelihood parameter estimation for all the fitted models, jointly with their 95% confidence intervals. It also contains the values of the log-likelihood and the AIC. The AIC value confirms that the worst models are Zipf and Zipf-Polylog. In contrast, the Zipf-PE distribution provides the best fit, followed by MOEZipf and the positive Zipf-PSS. These three models have linear tails, which is not the case for the Zipf-Polylog. The DGX gives a better fit than the Zipf and the Zipf-Polylog, but it is worse than the two models presented in this paper.

Table 6: The parameter estimations for each analyzed distribution, their confidence intervals, log-likelihood, and AIC goodness-of-fit measures.

| **Distribution** | $param_1$ | $CI_{param_1}$ | $param_2$ | $CI_{param_2}$ | **Log-like** | **AIC** |
|---|---|---|---|---|---|---|
| **Zipf-PE** | 2.3241 | (2.305, 2.3432) | 4.8585 | (4.7169, 5.0001) | **-49429.9595** | **98863.9191** |
| MOEZipf | 2.5575 | (2.5313, 2.5837) | 9.3057 | (8.8438, 9.7676) | -49518.0492 | 99040.0984 |
| zt-Zipf-PSS | 2.1698 | (2.1517, 2.1878) | 1.6747 | (1.6377, 1.7116) | -49563.1824 | 99130.3648 |
| DGX | 0.9308 | (0.9054, 0.9563) | 1.1616 | (1.1424, 1.1807) | -49818.0469 | 99640.0939 |
| Zipf-Polylog | 1.0091 | (0.9859, 1.0322) | 0.9454 | (0.9425, 0.9483) | -50816.0231 | 101636.0462 |
| Zipf | 1.6174 | (1.6086, 1.6262) | - | - | -53085.1701 | 106172.3402 |

Figure 13 illustrates the fits obtained for each considered model. Observe the Zipf's model lacks of flexibility in adapting the top-concave pattern and the curvature drawn by the Zipf-Polylog in the middle range of the data. This highlights its lack of fit. On the other hand, the models Zipf-PE, MOEZipf, zt-Zipf-PSS and DGX seem to provide a quite accurate fit. With respect to Clauset's methodology, it establishes a cut-off equal to 4, from which the distribution is fitted. With this cut-off, the method loses approximately 61% of the data.

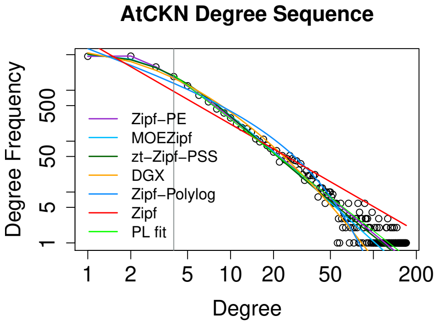

Fig. 13: Degree sequence of the PPI network and the fit obtained by each considered model. In addition, the fit obtained using the methodology proposed by Clauset et al. (2009) is also included.

Observe that both RSEDs agree with modeling the data in terms of maximums. This is because, it makes sense to assume that a protein must interact with the maximum number of elements required if it is going to produce the biological organism being modeled by the interaction network.

In this example $N$ is interpreted as the number of times that a protein is active. Variable $X$ describes the number of interactions performed every time that it is active. As in the previous example, it has sense to assume that the maximum of the number of interactions performed each time that a protein is active, is a good estimation of the total number of interactions. Based on the estimated parameters of the Zipf-PE distribution, we can say that, in average, a given protein is expected to be active approximately $\widehat{E[N]} = 4.89$ times. Moreover, every time a protein is active, we estimate its number of interactions to be equal to $\widehat{E[X]} = 2.62$. In general, the number of expected interactions of a given protein in the network is estimated by $\widehat{E[Y]} = 6.87$. According to Grigoriev (2003), the average interacting partners per protein in the proteome of a yeast (*Saccharomyces cerevisiae*) is about five; the estimates obtained from the Zipf-PE distribution - the best model - agree with the results of their paper.

Applying the LRT to compare $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$, we see that the critical value is equal to $\chi^2_{0.95,1} = 3.84$, and the likelihood ratio statistic for this degree sequence is equal to $-2[-53085.17 - (-49429.96)] = 7310.42$. By comparing the values, and given that $7310.42 \geq 3.84$, the null hypothesis is clearly rejected, and we conclude that the Zipf-PE distribution provides a better fit than the classical Zipf distribution.

Table 8 in Appendix A contains, for the first 15 integer values, their relative frequencies jointly with the theoretically probabilities of the Zipf$(\hat{\alpha})$ and Zipf-PE$(\hat{\alpha}, \hat{\beta})$ distributions. It also shows the ratio of two consecutive frequencies as well as the ratio of two consecutive probabilities of the already mentioned distributions. As in the previous example, this table allows us to observe that the probabilities of the Zipf-PE are closer to the observed frequencies than those of the Zipf. Also, it shows that while the ratios of the Zipf distribution are always decreasing, those of the Zipf-PE distribution can increase and later decrease when the data show this pattern.

## 7 The RSEZipf distributions in Synthetic Data Generation

It is very important to be able to synthetically generate graphs that mimic the characteristics of the real ones. This is because, researches are not always able to have as many real graphs as desired to meet their objectives, either for privacy or economic reasons.

At present we are working on defining a random graphs generator that ensures a degree distribution that follows the models presented in this paper. This is an early stage work that requires more time to be improved. Nevertheless, the first results are encouraging. Figure 14 shows the flow chart of an algorithm used for randomly generating graphs, whose degree sequence follows a Zipf-PE distribution defined in terms of a maximum ($\beta > 0$). In the chart, the list of *candidate nodes (CN)* refers to those nodes that have not being processed yet.

The algorithm presented takes into account two important concepts that appear in the paper by Barabási and Albert (1999) which are the *preferential attachment* (PA) and the *growing* of the network. PA describes the fact that nodes that are more connected have more probability to be connected ("the richer get richer"). The growth of the graph is based on the addition of new nodes and edges based on the distribution considered. In a similar way, it is possible to generate random graphs with a MOEZipf distribution, which just requires to change the stopping distribution.

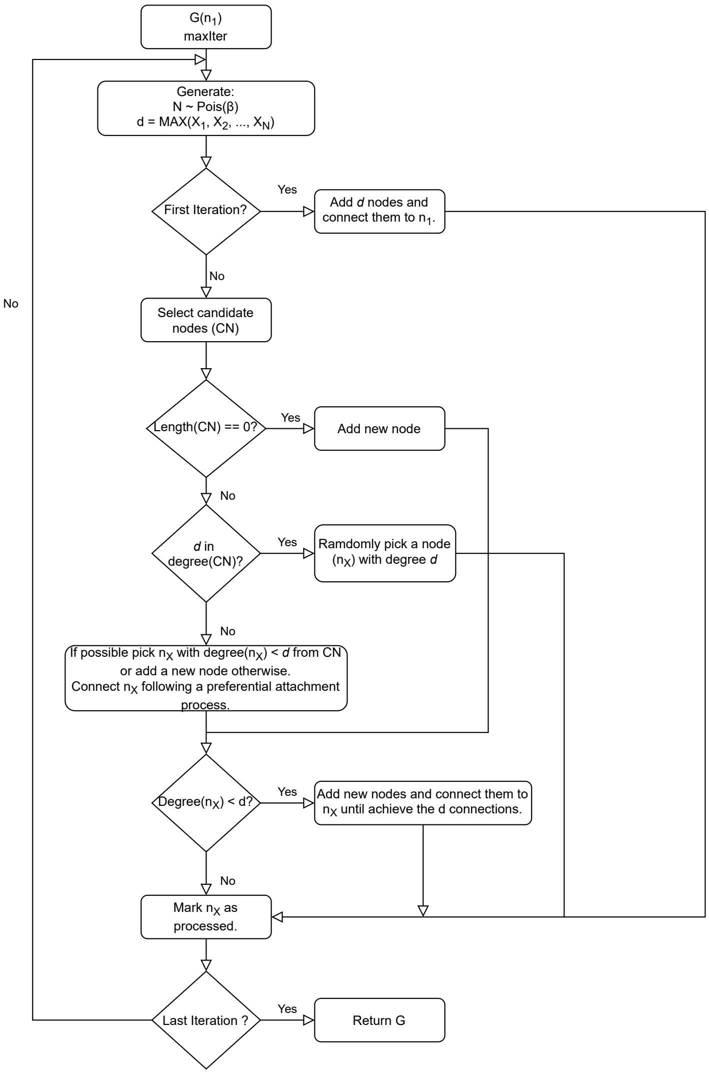
G($n_1$)
maxIter
Generate:
N ~ Pois(β)
d = MAX($X_1$, $X_2$, ..., $X_N$)
First Iteration?
Yes
Add *d* nodes and connect them to $n_1$.
No
No
Select candidate nodes (CN)
Length(CN) == 0?
Yes
Add new node
No
*d* in degree(CN)?
Yes
Ramdomly pick a node ($n_X$) with degree *d*
No
If possible pick $n_X$ with degree($n_X$) < *d* from CN or add a new node otherwise. Connect $n_X$ following a preferential attachment process.
Degree($n_X$) < d?
Yes
Add new nodes and connect them to $n_X$ until achieve the d connections.
No
Mark $n_X$ as processed.
Last Iteration ?
Yes
Return G

Fig. 14: Flow chart of an algorithm used for randomly generating graphs whose degree sequence follows a Zipf-PE distribution defined in terms of a maximum ($\beta > 0$).

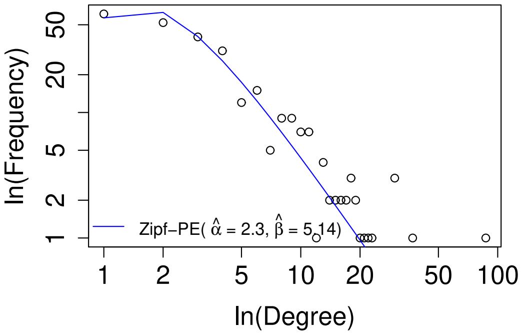

Fig. 15: Degree distribution, in log-log scale, of a synthetically generated graph from a Zipf-PE(3, 5) distribution and 500 iterations, jointly with the theoretical probabilities of the Zipf-PE at the MLE.

Figure 15 shows, in log-log scale, the degree sequence of a graph generated following the algorithm just described for $\alpha = 3$ and $\beta = 5$ and with 500 iterations, jointly with the theoretical probabilities of the Zipf-PE at the MLE. The MLE parameter estimators and their corresponding 95% confidence intervals are respectively equal to $\hat{\alpha} = 2.3$ with $CI_{\alpha} = (2.14, 2.45)$ and $\hat{\beta} = 5.14$ with $CI_{\beta} = (3.91, 6.37)$. It is important to see that the Zipf-PE$(2.3, 5.14)$ fits the generated data reasonably well. Although the theoretical value for the parameter $\alpha$ is not in the confidence interval for $\alpha$, the theoretical value for parameter $\beta$ is.

## 8 Conclusions

The Zipf distribution is widely used in many different disciplines to fit empirical data. Notwithstanding, it has important limitations such as its lack of flexibility, which allows practitioners to fit the distribution only in the data tails. The two RSEZipf distributions proposed in this work have been proven to allow for not only top-concavity and top-convexity when plotting the probabilities as a function of the size in log-log scale, but they also maintain the linearity in the tail. As a consequence, they can fit the data in all its range.

In addition, the parameters of the presented models allow in some cases for better understanding of the data generation process. The suitability of the models for fitting real data has been shown by fitting two degree sequences: a collaboration and a protein-protein interaction network. To do so, the R-package *zipfextR* has been used because it implements the MOEZipf and the Zipf-PE families. Finally, the authors have included the flow chart of an algorithm for randomly generating graphs with a degree sequence that follows the models presented. This work is in an early step, but the results are encouraging.

**Acknowledgements** The authors thank Živa Ramšak, PhD. from the National Institute of Biology, Slovenia, for sharing the data set used in the second application, and her helpful comments in the interpretation of the results obtained. We also thank Josep. Ll. Larriba-Pey from the DAMA-UPC research group from his interesting comments that have improved the Applications section. Finally, we acknowledge the support of Editor-in-Chief and the unknown Associate Editor and Referees for their constructive and interested suggestions.

## Conflict of interest

The authors declare that they have no conflict of interest.

## Appendix A Supplementary tables for Section 6

Table 7: First 15 degree values of the Collaboration network, jointly with their associated frequencies and estimated probabilities from the Zipf and Zipf-PE distributions (columns 1-4). The last three columns contain the ratio of two consecutive frequencies and the ratio of two consecutive probabilities for the mentioned distributions.

| Degree | Freq. | Zipf($\hat{\alpha}$) | Zipf-PE($\hat{\alpha},\hat{\beta}$) | $Freq_i/Freq_{i+1}$ | $P(X=x_i)/P(X=x_{i+1})$ | $P(Y=x_i)/P(Y=x_{i+1})$ |
|---|---|---|---|---|---|---|
| 1 | 0.0991 | 0.3829 | 0.0903 | 0.6 | 2.83 | 0.46 |
| 2 | 0.1649 | 0.1353 | 0.1968 | 0.98 | 1.84 | 1.22 |
| 3 | 0.169 | 0.0737 | 0.1616 | 1.32 | 1.54 | 1.41 |
| 4 | 0.1284 | 0.0478 | 0.1146 | 1.45 | 1.4 | 1.41 |
| 5 | 0.0887 | 0.0342 | 0.0813 | 1.34 | 1.32 | 1.37 |
| 6 | 0.0663 | 0.026 | 0.0592 | 1.32 | 1.26 | 1.33 |
| 7 | 0.0504 | 0.0207 | 0.0445 | 1.29 | 1.22 | 1.3 |
| 8 | 0.0391 | 0.0169 | 0.0343 | 1.33 | 1.19 | 1.27 |
| 9 | 0.0295 | 0.0142 | 0.027 | 1.29 | 1.17 | 1.24 |
| 10 | 0.0229 | 0.0121 | 0.0218 | 1.46 | 1.15 | 1.22 |
| 11 | 0.0157 | 0.0105 | 0.0178 | 1.18 | 1.14 | 1.2 |
| 12 | 0.0133 | 0.0092 | 0.0148 | 1.1 | 1.12 | 1.18 |
| 13 | 0.0121 | 0.0082 | 0.0125 | 1.12 | 1.12 | 1.18 |
| 14 | 0.0108 | 0.0073 | 0.0106 | 1.16 | 1.11 | 1.16 |
| 15 | 0.0093 | 0.0066 | 0.0091 | | | |

## References

Adamic LA, Huberman BA (2002) Zipf's law and the Internet. Glottometrics 3(1):143–150

Apostol TM (1974) Mathematical analysis

Ausloos M, Nedic O, Fronczak A, Fronczak P (2016) Quantifying the quality of peer reviewers through Zipfs law. Scientometrics 106(1):347–368

Barabási AL, Albert R (1999) Emergence of scaling in random networks. science 286(5439):509–512

Table 8: First 15 degree values of the PPI network, jointly with their associated frequencies and estimated probabilities from the Zipf and Zipf-PE distributions (columns 1-4). The last three columns contain the ratio of two consecutive frequencies and the ratio of two consecutive probabilities for the mentioned distributions.

| Degree | Freq. | Zipf($\hat{\alpha}$) | Zipf-PE($\hat{\alpha},\hat{\beta}$) | $Freq_i/Freq_{i+1}$ | $P(X=x_i)/P(X=x_{i+1})$ | $P(Y=x_i)/P(Y=x_{i+1})$ |
|---|---|---|---|---|---|---|
| 1 | 0.2323 | 0.4478 | 0.2305 | 0.99 | 3.07 | 0.98 |
| 2 | 0.2343 | 0.1457 | 0.2342 | 1.63 | 1.93 | 1.62 |
| 3 | 0.1434 | 0.0755 | 0.1447 | 1.55 | 1.59 | 1.6 |
| 4 | 0.0927 | 0.0474 | 0.0907 | 1.51 | 1.44 | 1.51 |
| 5 | 0.0613 | 0.033 | 0.0602 | 1.54 | 1.34 | 1.43 |
| 6 | 0.0398 | 0.0246 | 0.0422 | 1.35 | 1.29 | 1.37 |
| 7 | 0.0296 | 0.0191 | 0.0309 | 1.38 | 1.24 | 1.32 |
| 8 | 0.0214 | 0.0154 | 0.0234 | 1.2 | 1.21 | 1.29 |
| 9 | 0.0179 | 0.0127 | 0.0182 | 1.25 | 1.19 | 1.25 |
| 10 | 0.0143 | 0.0107 | 0.0146 | 1.32 | 1.16 | 1.24 |
| 11 | 0.0109 | 0.0092 | 0.0118 | 1.18 | 1.15 | 1.2 |
| 12 | 0.0092 | 0.008 | 0.0098 | 1.06 | 1.14 | 1.2 |
| 13 | 0.0087 | 0.007 | 0.0082 | 1.15 | 1.13 | 1.17 |
| 14 | 0.0075 | 0.0062 | 0.007 | 1.36 | 1.11 | 1.17 |
| 15 | 0.0055 | 0.0056 | 0.006 | | | |

Barabási AL, Pósfai M (2016) Network science. Cambridge university press

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 AD, Clauset A (2019) Scale-free networks are rare. Nature communications 10(1):1–10

Cancho VG, Louzada-Neto F, Barriga GDC (2011) The Poisson-exponential lifetime distribution. Comput Statist Data Anal 55(1):677–686, DOI 10.1016/j.csda.2010.05.033, URL https://doi.org/10.1016/j.csda.2010.05.033

Caron Y, Makris P, Vincent N (2007) Use of power law models in detecting region of interest. Pattern recognition 40(9):2521–2529

Casellas A (2013) La distribució Zipf Estesa segons la transformació Marshall-Olkin. Master's thesis, Universitat Politècnica de Catalunya

Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661–703, DOI 10.1137/070710111, URL https://doi.org/10.1137/070710111

Drees H, Janßen A, Resnick SI, Wang T (2020) On a minimum distance procedure for threshold selection in tail analysis. SIAM Journal on Mathematics of Data Science 2(1):75–102

Duarte-López A, Pérez-Casany M (2018) zipfextR: Zipf Extended Distributions. URL https://CRAN.R-project.org/package=zipfextR, r package version 1.0.1

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

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, DOI 10.1016/j.csda.2019.106838, URL https://doi.org/10.1016/j.csda.2019.106838

Ectors W, Kochan B, Janssens D, Bellemans T, Wets G (2018) Exploratory analysis of Zipfs universal power law in activity schedules. Transportation pp 1–24

Girvan M, Newman ME (2002) Community structure in social and biological networks. Proceedings of the national academy of sciences 99(12):7821–7826

Gómez-Déniz E (2010) Another generalization of the geometric distribution. TEST 19(2):399–415, DOI 10.1007/s11749-009-0169-3, URL https://doi.org/10.1007/s11749-009-0169-3

Grigoriev A (2003) On the number of protein–protein interactions in the yeast proteome. Nucleic acids research 31(14):4157–4161

Gulisashvili A (2012) Analytically tractable stochastic stock price models. Springer Finance, Springer, Heidelberg, DOI 10.1007/978-3-642-31214-4, URL https://doi.org/10.1007/978-3-642-31214-4

Güney Y, Tuaç Y, Arslan O (2017) Marshall-Olkin distribution: parameter estimation and application to cancer data. J Appl Stat 44(12):2238–2250, DOI 10.1080/02664763.2016.1252730, URL https://doi.org/10.1080/02664763.2016.1252730

Hill BM (1975) A simple general approach to inference about the tail of a distribution. The annals of statistics pp 1163–1174

Jessen AH, Mikosch T (2006) Regularly varying functions. Publications de l'Institut Mathematique (94)

Johnson NL, Kemp AW, Kotz S (2005) Univariate discrete distributions, 3rd edn. Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, DOI 10.1002/0471715816, URL https://doi.org/10.1002/0471715816

Kuş C (2007) A new lifetime distribution. Comput Statist Data Anal 51(9):4497–4509, DOI 10.1016/j.csda.2006.07.017, URL https://doi.org/10.1016/j.csda.2006.07.017

Lee K, Thorneycroft D, Achuthan P, Hermjakob H, Ideker T (2010) Mapping plant interactomes using literature curated and predicted protein–protein interaction data sets. The Plant Cell 22(4):997–1005

Louzada F, Bereta EM, Franco MA (2012) On the distribution of the minimum or maximum of a random number of iid lifetime random variables. Applied Mathematics 3(4):350–353

Manaris B, Romero J, Machado P, Krehbiel D, Hirzel T, Pharr W, Davis RB (2005) Zipf's law, music classification, and aesthetics. Computer Music Journal 29(1):55–69

Marshall AW, Olkin I (1997) A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. Biometrika 84(3):641–652, DOI 10.1093/biomet/84.3.641, URL https://doi.org/10.1093/biomet/84.3.641

McKelvey B, et al. (2018) Using maximum likelihood estimation methods and complexity science concepts to research power law-distributed phenomena. In: Hand-book of Research Methods in Complexity Science, Edward Elgar Publishing

Molontay R, Nagy M (2019) Two decades of network science: as seen through the co-authorship network of network scientists. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp 578–583

Pérez-Casany M, Casellas A (2013) Marshall-Olkin Extended Zipf Distribution. arXiv preprint arXiv:13044540

Pérez-Casany M, Valero J, Ginebra J (2016) Random-Stopped Extreme distributions. International Conference on Statistical Distributions and Applications. Niagara Falls, Canada. http://people.cst.cmich.edu/lee1c/icosda2016/ProgramBrochure/ProgramBrochure_ICOSDA2016_10-20-16.pdf#page=52

Ramos PL, Dey DK, Louzada F, Lachos VH (2018) An extended poisson family of life distribution: A unified approach in competitive and complementary risks. arXiv preprint arXiv:180507672

Ramšak Ž, Coll A, Stare T, Tzfadia O, Baebler Š, Van de Peer Y, Gruden K (2018) Network Modeling Unravels Mechanisms of Crosstalk between Ethylene and Salicylate Signaling in Potato. Plant physiology 178(1):488–499

Tahir MH, Cordeiro GM (2016) Compounding of distributions: a survey and new generalized classes. Journal of Statistical Distributions and Applications 3(1):13

Valero J, Pérez-Casany M, Duarte-López A (2020) The Zipf as a Mixture Distribution and its Polylogarithm Generalization (Submmited, 2020)

Watts DJ, Strogatz SH (1998) Collective dynamics of small-worldnetworks. nature 393(6684):440

Zipf GK (1949) Human Behaviour and the Principle of Least-Effort. Cambridge MA edn