Zipf Extensions and Their Applications for Modeling the Degree Sequences of Real Networks

By Ariel Duarte-López

Supervisor: Prof. Marta Pérez-Casany, PhD.

Department of Statistics and Operation Research
Technical University of Catalonia

This dissertation is submitted for the degree of
Doctor of Philosophy

November 2020

En memoria de Esther D. García Alonso, la mejor abuela del mundo, aunque no estás presente siento que me cuidas cada día.

In loving memory of Esther D. García Alonso, the best grandmother in the world, who even without being present, every day takes care of me.

Agradecimientos

...Al final llegó el final...
Joaquin Sabina

Parece mentira pero es una realidad que se materializará en breve. Este es el resultado de años de trabajo y esfuerzo. Ha sido un proceso emocionante, a veces divertido, otras no tanto... pero siempre enriquecedor y edificante. Lo que es una verdad inamovible es que no hubiera podido llevarlo a cabo sin la ayuda de muchas personas. Mientras escribo estas notas me corroe la culpa de dejarme algún nombre fuera de la lista de agradecimientos, mil perdones de antemano.

En primer lugar, no podría ser de otra manera, GRACIAS (y no solo ahora, toda mi vida les agradeceré una y mil veces) Carles Miquel Magriñá y María José Ruzafa Mateo, mi familia catalana; sin su ayuda, ni el doctorado, ni ninguna de las oportunidades que he tenido en los últimos 8 años de mi vida hubieran sido posible. Han confiado en mí con los ojos cerrados y eso se los agradeceré siempre. Gracias por la acogida, por la paciencia, por abrirme las puertas de su casa y por hacerme sentir un miembro más de la familia.

A mi madre Sarah E. López García (Dra!!!, ya somos dos!), a mi tío y al resto de mi familia, aunque estemos lejos, cada palabra de ánimo, cada consejo han visto su fruto a lo largo de estos años, y esto es parte del resultado. ¡No podía haber tenido una familia mejor!

La primera persona que conocí al llegar a la UPC fue al Prof. Argimiro Arratia, Ph.D., a él también le agradezco sus consejos, su amistad y las conversaciones que hemos tenido a lo largo de estos años. Gracias por confiar en mí y en mi trabajo. También mi agradecimiento al Prof. Josep L. Larriba Pey, Ph.D., por aceptarme como parte de su grupo de investigación DAMA-UPC, por su cercanía y apoyo en todo momento. Gracias a todos los miembros de DAMA-UPC y Sparsity, a los actuales y a los que han pasado por ahí desde mi llegada en 2014.

Más de mil gracias al Prof. Jordi Valero, Ph.D., sus contribuciones a lo largo de esta tesis han sido invaluables. Gracias por compartir conmigo esa manera tan especial que tienes de ver las matemáticas.

Cuando a principios de 2016 cambié mi despacho al Dpto. de Estadística e Investigación Operativa del ETSEIB, tenía mis dudas, no lo voy a negar. Pero la acogida que me brindaron todos desde el primer momento hizo desaparecer cualquier tipo de incertidumbre. He pasado mucho tiempo en mi despacho "6-45", y he vivido muchos momentos importantes junto a ustedes. ¡Gracias por su amistad y cercanía siempre! Gracias Xavi Puig, Lesly, Lourdes, Lluis, Ignasi, Pau, Jan, Josep Ginebra, Pere, Xavier Tort, Sara, Dani, Martí y a mis compañeros de despacho Rafael Xampeny y Toni Font. Tampoco puedo dejar de mencionar a los becarios que han pasado por el despacho en estos años Carla, Ferran e Ignasi. No, no me he olvidado de ti, pero tú mereces un apartado especial, Josep A. Sánchez, gracias por tu amistad, por tu confianza, por cada palabra de ánimo. Además por todo lo que me has enseñado, eres un pilar fundamental en mi formación!

Gracias también a mis compañeros de AcuityTrading. Andrew Lane y Nikki Carpenter, agradezco todo el apoyo y la confianza de los últimos años.

También gracias a los dos revisores externos, Prof. Peter Zörning, Ph.D. y el Prof. Emilio Gómez-Déniz, Ph.D. sus comentarios y sugerencias han sido de gran utilidad en la elaboración de este documento.

El listado de personas importantes que han aparecido en mi vida en esta etapa es largo y a todos les agradezco que de una forma u otra me hayan hecho crecer, y aprender a ser mejor persona. Gracias David Ruiz, tu apoyo en cada momento ha sido fundamental, y eso que te ha tocado aguantar la cara menos agradable de este trabajo: la frustración y los momentos más estresantes. Gracias a los doctores, Rafael Mena y Sicong Zhuang, mis amigos que se me han adelantado en presentar la tesis, su compañía en esta aventura ha sido esencial. Lo dicho, el listado es largo, Alexis, Xavi, Marcia Díaz Gallardo, Ana Lourdes, Gemma, Sara Antoja, Mon, Jordi, Itziar, Borja, Oscar, Sandra, Carles, Eli. ¡Gracias a todos por su apoyo! Beñat no me he olvidado, mil gracias por tu ayuda con los diseños de posters, y la portada de la tesis. Y Jordi Aguilar, gracias por tu amistad, los cafés y tu ayuda en la realización del video del póster. También gracias a mis primos por el mundo que siempre están pendientes del avance de la tesis, ya nos reuniremos a celebrarlo. A mis amigos del otro lado del charco, Indira, Josué, Sarah-Liz, Yari, Fredys, Sheylah mil gracias por su cercanía, apoyo y amistad, que a pesar de la distancia y el tiempo siguen como el primer día.

Quiero terminar expresando mi eterno agradecimiento a mi directora de tesis Prof. Marta Pérez-Casany, Ph.D. No es un secreto, sin ti no hubiera sido posible, en estos años he aprendido mucho a tu lado. Nunca olvidaré la paciencia que has tenido y lo mucho que me has empujado para que esta tesis se convirtiera en una realidad. Gracias por todas las horas invertidas en mí y en este trabajo, gracias por tu dedicación sin tener en cuenta que fuera fin de semana o las 9:00 p.m. Gracias por la confianza que pusiste en mí, aún en los momentos en que la entropía era grande. Para ti siempre, mi respeto, admiración y agradecimiento.

Ariel Duarte-López
Barcelona, November 17, 2020

Acknowledgements

...Al final llegó el final...
Joaquin Sabina

It seems unreal but it is a reality that will materialize soon enough. This is the result of years of hard work and endeavor. It has been an exciting process, sometimes fun, and on other occasions not so much... but always enriching and fulfilling. One thing for sure, is that I could not have carried it out without the help of many people. As I am writing these notes, a sense of guilt creeps up on me in case I do not acknowledge someone's name, a thousand apologies in advance.

First of all, it could not be otherwise, THANK YOU (and not only now, but for the rest of my life I will be thanking you) Carlos Miquel Magriñá and María José Ruzafa Mateo, my Catalan family; without your help, neither my doctorate nor any of the opportunities that I have had in the last 8 years of my life would have been possible. Both of you have taken a leap of faith and trusted me with all your heart and I will always be grateful for that. Thank you for the welcome, your patience, for opening the doors of your house to me, and for making me feel like a member of the family.

I would like to thank my mother Sarah E. López García (Dra.!!!, now there will be two Doctors in the family!), my uncle and the rest of my family, although we are half a world apart, each word of encouragement, each piece of advice has paid off throughout these years, culminating in this achievement. I could not have had a better family!

The first person that I met when I arrived at the UPC was Prof. Argimiro Arratia, Ph.D., I also wish to thank him for offering me his advice and friendship; and the conversations we have had over the years. Thankyou for believing in me and my work. I would like to also acknowledge Prof. Josep L. Larriba Pey, Ph.D., for accepting me as part of your research group DAMA-UPC, for your closeness and support at all times. I would also like to thank my colleagues of DAMA-UPC and Sparsity, both current ones and to those who have passed through there since I started in 2014.

More than a thousand thanks to Prof. Jordi Valero, Ph.D., his contributions throughout this thesis have been invaluable. Thank you for sharing with me your unique mathematical viewpoints.

At the beginning of 2016, when I changed my office to the Department of Statistics and Operations Research of the ETSEIB, I will admit that I was nervous. But the welcome that everyone gave me from the get-go made all my doubts disappear. I have spent a lot of time in my office “6-45”, and I have shared many important moments of my life alongside all of you. Thank you for your friendship and companionship! Thanks, Xavi Puig, Lesly, Lourdes, Lluis, Ignasi, Pau, Jan, Josep Ginebra, Pere, Xavier Tort, Sara, Dani, Martí, and to my office colleagues Rafael Xampeny y Toni Font. Nor can I fail to mention some of the interns who have worked in my office in these years Carla, Ferran e Ignasi. No, I have not forgotten about you, but you deserve a special mention, Josep A. Sánchez, thank you for your friendship, for your trust, for every word of encouragement. Also for everything you have taught me, you are a indispensable figure in my training!

I would like to thank my colleagues at AcuityTrading. Andrew Lane and Nikki Carpenter, I appreciate all your support and trust in recent years.

Also thanks to the external reviewers that have made the final revision of my thesis. Prof. Peter Zörning, Ph.D. and Prof. Emilio Gómez-Déniz, Ph.D. your comments and suggestions have been very helpful in preparing this document.

The list of outstanding people who have appeared in this phase of my life is long and I am grateful to all of them. In one way or another they have made me grow and made me a better person. I would like to thank David Ruiz, your support at all times has been essential, and most of the time you had to put up with the less pleasant side of this job: frustration and the most stressful moments. Thanks to my other Ph.D. colleagues, Rafael Mena y Sicong Zhuang, my friends who earlier than me have presented their thesis, your company in this adventure has been crucial. As I have said earlier, the list is long, Alexis, Xavi, Marcia Díaz Gallardo, Ana Lourdes, Gemma, Sara Antoja, Mon, Jordi, Itziar, Borja, Oscar, Sandra, Carles, Eli, thanks to everyone for your support! Beñat I have not forgotten you, thank you for your help with the design of the poster and cover of the thesis. And Jordi Aguilar, thank you for your friendship, the coffee breaks, and your help in making the video for the poster. To my cousins around the world who are always concerned about the progress of the thesis, we will celebrate together. To my friends on the other side of the world, Indira, Josué, Sarah-Liz, Yari, Fredys, Sheylah, a thousand thanks for their confidence, help, and friendship, which despite the distance and time continue to flourish.

I would like to end by expressing my eternal gratitude to my supervisor Prof. Marta Pérez-Casany, Ph.D. It is no secret that without you none of this would have been possible.

In these years I have learned plenty alongside you. I will never forget the patience you have had and how much you have encouraged me to make this thesis come true. Thank you for all the hours you have dedicated to me and this research. I appreciate your dedication regardless of whether it was the weekend or 9:00 pm. Thank you for the trust you placed in me, even at times when it all seemed to be falling apart. You will always have my respect, admiration, and gratitude.

Ariel Duarte-López
Barcelona, November 17, 2020

Abstract

The Zipf distribution, also known as discrete Pareto distribution, attracts considerable attention because it helps describe skewed data from many natural as well as man-made systems. Under the Zipf distribution, the frequency of a given value is a power function of its size. Consequently, when plotting the frequencies versus the size in log-log scale for data following this distribution, one obtains a straight line. Nevertheless, for many data sets the linearity is only observed in the tail and when this happens, the Zipf is only adjusted for values larger than a given threshold. This procedure implies a loss of information, and unless one is only interested in the tail of the distribution, the need to have access to more flexible alternatives distributions is evidenced.

The work conducted in this thesis revolves around four bi-parametric extensions of the Zipf distribution. The first two belong to the class of Random Stopped Extreme distributions. The third extension is the result of applying the concept of Poisson-Stopped-Sum to the Zipf distribution and, the last one, is obtained by including an additional parameter to the probability generating function of the Zipf. An interesting characteristic of three of the models presented is that they allow for a parameter interpretation that gives some insights about the mechanism that generates the data. In order to analyze the performance of these models, we have fitted the degree sequences of real networks from different areas as: social networks, protein interaction networks or collaboration networks. The fits obtained have been compared with those obtained with other bi-parametric models such as: the Zipf-Mandelbrot, the discrete Weibull or the negative binomial. To facilitate the use of the models presented, they have been implemented in the *zipfextR* package available in the Comprehensive R Archive Network.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AIC	Akaïke Information Criterion
AtCKN	*Arabidopsis thaliana* comprehensive knowledge network
BA	Barabási-Albert
CDF	Cumulative density function
CRAN	Comprehensive R Archive Network
DGX	Discrete Gaussian exponential
DW	Discrete Weibull
ER	Erdös-Rényi
KS	Kolmogorov-Smirnov
LR	Likelihood ratio
LRT	Likelihood ratio test
MLE	Maximum likelihood estimation
MOEZipf	Marshall-Olkin extended Zipf
MO	Marshal-Olkin transformation
MP	Mixed Poisson
MZTP	Mixed zero-truncated Poisson
NB	Negative binomial

PGF	Probability generating function
PL	Power law
PMF	Probability mass function
PPI	Protein Protein interaction
PSS	Poisson stopped sum
RGF	Random group formation
RSED	Randon stopped extreme distribution
RSEZipf	Randon stopped extreme Zipf distribution
RVF	Regularly varying function
SF	Survival function
Zipf-PE	Zipf-Poisson extreme
Zipf-PSS	Zipf-Poisson stopped sum
ZM	Zipf Mandelbrot
ZTMP	Zero-truncated mixed Poisson

Introduction

Discrete Power Law (PL) distributions are those families of distributions such that the probabilities are inversely proportional to a positive power of the value itself. The history of PLs can be traced back to the late XIX century, where the economist Vilfredo Pareto developed the Pareto distribution [Mitzenmacher, 2004; Vilfredo, 1896], this is why it is also known as discrete Pareto distribution. The PL distribution defined in a finite support is well known by its associated "Pareto rule" or "80-20 rule" that says that 80% of the abundance of a country is held by 20% of the population, or that 80% of the work is done in 20% of the workday. Later on, Auerbach [Auerbach, 1913], Estoup [Estoup, 1916] and Zipf [Zipf, 1935, 1949] used particular cases of PL distributions to model variables such as: the population in cities, the frequency of words in a text or the length of intervals between repetitions in a Mozart's *concerto*. Even though part of the work of Auerbach and Estoup overlaps in some aspects, with that by Zipf, it is the last one who popularizes the "Zipf's law" as the law of "least-effort". This is because by analyzing the frequency of words in texts, he observed that usually there are many words that appear very few times and, at the same time, very few words that appear a very large (sometimes huge) number of times. This results in highly skewed data sets.

This "least-effort" phenomenon also appears in Network Science, where the Zipf distribution is usually assumed to be the distribution of the number of connections of a node in a network. This is because in a real network, there usually exist many nodes exhibiting very few connections, and very few nodes highly connected, the last ones denoted as *hubs*. A network with this characteristic is usually known as an "scale-free network", see for instance Csermely [2009] or Barabási and Pósfai [2016].

Network Science, in particular, has been the core motivation of this thesis. Its initial part was performed in collaboration with the researchers Josep Larriba-Pey and Arnau Prat-Pérez from the DAMA-UPC research group [1]. They pointed out that the PL assumption for the degree sequence of a real network (the sequence of the number of connections of the nodes) was not as general as it was thought. They also pointed out that generating graphs similar to

[1] https://www.dama.upc.edu/

those observed in reality was becoming very important, to be able to synthetically generate graphs similar to those observed in reality, and that one first step to do so, was to be able to generate data from distributions more flexible than the Zipf distribution.

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 an small percentage of those are what they denote as “pure scale free”.

The main reason for the real degree sequences to deviate from PL behavior is that, when the probabilities of a PL 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. The work by Newman [2005] proposes to face this problem by first determining a cut-off value, and then by fitting a PL distribution on the tail. The cut-off value determination was quite arbitrary since it requires a visual inspection. Later on, Clauset et al. [2009] proposed a methodology to determine a good cut-off value, based on imposing the distance between the theoretical and the empirical distribution to minimum. This approximation has been applied in many recent research papers such as, Cho et al. [2011] and Gomez-Lievano et al. [2012]. Nevertheless, this approximation requires the suppression of some observed values that may contain important information. Thus, unless the person is only interested in modeling the tail of the distribution, alternatives to the Zipf distribution that allow to model the data in its entire space are required, and this is the main objective of this thesis.

In this work, the Zipf distribution is the discrete PL family of distributions with support the integer values strictly larger or equal than one, and four extensions of this family of distributions are proposed. Two of them allow for top-concavity as well as top-convexity while maintaining the linearity in the tail. The other two only adapt top-concavity which is the most common situation. In three of the four models proposed, the parameter interpretation provides information about the mechanism that generates the data. For all the models, a mechanism that randomly generates data from the distributions is established, and its performance is tested by means of the Kolmogorov-Smirnov (KL) test. In order to promote the use of the models presented, the *zipfextR* package [Duarte-López and Pérez-Casany, 2020] has been created, and it is available in the Comprehensive R Archive Network (CRAN). The suitability of the models presented to fit real-world data has been tested, by modeling degree sequences of networks that come from many different areas.

This thesis is organized as follows. In Chapter 1 the Zipf distribution is introduced as well as its main characteristics and limitations. Some alternatives that appear in the scientific literature are also introduced, because they are used to compare the performance of the models presented in this thesis. We have obtained two new theoretical results that allow to see the Zipf distribution as a mixture distribution, and those are also included in this Chapter.

Chapter 2 contains the first two Zipf's generalizations: the Marshall-Olkin extended Zipf distribution (MOEZipf) and the Zipf-Poisson Extreme distribution (Zipf-PE). The work done with respect to the first one has been the continuation of what was done in the master thesis of Aina Casellas (see Casellas [2013] and Pérez-Casany and Casellas [2013]). This generalization is the result of applying the Marshall Olkin transformation (MO) to the Zipf distribution. The MO has been largely used to generalize continuous distributions see, for instance, Yeh [2004], Ghitany et al. [2005], Ghitany et al. [2007] or Jose [2011]. With respect to discrete distributions, the paper by Gómez-Déniz [2010] applies it to generalize the geometric distribution. Given that, the result of applying the MO transformation to a given random probability distribution is a random stopped extreme distribution with a geometrical stopping distribution, this drives us to investigate the result of changing the stopping distribution and to consider a Poisson instead of a geometrical. This gave place to the Zipf-PE distribution. The results obtained with these two extensions appear in the paper by Duarte-López et al. [2020a].

Chapter 3 generalizes the Zipf by means of the concept of Poisson-stopped-sum (PSS) [Johnson et al., 2005], giving place to the Zipf-Poisson-stopped sum distribution (Zipf-PSS). The PSS mechanism has been widely used for generalizing families of distributions, and some examples are the negative binomial, the Generalized-Inverse-Gaussian-Poisson or the Pòlya-Aeppli. Of all the models presented, the Zipf-PSS is the only one that has the zero value in its domain and thus, in the particular case of Network Analysis, allows to model networks with isolated nodes. The main results related to the Zipf-PSS are published in the paper by Duarte-López et al. [2020b].

Chapter 4 is devoted to the Zipf-Polylog extension, which is obtained by adding an extra parameter to the Zipf probability generating function. After defining it, we saw that it had already been proposed in the literature with the name of *PL distribution with exponential cutoff*, see Clauset et al. [2009], or *hybrid geometric/power model*, see Visser [2013]. Nevertheless, apart from using it to fit data, there is not a profound study of its properties. For that reason, in this chapter we have stated and proved some of its main characteristics. In particular, this is the only extension that is a two-parameter exponential family and that has moments of any order regardless of the parameter values. The results related to this family appear in Valero et al. [2020].

Chapter 5 is devoted to the analysis of the degree sequence of networks coming from different domains. The fits obtained with the best of our models are compared, whenever possible, with those obtained with the Zipf distribution by means of the likelihood ratio test (LRT). By means of the Akaïke's Information Criterion (AIC) our models are also compared with the other bi-parametric ones mentioned in Chapter 1.

We finish the document with the conclusions and the future work.

Appendixes A and B are devoted, respectively, to perform a comparative study of the models presented, and to introduce the *zipfextR* package [Duarte-López and Pérez-Casany, 2020]. Both tools are done to facilitate the use of our models to practitioners. Appendix C contains the plots of the stopping distribution considered in Table 2.2 of Chapter 2. Finally, Appendix D contains the implementation of the KS test, and the source code necessary to analyze the data sets presented in Chapter 5.

This thesis has been funded by grants TIN2017-89244-R from MINECO (Ministerio de Economía, Industria y Competitividad), Spain; the recognition 2017SGR-856 (MACDA) and 2018_FI_B2_00064 from AGAUR (Generalitat de Catalunya), Spain. It has also been supported by projects J-01914 (Use of Graph Databases in Industrial Environments) and EU FP7 (Linked Data Benchmark Council). Borses Ferran Sunyer i Balaguer, from l'Institut d'Estudis Catalans, funded the one month stay of Ariel Duarte-López at the National Institute of Biology in Slovenia in 2018.

Chapter 1

The Zipf distribution

This chapter introduces the Zipf distribution [Zipf, 1949], which forms the basis of the work developed throughout this thesis. We begin by defining the Zipf distribution and analyzing its main properties. One of the most surprising characteristics of this distribution is its emergence in dissimilar and unconnected fields, which makes some authors consider it to be a pervasive distribution [Adamic, 2011, and references therein]. We present two of the most important Zipf generative models that appear in the literature. In addition, we discuss the main properties as well as the main limitations of this family of distributions. These limitations have motivated this thesis, and before introducing our work we would like to comment on several alternatives already introduced in the literature.

The chapter concludes by stating the first contribution of this thesis: we prove that the Zipf model is a mixture of geometric distributions as well as a mixture of zero-truncated Poisson distributions (MZTP). This result relates the Zipf to two classical distributions and points out that it arises when the parameter of the geometrical and Poisson distributions follow a given probability distribution. What is more, based on the results that appear in Valero et al. [2010], we prove that the Zipf distribution is not the zero-truncation of a mixed Poisson distribution (ZTMP).

1.1 Definitions, properties and applications

Before introducing the Zipf distribution we present a brief overview of a more general family of distributions, the PL, where the Zipf family is included.

PL distributions are very common in the literature and they can emerge in a large variety of unrelated scenarios. Maybe one of the particular characteristics of these family of distributions is that they tend to concentrate the probabilities in the small values in a large range of its parameter space. The PLs can be defined as continuous or discrete distribution families. In a discrete PL distribution the probabilities change inversely as a power of the values. One of the most used continuous PL is the Pareto Distribution [Mitzenmacher, 2004; Vilfredo, 1896]. In his work Pareto states that the distribution of the incomes follow a PL, and define what it is known as the "80/20 rule", which says that, the 80% of the incomes is earned by the 20% of the population. The probability density function of this distribution is defined as follows:

$$P(X=x) = \begin{cases} \frac{\alpha-1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha} & \forall x \geq x_{min} \\ 0 & \text{otherwise}, \end{cases} \tag{1.1.1}$$

where $\alpha > 1$, $x_{min} > 0$ and its support is the values such that $x \in [x_{min}, +\infty)$.

The discrete version of the Pareto distribution is defined as:

$$P(X=x) = \begin{cases} \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})} & \forall x \geq x_{min} \\ 0 & \text{otherwise}, \end{cases} \tag{1.1.2}$$

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 (1.1.2) the Zipf distribution is obtained. This is why the Zipf distribution is a particular case of a discrete PL distribution.

Thus, it is said that a random variable (r.v.) X follows a Zipf distribution with parameter $\alpha > 1$ if, and only if, its PMF is equal to:

$$P(X=x) = \frac{x^{-\alpha}}{\zeta(\alpha)}, \; x = 1, 2, ..., \; \alpha > 1, \tag{1.1.3}$$

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 also known as the *Zeta* distribution and, as proved in Zörnig and Altmann [1995], it is a particular case of the Lerch distribution. It 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 PL 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. Assuming that n is the counting of objects which belong to a certain group, r stands for their rank and $n(r)$ is the number of objects of a given group with rank r, the formulation of the Zipf law establishes that:

$$n(r) \propto r^{-\gamma},$$

where the symbol $\propto$ denotes proportionality.

For frequencies of frequency data, one understands data that are frequency tables of counts. For instance, assuming that the number of followers that each Instagram account has is known, 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 category.

The data sets considered in Chapter 5 of this thesis are frequencies of frequency data. Notwithstanding, they could also be analyzed in terms of ranks.

In fact, the works by Adamic and Huberman [2002] and Corral et al. [2019] relates the exponent of the two data formulations. If α is the exponent when the data are seen as frequencies of frequency and γ is the exponent when they are seen as ranked data, it is verified that:

$$\alpha = 1 + \frac{1}{\gamma}.$$

The reader is encouraged to review the last cited work for a detailed analysis on the relationship between the two approaches.

In what follows we point out the main characteristics of the Zipf distribution.

By taking logarithm in both sides of (1.1.3) 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 1.1 shows the probabilities of the Zipf for different values of the α parameter. On the left-hand side the probabilities are shown in standard scale, and on the right-hand side in log-log scale. Observe that when the α parameter increases, the probabilities concentrates at the low values.

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

$$\overline{F}_{\alpha}(x) = P(X > x) = \frac{1}{\zeta(\alpha)} \sum_{i=x+1}^{+\infty} i^{-\alpha} = \frac{\zeta(\alpha, x+1)}{\zeta(\alpha)},\ \alpha > 1, \qquad (1.1.4)$$

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

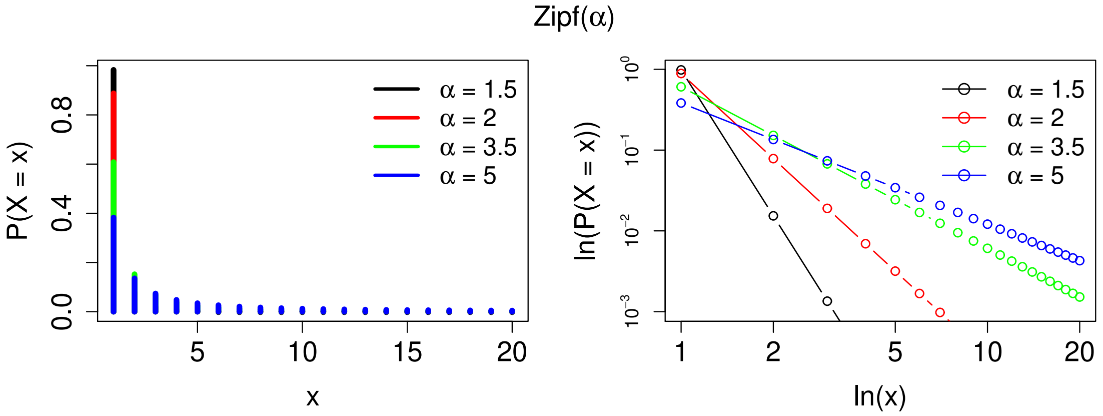

Fig. 1.1 PMFs of the Zipf distribution for $\alpha = 1.5, 2, 3.5$ and 5. On the left-hand side: normal scale. On the right-hand side: log-log scale.

$$F_\alpha(x) = 1 - \frac{\zeta(\alpha, x+1)}{\zeta(\alpha)} = \frac{\zeta(\alpha) - \zeta(\alpha, x+1)}{\zeta(\alpha)}, \alpha > 1. \tag{1.1.5}$$

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

$$E[X^k] = \sum_{x=1}^{+\infty} \frac{x^k x^{-\alpha}}{\zeta(\alpha)} = \frac{\zeta(\alpha - k)}{\zeta(\alpha)}, \tag{1.1.6}$$

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:

$$E[X] = \frac{\zeta(\alpha - 1)}{\zeta(\alpha)}, \alpha > 2. \tag{1.1.7}$$

Since the variance depends on the second moment of the distribution, it is finite if, and only if, $\alpha > 3$ and when it exists, it is equal to:

$$Var[X] = E[X^2] - (E[X])^2 = \frac{\zeta(\alpha - 2)\,\zeta(\alpha) - \zeta(\alpha - 1)^2}{\zeta(\alpha)^2}, \alpha > 3. \tag{1.1.8}$$

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

$$\mathscr{L}(\alpha; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{x_i^{-\alpha}}{\zeta(\alpha)} = \frac{(\prod_{i=1}^{n} x_i)^{-\alpha}}{(\zeta(\alpha))^n}, \tag{1.1.9}$$

and its logarithm is equal to:

$$\ell(\alpha;x_1,x_2,\ldots,x_n) = -\alpha\sum_{i=1}^{n}\log(x_i) - n\log(\zeta(\alpha)). \tag{1.1.10}$$

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

$$-\sum_{i=1}^{n}\log(x_i) - n\frac{\zeta'(\alpha)}{\zeta(\alpha)} = 0, \tag{1.1.11}$$

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.

In Visser [2013], it is proved that the Zipf distribution is the discrete uni-parametric distribution with support on the strictly positive integer values that has maximum Shanon entropy, for a fixed value of $\overline{\log(x)}$. The probability distribution with maximum entropy is assumed to be the best distribution since it is the one maximizing the amount of information or the uncertainty about what is unknown [Floudas, 2009, p. 1779]. In Chapter 4 is introduced the Zipf-Polylog distribution which is a two-parameter extension of the Zipf with maximum Shannon entropy once $(\bar{x},\overline{\log(x)})$ are fixed.

The probability generating function (PGF) of a Zipf distributed r.v., is equal to:

$$G_X(z) = E(z^X) = \sum_{x=1}^{+\infty}\frac{z^x x^{-\alpha}}{\zeta(\alpha)} = \frac{Li_\alpha(z)}{Li_\alpha(1)},\ |z| < 1 \text{ and } \alpha > 1, \tag{1.1.12}$$

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

$$Li_\alpha(z) = \sum_{x=1}^{+\infty}\frac{z^x}{x^\alpha}. \tag{1.1.13}$$

The *Li* function of order α is defined for any complex number α and any complex number z, such that $|z| < 1$. Nevertheless, using analytic prolongation, the *Li* function is defined throughout the whole complex plane. For $Re(\alpha) > 0$, and all z except for z that are real and larger or equal than one, the polylogarithm function may be expressed in terms of the integral of Bose-Einstein distribution as follows:

$$Li_{\alpha}(z) = \frac{1}{\Gamma(\alpha)} \int_{0}^{\infty} \frac{t^{\alpha-1}}{\frac{e^{t}}{z}-1} dt. \qquad (1.1.14)$$

The last equality can be checked by computing the Taylor expansion of the integrand and integrating termwise Lee [1997]. Concerning to this thesis, in the instances where the *Li* function is required, 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.

There are countless examples where researchers have argued the suitability of the Zipf distribution for modeling the data associated with their research. After the work by Zipf [1949] which use it for fitting the frequency of the words in texts, this distribution has keep being quite popular in linguistics see, for instance, Ferrer-i Cancho and Vitevitch [2018]. Several examples of applications in other areas appear in the work by Newman [2005], who uses a general PL distribution to fit the tail of data sets related to: the number of copies of books sold in the US from 1895 to 1965; the populations of US cities; or earthquake magnitudes among others. Apart from that, others researchers have focus on the application of this distributions in studies related to the human behavior. For example, the analysis performed by Gomez-Lievano et al. [2012] suggests that the total number of homicides in Colombia, Mexico and Brazil can be described by a PL, while Krumme et al. [2013] use this distribution for predicting consumer visitation patterns. The last work evidences that, independently of shopper preferences, the Zipf distribution can be used to describe how frequently a client visits a store. More recently, Ectors et al. [2018] have shown that the Zipf distribution also emerges in the frequency of occurrence of daily people's 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 a large set of metrics based of the Zipf's law are used as input in a music classification problem. In general, these metrics are created to measure the proportion or distribution of several music parameters, such as: pitch, duration, melody intervals, among others. A total of 40 metrics are generated and used to feed a neural network aimed to perform various classification tasks like: author attribution, style identification and pleasantness prediction.

Also, it has been used for automatic detection of regions of interest in digital images [Caron et al., 2007]. To that aim, the Zipf's law and the Inverse Zipf's law (the last one out of the scope of this thesis) are used to model the frequency of appearance of patterns contained in images. These type of models allows to describe the structural complexity of images textures. At the same time, this measure of complexity suggests a salient region in the image.

Additional examples from computer science are presented in the review by Mahanti et al. [2013], where the authors use a PL distribution to adjust several Internet measures such as YouTube video popularity and web access, among others. A more recent work by Wang et al. [2017] on user authentication in cybersecurity shows that the “vulnerable portion of user-chosen passwords" can also be adjusted by this distribution. Also, Malone and Maher [2012] analyze its suitability for describing the frequency of chosen passwords.

In academia, the distribution has been used in the work by Ausloos et al. [2016] to assess the quality of peer reviewers process. In particular the data analyzed came from peer review reports of the Journal of the Serbian Chemical Society. Every report is analyzed considering the quantity, variation and distribution of the words and compared with respect to the whole set of reports. The authors state that the Zipf exponent seem to indicate a specific reviewer.

Finally, in Network Analysis, the Zipf distribution is considered a reasonable distribution for fitting the degree sequence of real network. For example, the work by Adamic and Huberman [2002] has shown that it provides the best fit to the connections of the Internet routers as well as the number of links on the Internet sites. For a PL application to model the degree sequence of a network representation of a stock market, see Boginski et al. [2005]. Usually, in the area of Network Science, researchers assume that the degree sequence of real networks are PL distributed. The networks whose degree distribution follows a PL are also known as *scale-free networks* [Barabási and Pósfai, 2016]. Motivated by its applicability in this field and for the results obtained in our first work [Duarte-López et al., 2015], all the examples in Chapter 5 use this type of data for illustrating the suitability of the proposed models.

We can assume the Zipf distribution as a plausible candidate in scenarios where small observations take place with a high frequency and large observations are less frequent. Retaking the example about the links on Internet sites, is evinced the existence of a high number of sites having a few number of links (i.e. personal or small business websites) and a few of sites highly linked (i.e. Google, Microsoft, Amazon, etc.).

The work by Zörnig [2015] studies the probability distribution of a rank-frequencies of random sequences of numbers. The authors define a *sufficient fitting criterion* that states a condition under which the right-truncated Zipf distribution properly fits the rank-frequency vector. Moreover, they prove that when the sequence is generated with the same probability for all the numbers, the criterion is fulfilled with a high probability. On the contrary, if the selecting probabilities differ a lot, the probability of verifying the criterion becomes smaller.

In particular, when generating random texts assuming that the symbols are independent, and with equal probabilities, it results in a Zipf distribution for the rank-frequency vector. In previous work by Günther et al. [1996] it is studied the *evolutionary model* presented by Schapiro [1994], which has the form of a non-stationary Markov branching process, and it is used to model evolutionary complex systems. The authors prove that under general transition probabilities the ranking procedure leads to a Zipf's law.

Next section introduces two popular mechanisms that explain the emergence of the Zipf's law; one is related to the field of Network Analysis and the other one is a general model that can be applied to any knowledge domain. Thus, it allows to understand the pervasiveness of the Zipf distribution.

1.2 Genesis

As stated above, Zipf's law emerges in many unconnected areas, which is why this section introduces two mechanisms that explain the emergence of this distribution. Even though there are a wide variety of methods that try to explain how Zipf's law arises, they are usually linked to the domain in which the phenomenon takes place [Newman, 2005]. Throughout this section, we present two mechanisms that induce Zipf's law. The first one is the Barabási-Albert model [Barabási and Albert, 1999], which is quite popular in the field of Network Analysis. The second one is the Random Group Formation (RGF) [Baek et al., 2011], which proposes a general framework for explaining the origin of Zipf's law, independently of the research domain. An introduction to these two mechanisms can be found in the work by Adamic [2011].

The *Barabási-Albert model* [Barabási and Albert, 1999] appeared in the late 90's as a mechanism for generating random *scale-free networks*. As Barabási and Pósfai [2016] state in their book, the idea of scale-free captures the lack of an internal scale, which is caused by the co-existence of nodes with a large difference in their degrees. That is to say, the degree of a randomly chosen node in a scale-free network is equal to $k = \overline{k} + \sigma_k$, with $\overline{k}$ being the average degree and σ_k the standard deviation. Assuming a power law exponent α that is strictly smaller than 3 implies that the second moment of the distribution diverges (see 1.1.8). Thus, one can expect the largest fluctuation around the average. As a consequence, the expected degree of the node could be small or arbitrarily large. See Barabási and Bonabeau [2003] for a review on the characteristics of the scale-free networks.
Before the creation of the Barabási-Albert (BA) model, the most applied methodology for random network generation was the Erdös-Rényi (ER) model [Erdős and Rényi, 1960]. Formally, the model is defined as $G_{N,p}$, where N is the number of nodes of the network and p

is the probability of creating an edge between two nodes. The ER random generation process randomly selects a pair of nodes, and an edge is created between them with probability p. In other words, each possible edge is included in the graph with the same probability. Increasing (decreasing) the probability implies denser (sparser) networks. The degree distribution of the networks constructed by this methodology are expected to be Poisson distributed, which is contrary to what one observes in real networks.

One of the advantages of the BA model is its ability to generate networks with a degree sequence that is distributed similarly to most of those observed in the reality. This methodology takes into account two properties of the networks that are not considered in the ER model: the *growth of the network* and the *preferential attachment* principle.

The growth of the network appears naturally in various scenarios, such as the addition of new followers on Instagram, Facebook or other social platforms. Note that it is not restricted only to social networks, as it also appears in citations, communications networks, and generally, in any network that changes with time.
On the other hand, the preferential attachment property considers that the probability of a new node connecting to existing ones is not constant, meaning that the probability is higher when connecting to a node that already has large number of connections. Consequently, it evolves with time. Preferential attachment is a reasonable assumption because highly connected people have a higher probability of meeting new people in a social event, just as the oldest and most famous companies are more likely to sign new contracts than are the newest ones. In the authors' words, "a rich-get-richer" effect can easily be detected in real networks.

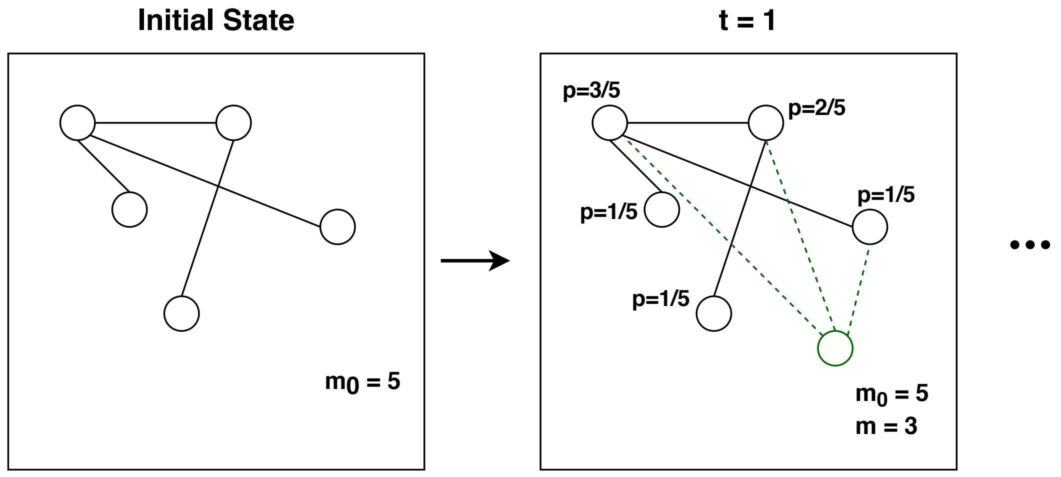

Fig. 1.2 Illustrative example of the firsts steps of the BA model.

The BA model starts by assuming that the graph initially has a small number of nodes, denoted by m_0, and a small number of edges equal to m_0 that are randomly generated with the same probability. At each step, a new node is added to the graph jointly with m edges, being $m \leq m_0$. If i is a node with degree k_i, and we denote by $p(k_i)$ the probability that the new node is connected to a node of degree k_i, then the BA random graph generator assumes that:

$$P(k_i) = \frac{k_i}{\sum_j k_j},$$

which gives a greater probability of connecting to the nodes that already have more connections. At each step, the procedure stops when m connections are obtained. After t steps, the total number of nodes will be equal to $m_0 + t$, and the total number of edges will be equal to $m_0 + m * t$. The authors prove that, independently of the initial m_0 value, the system converges to a stationary state that results in a graph whose degree distribution is a Zipf distribution with parameter $\alpha = 2.9$. Figure 1.2 illustrates the initial steps of the BA methodology.

The second methodology presented here is the random group formation (RGF) [Baek et al., 2011]. By means of this procedure, the authors look for a global explanation of the PL's distribution ubiquity. The methodology proposed is based on placing M numbered objects in N boxes and, thus, it is not tied to any particular domain.

It is said that a box has size k if it has exactly k slots, and k_{max} denotes the maximum box size. The model assumes $k_1 + k_2 + \cdots + k_N = M$ (see Figure 1.3). It also assumes that the probability of finding a given object in a particular place is constant and independent of the place, i.e., it is equal to $p = 1/M$. Denoting by $N(k)$ the number of boxes with size k, the probability of placing an object in a box of size equal to k is equal to $p(k) = N(k)/N$. The information for locating an object with no additional knowledge is $I_{total} = \ln(M)$ nats. The objective is to assign the objects to boxes with-in the *minimum information cost*, which is defined as: the additional information which on average is required for locating a ball if you know the box size.

The output of this methodology is a PL distribution with an exponential cut-off (see Chapter 4). Under certain restrictions placed on M, N and k_{max}, the distribution becomes a pure PL.

Several years before the RGF was introduced, Hill and Woodroofe [1975] presented an approach that proved the Zipf distribution is the limit on the proportion of classes that have exactly x units in a classification problem of N units in M categories, specifically when the number of units to be classified tends to infinity.

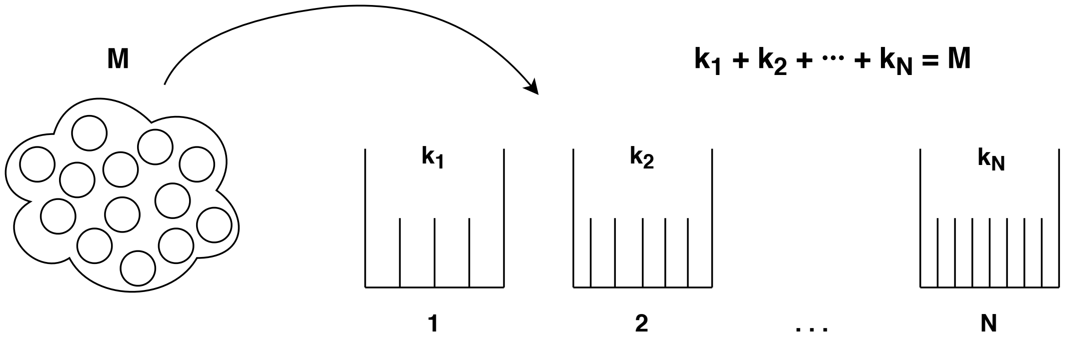

Fig. 1.3 RGF schema.

The next section states the main limitation of the Zipf distribution when used to fit real data. In addition, we include some of the popular alternatives found in the literature on this models.

1.3 Limitations and existing alternatives

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. 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 1.4 shows several plots in log-log scale of the degree distributions of real networks. These plots illustrate a clear deviation from pure PL 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 Chapter 5, these three data sets are fitted by means of the Zipf extensions proposed in this thesis.

According to McKelvey et al. [2018], in just a few scenarios the PL 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 PL.

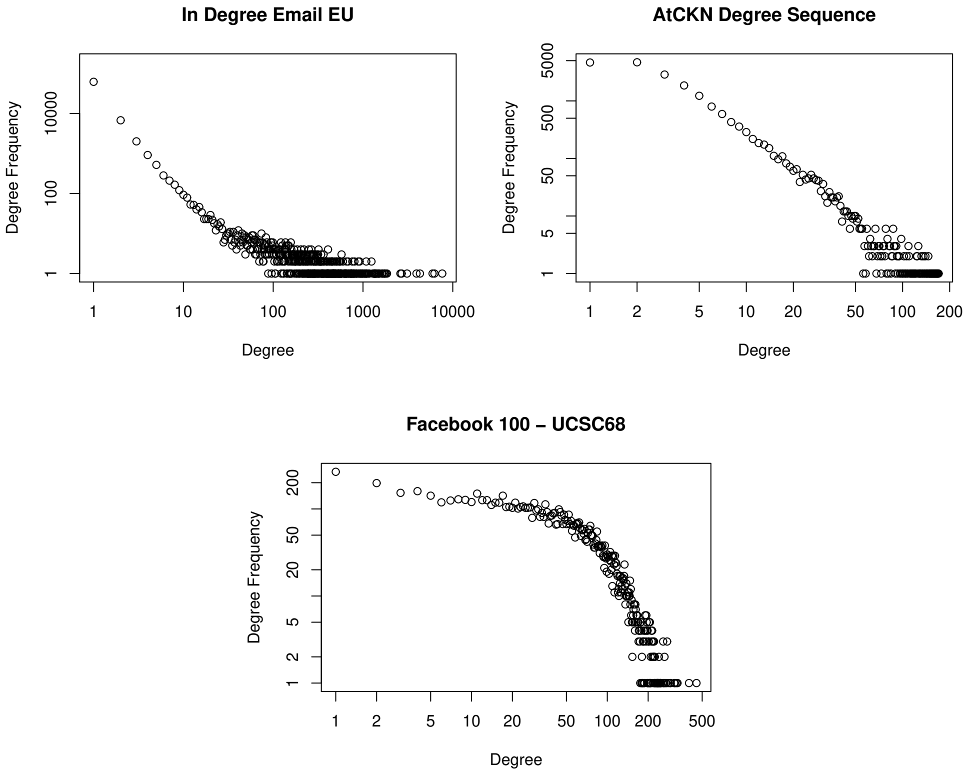

Fig. 1.4 Examples of degree sequences of real networks plotted in log-log scale. On the upper left-hand side is the in-degree sequence of a communication network representing emails exchanged in a European institution. The upper right-hand side shows the degree sequence of the *Arabidopsis thaliana* comprehensive knowledge network. Finally, the example at the bottom corresponds to the Facebook network of the University of California, Santa Cruz in 2005.

This implies that fitting a PL to many data sets requires the selection of a plausible cut-off point, x_{min}.

As mentioned previously, the work by Newman [2005] shows twelve examples in which the corresponding data sets follow a PL in at least a part of their range (i.e., word frequency, copies of books sold, magnitude of earthquakes, etc.). Notwithstanding, the authors believes that in some of the analyzed cases the PL behavior remains unconfirmed. Since the deviation from the PL occurs in values below a certain x_{min} point, the author proposes truncating the data for those values below x_{min}. Consequently, it is assumed that $P(X = x) = 0$ for $x < x_{min}$. The difficulty of this approach is the selection of an appropriate x_{min} value. Unfortunately, an approach for selecting the cut-off point is missing in this work, and the value is obtained through a visual analysis, which may bias the obtained results.

Note that the main consequence of the visual selection lies in selecting a low value of x_{min}, which obtains a bias estimation of $\hat{\alpha}$ because the procedure will try to fit non-power-law data. In contrast, choosing a high x_{min} value discards valuable information.

After choosing the cut-off point, the exponent of the distribution can be calculated by using the following expression:

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1}, \tag{1.3.1}$$

where n is the sample size, and x_{min} is the minimum value of x from which the PL holds. For a complete derivation of the equation, see Appendix B in Newman [2005].

Some years later, Clauset et al. [2009] extended the previous methodology and, instead of seeking the x_{min} value through a visual inspection, the authors proposed: "to choose the value of $\hat{x}_{min}$ that makes the empirical probability distribution and the best-fit PL model as similar as possible above $\hat{x}_{min}$". Basically, the authors are seeking for an x_{min} that minimizes the distance between the CDF of the data that includes observations larger than or equal to x_{min} and the CDF of the PL that provides the best fit for the data in the range $x > x_{min}$.

The KS statistic is the measure used for quantifying the distance between the two distributions, and (1.3.1) is used for estimating the $\hat{\alpha}$ parameter. First, the authors compute the KS_0 statistic for the initial sample. Then, they simulate M samples of size n from a PL with parameters $\hat{\alpha}$ and $\hat{x_{min}}$. For each synthetic generated sample, the KS statistic is computed and denoted by KS_i, $i = 1 \div M$. The p-value is computed as $\#KS_i/M \geq KS_0$. The hypothesis test is:

$$H_0 : PL$$

$$H_1 : \neg PL,$$

and it is performed by rejecting H_0 if p-value ≤ 0.1.
Nowadays, Clauset's methodology is still used for fitting data that shows a PL behavior in the tail. Recently, the work by Bhattacharya et al. [2020] has proved the consistency of this method. Notwithstanding, setting a cut-off value results in important information not being considered, which may bias the analysis. The goal of this thesis is to develop new alternative distributions that lead to more accurate results without suffering a lack of information.

In what follows, we describe several probability distributions that have proved to be useful alternatives to the Zipf family. Some of them are extensions of the Zipf distribution or they contain this distribution for particular values of their parameters. This is the case for: the Zipf-Mandelbrot, the discrete Gaussian exponential and the double Pareto-Lognormal distributions. Others, such as the discrete Weibull and the negative binomial distributions, are counting distributions that are shown to be appropriate when fitting skewed data.

The *Zipf-Mandelbrot distribution* (ZM) [Mandelbrot, 1965] is perhaps the most well-known extension of Zipf's law. Aware of the difference between Zipf's law and the first ranked values in real linguistic data, Mandelbrot related the frequency of the words to its rank x in such a way that:

$$P(X=x) = C(V,\alpha)\,(x+V)^{-\alpha},\ x=1,2,\dots, \tag{1.3.2}$$

where $V \in [0,+\infty)$, and $C(V,\alpha)$ is the normalizing constant. Observe that, for the particular case when $V=0$, the Zipf distribution is obtained.

The *Discrete Gaussian Exponential distribution* (DGX) defined by Bi et al. [2001] is the discrete version of the log-normal distribution. The work by Mitzenmacher [2004] examines the intrinsic connection between the PLs and the log-normal families. It is a bi-parametric distribution defined in the strictly positive integers, where the parameter $\mu \in (-\infty;+\infty)$ represents the mean and the parameter $\sigma > 0$ corresponds to the standard deviation. The authors show that, when $\mu \to -\infty$, the distribution reduces to Zipf's law with slope $1-\mu/\sigma^2$. The PMF of the DGX distribution is defined as follows:

$$P(X=x) = \frac{A(\mu,\sigma)}{x}\, e^{\left[-\frac{(ln(x)-\mu)^2}{2\sigma^2}\right]},\ x=1,2,\dots, \tag{1.3.3}$$

where $A(\mu,\sigma) = \{\sum_{x=1}^{\infty} x^{-1}\, e^{\left[-\frac{(ln(x)-\mu)^2}{2\sigma^2}\right]}\}^{-1}$ is the normalizing constant.

The *Discrete Weibull distribution* (DW) [Nakagawa and Osaki, 1975] is the discrete version of the Weibull family. It is a bi-parametric family of distributions with support on the non-negative integers, which is proposed for modeling failure time when life is measured by means of blows or revolutions. Even though it does not provide a straightforward interpretation of its parameters or have closed expressions for the mean and the variance, some authors have proved that it is useful for modeling phenomena such as the abundance of proteins [Koziol et al., 2013] or microbial counts in water [Englehardt and Li, 2011]. Its PMF is equal to:

$$P(X=x) = q^{x^{\beta}} - q^{(x+1)^{\beta}}, x=0,1,2,\dots, 0<q<1, \beta>0. \tag{1.3.4}$$

The *negative binomial distribution* (NB) [Johnson et al., 2005, p. 209] is a counting distribution that is widely known as the classical over-dispersed two-parameter Poisson alternative. It is also a PSS distribution, with the log-series(q) as the primary distribution and

the Poisson with $\lambda = -k\log(1-q)$ as the secondary distribution. The definition of PSS can be found in Section 3.1 of Chapter 3.
One may find several parametrizations of this distribution in the literature. The one used in this thesis is the following:

$$P(X=x) = \binom{k+x-1}{k-1} q^x (1-q)^k,\ x = 0,1,2,\dots\ ,0<q<1,\ k>0. \tag{1.3.5}$$

Another family of distributions that needs to be considered is the *Double Pareto-Lognormal* [Reed and Jorgensen, 2004]. It is a four-parametric model, composed of: two PL parameters representing the left-hand and right-hand PL tails; and two parameters that correspond to the lognormal distribution. Several researchers have argued in favor of its suitability for modeling the degree sequence of social networks such as Facebook [Sala et al., 2010] or for measuring different aspects of a mobile call graph [Seshadri et al., 2008].

In Chapter 5, the fits obtained with the distributions proposed in this thesis are compared to those obtained by means of the first four mentioned distributions. The last family of distributions is excluded, since it has four parameters and we focus on analyzing bi-parametric distributions.

1.4 Our contribution: the Zipf as a mixture distribution

Given a parametric probability distribution, one way to generalize it is by assuming that one of its parameters follows a given r. v. instead of being constant, which gives place to what is called a *mixture distribution*. For example, if one assumes that the p probability of a Binomial distribution follows a Beta distribution, the resulting mixing distribution is the Beta-Binomial distribution. Similarly, by assuming that the λ parameter of the Poisson distribution follows a Gamma distribution one obtains the Negative-Binomial distribution. A mixing distribution is required, for instance, to adapt the heterogeneity that usually exists among experimental units. One property of any mixed distribution is that its variance is always larger than the variance of the initial distribution with the same mean. See Chapter 8 of Johnson et al. [2005] for more information about mixing distributions of discrete r.v.'s. In what follows we define the general formulation of a mixed Zipf distribution with a continuous mixing distribution.

Definition 1. *A r.v. M follows a mixed Zipf distribution with parameter $\theta \in \Theta$ if, and only if,*

$$P(M=x;\theta) = \int_1^{+\infty} \frac{x^{-\alpha}}{\zeta(\alpha)} f(\alpha;\theta)\,d\alpha,$$

being $f(\alpha;\theta)$ the density function of the mixing distribution, which is defined in $(1,+\infty)$.

Theorem 1 of Hill and Woodroofe [1975] proves that under certain regularity assumptions, the Zipf distribution is asymptotically the limit of a mixing geometric distribution for different mixing distributions. Their result can be explained in terms of a double classification problem and assumes that the number of objects to be classified tends to infinity. However, their mixing expression is true only in the tail of the distribution, that is for large values of x.

This section proves that the Zipf distribution is a mixture of geometric distributions. We also show that it is MZTP distribution. In both cases, the mixing distribution is specified analytically. As a consequence of being a MZTP distribution, its variance is larger than the one of a zero-truncated Poisson distribution with the same mean. Important to note that our results are not asymptotic, and thus do not require a large value of x.

We start by proving that the geometric distribution with support in the integer numbers strictly larger than one is a MZTP distribution. Also, we want to point out that if N^{zt} denotes the zero-truncated version of a r.v. N, then its PGF is equal to:

$$h_{N^{zt}}(z;\theta) = \frac{h_N(z;\theta) - h_N(0;\theta)}{1 - h_N(0;\theta)}. \tag{1.4.1}$$

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

$$h_{N^{zt}}(z;\lambda) = \frac{e^{\lambda z} - 1}{e^{\lambda} - 1}. \tag{1.4.2}$$

As a consequence of the fact that $\lim_{\lambda \to 0} h_{N^{zt}}(z;\lambda) = z$, 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.

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

$$f(\lambda;p) = \frac{p}{(1-p)^2} e^{-\lambda/(1-p)}(e^{\lambda} - 1),\ \lambda \in (0,+\infty). \tag{1.4.3}$$

Proof. The PGF of the geometric(p) distribution, with support in the positive integers that are equal to or larger than one, is equal to $pz/(1-qz)$, where $q = 1-p$. Moreover, the PGF of the zero-truncated Poisson distribution is equal to (1.4.2). Given that the PGF of a 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:

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

with $f(\lambda;p)$ defined as in (1.4.3). Substituting $f(\lambda;p)$ in the previous equation for its corresponding expression and taking into account that $z-1/(1-p)<0$ because $z<1$ and $p\in(0,1)$, we have:

$$\begin{aligned}\int_0^{+\infty} \frac{e^{\lambda z}-1}{e^{\lambda}-1} f(\lambda;p)\,d\lambda &= \frac{p}{(1-p)^2}\int_0^{+\infty}\left[e^{\lambda(z-1/(1-p))}-e^{-\lambda/(1-p)}\right]d\lambda = \\ &= \frac{p}{(1-p)^2}\left[\left.\frac{e^{\lambda(z-1/(1-p))}}{z-1/(1-p)}\right|_0^{+\infty}+\left.\frac{e^{-\lambda/(1-p)}}{1/(1-p)}\right|_0^{+\infty}\right]= \\ &= \frac{-p(1-p)}{(1-p)^2}\left[\frac{1}{z(1-p)-1}+1\right]=\frac{pz}{1-(1-p)z}. \qquad (1.4.4)\end{aligned}$$

□

Next two theorems show the Zipf distribution as a mixture distribution.

Theorem 1. *The Zipf(α) distribution is a mixture of geometric distributions with domain $\{1,2,3,\cdots\}$ and parameter $s=-log(1-p)$, with mixing distribution:*

$$f(s;\alpha)=\frac{s^{\alpha-1}}{(e^s-1)\zeta(\alpha)\Gamma(\alpha)},\ s>0, and\ \alpha>1. \qquad (1.4.5)$$

Proof. The PGF of the geometric distribution defined in strictly positive integers, as a function of the parameter s is equal to:

$$\frac{pz}{1-qz}=\frac{(1-e^{-s})z}{1-e^{-s}z}=\frac{(e^s-1)z}{e^s-z}.$$

Observe that, with the new parametrization, the parameter space is $(0,+\infty)$ instead of $(0,1)$. Thus, taking into account the integral expression of the Li function that appears in (1.1.14), its mixture distribution with mixing distribution as defined in (1.4.5) is equal to:

$$\begin{aligned}&\int_0^{+\infty}\frac{(e^s-1)z}{e^s-z}\frac{s^{\alpha-1}}{(e^s-1)\zeta(\alpha)\Gamma(\alpha)}ds=\frac{z}{\zeta(\alpha)\Gamma(\alpha)}\int_0^{+\infty}\frac{s^{\alpha-1}}{e^s-z}ds \\ &=\frac{1}{\zeta(\alpha)}\int_0^{+\infty}\frac{s^{\alpha-1}}{\frac{e^s}{z}-1}ds=\frac{Li_\alpha(z)}{Li_\alpha(1)}, \qquad (1.4.6)\end{aligned}$$

which is the PGF of the Zipf(α) distribution (see (1.1.12)). □

Figure 1.5 contains the plot of the mixing distribution of Theorem 1, as a function of s (on the left-hand side) and as a function of p (on the right-hand side), for different α values.

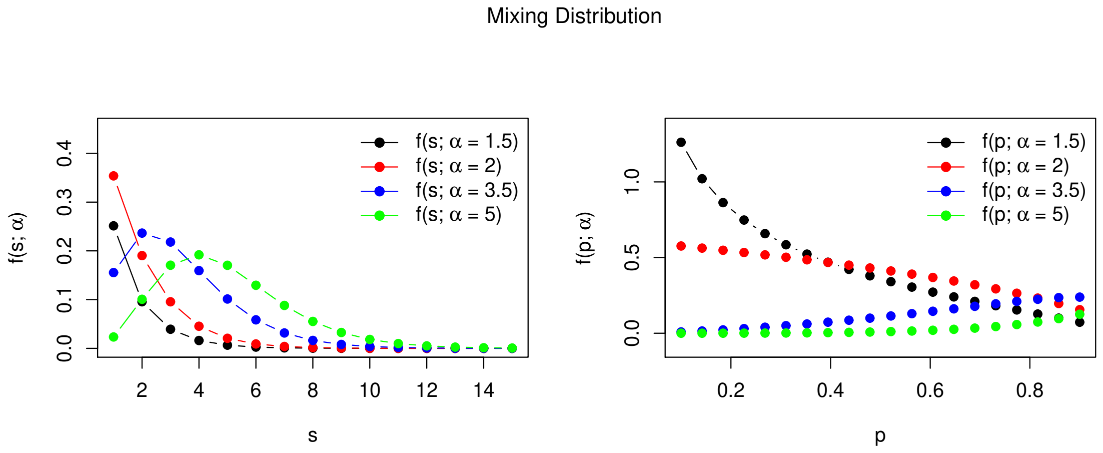

Fig. 1.5 The mixing distribution of Theorem 1, as a function of s (on the left-hand side) and as a function of p (on the right-hand side), for different α values.

As mentioned in the previous section, Krumme et al. [2013] proves that the Zipf distribution is useful to describe how frequently a client visits a store. Based on the mixture interpretation of the Zipf distribution proved in Theorem 1, we can interpret that the number of customer's visits to a store follows a geometric distribution with a s value that depends on the customer and comes from the (1.4.5) distribution. This has sense if one assumes that the customer goes to the store until she/he gets the desired product.

Theorem 1 of Valero et al. [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) $\lim_{z\to-\infty} h(z) = -L$, 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. The next theorem establishes that the Zipf belongs to the MZTP class, but not to the ZTMP class.

Theorem 2. *The Zipf(α) distribution verifies that:*

a) it is an MZTP distribution with mixing distribution equal to:

$$f(\lambda;\alpha) = \frac{(e^{\lambda}-1)\int_0^{+\infty} e^{s-\lambda e^{s}} s^{\alpha-1} ds}{\Gamma(\alpha)\zeta(\alpha)},\ \lambda > 0, \tag{1.4.7}$$

b) it is not a ZTMP distribution.

Proof. Taking into account (1.1.12), proving a) is equivalent to seeing that

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

with $f(\lambda;\alpha)$ defined as in (1.4.7). First observe that, as a consequence of Theorem 1, from (1.4.6) we have:

$$\frac{Li_\alpha(z)}{Li_\alpha(1)} = \int_0^{+\infty} \frac{(e^{s}-1)z}{e^{s}-z} \frac{s^{\alpha-1}}{(e^{s}-1)\zeta(\alpha)\Gamma(\alpha)} ds. \tag{1.4.8}$$

Now, rewriting (1.4.4) with the s parametrization we also have:

$$\frac{(e^{s}-1)z}{e^{s}-z} = \int_0^{+\infty} \frac{e^{\lambda z}-1}{e^{\lambda}-1} f^{*}(\lambda;s) d\lambda, \tag{1.4.9}$$

where

$$f^{*}(\lambda;s) = f(\lambda;1-e^{-s}) = e^{s}(e^{s}-1)e^{-\lambda e^{s}}(e^{\lambda}-1).$$

Substituting (1.4.9) in (1.4.8) gives that:

$$\begin{aligned}
\frac{Li_\alpha(z)}{Li_\alpha(1)} &= \int_0^{+\infty} \Big[\int_0^{+\infty} \frac{e^{\lambda z}-1}{e^{\lambda}-1} f^{*}(\lambda;s)\, d\lambda\Big] \frac{s^{\alpha-1}}{(e^{s}-1)\zeta(\alpha)\Gamma(\alpha)} ds \\
&= \int_0^{+\infty} \frac{e^{\lambda z}-1}{e^{\lambda}-1} \int_0^{+\infty} \Big[f^{*}(\lambda;s) \frac{s^{\alpha-1}}{(e^{s}-1)} \frac{1}{\zeta(\alpha)\Gamma(\alpha)}\Big] ds d\lambda \\
&= \int_0^{+\infty} \frac{e^{\lambda z}-1}{e^{\lambda}-1} \Big[\frac{e^{\lambda}-1}{\zeta(\alpha)\Gamma(\alpha)} \int_0^{+\infty} e^{s-\lambda e^{s}} s^{\alpha-1} ds\Big]\, d\lambda,
\end{aligned}$$

which proves $a)$. To prove $b)$, it is necessary to see that condition d) of Theorem 1 of Valero et al. [2010] is not satisfied. But, this is the case since:

$$\begin{aligned}\lim_{z\to-\infty}\frac{Li_\alpha(z)}{Li_\alpha(1)} &= \lim_{z\to-\infty}\frac{z}{\Gamma(\alpha)Li_\alpha(1)}\int_0^{+\infty}\frac{t^{\alpha-1}}{e^t-z}dt \\ &= \lim_{z\to-\infty}\frac{1}{\Gamma(\alpha)\zeta(\alpha)}\int_0^{+\infty}\frac{t^{\alpha-1}}{\frac{e^t}{z}-1} = -\infty. \qquad (1.4.10)\end{aligned}$$

□

The interpretation of Theorem 2 is as follows: if we take an MP distribution and we truncate it at zero, in order that it takes values from one to infinity, we will never obtain the Zipf distribution. However, if we first truncate at zero a Poisson distribution, and then the parameter of the zero-truncated Poisson is considered to follow the probability distribution defined at (1.4.7), the result is the Zipf distribution. Applying this reasoning to the context to model the degree sequence of graphs that do not contain isolated nodes (see Chapter 5 for the definition of the node, degree sequence, and isolated node), it means that if the Zipf distribution is a suitable distribution to fit a degree sequence, it is because the number of connections of a node follows a zero-truncated Poisson distribution with a λ parameter that varies from node to node with probability density function as in (1.4.7).

Chapter 2

Random stopped extreme Zipf extensions

This chapter extends the Zipf distribution by means of the concept *Random Stopped Extreme distribution* (RSED), which covers those families defined as the minimum or maximum of a random number of i.i.d. r.v.'s. The name RSED was introduced by Pérez-Casany et al. [2016] in the conference ICOSDA 2016. However, these kinds of distributions have been widely studied in the literature (see, for instance, Cancho et al. [2011]; Kuş [2007] and Ramos et al. [2019]). We start by reviewing the concept of RSED, and introducing two new properties related to this class of distribution, of which the later are part of the contribution of this thesis. Next, we present the notion of *regularly varying function* (RVF), followed by an analysis of a particular set of RSED known as the *Random Stopped Extreme Zipf Distribution* (RSEZipf). The last two sections are devoted to the analysis of the Marshall-Olkin extended Zipf (MOEZipf) and the Zipf-Poisson extreme (Zipf-PE) distributions, which are particular cases of RSEZipf distributions. The MOEZipf distribution was defined by A. Casellas in her Master's thesis (see, Casellas [2013]) under the guidance of Prof. M. Pérez-Casany. The main results obtained in that work appear in Pérez-Casany and Casellas [2013]. In fact, the work of A. Casellas and M. Pérez-Casany constitutes the starting point of this research work. With respect to the MOEZipf family of distributions, in this chapter we extend some of the properties of Pérez-Casany and Casellas [2013] and we prove new ones. The definition and analysis of the Zipf-PE distribution pertains entirely to this thesis. A more reduced version of all the work contained in this chapter appears in the paper *"Random Stopped Extreme Zipf Extensions"* [Duarte-López et al., 2020a], which has been submitted for publication.

2.1 Background

In this section we start by defining the general concept of RSED and by stating two new results related to these families of distributions. Later, we introduce the concept of RVF, because it is essential to prove some of the properties of the Zipf's extensions proposed in this chapter.

2.1.1 Random 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] where the author introduces the exponential-Poisson distribution as the $min(W_1, W_2, \ldots, W_N)$, where the i.i.d. r.v.'s W are exponentially distributed, and independent of the r.v. N which follows a zero-truncated Poisson distribution. Four years latter, Cancho et al. [2011] defined the Poisson-exponential distribution which, contrary to the first example, it is defined in terms of maximums, i.e. $max(W_1, W_2, \ldots, W_N)$, where the i.i.d. r.v.'s W follow an exponential distribution and N is zero-truncated Poisson distributed. The survey by Tahir and Cordeiro [2016] reviews a large amount of RSEDs. Most of the models considered in their work, correspond to families where X follows a continuous distribution. The authors denote the RSEDs as "compound distributions", but this name is avoided here because some authors use it to refer to mixture distributions while others used it to refer to stopped-sum distributions. Thus, the use of this term may give place to some misunderstandings.

In what follows, we formally define the concept of RSED and then, we introduce our firsts two contributions to this Chapter.

Definition

Let X be an r.v. with parameter vector α and 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 PGF $h_N(t;\theta)$, with θ being the parameter vector. The r.v.'s defined as:

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

where X_i are i.i.d. copies of X, have CDF and SF, respectively, equal to:

$$F_{Y_{X;N}^{max}}(x;\alpha,\theta) = h_N(F_X(x;\alpha),\theta) \quad \text{and} \quad S_{Y_{X;N}^{min}} = h_N(S_X(x;\alpha),\theta), \tag{2.1.1}$$

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, RSEDs, 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].

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 a currency i that is 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 in that period. It is important to observe that X is associated with the phenomena under study, in the case of the example the currency price, while N is related to the number of observations of X that one has in a given period. RSEDs appear in real situations when one observes only the variable of interest when it is larger (smaller) than a given upper (lower) bound.

The distributions of the r.v.'s X and N are, respectively, denoted by *stopped* and *stopping* distributions. This allows for a parallelism between RSEDs 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 (see Subsection 3.1 of Chapter 3). The stopped distribution of an RSED serves as the secondary distribution of a stopped sum, while the stopping distribution represents the primary distribution. Random Stopped Extreme and Stopped-Sum are two mechanisms that allow us to generalize the distribution of X. Both transformations help us better understand the mechanism that generates the data. Based on (2.1.1), 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.
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.

Table 2.1 contains the name of five probability distributions that may be considered as stopping distribution, jointly with its mean and PGF (second and third columns) and their parameter space (fourth column). The geometric, the one that takes the strictly positive integer values, and the log-series distributions do not require to be zero-truncated since they give probability zero at zero. On the contrary, the Poisson, the Hermite, and the negative binomial distributions require to be zero-truncated. The PFG of the positive Poisson appears in (1.4.2). With respect to the other two, one has that the PGF of the Hermite distribution is equal to:

$$h_N(z;\theta,\beta) = e^{\theta(z-1)+\beta(z^2-1)} \qquad \theta,\beta > 0,$$

and the PGF of the negative binomial is equal to:

$$h_N(z;\theta,\beta) = \Big(\frac{1-\theta}{1-\theta z}\Big)^{\beta} \qquad 0 < \theta < 1,\ \beta > 0.$$

Applying (1.4.1), one obtains that the PGF of the positive Hermite distribution is equal to:

$$\frac{e^{\theta z+\beta z^2}-1}{e^{\theta+\beta}-1}$$

and the PGF of the positive negative binomial distribution is equal to:

$$\frac{(\frac{1-\theta}{1-\theta z})^{\beta}-(1-\theta)^{\beta}}{1-(1-\theta)^{\beta}}.$$

Taking into account (2.1.1) 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:

$$F_{Y^{max}_{X;N}}(x;\alpha,\theta)=h_{N^{zt}}(F_X(x;\alpha),\theta)=\frac{h_N(F_X(t;\alpha);\theta)-h_N(0;\theta)}{1-h_N(0;\theta)}, \tag{2.1.2}$$

and

$$S_{Y^{min}_{X;N}}(x;\alpha,\theta)=h_{N^{zt}}(S_X(x;\alpha),\theta)=\frac{h_N(S_X(t;\alpha);\theta)-h_N(0;\theta)}{1-h_N(0;\theta)}. \tag{2.1.3}$$

Observe that (2.1.2) and (2.1.3) generalize the two equalities that appear in (2.1.1), because if N gives probability zero to the zero value, then $h_N(0;\theta)=0$. That is why from now on we only work with (2.1.2) and (2.1.3).

Two new results on RSED

Here we prove two new theorems related to the general theory of RSEDs. 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 3. *If N is defined in the strictly positive integer values and a value θ_0 exists in the parameter space, such that $h_N(z;\theta_0)=z$, then the distribution of X belongs to both sets of maximum and minimum stopped extreme distributions. Consequently, the families of maximums as well as the family of minimums extend the initial family of distributions.*

Proof. Given that $h_N(z;\theta_0)=z$, from (2.1.1) one has that:

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

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

Table 2.1 Some possible stopping distributions together with their PGFs, parameter spaces and the parameter values that gives the family of distributions of X.

and that:

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

which proves the theorem. □

Observe that saying $h_N(z;\theta_0) = z$ is equivalent to saying that the family contains the degenerate distribution at one, as a particular case. This is the case of the zero-truncated Poisson distribution as already mentioned, and the rest of stopping distributions considered in the first column of Table 2.1. The parameter values which allow obtaining the initial family of distributions appear in the last column of the mentioned table.

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

Theorem 4. *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 $u' = h_N^{-1}(u(1-h_N(0;\theta))+h_N(0;\theta);\,\theta)$, and* *b) if Y is a minimum, then $u' = 1 - h_N^{-1}(1 - u(1 - h_N(0,\theta));\theta)$;*

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 (2.1.2), 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, with a little bit of algebra, is equivalent to saying that:

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

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

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 (2.1.3), 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 $u' = 1 - h_N^{-1}(1 - u(1 - h_N(0,\theta));\theta)$. □

This results has been very useful when implementing the two Zipf extensions of Sections 2.3 and 2.4 of this chapter in the R-package *zipfextR* [Duarte-López and Pérez-Casany, 2020] (see Appendix B).

2.1.2 Regularly varying functions

A *regularly varying function* can roughly be defined as a function that performs asymptotically as a power function. This concept will be necessary to prove some of the properties of the distributions families introduced in this PhD thesis. The formal definitions below can be found in the book by Gulisashvili [2012, p. 201].

Definition 2. *Let f be a positive measurable function on $[a, +\infty)$ with $a > 0$, and let α be a real number. Then, the function f is a regularly varying function of index α at infinity, if for any $\lambda > 0$,*

$$\lim_{x \to \infty} \frac{f(\lambda x)}{f(x)} = \lambda^{\alpha}. \tag{2.1.4}$$

The class of all regularly varying functions with index α is denoted by R_{α}.

Definition 3. *A slowly varying function at infinity is a function L which satisfies that for all $\lambda > 0$,*

$$\lim_{x \to \infty} \frac{L(\lambda x)}{L(x)} = 1. \tag{2.1.5}$$

The class of slowly varying at infinity functions is denoted by R_0. Observe that $f \in R_{\alpha}$ if, and only if, $f(x) = x^{\alpha} L(x)$, where the function L is slowly varying at infinity.

According to Gulisashvili [2012, p. 220], Pareto-type distributions (and in particular, the Zipf distribution) are camouflaged versions of regularly varying functions. Basically, a function f belongs to the class of Pareto-type distribution if it is asymptotically equivalent to a regularly varying function. This implies that f is also a regularly varying function. Thus, the Zipf distribution is, by definition, a regularly varying function.

2.2 Random stopped extreme Zipf extensions

By random stopped extreme Zipf distribution (RSEZipf), we denote the RSEDs that emerge when it is assumed that X follows a Zipf distribution. Table 2.2 contains the CDFs of the maximums as well as the SFs of the minimums of the RSEZipfs obtained by considering the stopping distributions that appear in the first column of Table 2.1. They were obtained by compounding the PGF of the stopping distribution (that appears in the second column of Table 2.1) with: the CDF of the Zipf (1.1.5) (in the case of maximums); and the SF of the Zipf (1.1.4) (in the case of minimums). Appendix C contains the figures of the CDF of maximums, as well as of minimums, associated with the stopping distributions that appear in Table 2.2. It is observed that independently of the stopping distribution, the cumulative distribution function of minimums at small values as 4 or 5 is already very close to one. On the contrary, the support of the r.v's defined in terms of maximums is much larger as it has sense to be. In all the cases, for a fixed value of α, the parameter θ controls the growth of the probabilities in a way that depends on the stopping distribution.

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 2.2 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.

The next theorem states the asymptotically relation of the tails of the Zipf and any RSEZipf distribution with the same parameter α. The results obtained are a direct consequence of the work by Jessen and Mikosch [2006, p. 187–188], and they hold because the Zipf is a regularly varying function as mentioned before. Here $f(x) \sim g(x)$ 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 5. *The tail of an r.v $Y \sim$ RSEZipf(α,β) is asymptomatically related to the tail of an r.v. $X \sim$ Zipf(α). 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 and $E[N] < +\infty$, then $P(Y > x) \sim E[N]P(X > x)$.

Proof. The proof is straightforward from results that appear in sections 5.3 and 5.4 of the work by Jessen and Mikosch [2006, p. 187–188]. An important consequence of Theorem 5 is that any extension of the Zipf distrbution obtained by RSED mechanism has a linear tail, in log-log scale, if the stopping distribution has support the strictly positive integer numbers.

This is because in that case $n_0 = 1$, and one has that:

$$P(Y = x) = P(Y > x-1) - P(Y > x) = \begin{cases} P(N = 1) \times P(X = x) & \text{if } Y \text{ is a minimum,} \\ E[N] \times P(X = x) & \text{if } Y \text{ is a maximum.} \end{cases}$$

□

The next two sections are devoted to two particular Zipf's extensions.

2.3 The Marshall-Olkin extended Zipf distribution

This section focuses on the MOEZipf distribution which is obtained by assuming the geometric as stopping distribution. This family of distributions is also obtained as the result of applying the Marshall-Olkin transformation (MO) [Marshall and Olkin, 1997] to the Zipf distribution. As already said, it was originally defined by A. Casellas in her master thesis [Casellas, 2013] and it constitutes the starting point of this work.

2.3.1 Definition

The MO transformation allows to extend a family of probability distributions by adding an extra parameter. In their work the authors prove that the extended family is what they call *geometric extreme stable*, which means that any of their distributions can be interpreted as the minimum (maximum) of a geometric number of i.i.d. r.v.'s from the initial family [Marshall and Olkin, 1997, p. 646]. This is true for a geometric distribution supporting the strictly positive integer values. The MO transformation is applied to the SF of a given distribution in the initial family and gives place to the SF of a distribution in the extended family. It is defined as:

$$G(x;\beta) = \frac{\beta \overline{F}(x)}{1 - \overline{\beta}\overline{F}(x)},$$

for $\beta > 0$ where $\overline{\beta} = 1 - \beta$. Note that for $\beta = 1$ the initial distribution is obtained.

Applying the MO transformation to the SF of the Zipf, that appears in (1.1.4), results in the SF of the MOEZipf distribution, that it is equal to:

$$\overline{F}_{\alpha,\beta}(x) = \frac{\beta \overline{F}_\alpha(x)}{1 - \overline{\beta}\overline{F}_\alpha(x)} = \frac{\beta \zeta(\alpha, x+1)}{\zeta(\alpha) - \overline{\beta}\zeta(\alpha, x+1)}, \alpha > 1, \beta > 0. \tag{2.3.1}$$

Based on the definition, the MOEZipf distribution has support on the strictly positive integer values, and its parameters are the α parameter of the Zipf distribution and the β parameter of the geometric distribution. Thus, the parameter space is $(1,+\infty) \times (0,+\infty)$. The PMF of the MOEZipf distribution, for any $x \geq 2$, can be derived from (2.3.1) by computing $\overline{F}_{\alpha,\beta}(x-1) - \overline{F}_{\alpha,\beta}(x)$ and it is equal to:

$$P(Y=x) = \frac{x^{-\alpha}\beta\,\zeta(\alpha)}{[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)]\,[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)]}, x = 2,3,4,\ldots. \tag{2.3.2}$$

For $x = 1$ one has that,

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

Observe that $P(Y=1)$ is equal to (2.3.2) at $x = 1$; and thus (2.3.2) is the PMF in the entire support.

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

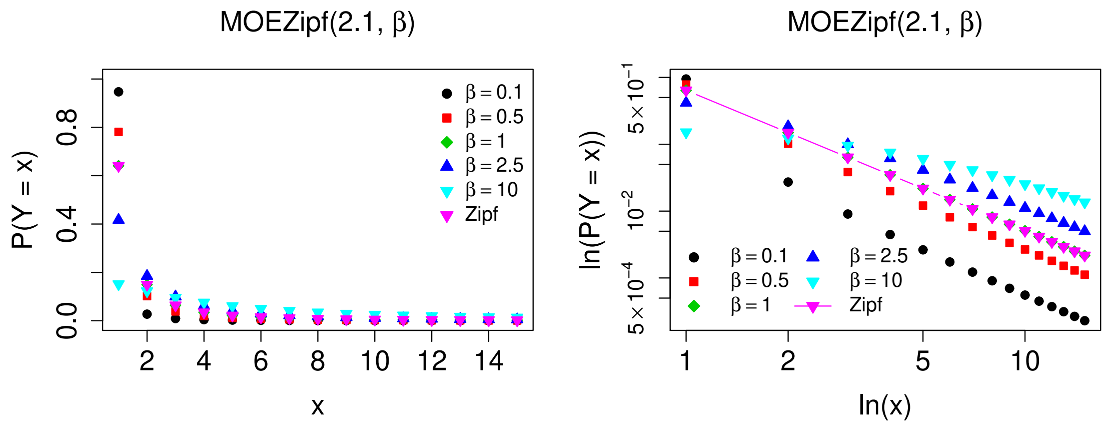

Fig. 2.1 PMFs of the MOEZipf distribution for $\alpha = 2.1$ and $\beta = 0.1, 0.5, 1, 2.5$ and 10. On the left-hand side: normal scale. On the right-hand side: log-log scale.

2.3.2 Properties

In order to clarify which results are established in the works by Pérez-Casany and Casellas [2013] and Casellas [2013], and which are part of this PhD thesis, in what follows we summarize the main results in the aforementioned papers. After defining the distribution they have proved that:

a) the parameter β is obtained as the limit of the ratio of the MOEZipf and Zipf probabilities at x, when they have the same parameter α. As a consequence, $P(Y = x)$ is proportional to $P(X = x)$ being β the proportionality constant;

b) for large values of x, $\log(P(Y = x))$ is a linear function of the $\log(x)$. Thus, the extended family is also linear in the tail in log-log scale. At this point, it is necessary to say that in the original papers this result only was mentioned and illustrated, but it was not proved analytically. The prove of this result corresponds to Theorem 6 of this chapter;

c) the *k-th* moment of the MOEZipf distribution exists if, only if, $\alpha > k+1$ [see, Casellas, 2013, p. 40–41]. Figure 2.2 illustrates the behavior of the mean as: a function of α for $\beta = 0.5, 1, 1.5$ and 3 (left-hand side); and as a function of β for $\alpha = 2.5, 4.8, 7.5$ and 20 (right-hand side). In the same way, Figure 2.3 shows the behavior of the variance. On the left-hand side, as a function of α for $\beta = 0.5, 1, 1.5$ and 3; and on the right-hand side as a function of β for $\alpha = 3.5, 4.8, 7.5$ and 20. Observe that both, the E[Y] and the VAR[Y], are decreasing functions of α which decrease faster as β becomes small. On the contrary, the E[Y] and the VAR[Y] are increasing functions of β whose slope decreases as α increases.

d) the ratio of two consecutive MOEZipf probabilities is greater (smaller) than that of probabilities coming from a Zipf distribution with the same α, if $\beta > 1 (\beta < 1)$. For $\beta = 1$ the two distributions are the same and consequently, the two ratios are equal;

e) they compare $P(Y = x)$ with respect to $P(X = x)$ obtaining an inequality between them. This result has been extended in Proposition 4.

f) two methods are proposed for estimating the value of the parameters α y β given sequence of values $x_1, x_2, \ldots, x_n$. The first one corresponds to solve the system of equations that comes from equating the empirical probability at one to the theoretical probability at one, and the sample mean to the mean. The second one is the MLE which is obtained by maximizing the logarithm of the likelihood. The log-likelihood

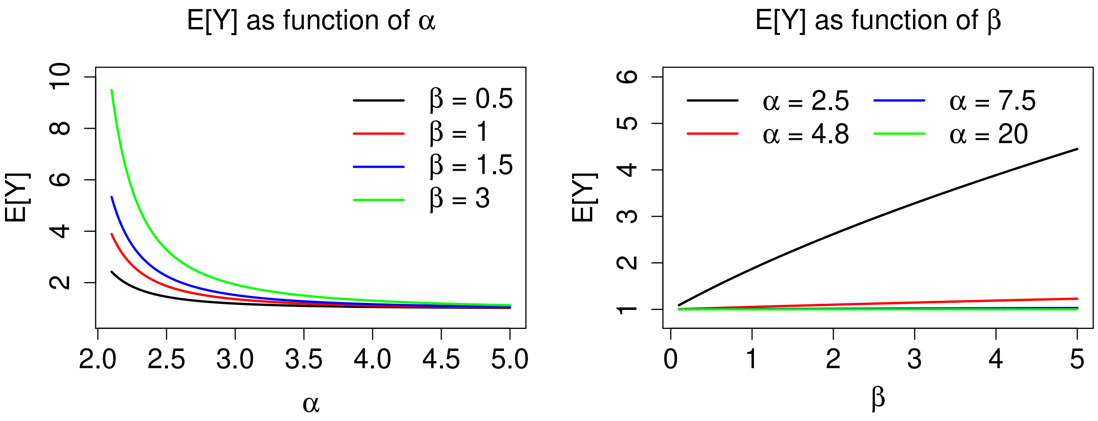

Fig. 2.2 Mean values of a MOEZipf(α,β) distribution. On the left-hand side: as a function of α for $\beta = 0.5, 1, 1.5$ and 3. On the right-hand side: as a function of β for $\alpha = 2.5, 4.8, 7.5$ and 20.

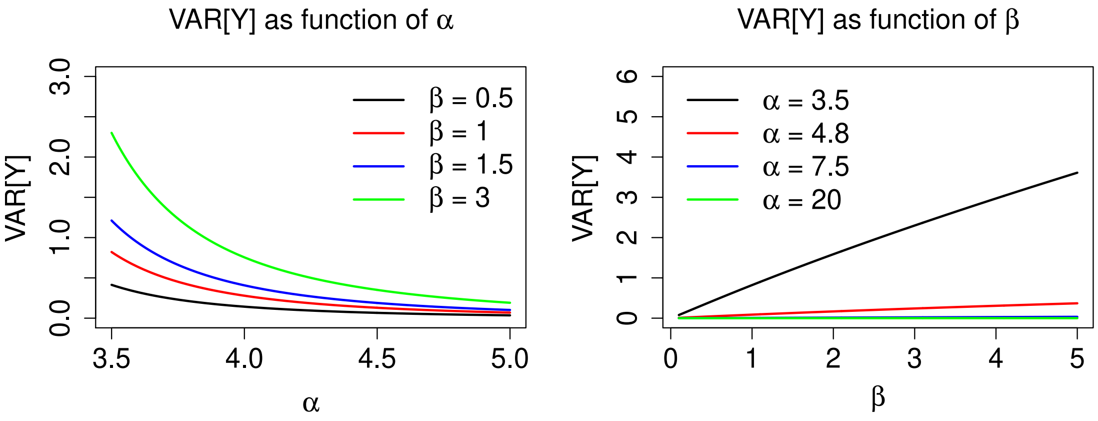

Fig. 2.3 Variance values of a MOEZipf(α,β) distribution. On the left-hand side: as a function of α for $\beta = 0.5, 1, 1.5$ and 3. On the right-hand side: as a function of β for $\alpha = 3.5, 4.8, 7.5$ and 20.

function for the MOEZipf with parameters α and β is equal to:

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

$$-\sum_{i=1}^{n}\log(\zeta(\alpha) - \overline{\beta}\zeta(\alpha,x_i)) - \sum_{i=1}^{n}\log(\zeta(\alpha) - \overline{\beta}\zeta(\alpha,x_i+1));$$

In what follows we expose the results obtained on the MOEZipf distribution in this PhD thesis.

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. This proposition is a consequence of the results that appear in Section 5 of Marshall and Olkin [1997].

Proposition 2. *Let Y be a MOEZipf distributed r.v. with parameters (α,β). Then:*

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

iii) If $\beta = 1$, Y follows a Zipf(α) distribution and may be seen as a maximum as well as a minimum of i.i.d. Zipf(α) 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 (2.3.1) 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 2.2 shows that this corresponds to a maximum of i.i.d. Zipf(α) r.v.'s, with a geometric stopping distribution with parameter $\theta = 1/\beta$, which proves (i).
By dividing the SF of the MOEZipf(α,β) that appears in (2.3.1) 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 2.2, corresponds to the SF of an RSED, with a Zipf(α) distribution as the secondary distribution and a geometric distribution with parameter $\theta = \beta$ as the primary distribution, which proves (ii).
Using Theorem 3 when $\beta = 1$, (2.3.1) is equal to the SF of a Zipf(α) 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, to 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 of its PMF. For any $\alpha > 1$ and $\beta > 0$, let us define the function

$$h_{(\alpha,\beta)}(x) = \zeta(\alpha) - (1-\beta)\zeta(\alpha,x+1),\ x \geq 1. \tag{2.3.3}$$

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:

$$h'_{(\alpha,\beta)}(x) = \alpha(1-\beta)\zeta(\alpha+1,x+1), \text{ and}$$
$$h''_{(\alpha,\beta)}(x) = -\alpha(\alpha+1)(1-\beta)\zeta(\alpha+2,x+1). \tag{2.3.4}$$

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. □

Figure 2.4 illustrates the results stated in Lemma 1. On the left-hand side the figure contains the plot of $h_{(\alpha,\beta)}(x)$ for $\alpha = 2.1$ and $\beta = 0.76$ and, on the right-hand side for $\alpha = 2.1$ and $\beta = 3.5$.

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 3. *Let Y be an r.v., such that Y ∼ MOEZipf(α,β), with $\beta \in (0,1]$. Then, Y has a log-convex distribution.*

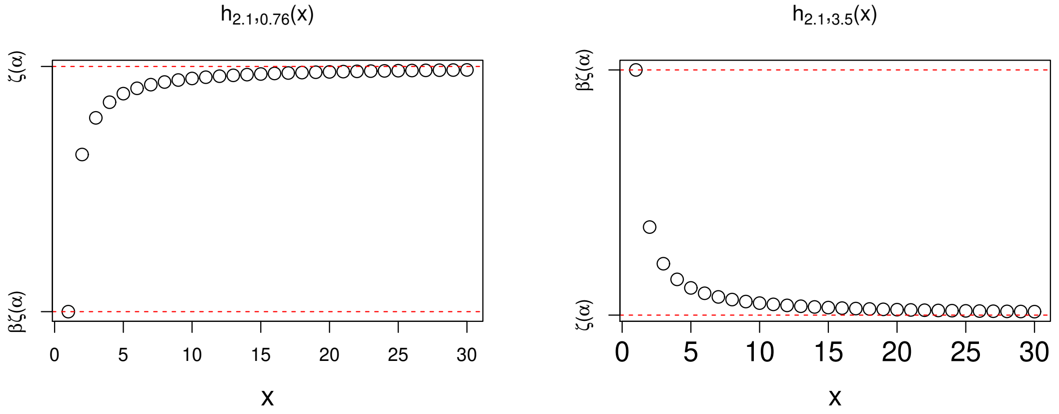

Fig. 2.4 Function $h_{(\alpha,\beta)}(x)$ for $\alpha = 2.1$. On the left-hand side, for $\beta = 0.76$ and, on the right-hand side, for $\beta = 3.5$. The function limits are represented by a dash line.

Proof. As stated in Johnson et al. [2005], a discrete distribution is said to be log-convex when

$$\frac{P(Y=x)P(Y=x+2)}{(P(Y=x+1))^2} \geq 1, \ \forall x. \tag{2.3.5}$$

Thus, it is necessary to prove that (2.3.5) holds for any $\beta \in (0,1]$. From (2.3.2), one has that (2.3.5) is equivalent to:

$$\frac{P(Y=x)P(Y=x+2)}{(P(Y=x+1))^2} = \left(\frac{x(x+2)}{(x+1)^2}\right)^{-\alpha} \left(\frac{\frac{h_{\alpha,\beta}(x+1)}{h_{\alpha,\beta}(x)}}{\frac{h_{\alpha,\beta}(x+3)}{h_{\alpha,\beta}(x+2)}}\right) \geq 1. \tag{2.3.6}$$

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 appears in (2.3.6) is always larger than one. Thus, to prove the proposition 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 of the product on the right-hand side of the equality is equal to $g(x)/g(x+2)$. Observe that by (2.3.3),

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

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 2.5 shows the behavior of the ratio that appears on the left-hand side of equation (2.3.6) for $\alpha = 2.34$ (left-hand side) and $\alpha = 5$ (right-hand side). 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 β. However, for $\beta > 1$, the function can be log-convex or log-concave.

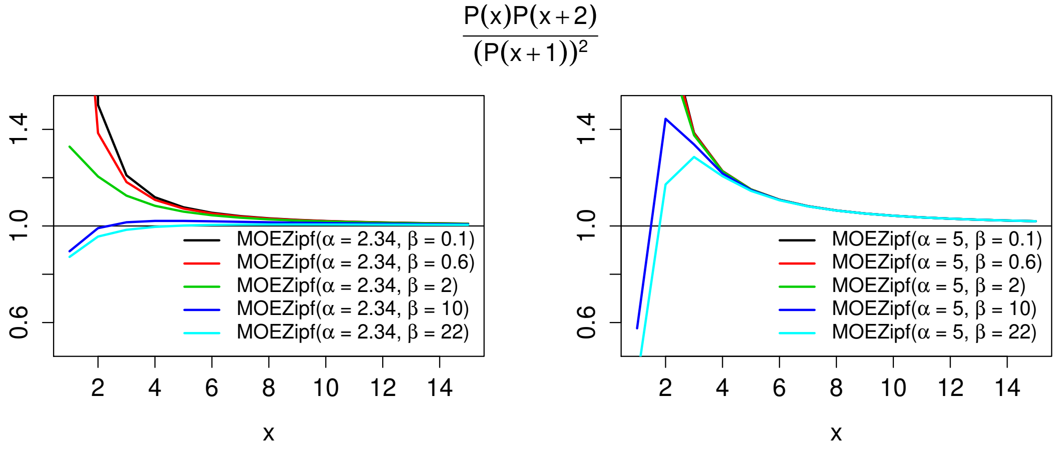

Fig. 2.5 Behavior of the ratio that appears on the left-hand side of equation (2.3.6). On the left-hand side, for $\alpha = 2.34$ and, on the right-hand side, for $\alpha = 5$. In both cases, $\beta = 0.1, 0.6, 2, 10$ and 22.

The next proposition establishes the relationship between the probability values of a MOEZipf and a Zipf distribution with the same α parameter. This proposition extends Proposition 3.3 by Pérez-Casany and Casellas [2013, p. 6], where only the lower bounds are stated.

Proposition 4. *Let Y and X be two r.v.'s, such that $Y \sim$ MOEZipf(α,β) and $X \sim$ Zipf(α). 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

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

which proves $b)$. 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 α. It is a consequence of Theorem 5 of Section 2.2.

Theorem 6. *Let Y and X be two r.v.'s, such that $Y \sim$ MOEZipf(α,β) and $X \sim$ Zipf(α). The tail of Y is asymptotically equivalent to β times the tail of X, $\forall \beta > 0$.*

Proof. 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 5 a), one has that:

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

which implies that

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

$$P(Y=x) \sim \beta P(X=x).$$

If Y is a maximum, from Proposition 2, one has that $E[N] = 1/(1/\beta) = \beta$. And thus, from Theorem 5 b), one has that:

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

which implies that

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

□

Figure 2.6 illustrates the results stated in Theorem 6 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.

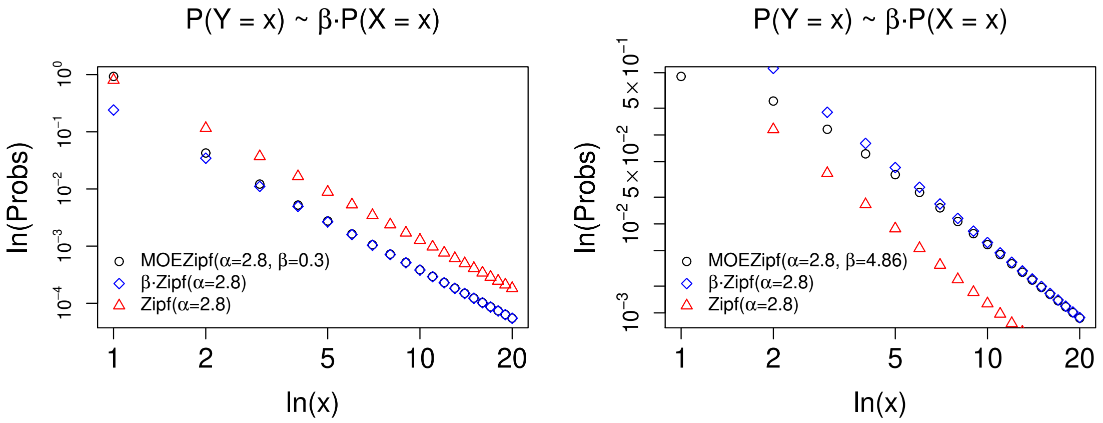

Fig. 2.6 Probabilities of the Zipf and the MOEZipf distributions with the same α parameter in log-log scale, jointly with β times the probability of the Zipf. The MOEZipf on the left-hand side is defined in terms of minimums and, on the right-hand side, it is defined in terms of maximums.

The next subsection is devoted to analyze the performance of the random number generator mechanism stated in Theorem 4, for the MOEZipf distribution. At the beginning of this PhD thesis, the MOEZipf distribution was implemented in the R-package *moezipfR* [Duarte-López et al., 2017] that, in particular, contains the random number functionality. Later on, a new package with name *zipfextR* [Duarte-López and Pérez-Casany, 2020] was created to include the four extensions defined in this thesis. The functionalities of the package are described in detail in Appendix B.

2.3.3 Random data generation

Based on Theorem 4 in Section 2.1.1, it is possible to generate data from any MOEZipf distribution if one knows how to generate data from a Zipf distribution. For the particular case of the MOEZipf, this theorem takes the form of Proposition 5. Observe that the MOEZipf distribution does not require distinguishing whether the distribution comes from a maximum or a minimum. The Zipf random number generator used throughout this thesis is the one implemented in the R-package *tolerance* (see Young et al. [2010]).

Proposition 5. *Generating random data from a MOEZipf(α,β) distribution can be done by applying the inversion method to the Zipf(α) distribution and using the u' value equal to:*

$$u' = \frac{u\beta}{1+u(\beta-1)},$$

regardless of whether the distribution is defined in terms of maximums or minimums, where $u \in (0,1)$ comes from an Uniform distribution.

Proof. Applying the results of Theorem 4 to the particular case of the geometric distribution, one has the following:

a) if Y comes from a maximum family ($\beta > 1$), point 2a) of Theorem 4 give us:

$$u' = h_N^{-1}(u(1-h_N(0;\theta)) + h_N(0;\theta);\, \theta) = h_N^{-1}(u(1-0)+0;\, \theta) = h_N^{-1}(u;\, \beta) \Leftrightarrow$$

$$u' = \left(\frac{\beta\, u}{1-(1-\beta)\, u}\right)^{-1} \Leftrightarrow \frac{\beta\, u'}{1-(1-\beta)\, u'} = u \Leftrightarrow \beta\, u' = u-(1-\beta)\, u u' \Leftrightarrow$$

$$u' = \frac{u}{\beta - u(\beta-1)}.$$

b) if Y comes from a minimum family ($\beta < 1$), point 2b) of Theorem 4 give us:

$$u' = 1-h_N^{-1}(1-u(1-h_N(0,\theta));\theta) = 1-h_N^{-1}(1-u(1-0);\theta) \Leftrightarrow$$

$$u' = 1-h_N^{-1}(1-u;\theta) = 1-\left(\frac{\beta\,(1-u)}{1-(1-\beta)\, u}\right)^{-1} \Leftrightarrow 1-\frac{\beta-\beta\, u}{u-\beta\, u+\beta} \Leftrightarrow$$

$$u' = \frac{u}{\beta-u(\beta-1)},$$

which proves the proposition. □

The Kolmogorov-Smirnov (KS) test for discrete distributions is used to check the performance of our random number generator. The KS is a non-parametric test for assessing whether a given sample comes from a particular hypothesized distribution. This test has been widely used for continuous distributions because the distribution of its statistic does not depend on the distribution that is tested. Nevertheless, this is not the case when one is testing a particular discrete distribution. Several papers adapt the KS test for continuous distributions to the discrete case (see, for instance, Conover [1972] or Arnold and Emerson [2011]). The version considered in this work is the version implemented in the R-package *dgof*, which appears in the paper by Arnold and Emerson [2011].

The hypotheses associated with the test are:

$$H_0 : F_n(x) = F(x)$$

$$H_1 : F_n(x) \neq F(x),$$

where $F_n(x)$ is the empirical cumulative distribution of the observed data, and $F(x)$ is the hypothesized cumulative distribution. In this section, $F(x)$ is a MOEZipf distribution with fixed α and β. Later on, the other Zipf generalization introduced in this PhD thesis is analyzed in a similar way.

In general, this test involves two steps. The first one consists of calculating the test statistic, and the second one computes the p-value for the associated statistic. The p-value can be obtained by using either the reference distribution or by simulation.

The KS statistic for the discrete version implemented in the *dgof* R-package is computed as follows:

$$D = \sup_x |F_n(x) - F(x)| = \max_i(|F(x_i) - F_n(x_i)|, |F(x_i - \varepsilon) - F_n(x_{i-1})|),$$

where ε is a positive value such that the discontinuities in F are more than some distance ε (see [Arnold and Emerson, 2011, p. 35]).

Thus, to test if our MOEZipf random number generator has a good performance, we have taken the following steps:

1) We have fixed the values of the parameters to be equal to $\alpha = 1.25, 2, 3.5$ and 5, and $\beta = 0.1, 0.25, 0.5, 1, 1.75, 2.5, 3.5$ and 10.

2) For a given combination $(\alpha_0, \beta_0) \in 1.25, 2, 3.5, 5 \times 0.1, 0.25, 0.5, 1, 1.75, 2.5, 3.5, 10$, we have generated 500 samples of size $n = 100$ and $n = 1000$, respectively.

3) For each sample, we have computed the value D and concluded that the sample is not MOEZipf(α_0, β_0) distributed when the corresponding p-value is smaller than the significance level, which is set to be equal to 0.05.

The p-value has been computed in two different ways: first, by means of the classical KS test; second, by means of simulations. The number of times that the null hypothesis has not been rejected for each configuration and sample size appears in Table 2.3 for the classical KS test, while Table 2.4 shows this number after adapting to discrete distributions, which requires simulations.

Table 2.3 shows that, with the exception of three parameter configurations, the null hypothesis has not been rejected at least 95% of the times, meaning that the random number generator performs well. The configurations where the null hypothesis is rejected a number of times different to the nominal value correspond to $\beta = 3.5$ and $n = 1000$ and $\beta = 10$ and the two sample sizes. For those configurations, the probabilities are generally very small and increase very slowly, which makes difficult to generate the data (observe that those configurations have less than 500 samples). Also, for $\alpha = 1.25$, it was not possible to generate the 500 samples. Since the value of α is close to the lower bound of the parameter space, we have found some numerical problems: the low increase in the probabilities caused the generation process to exceed the defined time (30 min) and consequently abort the generation process.

The results in Table 2.4 show that the classical KS test is more conservative than the one adapted to discrete distributions, which is known and has been pointed out in many research papers (see, for instance, Arnold and Emerson [2011]). Nevertheless, with the exception of the same configurations mentioned above, the null hypothesis is not rejected within the range of 94% to 100%. In this case, the probability a type I error is in general very close to the nominal value.

Appendix D.1 contains the R scripts used for generating the random sequences and computing the computation of the KS test.

Table 2.3 Number of random sequences generated from the MOEZipf distribution, and the associated percentage of sequences for which the null hypothesis has not been rejected (classical KS test).

Distribution	N	α	β							
			0.1	0.25	0.5	1	1.75	2.5	3.5	10
MOEZipf	100	1.25	500 (100 %)	500 (99 %)	500 (99 %)	500 (97 %)	500 (97 %)	500 (96 %)	434 (95 %)	389 (75 %)
	1000		476 (100 %)	500 (99 %)	461 (98 %)	482 (96 %)	481 (96 %)	348 (95 %)	388 (20 %)	204 (0 %)
	100	2	500 (100 %)	500 (100 %)	500 (100 %)	500 (99 %)	500 (99 %)	500 (98 %)	500 (98 %)	500 (98 %)
	1000		500 (100 %)	500 (100 %)	500 (100 %)	500 (99 %)	500 (99 %)	500 (98 %)	500 (99 %)	500 (98 %)
	100	3.5	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (99 %)	500 (99 %)
	1000		500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (99 %)
	100	5	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (99 %)
	1000		500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)

Table 2.4 Number of random sequences generated from the MOEZipf distribution, and the associated percentage for which the null hypothesis has not been rejected (p-values of the KS test obtained by simulations).

Distribution	N	α	β							
			0.1	0.25	0.5	1	1.75	2.5	3.5	10
MOEZipf	100	1.25	500 (96 %)	500 (95 %)	500 (96 %)	500 (96 %)	500 (97 %)	500 (97 %)	434 (97 %)	389 (100 %)
	1000		476 (95 %)	500 (97 %)	461 (95 %)	500 (97 %)	500 (97 %)	354 (99 %)	388 (93 %)	231 (0 %)
	100	2	500 (94 %)	500 (94 %)	500 (96 %)	500 (95 %)	500 (95 %)	500 (95 %)	500 (95 %)	500 (97 %)
	1000		500 (95 %)	500 (94 %)	500 (93 %)	500 (95 %)	500 (96 %)	500 (94 %)	500 (96 %)	500 (95 %)
	100	3.5	500 (96 %)	500 (93 %)	500 (96 %)	500 (96 %)	500 (95 %)	500 (96 %)	500 (95 %)	500 (96 %)
	1000		500 (96 %)	500 (93 %)	500 (96 %)	500 (96 %)	500 (95 %)	500 (96 %)	500 (95 %)	500 (96 %)
	100	5	500 (96 %)	500 (96 %)	500 (97 %)	500 (97 %)	500 (97 %)	500 (94 %)	500 (95 %)	500 (95 %)
	1000		500 (96 %)	500 (96 %)	500 (94 %)	500 (95 %)	500 (95 %)	500 (97 %)	500 (95 %)	500 (95 %)

2.4 The Zipf-Poisson extreme distribution

The paper by Ramos et al. [2019] 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 the applications that appear in the work by Ramos et al. [2019], the authors focus on extending the following 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.

2.4.1 Definition and properties

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 α parameter of the Zipf, and a β parameter that when positive is the positive Poisson parameter, and when is negative, it corresponds to the minus positive Poisson parameter. Thus, the parameter space is equal to $(1,+\infty)\times(-\infty,+\infty)$.

The following proposition states the conditions under which any Zipf-PE distribution can be interpreted in terms of maximums or minimums. In addition, it also agrees with the results shown in Theorem 3 where we prove that each distribution in the maximum family has a dual distribution in the minimum family.

Proposition 6. *Let Y be a Zipf-PE distributed r.v. with parameters α and β. Then:*

i) If $\beta > 0$, Y corresponds to a maximum of i.i.d Zipf(α) r.v.'s, where N follows a positive Poisson with parameter β.

ii) If $\beta < 0$, Y corresponds to a minimum of i.i.d Zipf(α) r.v.'s, where N follows a positive Poisson with parameter $-\beta$.

iii) If $\beta = 0$, Y follows a Zipf(α) r.v.'s and, by Theorem 3, Y belongs to maximum as well as to the minimum set of stopped extreme distributions, where N is the degenerate distribution at one.

Proof. Considering Y as an r.v. with a Zipf-PE distribution, the third row of Table 2.2 shows the CDF of Y. After applying some algebra we have:

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

which proves a).

In order to prove b) we have that:

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

Point c) is a direct consequence of Theorem 3. □

Thus, the CDF of any Zipf-PE is equal to:

$$F_{(\alpha,\beta)}(x) = \begin{cases} \frac{e^{\beta\left(\frac{\zeta(\alpha)-\zeta(\alpha,x+1)}{\zeta(\alpha)}\right)}-1}{e^{\beta}-1}, & \beta \in \mathbb{R}\backslash\{0\}, \\ 1 - \frac{\zeta(\alpha,x+1)}{\zeta(\alpha,x)}, & \beta = 0, \end{cases} \tag{2.4.1}$$

where positive values of β correspond to maximums of a positive Po(β) number of copies, and negative values correspond to minimums of a positive Po$(-\beta)$ number of copies. Moreover, as mentioned in Subsection 1.4 of Chapter 1, 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 (2.4.1) and, in this case, Y follows the baseline distribution, that is, the Zipf(α) distribution.

From (2.4.1) $\forall\alpha > 1$ and $x \geq 2$, one can obtain the PMF of Y as follows:

$$\begin{aligned} P(Y = x) &= F_{(\alpha,\beta)}(x) - F_{(\alpha,\beta)}(x-1) = \\ &= \frac{e^{\beta\left(\frac{\zeta(\alpha)-\zeta(\alpha,x+1)}{\zeta(\alpha)}\right)} - e^{\beta\left(\frac{\zeta(\alpha)-\zeta(\alpha,x)}{\zeta(\alpha)}\right)}}{e^{\beta} - 1} \\ &= \begin{cases} \frac{e^{\beta} e^{\frac{-\beta\zeta(\alpha,x)}{\zeta(\alpha)}}\left(e^{\frac{\beta x^{-\alpha}}{\zeta(\alpha)}}-1\right)}{e^{\beta}-1}, & \beta \in \mathbb{R}\backslash\{0\}, \\ \frac{x^{-\alpha}}{\zeta(\alpha)}, & \beta = 0. \end{cases} \end{aligned} \tag{2.4.2}$$

For $x = 1$,

$$P(Y = 1) = F_{(\alpha,\beta)}(1) = \begin{cases} \frac{e^{\frac{\beta}{\zeta(\alpha)}}-1}{e^{\beta}-1}, & \beta \in \mathbb{R}\backslash\{0\}, \\ \frac{1}{\zeta(\alpha)}, & \beta = 0. \end{cases} \tag{2.4.3}$$

Observe that (2.4.3) is equal to (2.4.2) at $x = 1$. Thus (2.4.3) is the PMF in the entire support.

Figure 2.7 shows the PMFs of the Zipf-PE distribution for $\alpha = 2.1$ and different values of the β parameter. On the left-hand side, the probabilities are at the standard scale and, on the right-hand side, they are plotted in log-log scale. Note that the β 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.

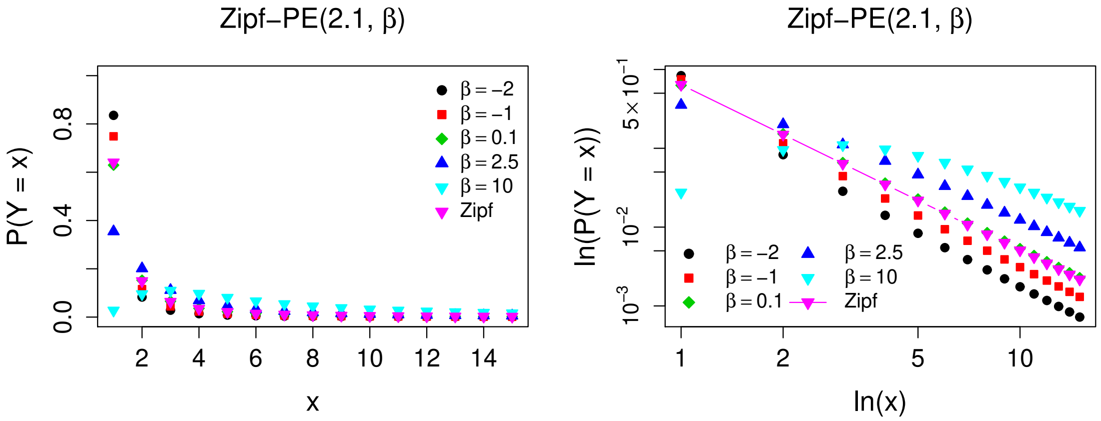

Fig. 2.7 PMFs of the Zipf-PE distribution for $\alpha = 2.1$ and $\beta = -2, -1, 0.1, 2.5$ and 10. On the left-hand side, the probabilities are shown at standard scale and, on the right-hand side, they are plotted in log-log scale.

The next proposition states that the probability at one of an r.v. with a Zipf-PE(α, β) distribution is always smaller (larger) than the probability at one of a Zipf distribution with the same parameter α, depending on the sign of β. Negative values of β 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 7. *Let Y and X be two r.v.'s, such that $Y \sim$ Zipf-PE(α, β) 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 (2.4.3) 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 it exists a value $x_0 \in (0,1)$, such that $g'(x_0)=0$. Differentiating, one has:

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

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, $g(\frac{1}{\zeta(\alpha)}) \geq 0$. □

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 8. *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(α,β) and $X \sim$ Zipf(α). As mentioned in Section 1.1, 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(α,β) distribution converges if, and only if, the *k-th* moment of the Zipf(α) converges, that is, when $\alpha > k+1$. □

Figure 2.8 shows the behavior of the mean as: a function of α for $\beta=-1,-0.5,1.5$ and 3 (on the left-hand side); and as a function of β for $\alpha=2.5,4.8,7.5,20$ (on the right-hand side). A similar plot for the variance appears in Figure 2.9: on left-hand side as a function of α for $\beta=-1,-0.5,1.5$ and 3; and on the right-hand side as a function of β 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 α, but they decrease faster as β becomes smaller. On the right-hand side of both figures can be observed that the $E[Y]$ and $VAR[Y]$ are increasing functions of β, with a slope that decreases when α increases.

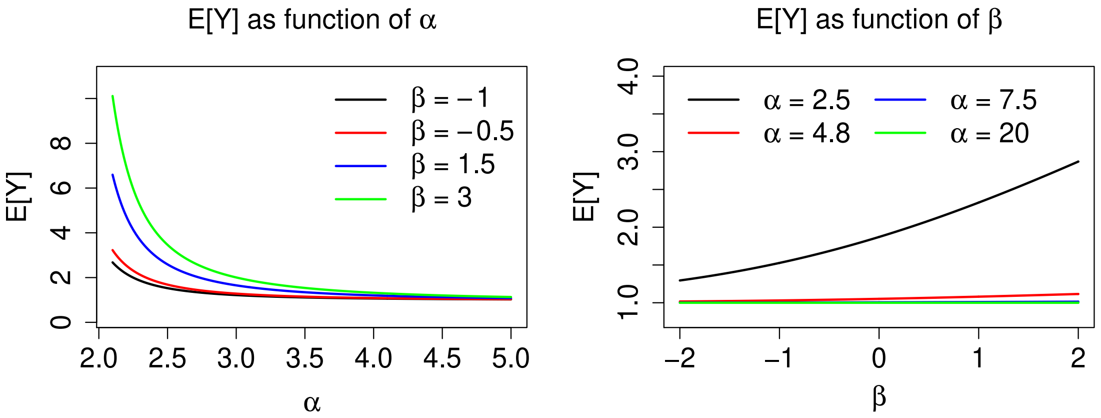

Fig. 2.8 Mean values of a Zipf-PE(α,β) distribution. On the left-hand side: as a function of α for $\beta=-1,-0.5,1.5$ and 3. On the right-hand side: as a function of β for $\alpha=2.5,4.8,7.5$ and 20.

Proposition 9. *Let Y and X be two r.v.'s, such that $Y\sim$ Zipf-PE(α,β) and $X\sim$ Zipf(α). Then, the ratio of two consecutive probabilities of Y is equal to:*

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

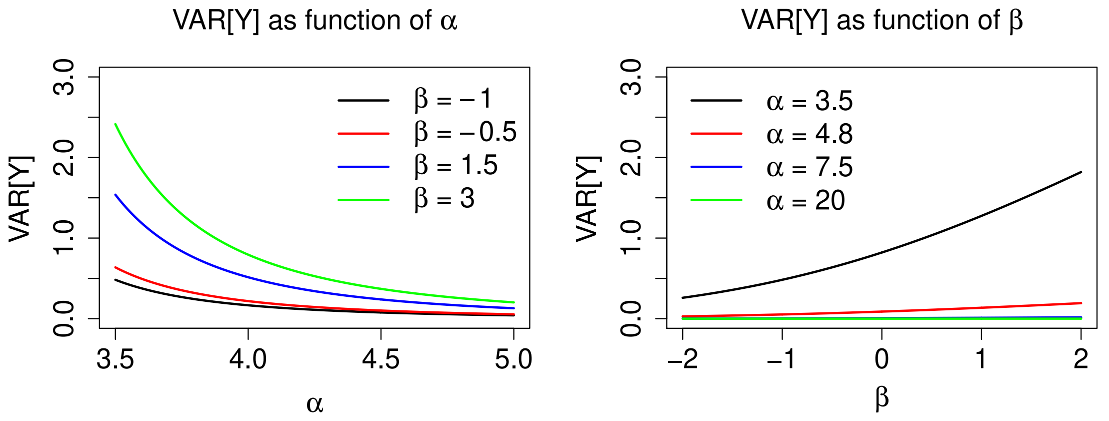

Fig. 2.9 Variance values of a Zipf-PE(α,β) distribution. On the left-hand side: as a function of α for $\beta = -1, -0.5, 1.5$ and 3. On the right-hand side: as a function of β for $\alpha = 3.5, 4.8, 7.5$ and 20.

Proof. From (2.4.2) 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 2.10 shows the behavior of this ratio for $\alpha = 2.1$ and $\beta = -3$ and 3. The ratio of the Zipf(α) is also included in order to facilitate comparison between both distributions. Note that, when $\beta > 0$, the ratio associated with the Zipf-PE(α,β) 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 β value, those values in the tail of the distribution behave similarly to those of the Zipf distribution, which is proven in the Theorem 7. This theorem, based on the results stated in Theorem 5, establishes the relationship between the tail of the Zipf-PE and the tail of the Zipf distributions.

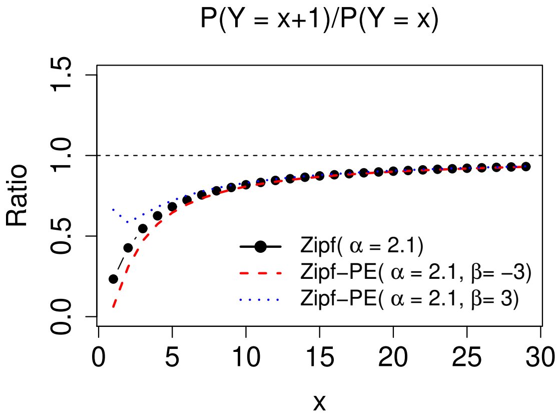

Fig. 2.10 Ratio of two consecutive Zipf-PE probabilities for $\alpha = 2.1$, with $\beta = -3$ and 3, respectively.

Theorem 7. *The tail of an r.v. $Y \sim$ Zipf-PE(α, β) is asymptotically related to the tail of an r.v. $X \sim$ Zipf(α), in such a way that:*

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

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

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 5 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:

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

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

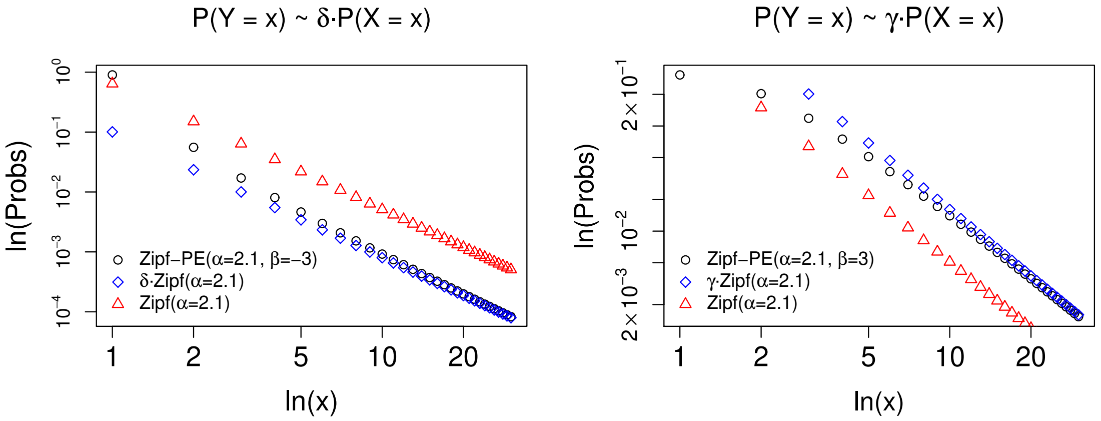

Fig. 2.11 The probabilities of the Zipf and Zipf-PE distributions with the same α parameter. On the left-hand side, jointly with $\delta = -\beta\, e^{\beta}/(1-e^{\beta})$ times the probability of the Zipf. On the right-hand side, jointly with $\gamma = \beta/(1-e^{\beta})$ times the probability of 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.

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 2.11 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 $x \geq 10$.

In what follows, we introduces the methodology used for generating random data following a Zipf-PE(α, β) distribution.

2.4.2 Random data generation

The next proposition shows how to generate data from a Zipf-PE distribution. The approach applied to the Zipf-PE distribution is similar to as the one adopted in Subsection 2.3.3. The performance of the proposed methodology is tested later by means of the KS goodness-of-fit test.

Proposition 10. *Generating random data from a Zipf-PE(α, β) distribution is done by applying the inversion method to the Zipf(α) distribution using a u' value equal to:*

$$u' = \frac{\log(u(e^{\beta} - 1) + 1)}{\beta},$$

independently of whether the distribution family is defined in terms of maximums or minimums, and where $u \in (0, 1)$ has been randomly selected from an Uniform distribution.

Proof. By means of Theorem 4 and the h_N that appears in the second row of Table 2.1, we have the following:

- if the data has to be generated for a distribution in the minimum family $\beta > 0$, then,

$$\begin{aligned} & u' = h_N^{-1}(u(1 - h_N(0; \theta)) + h_N(0; \theta); \theta) = h_N^{-1}(u; \theta) \\ & \Leftrightarrow \frac{e^{\beta u'} - 1}{e^{\beta} - 1} = u \Leftrightarrow e^{\beta u'} = u e^{\beta} - u + 1 \Leftrightarrow u' = \frac{\log(u(e^{\beta} - 1) + 1)}{\beta}; \end{aligned}$$

- if the data has to be generated for a distribution in the minimum family $\beta < 0$, then,

$$\begin{aligned} & u' = 1 - h_N^{-1}(1 - u(1 - h_N(0, \theta)); \theta) = 1 - h_N^{-1}(1 - u; \theta) \\ & \Leftrightarrow 1 - \frac{e^{\beta(1-u')} - 1}{e^{\beta} - 1} = u \Leftrightarrow e^{\beta(1-u')} = ue^{\beta} - u + 1 \\ & \Leftrightarrow u' = 1 - \left(1 - \frac{\log(u(e^{\beta} - 1) + 1)}{\beta}\right) \Leftrightarrow u' = \frac{\log(u(e^{\beta} - 1) + 1)}{\beta}. \end{aligned}$$

As stated in Proposition 6, the positive Poison parameter is $\beta = -\beta$.

□

For ensuring the quality of our data generation process, we take a similar approach to the one followed for the MOEZipf(α, β) distribution and have generated 500 samples with sizes 100 and 1000, respectively, using all the possible combinations of the parameter values $\alpha = 1.25, 2, 3.5$ and 5, and $\beta = -5, -2.25, -1, -0.25, 0.25, 1, 2.5$ and 5. Then, we applied the KS test as described in Subsection 2.3.3. Tables 2.5 and 2.6 summarize: the number of sequences generated for each combination of parameters; and the percentages of them that did not reject the null hypothesis that the data come from a Zipf-PE(α, β), with a significance level of $\alpha = 0.05$. The first table mentioned presents the results of the classical KS test and the second one, those obtained with the discrete version of the test using a Monte Carlo approach for computing the p-value.

In general, the results in both cases are good, with the classical KS test being, again, more conservative. The random generation process presents numerical problems only when $\alpha = 1.25$ and $\beta = 5$. This is a consequence of the fact that, for large βs, the probabilities increase very slowly.

Table 2.5 Number of random sequences generated from the Zipf-PE distribution, as well as the associated percentage of the sequences that have passed the classical Kolmogorov-Smirnov test.

Distribution	N	α	β							
			-5	-2.25	-1	-0.25	0.25	1	2.5	5
Zipf-PE	100	1.25	500 (100 %)	500 (99 %)	500 (97 %)	500 (97 %)	500 (97 %)	500 (97 %)	500 (97 %)	500 (97 %)
	1000		500 (100 %)	500 (99 %)	471 (96 %)	494 (95 %)	477 (96 %)	417 (96 %)	421 (95 %)	315 (0 %)
	100	2	500 (100 %)	500 (100 %)	500 (100 %)	500 (99 %)	500 (99 %)	500 (99 %)	500 (99 %)	500 (98 %)
	1000		500 (100 %)	500 (100 %)	500 (100 %)	500 (99 %)	500 (100 %)	500 (99 %)	500 (99 %)	500 (98 %)
	100	3.5	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)
	1000		500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)
	100	5	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)
	1000		500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)

Table 2.6 Number of random sequences generated from the Zipf-PE distribution, as well as the associated percentage of the sequences that have passed the Kolmogorov-Smirnov test for discrete distributions.

Distribution	N	α	β							
			-5	-2.25	-1	-0.25	0.25	1	2.5	5
Zipf-PE	100	1.25	500 (95 %)	500 (96 %)	500 (96 %)	500 (96 %)	500 (96 %)	500 (96 %)	500 (98 %)	500 (99 %)
	1000		500 (95 %)	500 (96 %)	471 (96 %)	494 (94 %)	456 (96 %)	417 (97 %)	401 (100 %)	302 (28 %)
	100	2	500 (98 %)	500 (95 %)	500 (95 %)	500 (97 %)	500 (96 %)	500 (94 %)	500 (96 %)	500 (96 %)
	1000		500 (96 %)	500 (94 %)	500 (96 %)	500 (96 %)	500 (96 %)	500 (96 %)	500 (93 %)	500 (95 %)
	100	3.5	500 (97 %)	500 (95 %)	500 (94 %)	500 (95 %)	500 (95 %)	500 (95 %)	500 (96 %)	500 (97 %)
	1000		500 (94 %)	500 (95 %)	500 (96 %)	500 (96 %)	500 (94 %)	500 (95 %)	500 (95 %)	500 (94 %)
	100	5	500 (100 %)	500 (97 %)	500 (94 %)	500 (96 %)	500 (97 %)	500 (94 %)	500 (97 %)	500 (95 %)
	1000		500 (96 %)	500 (96 %)	500 (96 %)	500 (94 %)	500 (96 %)	500 (95 %)	500 (95 %)	500 (97 %)

Chapter 3

The Zipf Poisson-stopped-sum extension

In this chapter we define and develop the Zipf-Poisson-stopped-sum (Zipf-PSS) family of distributions which is obtained by extending the Zipf distribution according to the PSS definition. In the first part of this chapter we review the concept of PSS. The second part of this chapter is devoted to the definition and the analysis of the main properties of the Zipf-PSS family of distributions. The results included here appear in the paper: *"The Zipf-Poisson-stopped-sum distribution with an application for modeling the degree sequence of social networks"* [Duarte-López et al., 2020b].

3.1 Poisson-stopped-sum distributions

PSSs [Johnson et al., 2005] appear to be the distribution of an r.v. Y in relation to a branching process, where we initially assume to have N individuals or experimental units, and that each individual gives rise to X_i new individuals in a given period of time, whereby the total number of first-generation individuals is equal to:

$$Y = X_1 + X_2 + ... + X_N, \tag{3.1.1}$$

where N is assumed to be a Poisson r.v. and X_i, $i = 1, \ldots, n$, are i.i.d r.v.'s with a given distribution that may be either continuous or discrete. It is assumed that when $N = 0$, $Y = 0$, meaning that if there are no initial individuals, then no new ones are generated. If X_i has a distribution with parameter vector θ and parameter space $\Theta \subseteq \mathbb{R}^n$, then the distribution of Y has parameter space $\{(\lambda, \theta) | \lambda \in (0, +\infty) \text{ and } \theta \in \Theta\}$. The r.v. X is obtained when $N = 1$, which takes place with probability $\lambda e^{-\lambda}$, and the Poisson distribution is obtained when X_i has a degenerate distribution at one. The distributions of N and X are called *primary* and *secondary* distributions, respectively. Denoting by $G_x(z)$ the PGF of X, one has that the PGF of Y defined as in (3.1.1) is equal to:

$$G_Y(z) = e^{\lambda\,(G_X(z)-1)},\, \lambda > 0, \tag{3.1.2}$$

and it is defined at least for $|z| \leq 1$. Since the PGF of a PSS is the composition of the PGF of the Poisson and the PGF of the distribution of X, these families are also know as *compound distributions*.

A *discrete compound Poisson distribution* (DCP) is defined as a PSS with a discrete secondary distribution [Feller, 1971; Zhang and Li, 2016]. The notation for an r.v. with a DCP distribution is $Y \sim DCP(\alpha_1\,\lambda,\, \alpha_2\,\lambda, \ldots)$, with λ being the Poisson parameter and $\alpha_i = P(X_i = i)$. Its approximation of order r, Y_r, is defined by approximating the Taylor's expansion of $G_X(z)$ at $z = z_0$, by the first terms up to order r at a given point $z = z_0$, and its distribution is denoted by: $DCP(\alpha_1\,\lambda,\, \alpha_2\,\lambda, \ldots,\, \alpha_r\,\lambda)$.
One of the most studied approximations is the second-order approximation, which is the Hermite distribution. Several works on the third and fourth approximations can be found in the literature under the name *sttutering Poisson distributions* [Patel, 1976] or *3rd and 4th order Hermite distribution* [Puig and Valero, 2007]. In addition, Puig and Valero [2006] have proved that under certain regularity conditions, a bi-parametric family of discrete distributions is partially closed under addition if, and only if, it is a PSS. By partially closed under addition one understands that, the sum of N independent copies of a r.v. X with distribution in a given family, also belongs to the same family. More recently, Valero et al. [2013] have characterized those PSS that are also mixed Poisson models.

PSS are widely applied in dissimilar fields. For example, Podur et al. [2010] uses PSS to model the annual area burned by forest fires in the Canadian province of Ontario; and the work by Low et al. [2016] compares the performance of several PSS used to model citation data. For its application to insurance data, see Meng and Gao [2018]. The PSS distributions appear naturally in many data generation processes. Thus, the parameter estimates provide important insights about the data generation mechanism.

Next section formally defines the Zipf-PSS family of distributions.

3.2 Definition

The Zipf-PSS model is obtained by assuming that the distribution of the r.v.'s X_i in (3.1.1) is the Zipf(α) distribution. By substituting (1.1.12) in (3.1.2), the PGF of r.v. Y with a

Zipf-PSS(α,λ) distribution is set to be equal to:

$$G_Y(z) = e^{\lambda\left(\frac{Li_\alpha(z)}{\zeta(\alpha)}-1\right)}, \lambda > 0, \alpha > 1, |z| \leq 1. \tag{3.2.1}$$

This family of distributions is a bi-parametric model with parameter space equal to $\{(\alpha,\lambda) \in (1,+\infty)\times(0,+\infty)\}$, and it belongs to the class of DCP distributions with parameters:

$$\left(\frac{\lambda}{\zeta(\alpha)}, \frac{\lambda}{2^\alpha\zeta(\alpha)}, \frac{\lambda}{3^\alpha\zeta(\alpha)}, \ldots\right).$$

The probabilities of a Zipf-PSS(α,λ) distribution may be computed using the generalization of the Panjer recursion that appears in Panjer [1981, p. 24-25] and in Sundt and Jewell [1981, p. 38], which says that:

$$P(Y=0) = \begin{cases} \frac{1-aP(X=0)}{1-a}, & \text{if } a \neq 0 \\ e^{-b[1-P(X=0)]}, & \text{if } a = 0, \end{cases} \tag{3.2.2}$$

and, for $x = 1, 2, \ldots$

$$P(Y=x) = \frac{1}{1-aP(Y=0)} \sum_{s=1}^{x} \left(a + \frac{bs}{x}\right) P(X=s)P(Y=x-s), \tag{3.2.3}$$

for given constants a and b. The work of Panjer [1981] also shows that if N has a Poisson distribution, then $a = 0$ and $b = \lambda$.

Thus, if $X \sim Zipf(\alpha)$ and $Y \sim$ Zipf-PSS(α,λ), from (3.2.2) one has that

$$P(Y=0) = e^{-\lambda}, \tag{3.2.4}$$

and from (3.2.3) that

$$P(Y=x) = \frac{\lambda}{\zeta(\alpha)x} \sum_{s=1}^{x} s^{1-\alpha} P(Y=x-s), x \geq 1. \tag{3.2.5}$$

Figure 3.1 shows the probabilities of the Zipf-PSS distribution for different values of α and λ. On the left-hand side, the plots are in normal scale while on the right-hand side are shown in log-log scale. The probabilities obtained for $\alpha = 2.3$ and different values of λ can be observed on the top-side of the figure, while the bottom-side contain the probabilities for $\lambda = 5$ and different values of α. We observe that the highest probabilities are obtained at the initial values when α and λ are small. Additionally, looking at top-right part of the figure, it is possible to compare the behavior of the Zipf-PSS probabilities with those achieved by a Zipf distribution with the same α parameter. Note that even though it looks like a straight line for smaller λ values, the probabilities obtained are different from those of the Zipf. In comparison, the larger the value of λ becomes, more curvature is shown by the PMF. This bring us to consider the λ parameter as a measure of departure from the Zipf distribution.

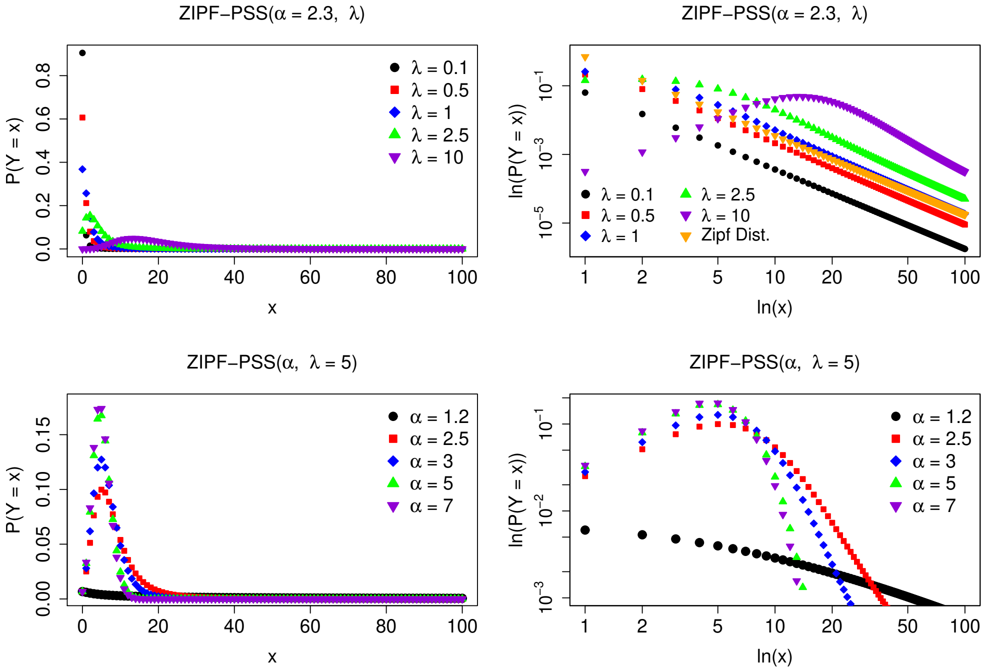

Fig. 3.1 PMF of the Zipf-PSS(α,λ) distribution. For $\alpha = 2.3$ and $\lambda = 0.1, 0.5, 1, 2.5$ and 10, (top-left) in normal scale and (top-right) in log-log scale. For $\lambda = 5$ and $\alpha = 1.5, 2.5, 3, 5, 7$, (bottom-left) in normal scale and (bottom-right) in log-log scale. In normal scale $x = 0, \dots, 100$, and $x = 1, \dots, 100$ in log-log scale.

3.3 Properties

3.3.1 Moments and index of dispersion

In Satterthwaite [1942] and Johnson et al. [2005] it is proved that the moments of any PSS are functions of the moments of the underlying secondary distribution which, in our case, is the Zipf distribution. These results give rise to the following propositions.

Proposition 11. *The k-th moment of the Zipf-PSS*(α,λ) *distribution is finite if, and only if,* $\alpha > k+1$.

Proof. Let $M(t), t \in \mathbb{R}$ be the moment generating function (MGF) of the Zipf-PSS distribution. According to equation (3.2.1), it is equal to:

$$M(t) = G_Y(e^t) = e^{\lambda\left(\frac{Li_\alpha(e^t)}{\zeta(\alpha)} - 1\right)}. \tag{3.3.1}$$

Given that $E[Y^k] = M^{(k)}(t)|_{t=0}$, the $k-th$ moment depends only on the derivatives of the PGF of the secondary distribution. Satterthwaite [1942] establishes the relationship between the moments of any PSS distribution and those of the underlying secondary distribution. Since the existence of the Zipf-PSS moments depends on the existence of Zipf moments, the moment of order k of the Zipf-PSS exists if, and only if, $\alpha > 1$. □

In what follows, we obtain the exact values for the mean and the variance.

Proposition 12. *The mean and the variance of a Zipf-PSS*(α, λ) *distribution are respectively equal to:*

$$E[Y] = \lambda\, E[X] = \lambda\, \frac{\zeta(\alpha-1)}{\zeta(\alpha)}, \alpha > 2, \tag{3.3.2}$$

and

$$Var[Y] = \lambda\, \frac{\zeta(\alpha-2)}{\zeta(\alpha)}, \alpha > 3, \tag{3.3.3}$$

Proof. The expectation of the Zipf-PSS, as well as its variance expression, can be derived either by means of the derivatives of the moment generating function or, what is more straightforward, from the *Law of Total Expectation* and the *Law of Total Variance*, which state that:

$$\begin{aligned} E[Y] &= E[N]\,E[X], \text{ and} \\ Var[Y] &= E[N]\,Var[X] + E^2[X]\,Var[N]. \end{aligned} \tag{3.3.4}$$

Let us take into account that if N is Poisson distributed, $E[N] = Var[N] = \lambda$, and that $E[X]$ and $Var[X]$ appear in (1.1.7) and (1.1.8), respectively. Then, in a simple substitution, we have:

$$Var[Y] = \lambda\, \frac{\zeta(\alpha-2)\,\zeta(\alpha) - \zeta(\alpha-1)^2}{\zeta(\alpha)^2} + \left(\frac{\zeta(\alpha-1)}{\zeta(\alpha)}\right)^2 \lambda = \lambda\, \frac{\zeta(\alpha-2)}{\zeta(\alpha)}. \qquad \square$$

Figure 3.2 shows the behavior of the mean of the Zipf-PSS distribution. On the left-hand side as a function of α, for $\lambda = 0.5, 1, 2.5$ and 3; and, on the right-hand side as a function of λ, for $\alpha = 2.5, 3, 5$ and 10. Observe that the expected value of a Zipf-PSS is a decreasing function of α for any value of λ. This makes sense, since $E[Y] = \lambda\, E[X]$ and the mean of a

Zipf distribution is a decreasing function of α because increasing α leads to the probability concentrating in the first values. On the left-hand side, one can observe that the mean is an increasing function of λ for any α. This is a consequence of the fact that increasing λ increases the number of terms in (3.1.1) and, consequently, the value of Y.

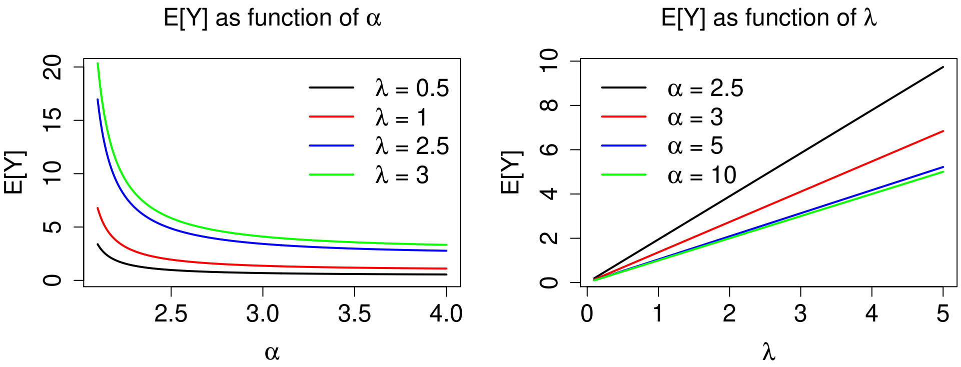

Fig. 3.2 Expected values of a Zipf-PSS(α,λ) distribution. On the left-hand side as a function of α, for $\lambda = 0.5, 1, 2.5$ and 3; and, on the right-hand side as a function of λ, for $\alpha = 2.5, 3, 5$ and 10.

A similar plot is presented in Figure 3.3 in regard to the variance of the Zipf-PSS distribution. On the left-hand side as a function of α for several values of λ, and on the right-hand side as a function of λ for several values of α. The variance of the distribution behaves quite similar to the mean. The variance is clearly a decreasing function of α, and it decreases faster as the λ value becomes smaller. From Proposition 12, the mean and the variance are linear functions of λ with a slope that decreases when α increases, which is observed on the right-hand side of figures 3.2 and 3.3, respectively.

Taking into account (3.3.2), (3.3.3) and that $\zeta(\alpha)$ is a decreasing function of α, one has that if $\alpha > 2$, then

$$ID = \frac{Var[Y]}{E[Y]} = \frac{\zeta(\alpha-2)}{\zeta(\alpha-1)} > 1.$$

Consequently, the Zipf-PSS is over-dispersed compared to a Poisson distribution with the same mean, which is consistent with the relationship between the ID of Y and the ID of X and N, pointed out in Johnson et al. [2005, p. 386].

In what follows, we provide a condition under which the variance of the Zipf-PSS is larger than the variance of a Zipf distribution with the same α parameter.

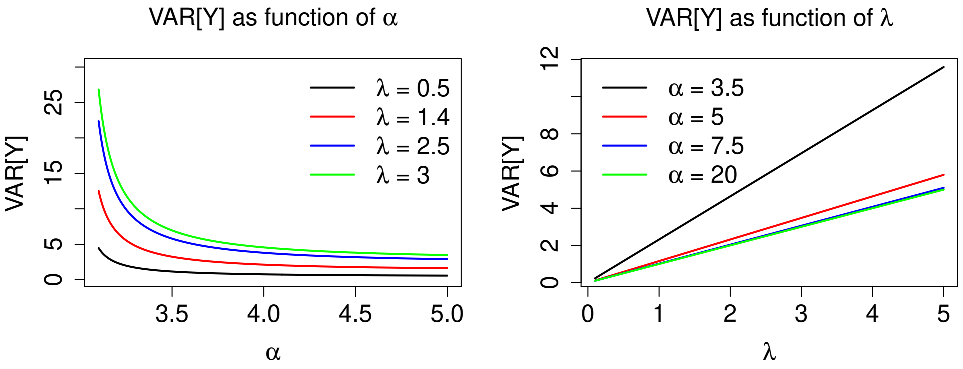

Fig. 3.3 Variance values of a Zipf-PSS(α,λ) distribution. On the left-hand side as a function of α, for $\lambda = 0.5, 1.4, 2.5$ and 3; and, on the right-hand side as a function of λ for $\alpha = 3.5, 5, 7.5$ and 20.

Proposition 13. *Let $Y \sim$ Zipf-PSS(α,λ) and $X \sim Zipf(\alpha)$. If $\alpha > 3$, then one has that $Var[Y] \geq Var[X]$ if, and only if,*

$$\lambda \geq 1 - \frac{\zeta^2(\alpha-1)}{\zeta(\alpha-2)\,\zeta(\alpha)}.$$

Proof. From (3.3.4), and given that $E[N] = Var[N] = \lambda$, one has that:

$$Var[Y] = \lambda\,(Var[X] + E^2[X]).$$

Thus,

$$Var[Y] \geq Var[X] \Leftrightarrow \lambda\,(Var[X] + E^2[X]) \geq Var[X] \Leftrightarrow \lambda \geq \frac{Var[X]}{Var[X] + E^2[X]} = \frac{1}{1 + \frac{E^2[X]}{Var[X]}}.$$

Taking into account (1.1.7) and (1.1.8), one has that:

$$1 + \frac{E^2[X]}{Var[X]} = \frac{\zeta(\alpha-2)\,\zeta(\alpha)}{\zeta(\alpha-2)\,\zeta(\alpha) - \zeta^2(\alpha-1)},$$

and thus,

$$Var[Y] \geq Var[X] \Leftrightarrow \lambda \geq \frac{\zeta(\alpha-2)\,\zeta(\alpha) - \zeta^2(\alpha-1)}{\zeta(\alpha-2)\,\zeta(\alpha)} = 1 - \frac{\zeta^2(\alpha-1)}{\zeta(\alpha-2)\,\zeta(\alpha)}.$$

Given that the Riemann zeta function is always positive, then if λ is larger than one (which is usually the case when fitting real data), the condition is always satisfied and the Zipf-PSS has a larger variance than the corresponding Zipf distribution. □

As already mentioned, according to Puig and Valero [2006, p. 333], all the PSSs distributions have the property of being partially closed under addition. For the particular case of the Zipf-PSS family of distributions, this property is established in the following proposition.

Proposition 14. *Let Y_i, $i = 1, \dots, n$ be n i.i.d. r.v.'s with a Zipf-PSS(α, λ) distribution. The r.v. Y defined as:*

$$Y = Y_1 + Y_2 + \ldots + Y_n$$

Follows a Zipf-PSS($\alpha, n\lambda$) distribution.

Proof. By (3.2.1) the PGF of the r.v. Y is equal to:

$$G_Y(z) = (G_{Y_i}(z))^n = e^{n\lambda(\frac{Li_\alpha(z)}{\zeta(\alpha)}-1)},$$

which proves the proposition. □

This property is specially interesting in order to analyze the evolution of a system, for instance a graph, over time. Figure 3.4 illustrates this property.

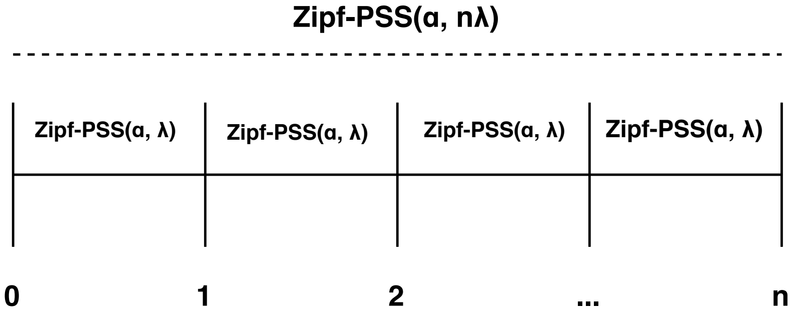

Fig. 3.4 Independents Zipf-PSS(α, λ) distributions with the same λ parameter.

3.3.2 Other important properties

Here, we first show the relationship between the parameters of the distribution and its modality. Later, we compare the ratio of two consecutive probabilities of a Zipf-PSS(α, λ)

with the one obtained from a Zipf distribution with the same α parameter. At the end of the subsection we prove that the Zipf-PSS is a MP distribution.

Proposition 15. *If $Y \sim$ Zipf-PSS(α,λ) and denoting by $\lambda_0 = 2\zeta(\alpha)(1-2^{-\alpha})$, one has that:*

i) if $\lambda \in (0,\zeta(\alpha)]$, then the distribution of Y is unimodal with a pseudo-mode at zero, and it is also log-concave;

ii) if $\lambda \in (\zeta(\alpha),\lambda_0]$, then the distribution of Y has a mode at one;

iii) if $\lambda \in (\lambda_0,+\infty)$, then the distribution of Y has a mode equal to or larger than two.

Proof. i) It is necessary to see that $P(Y=0) \geq P(Y=1)$

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

Hence, point i) holds when $\lambda \in (0,\zeta(\alpha)]$.
ii) Applying (3.2.5) for $x=1$ and $x=2$, one has that:

$$P(Y=2) \geq P(Y=1) \Leftrightarrow \frac{\lambda e^{-\lambda}}{2\zeta(\alpha)}\left[\frac{\lambda}{\zeta(\alpha)} + 2^{1-\alpha}\right] \geq \frac{\lambda e^{-\lambda}}{\zeta(\alpha)} \tag{3.3.5}$$

$$\Leftrightarrow \frac{1}{2}\left(\frac{\lambda}{\zeta(\alpha)} + 2^{1-\alpha}\right) \geq 1 \Leftrightarrow \lambda \geq 2\zeta(\alpha)(1-2^{-\alpha}). \tag{3.3.6}$$

To prove point ii), it is only necessary to see that for $\lambda \in (\zeta(\alpha), 2\zeta(\alpha)(1-2^{-\alpha}))$, $P(Y=1) \geq P(Y=0)$ and that $P(Y=1) \leq P(Y=2)$. Taking into account that:

$$P(Y=0) \leq P(Y=1) \Leftrightarrow e^{-\lambda} \leq \frac{\lambda e^{-\lambda}}{\zeta(\alpha)} \Leftrightarrow \zeta(\alpha) \geq \lambda, \tag{3.3.7}$$

and based on (3.3.5), one has that, if $\lambda \in (\zeta(\alpha),\, 2\zeta(\alpha)(1-2^{-\alpha}))$, there is a mode at one.
iii) This is a straightforward consequence of points i) and ii). □

Given that, $\zeta(\alpha) \to +\infty$ when $\alpha \to 1$, and that it tends to 1 when $\alpha \to +\infty$, one has that when $\alpha \to 1$, then $\lambda_0 \to +\infty$ and thus the parameter λ must be larger in order to have a mode larger than one. Also, when $\alpha \to +\infty, \lambda_0 \to 2$, and thus the distribution has a mode larger than one for any $\lambda \in (2,+\infty)$.

The next proposition establishes the relationship between the ratio of two consecutive probabilities of the Zipf-PSS(α,λ) and the Zipf(α) distributions.

Proposition 16. *The ratio of two consecutive probabilities of r.v. Y with a Zipf-PSS(α,λ) distribution is related to the same ratio of r.v. X with a Zipf(α) distribution by means of:*

$$\frac{P(Y=x+1)}{P(Y=x)}=\frac{P(X=x+1)}{P(X=x)}\left(\frac{x}{x+1}\right)^{1-\alpha}h(x;\alpha,\lambda),$$

where $h(x;\alpha,\lambda)$ is a ratio of two linear combinations of the probabilities $P(Y=i)$ for $i=0,\dots,x$.

Proof. From (3.2.5) one has that for $x=1,2,...$

$$P(Y=x+1)=\frac{\lambda}{\zeta(\alpha)\,(x+1)}\sum_{s=1}^{x+1}s^{1-\alpha}P(Y=x+1-s).$$

Dividing this expression by $P(Y=x)$, as in (3.2.5), one has that:

$$\frac{P(Y=x+1)}{P(Y=x)}=\frac{\frac{1}{x+1}}{\frac{1}{x}}\frac{\sum_{s=1}^{x+1}s^{1-\alpha}P(Y=x+1-s)}{\sum_{s=1}^{x}s^{1-\alpha}P(Y=x-s)}. \tag{3.3.8}$$

Thus, denoting by $h(x;\alpha,\lambda)$ the second ratio on the right-hand side of the equation one has that:

$$\frac{P(Y=x+1)}{P(Y=x)}=\frac{x}{x+1}h(x;\alpha,\lambda). \tag{3.3.9}$$

Taking into account that the ratio of two consecutive probabilities of the Zipf(α) distribution is equal to $(\frac{x}{x+1})^{\alpha}$, (3.3.9) is equivalent to:

$$\frac{P(Y=x+1)}{P(Y=x)}=\frac{P(X=x+1)}{P(X=x)}\left(\frac{x}{x+1}\right)^{1-\alpha}h(x;\alpha,\lambda). \qquad \square$$

Figure 3.5 shows the behavior of this ratio for $\alpha=2.3$ and two different values of the λ parameter. For comparison to the Zipf distribution, the ratio of the Zipf(α) probabilities is also included. When the λ value is close to zero, the ratio of the two consecutive probabilities of the Zipf-PSS(α,λ) quickly converges to the ratio of the same two consecutive Zipf probabilities. Otherwise, if the λ parameter is large, the ratio of the probabilities also converges to that of the Zipf, but the convergence takes longer. Moreover, both ratios tend to one when x tends to infinity. This implies that $h(x;\alpha,\lambda)$ also tends to one when x increases. In addition, this is another manner of observing the flexibility of the Zipf-PSS distribution, especially in the first integer values. Observe that increasing the value of x (i.e., considering those values in the tail of the distribution) leads to similar behavior as in the Zipf distribution, which is proved by Theorem 9 in the next subsection.

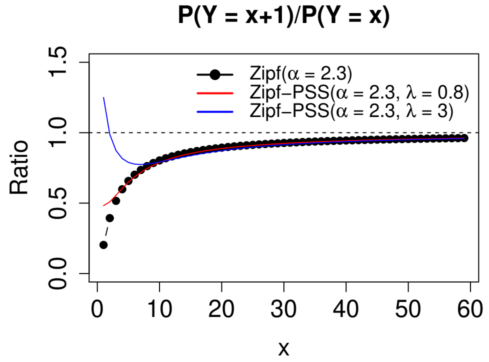

Fig. 3.5 Ratio of two consecutive Zipf-PSS probabilities for $\alpha = 2.3$ and $\lambda = 0.8$ and 3. Jointly with the same ratio of probabilities of the Zipf for $\alpha = 2.3$.

In Theorem 3 of Valero et al. [2013, p. 1833], it is proven that a non-negative integer PSS distribution with finite mean is an MP if, and only if, the zero-truncation of its secondary distribution is an MZTP distribution. As a consequence of this theorem, we have the following result:

Theorem 8. *Denoting by Zipf-PSS(α, λ) the distribution of parameters α and λ in the Zipf-PSS family, one has that for any $\alpha > 1$ and $\lambda > 0$, the distribution is an MP distribution.*

Proof. First observe that, by definition, the Zipf-PSS family has the Zipf as a secondary family of distributions. Moreover, given that any Zipf distribution does not contain the zero value in its support, it is equal to its zero-truncation version. Point a) of Theorem 2 in Chapter 1 states that the Zipf is an MZTP. Consequently, the Zipf-PSS has a secondary distribution that is an MZTP; and, thus, it is an MP, as a consequence of Theorem 3 of Valero et al. [2013, p. 1833]. □

It is known that any mixed distribution has a larger variance than the original distribution (see, Karlis and Xekalaki [2005], Section 2). Thus, this theorem agrees with the result proved in Section 3.3.1 that says that the ID of r.y. Y with a Zipf-PSS distribution is larger than one for $\alpha > 2$.

3.3.3 Linear tail behavior

In order to analyze the tail behavior of the Zipf-PSS distribution we have to use the notion of RVF introduced in Section 2.1.2 of Chapter 2. In addition, the analysis is based on the result achieved by Jessen and Mikosch [2006, p. 177] which is stated in the next Lemma.

Lemma 2. *Assuming that X is a regularly varying r.v. with index $\alpha > 0$, that $E[N] < +\infty$ and that $P(N > x) = o(P(X > x))$, then as $x \to +\infty$,*

$$P(Y > x) \sim E[N]P(X > x). \tag{3.3.10}$$

In Section 2.1.2 we have shown that the Zipf distribution is a RVF. The second condition of Lemma 2 holds because N is Poisson distributed. Next two lemmas are required to proof the third condition of Lemma 2.

Lemma 3. *Let $f(x)$ be defined as $f(x) = \frac{\lambda^x x^\alpha}{x!}$, for $x > 0$, then it verifies that:*

a) *it is a decreasing function of x for x large enough;*

b) $\lim_{x\to+\infty} f(x) = 0$.

Proof. Given that,

$$\frac{f(x+1)}{f(x)} = \frac{\lambda^{(x+1)}(x+1)^\alpha}{(x+1)!}\,\frac{x!}{\lambda^x x^\alpha} = \frac{\lambda}{(x+1)}\left(1+\frac{1}{x}\right)^\alpha,$$

one has that $\lim_{x\to+\infty} f(x+1)/f(x) = 0$ and thus, for large values of x, $f(x+1) \leq f(x)$, which proves a).

Taking into account that the Poisson distribution has moments of any order, and denoting by $\llcorner\alpha\lrcorner$ the integer part of α, we have that for any $\alpha \in \mathbb{R}$, and $\alpha > 1$,

$$0 \leq \sum_{x=0}^{+\infty} \frac{\lambda^x x^\alpha}{x!} \leq \sum_{x=0}^{+\infty} \frac{\lambda^x x^{\llcorner\alpha\lrcorner+1}}{x!} < +\infty,$$

and consequently, $\lim_{x\to+\infty} f(x) = 0$. □

For the next Lemma, a *regularly varying random variable with index $\alpha \geq 0$* is defined as r.v. X, such that

$$P(X > x) \sim px^{-\alpha}L(x) \quad \text{and} \quad P(X \leq -x) \sim qx^{-\alpha}L(x),$$

where $p + q = 1$, and L is a slowly varying function.

Lemma 4. *Let $N \sim Po(\lambda)$ and $X \sim Zipf(\alpha)$, N and X being independent r.v.'s. Then*

$$P(N > x) = o(P(X > x)).$$

Proof. We need to prove that $P(N>x)/P(X>x)$ has limit zero at infinity. However, taking into account point a) of Lemma 3, one has that for x large enough,

$$0 \leq \frac{P(N>x)}{P(X>x)} = \frac{\sum_{i=x+1}^{+\infty} \frac{e^{-\lambda}\lambda^i}{i!}}{\sum_{i=x+1}^{+\infty} \frac{i^{-\alpha}}{\zeta(\alpha)}} = e^{-\lambda}\,\zeta(\alpha)\,\frac{\sum_{i=x+1}^{+\infty} \frac{\lambda^i i^{\alpha} i^{-\alpha}}{i!}}{\sum_{i=x+1}^{+\infty} i^{-\alpha}}$$

$$\leq e^{-\lambda}\zeta(\alpha)\,\frac{\lambda^x x^{\alpha}}{x!}\,\frac{\sum_{i=x+1}^{+\infty} i^{-\alpha}}{\sum_{i=x+1}^{+\infty} i^{-\alpha}} \leq e^{-\lambda}\,\zeta(\alpha)\,\frac{\lambda^x x^{\alpha}}{x!}.$$

Taking limits in the inequality, and as a consequence of point b) of Lemma 3, one has that:

$$0 \leq \lim_{x\to+\infty} \frac{P(N>x)}{P(X>x)} \leq e^{-\lambda}\,\zeta(\alpha) \lim_{x\to+\infty} \frac{\lambda^x x^{\alpha}}{x!} = 0,$$

which proves the proposition. □

Based on Lemma 2 and supported by all the proofs conducted in Lemmas 3 and 4, we state the following Theorem.

Theorem 9. *The tail of r.v. Y with a Zipf-PSS*(α,λ) *distribution is asymptotically equivalent to* λ *times the tail of r.v. X with a Zipf*(α) *distribution.*

Proof. The proof is a consequence of Lemma 2, where $X \sim$Zipf(α) and $N \sim$Po(λ), and X and N are independent r.v.'s Thus, we have that $P(Y>x) \sim E[N]\,P(X>x)$, and given that

$$P(Y=x) = P(Y>x-1) - P(Y>x),$$

$E[N]=\lambda$, one has that:

$$P(Y=x) \sim \lambda\,[P(X>x-1) - P(X>x)] \Leftrightarrow P(Y=x) \sim \lambda\,P(X=x). \qquad \square$$

Figure 3.6 illustrates the result stated in Theorem 9. Observe that the larger the λ value becomes, the larger the x value must be to obtain the equivalence of the tails.

3.4 A limit distribution and approximations

This subsection establishes the conditions under which the Zipf-PSS converges in distribution to the Poisson. It also studies their DCPs approximations of order r for $r=1,2,3$ and 4.

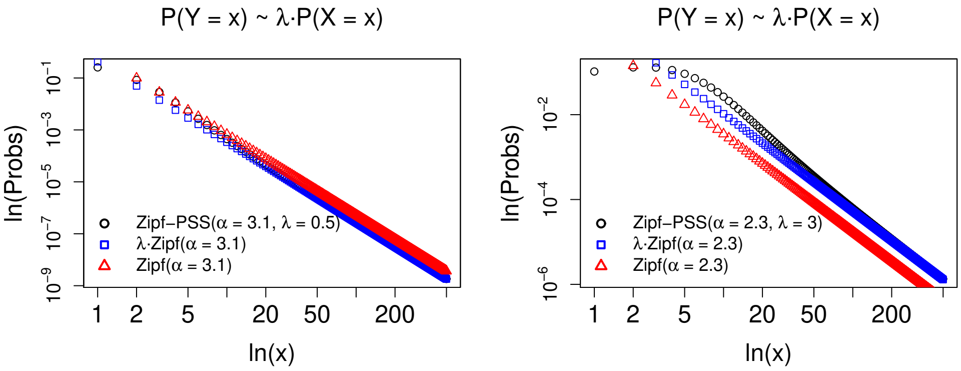

Fig. 3.6 Probabilities of the Zipf, λZipf, and the Zipf-PSS distributions. On the left-hand side with the same α parameter for $\alpha = 3.1$ and $\lambda = 0.5$; and, on the right-hand side for $\alpha = 2.3$ and $\lambda = 3$.

Proposition 17. *Let $Y_n \sim$ Zipf-PSS(α_n, λ), for $n \geq 1$, and Y_i is independent of Y_j. Also let $N \sim Po(\lambda)$, Y_i be independent of N $\forall i$. Then, if $\alpha_n \to +\infty$ when $n \to +\infty$,*

$$Y_n \xrightarrow[n\to+\infty]{D} N.$$

Proof. Given that, when $\alpha \to +\infty$, the Zipf distribution tends towards the degenerate distribution at one, we have that, for large values of α, the Zipf-PSS(α, λ) distribution is equal to the sum of N ones, which is equal to N. □

These results can be observed in the top right-hand side of Figure 3.1, where for $\alpha = 2.3$ and $\lambda = 10$, the PMF of the Zipf-PSS looks like the PMF of a Poisson.

In Subsection 3.2, we have introduced the approximation of order r of a DCP distribution. In what follows, we compute the firsts four approximations for the particular case of the Zipf-PSS.

Proposition 18. *The Zipf-PSS(α, λ) distribution has as:*

a) first order approximation, a Poisson distribution, with parameter $\lambda/\zeta(\alpha)$;

b) second order approximation, a Hermite distribution, with parameters $a_1 = \lambda/\zeta(\alpha)$ and $a_2 = a_1/2^\alpha$;

c) third order approximation, a 3^{rd} order Hermite distribution, with parameters $a_1 = \lambda/\zeta(\alpha)$, $a_2 = a_1/2^\alpha$ and $a_3 = a_1/3^\alpha$;

d) fourth order approximation, a 4^{th} order Hermite distribution, with parameters $a_1 = \lambda/\zeta(\alpha)$, $a_2 = a_1/2^\alpha$, $a_3 = a_1/3^\alpha$ and $a_4 = a_1/4^\alpha$.

Proof. From (3.2.1), the PGF of r.v. Y with a Zipf-PSS(α,λ) distribution may be written as:

$$G_Y(z) = e^{\frac{\lambda}{\zeta(\alpha)}(Li_\alpha(z)-\zeta(\alpha))}.$$

Taking into account that,

$$\begin{aligned}\frac{Li_\alpha(z)-\zeta(\alpha)}{\zeta(\alpha)} &= \frac{1}{\zeta(\alpha)}\left[\sum_{k=1}^{+\infty}\frac{z^k}{k^\alpha}-\sum_{k=1}^{+\infty}\frac{1}{k^\alpha}\right]=\frac{1}{\zeta(\alpha)}\sum_{k=1}^{+\infty}\frac{(z^k-1)}{k^\alpha}\\ &= \frac{(z-1)}{\zeta(\alpha)}+\frac{(z^2-1)}{2^\alpha\zeta(\alpha)}+\frac{(z^3-1)}{3^\alpha\zeta(\alpha)}+\dots,\end{aligned}$$

the first to fourth order approximations of the PGF of r.v. Y with a Zipf-PSS(α,λ) distribution are those mentioned in points a) to d), since they correspond to the Taylor's approximations of order one to four of $\frac{Li_\alpha(z)-\zeta(\alpha)}{\zeta(\alpha)}$. □

Figure 3.7 shows the behavior of the first four approximations. Observe that for small α values, a large λ value is required to properly approximate the distributions by a lower order approximation. In contrast, when α increases, the distribution is properly approximated by the first order approximation independently of the value of λ. Thus, the first order approximation is acceptable if α is large enough. For small values of α, it is better to consider the second or higher order approximations.

3.5 Random data generation

To generate data from a Zipf-PSS distribution, the inversion method is directly applied to the CDF of the distribution family. As it has been done in Chapter 2, some simulations have been performed in order to test if the Zipf-PSS random data generator works appropriately.

We have generated 500 samples of size 100 and 500 samples of size 1000. This has been done for all the possible pair of values (α,λ) obtained from the sequences $\alpha = 1.25, 2, 3.5$ and 5, and $\lambda = 0.1, 0.25, 0.5, 1, 1.75, 2.25, 3.5$ and 10.

Tables 3.1 and 3.2 contain the total number of generated sequences for each pair of parameters, and the percentage of them for which the null hypothesis of the KS test has not been rejected. Important to remark that the same definition of the KS test used with the MOEZipf and the Zipf-PE distributions has been applied here. Observe that for $\alpha = 1.25$, independently of the λ value, we cannot generate all the sequences. In particular, for $\lambda \geq 1$ we could not generate any. This is because, for these configurations the probabilities increase very slowly. Bear in mind that, as mentioned in the previous chapter, we have established a

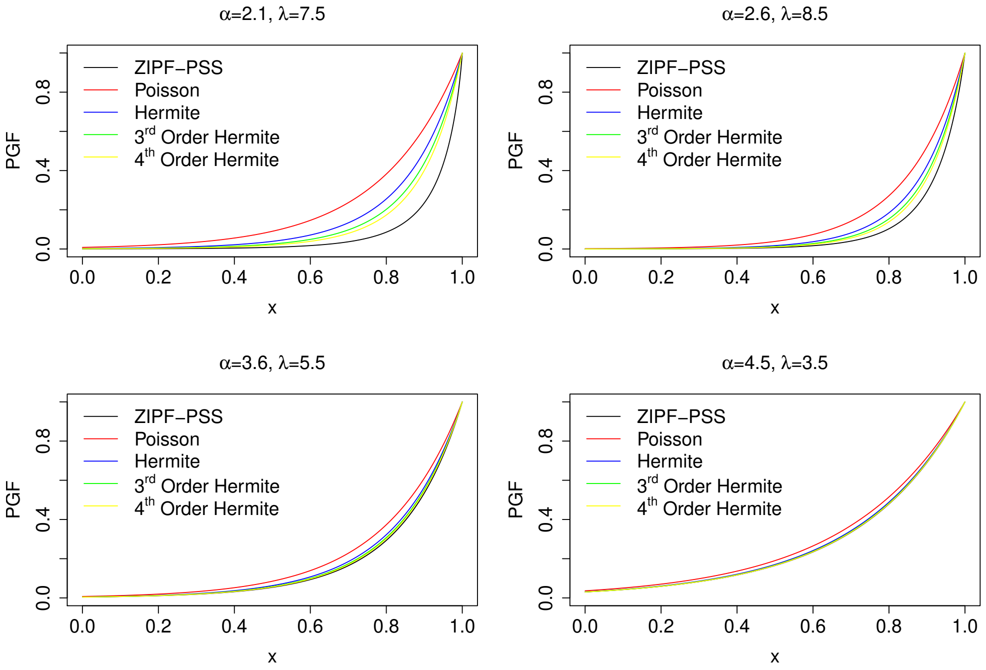

Fig. 3.7 Behavior of the PGF of the higher order approximations of the Zipf-PSS for $\alpha = 2.1$, $\lambda = 7.5$ (top-left), $\alpha = 2.6$, $\lambda = 8.5$ (top-right), $\alpha = 3.6$, $\lambda = 5.5$ (bottom-left) and $\alpha = 4.5$, $\lambda = 3.5$ (bottom-right).

timeout of 30 minutes for this procedure. However, for the remaining parameter configuration the results are very good.

For the more conservative KS test, the null hypothesis has not been rejected in a percentage that goes from 98% to 100% (see Table 3.1). For the KS test adapted to discrete distributions it is very close to 95% (see Table 3.2).

Table 3.1 Number of random sequences generated from the Zipf-PSS distribution, as well as the associated percentage of the sequences that have passed the classical Kolmogorov-Smirnov test.

Distribution	N	α	λ							
			0.1	0.25	0.5	1	1.75	2.25	3.5	10
Zipf-PSS	100	1.25	170 (100 %)	23 (100 %)	2 (100 %)	-	-	-	-	
	1000		-	-	-	-	-	-	-	-
	100	2	500 (100 %)	499 (100 %)	499 (100 %)	491 (98 %)	484 (98 %)	487 (98 %)	488 (98 %)	441 (98 %)
	1000		494 (100 %)	490 (100 %)	484 (99 %)	451 (99 %)	400 (98 %)	381 (97 %)	334 (98 %)	126 (94 %)
	100	3.5	500 (100 %)	500 (100 %)	500 (99 %)	500 (99 %)	500 (99 %)	500 (99 %)	500 (99 %)	500 (98 %)
	1000		500 (100 %)	500 (100 %)	500 (99 %)	500 (99 %)	500 (99 %)	500 (98 %)	500 (99 %)	500 (98 %)
	100	5	500 (100 %)	500 (100 %)	500 (100 %)	500 (99 %)	500 (100 %)	500 (99 %)	500 (99 %)	500 (99 %)
	1000		500 (100 %)	500 (100 %)	500 (100 %)	500 (99 %)	500 (99 %)	500 (99 %)	500 (98 %)	500 (98 %)

Table 3.2 Number of random sequences generated from the Zipf-PSS distribution, as well as the associated percentage of the sequences that have passed the Kolmogorov-Smirnov test for discrete distributions.

Distribution	N	α	λ							
			0.1	0.25	0.5	1	1.75	2.25	3.5	10
Zipf-PSS	100	1.25	170 (98 %)	23 (96 %)	2 (100 %)	-	-	-	-	-
	1000		-	-	-	-	-	-	-	-
	100	2	500 (95 %)	499 (95 %)	499 (96 %)	491 (93 %)	484 (95 %)	487 (95 %)	488 (94 %)	441 (96 %)
	1000		494 (94 %)	490 (94 %)	484 (96 %)	451 (95 %)	400 (96 %)	381 (93 %)	334 (95 %)	126 (92 %)
	100	3.5	500 (97 %)	500 (95 %)	500 (96 %)	500 (96 %)	500 (96 %)	500 (96 %)	500 (95 %)	500 (97 %)
	1000		500 (95 %)	500 (96 %)	500 (94 %)	500 (94 %)	500 (96 %)	500 (96 %)	500 (94 %)	500 (95 %)
	100	5	500 (96 %)	500 (95 %)	500 (96 %)	500 (96 %)	500 (95 %)	500 (97 %)	500 (94 %)	500 (96 %)
	1000		500 (96 %)	500 (96 %)	500 (97 %)	500 (96 %)	500 (96 %)	500 (93 %)	500 (93 %)	500 (96 %)

Chapter 4

The Zipf-Polylog extension

In this chapter we introduce what we have called the Zipf-Polylog generalization of the Zipf distribution, which has been naturally obtained from the PGF of the Zipf distribution by including an additional parameter. Nevertheless, after defining it, we realized that this family of distributions already appeared in the scientific literature under other names and other parameterizations. For instance Fenner et al. [2005], Clauset et al. [2009], and Smolinsky [2017] refer to it as the *PL distribution with an exponential cutoff*, and use it as an alternative to the Zipf distribution that allows to capture situations in which the Zipf exponent is positive but less than or equal to one. In Visser [2013] it is also mentioned as *the hybrid geometric/power model* and the authors prove that it is the two parameter discrete model defined in $1,2,\ldots$ that has maximum Shannon entropy, once the sample mean and the sample mean of the logarithm of the observations are fixed. At this point, we would like to mention that the Visser's paper contains a lack of precision when says that the additional parameter can take negative values. The extra parameter in the Visser parametrization can only be strictly positive, otherwise the odd negative integer values will have negative probabilities, which has no sense. A part from that, it is important to say that we have not been able to find in the literature an accurate analysis of the properties of this model, and it is because of that, that this chapter contains our definition as well as the properties that we have derived from it.

The chapter starts by defining the Zipf-Polylog family of distributions and then, by stating and proving some of its main properties. At the end we include a section devoted to the analysis of random data generation from the Zipf-Polylog distribution. Most of the work presented in this chapter is included in the paper *"The Zipf as a Mixture Distribution and Its Polylogarithm Generalization"* [Valero et al., 2020], which at the moment of writing this PhD thesis is under revision.

4.1 Definition

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

$$h_Y(z) = \begin{cases} \frac{Li_\alpha(\beta z)}{Li_\alpha(\beta)} & \text{if } \beta \neq 1 \text{ and } \alpha \in (-\infty,+\infty) \\ \\ \frac{Li_\alpha(z)}{Li_\alpha(1)} & \text{if } \beta = 1 \text{ and } \alpha > 1. \end{cases} \tag{4.1.1}$$

To see that (4.1.1) 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 (1.1.13), $h_Y(z)$ 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:

$$P(Y=x) = h_y^x(z)\Bigg|_{z=0} = \frac{\beta^x x^{-\alpha}}{Li_\alpha(\beta)}, x = 1,2,\ldots. \tag{4.1.2}$$

Observe that by defining $\gamma = -\log(\beta)$, the Zipf-Polylog distribution turns out to be the discrete version of the *PL distribution with exponential cut-off*, which appears in the paper by Clauset et al. [2009]. In the same way the work by Smolinsky [2017] uses the PL distribution with exponential cut-off in a comparative analysis between the mentioned distribution and the Lotka's law, where both distributions are used to measure the authors productivity in the chemistry area. 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 $\alpha = 1$, (4.1.2) is the PMF of the logarithmic-series distribution because $Li_1(\beta) = -\log(1-\beta)$, as observed in Section 1.1 of Chapter 1. Moreover, if $\alpha = 0$ we obtain the geometric distribution with support $\{1,2,3,\ldots\}$ and probability of success $p = 1-\beta$. Finally, if $\alpha = -1$, given that $Li_{-1}(\beta) = \beta(1-\beta)^{-2}$, (4.1.2) is equal to the PMF of a shifted negative binomial distribution with $r = 2$ successes and probability of success $p = 1-\beta$. Figure 4.1 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.

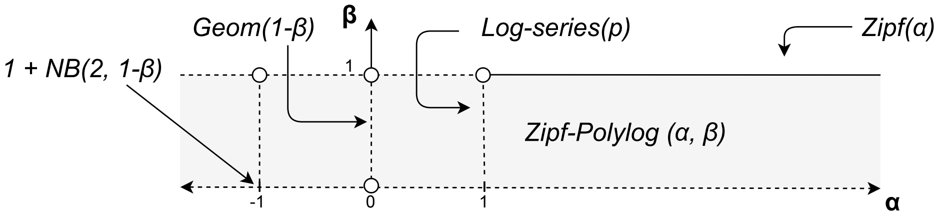

Fig. 4.1 Parameter space of the Zipf-Polylog family of distributions with the geometric, the log-series, the Zipf and the shifted negative binomial with $r = 2$ families as particular cases.

It is important to observe 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]. In what follows we see that the Zipf-Polylog family is also a particular case of the Lerch family. To that end, we consider the parametrization of the Lerch family that appears in Kemp [2010], p. 2257. With this parametrization, a distribution in the Lerch family with support $x = 1,2,\ldots$ is the one with PGF equal to:

$$G(z) = z\frac{\phi(\rho z, c, \nu)}{\phi(\rho, c, \nu)}, \tag{4.1.3}$$

being $\phi(\rho, c, \nu)$ the Lerch function that is defined for complex ρ, c and ν such that $|\rho| < 1, \nu \neq 0, -1, -2, \ldots$ as:

$$\phi(\rho, c, \nu) = \sum_{x=0}^{+\infty} (\nu + x)^{-c} \rho^x.$$

Observe that for $\rho = \beta, c = \alpha$ and $\nu = 1$, (4.1.3) is equal to $Li_\alpha(\beta z)/Li_\alpha(\beta)$ that, by (4.1.1) is the PGF of the Zipf-Polylog(α, β) distribution.

As a consequence of (4.1.2), 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 by Patil and Rao [1978]. According to these authors, a weighted distribution is needed when the probability of observing a value x depends on the size of the value.
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 et al. [2017] for a recent review on weighted distributions.

Figure 4.2 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 4.3 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.

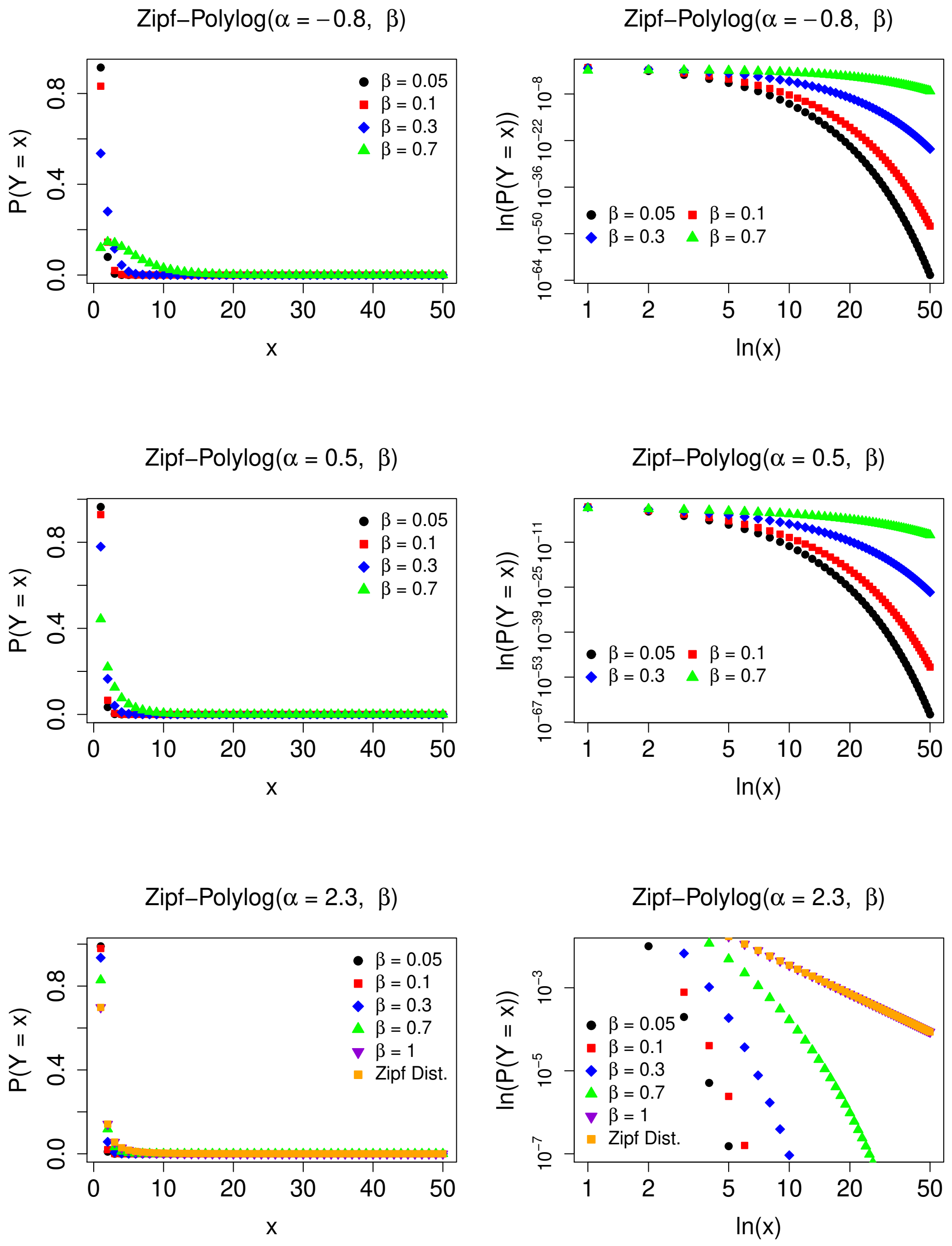

Fig. 4.2 PMF of the Zipf-Polylog(α,β). At the top are negative values of the α parameter and different values of β. At the bottom are positive α values and different values of β. Both cases include, respectively, the plots in normal scale and in log-log scale. The probabilities of the Zipf distribution are included when $\alpha > 1$ and $\beta = 1$.

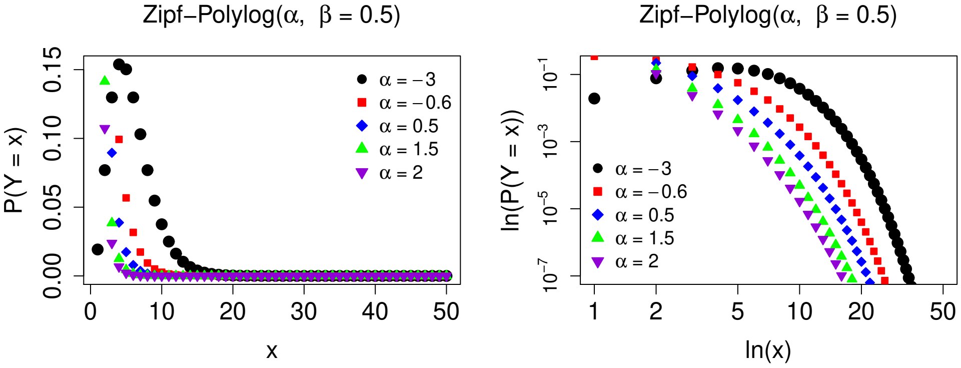

Fig. 4.3 PMF of the Zipf-Polylog(α,β) for $\alpha = -3, -0.6, 0.5, 1.5, 2$ and $\beta = 0.5$. The left-hand side shows the probabilities in normal scale; while the right-hand side shows the same probabilities in log-log scale.

4.2 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 a generalization of Theorems 1 and 2, both of them introduced in Section 1.4 of Chapter 1.

Theorem 10. *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:

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

$$P(Y = x) = \frac{x^{-\alpha}e^{-\gamma x}}{Li_\alpha(e^{-\gamma})} = \frac{e^{-\alpha\log(x)-\gamma x}}{Li_\alpha(e^{-\gamma})}. \tag{4.2.1}$$

At the right-hand side of (4.2.1), 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 $T(x) = (-\log(x), -x)$. □

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 $x_1, x_2, \ldots, x_n$ 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:

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

which has a unique solution if $t(x)$ belongs to the interior of the convex hull of $t(\mathscr{N})$, being $\mathscr{N}$ the space where takes values $t(x)$. Note that from (1.1.11), the second equation to be solved is the same as the one needed for finding the MLE for the Zipf distribution. The first equation corresponds to the Gauss Principle [see Teicher, 1961].

Proposition 19. *If Y ~Zipf-Polylog(α, β) 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 et al., 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:

$$\mu'_k = \left. \frac{\partial^k}{\partial z^k} h_Y(z) \right|_{z=1} = \frac{1}{Li_\alpha(e^{-\gamma})} e^{-k\gamma} Li_\alpha^{(k)}(ze^{-\gamma})|_{z=1} < +\infty, \qquad (4.2.2)$$

because $Li_\alpha(ze^{-\gamma})$ is analytical in $(-\infty, 1]$. □

Figure 4.4 shows the behavior of the expected value of the Zipf-Polylog distribution. On the left-hand side as a function of α, for $\beta = 0.25, 0.5, 0.75$ and 0.9. On the right-hand side as a function of β, for $\alpha = -1.5, -0.5, 2$ and 3.5. Observe that it is a decreasing function of α with a slope that increases for large β values. On the contrary, it is an increasing function of β where the largest values are obtained for small values of α.

In the same way Figure 4.5 illustrates the behavior of the variance of several Zipf-Polylog(α, β) distributions. On the left-hand side as a function of α, for $\beta = 0.25, 0.5, 0.75$ and 0.9. On the right-hand side as a function of β, for $\alpha = -1.5, -0.5, 2$ and 3.5. Note that the variance shows a similar pattern than the one achieved in the previous plots. In general it seems that the variability decreases as a function of α and increases as a function of β.

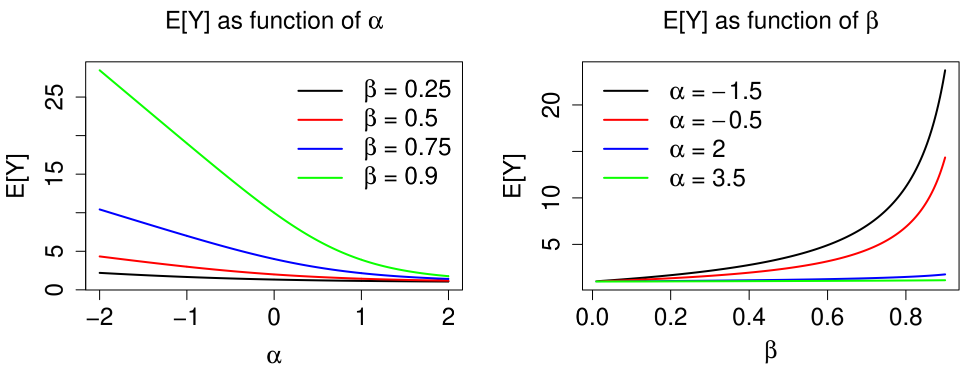

Fig. 4.4 Expected values of a Zipf-Polylog(α,β) distribution. On the left-hand side as a function of α, for $\beta = 0.25, 0.5, 0.75$ and 0.9. On the right-hand side as a function of β, for $\alpha = -1.5, -0.5, 2$ and 3.5.

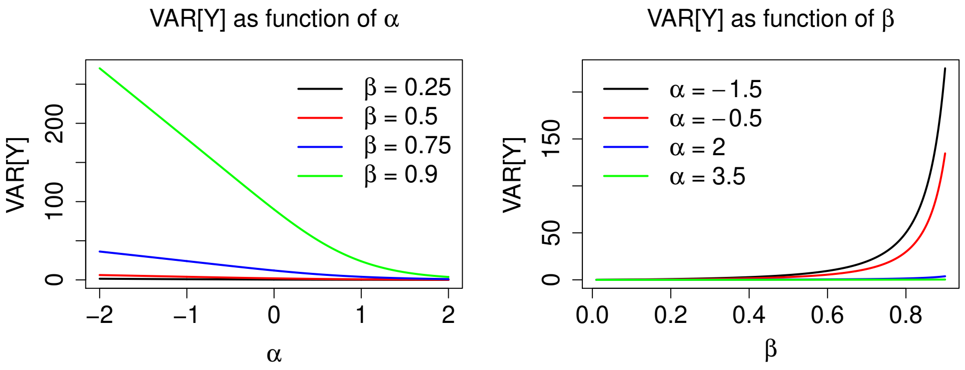

Fig. 4.5 Variance values of a Zipf-Polylog(α,β) distribution. On the left-hand side as a function of α, for $\beta = 0.25, 0.5, 0.75$ and 0.9. On the right-hand side as a function of β, for $\alpha = -1.5, -0.5, 2$ and 3.5.

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 20. *If $Y \sim$ Zipf-Polylog(α,β) with $\alpha > 1$, and X ~Zipf(α), 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 (4.1.2) 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 generalizes Theorem 1 of Section 1.4 of Chapter 1, and states that any Zipf-Polylog with a positive value of α is a mixture of geometric distributions.

Theorem 11. *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*

$$f(s;\alpha,\beta) = \frac{\frac{s^{\alpha-1}}{e^s-\beta}}{\int_0^{+\infty} \frac{t^{\alpha-1}}{e^t-\beta}dt} = \frac{\beta}{\Gamma(\alpha)Li_\alpha(\beta)}\frac{s^{\alpha-1}}{e^s-\beta}. \tag{4.2.3}$$

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

$$\frac{pz}{1-(1-p)z} = \frac{(1-\beta e^{-s})z}{1-\beta e^{-s}z} = \frac{(e^s-\beta)z}{e^s-\beta z}. \tag{4.2.4}$$

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

$$h_y(z) = \frac{Li_\alpha(\beta z)}{Li_\alpha(\beta)} = \int_0^{+\infty} \frac{(e^s-\beta)z}{e^s-\beta z} f(s;\alpha,\beta)\, ds, \tag{4.2.5}$$

with $f(s;\alpha,\beta)$ defined as in (4.2.3). By substituting $f(s;\alpha,\beta)$ for its expression and taking into account (1.1.14), we have:

$$\int_0^{+\infty} \frac{(e^s-\beta)z}{e^s-\beta z} \frac{\frac{s^{\alpha-1}}{e^s-\beta}}{\int_0^{+\infty} \frac{t^{\alpha-1}}{e^t-\beta}dt} ds = \frac{\beta z}{\beta \int_0^{+\infty} \frac{t^{\alpha-1}}{e^t-\beta}dt} \int_0^{+\infty} \frac{s^{\alpha-1}}{e^s-\beta z} ds = \frac{Li_\alpha(\beta z)}{Li_\alpha(\beta)}, \tag{4.2.6}$$

which proves the theorem. □

The next theorem generalizes Theorem 2 in Section 1.4 of Chapter 1 for any distribution in the generalized family.

Theorem 12. *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:

$$f(\lambda;\alpha,\beta) = \frac{e^\lambda - 1}{\beta\Gamma(\alpha)Li_\alpha(\beta)} \int_0^{+\infty} s^{\alpha-1} e^{s-\frac{\lambda}{\beta}e^s}\, ds, \tag{4.2.7}$$

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

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

where $f(\lambda;\alpha,\gamma)$ is defined as in (4.2.7). From Theorem 11 we have

$$\frac{Li_\alpha(\beta z)}{Li_\alpha(z)} = \int_0^{+\infty} \frac{(e^s-\beta)z}{e^s-\beta z} \frac{\beta}{\Gamma(\alpha)Li_\alpha(\beta)} \frac{s^{\alpha-1}}{e^s-\beta}\,ds. \tag{4.2.8}$$

Moreover, taking into account (4.2.4), by Proposition 1 we have:

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

where

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

Thus, we have:

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

Now, substituting the last equality in (4.2.8) gives:

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

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:

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

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 Proposition 1. Moreover, given that

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

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 $Li_\alpha(z)$ function for integer values of α verifies:

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

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

$$\frac{\partial Li_{-1}(z)}{\partial z} = \frac{1}{z}Li_0(z) = \frac{1}{z}\frac{z}{1-z} = \frac{1}{1-z}, \tag{4.2.9}$$

and it is negative when $z < -1$. For $\alpha = -n$, by applying (4.2.9) recursively n times, we have that for certain real values $a_1, a_2, \cdots, a_{n-2}$,

$$\frac{\partial Li_{-n}(z)}{\partial z} = \frac{z(z^{n-1} + a_{n-2}z^{n-2} + \cdots + a_1 z + 1)}{(1-z)^n},$$

from which we have $\forall\, n\ Li'_{-n}(0) = 0$. Moreover, given that $\lim_{z\to-\infty} Li_{-n}(z) = 0$, 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

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

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

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

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.

□

Figure 4.6 shows representations of the $Li_\alpha(z)$ function for $\alpha = -0.8, -1, -1.7$ and -2.55, and $z \in (-\infty, 0)$. Observe that the $Li_\alpha(\beta)$ is decreasing in some interval of the negative real line defined as defined in the proof of the last theorem.

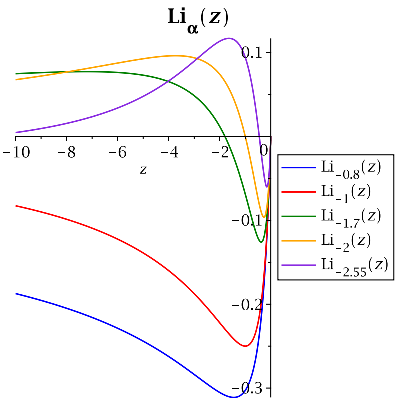

Fig. 4.6 Plots of the $Li_\alpha(z)$ function for $\alpha = -0.8, -1, -1.7$ and -2.55, and $z \in (-\infty, 0)$

4.3 Random data generation

The approach used for generating random data from a Zipf-Polylog distribution is similar to the one used in Chapter 3, that is to apply the inversion method to the CDF of the distribution family. However, we have introduced a sightly variation in its implementation with the aim of reducing the computational time required for generating the samples. Thus, the inversion method is implemented in such a way that it consist in choosing the first $x \geq 1$ verifying that:

$$u \cdot Li_\alpha(\beta) \leq \sum_{i=0}^{x} i^{-\alpha} \cdot \beta^i,$$

where $u \in (0,1)$ has been randomly selected form an Uniform distribution in $(0,1)$..

For validating the performance of the inversion method we have used all the pairs of values (α, β) obtained from $\alpha = -1.3, -0.05, 0.85, 1.25$ and 2, and $\beta = 0.35, 0.45, 0.55, 0.65, 0.75,$ 0.85 and 0.95. For each pair we have generated 500 samples of sizes 100 and 1000, respectively. Observe that we have avoided the use of values of $\beta < 0.35$, because for such values the probabilities basically concentrate in two points, as it can be seen in Table 4.1.

Tables 4.2 and 4.3 summarize the total number of sequences generated for each pair of parameters as well as the percentage of them that have not rejected the null hypothesis of

Table 4.1 First ten probabilities of the Zipf-Polylog distribution for several values of α and small values of β.

$\alpha = -1.3$	$\alpha = -0.05$	$\alpha = 0.85$	$\alpha = 1.25$
$\beta = 0.1$	$\beta = 0.25$	$\beta = 0.1$	$\beta = 0.25$
0.7722	0.7423	0.9436	0.8894
0.1901	0.1921	0.0524	0.0935
0.0322	0.049	0.0037	0.0141
0.0047	0.0124	0.0003	0.0025
0.0006	0.0031	0	0.0005
0.0001	0.0008	0	0.0001
0	0.0002	0	0
0	0.0001	0	0
0	0	0	0
0	0	0	0

the KS test. The definition of the KS test used corresponds to the one explained in Section 2.3.3 of Chapter 2. In particular Table 4.2 contains the results related to the continuous version of the KS test, while Table 4.3 summarizes the results associated with its discrete version. As already evinced in previous chapters, the continuous KS test tends to be more conservative than its discrete version where the p-values are obtained through simulations. In both scenarios the null hypothesis is not rejected in more than 92% of the cases, and most of the times is around the nominal value.

Table 4.2 Number of random sequences generated from the Zipf-Polylog distribution, as well as the associated percentage of sequences for which the null hypothesis has not been rejected (classical KS test).

Distribution	N	α	β						
			0.35	0.45	0.55	0.65	0.75	0.85	0.95
Zipf-Polylog	100	-1.30	500 (98 %)	500 (99 %)	500 (99 %)	500 (98 %)	500 (98 %)	500 (97 %)	500 (97 %)
	1000		500 (99 %)	500 (99 %)	500 (97 %)	500 (98 %)	500 (98 %)	500 (98 %)	500 (95 %)
	100	-0.05	500 (100 %)	500 (99 %)	500 (99 %)	500 (98 %)	500 (99 %)	500 (98 %)	500 (97 %)
	1000		500 (100 %)	500 (99 %)	500 (99 %)	500 (99 %)	500 (99 %)	500 (98 %)	500 (96 %)
	100	0.85	500 (100 %)	500 (100 %)	500 (99 %)	500 (99 %)	500 (99 %)	500 (99 %)	500 (98 %)
	1000		500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (99 %)	500 (99 %)	500 (98 %)
	100	1.25	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (99 %)	500 (99 %)
	1000		500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (99 %)	500 (99 %)
	100	2	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)
	1000		500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)	500 (100 %)

Table 4.3 Number of random sequences generated from the Zipf-Polylog distribution, as well as the associated percentage for which the null hypothesis has not been rejected (*p*-values of the KS test obtained by simulations).

Distribution	N	α	β						
			0.35	0.45	0.55	0.65	0.75	0.85	0.95
Zipf-Polylog	100	-1.30	500 (96 %)	500 (96 %)	500 (96 %)	500 (96 %)	500 (96 %)	500 (95 %)	500 (96 %)
	1000		500 (94 %)	500 (94 %)	500 (92 %)	500 (95 %)	500 (94 %)	500 (97 %)	500 (95 %)
	100	-0.05	500 (96 %)	500 (96 %)	500 (95 %)	500 (95 %)	500 (95 %)	500 (96 %)	500 (94 %)
	1000		500 (95 %)	500 (96 %)	500 (95 %)	500 (94 %)	500 (95 %)	500 (96 %)	500 (93 %)
	100	0.85	500 (96 %)	500 (95 %)	500 (95 %)	500 (96 %)	500 (95 %)	500 (97 %)	500 (95 %)
	1000		500 (93 %)	500 (95 %)	500 (96 %)	500 (96 %)	500 (95 %)	500 (97 %)	500 (95 %)
	100	1.25	500 (97 %)	500 (96 %)	500 (96 %)	500 (97 %)	500 (96 %)	500 (96 %)	500 (96 %)
	1000		500 (94 %)	500 (96 %)	500 (95 %)	500 (94 %)	500 (96 %)	500 (95 %)	500 (94 %)
	100	2	500 (95 %)	500 (97 %)	500 (96 %)	500 (96 %)	500 (95 %)	500 (96 %)	500 (98 %)
	1000		500 (96 %)	500 (95 %)	500 (95 %)	500 (94 %)	500 (95 %)	500 (95 %)	500 (97 %)

Chapter 5

Applications

This chapter aims to illustrate that real data can be modeled accurately by the Zipf extensions presented in the previous chapters of this thesis. To that end, we adjust several degree sequences of real networks in dissimilar and unrelated areas, and we interpret the model parameters in terms of the data generation mechanism.

In the first subsection, we define the basic concepts related to the graphs required for our data analysis. Even though the examples presented here are related to the field of Network Analysis, our contributions may be applied to other scenarios. For that reason, Appendix A proposes a guide that helps practitioners assess the suitability of our models in their particular areas of research.

In the next four subsections, we analyze data related to protein interactions, social networks like Facebook, email interaction networks, and work collaboration networks. The fits obtained with the best of our models (when possible) are compared with the fits obtained for the Zipf distribution using the Likelihood Ratio Test (LRT). This test allows us to determine whether or not the extra parameter is really necessary for a particular data set. Given that the Zipf distribution pertains to the interior of the parameter space of the MOEZipf and Zipf-PE distributions, the Likelihood Ratio (LR) statistic under the null hypothesis for those models will follow a χ^2_1 distribution. In the case of the Zipf-Polylog family, the fact that the Zipf model is placed on the boundary of the parameter space implies that, under the null hypothesis, the LR statistic follows a 50:50 mixture of χ^2_0 and χ^2_1 Self and Liang [1987].

We have also found it necessary to compare our fits with the ones obtained from the other two-parameter families of distributions that are widely used in the literature as alternatives to the Zipf. The ones that have been considered are: the Zipf-Mandelbrot, DGX, NB and DW, all of which were introduced in Section 1.3 of Chapter 1. To perform these comparisons, the AIC and the log-likelihood have been chosen as goodness-of-fit criteria. In some examples, the fit obtained with the methodology proposed by Clauset et al. [2009] (also explained in Chapter 1.3) has also been taken into account.

Finally, when necessary, the Zipf-PSS has been truncated at zero in order to adjust samples that do not contain the zero value.

5.1 Networks, basic concepts

Since the applicability of the Zipf extension models in the next sections are all related to the analysis of degree sequences in real networks, it is convenient to introduce some of the definitions used in this research area. We start by defining what a graph is, followed by some of its structural properties.

A graph is the mathematical structure used to represent a network or complex system. It is defined by means of nodes and edges. It is usually denoted by: $G = (N,E)$, where N comprises the set of nodes and E stands for the set of edges. The relationships among the nodes in a graph are represented by the edges. It is also said that two nodes are *neighbors* if they share at least one edge. Sometimes, the nodes in the graph have extra information about the objects that they are representing. This extra information is stored in what is known as the *attributes* of a node.

Usually a file containing a graph is composed of pairs of nodes indicating which nodes are related. This is perhaps the most common approach to storing a graph, and it is called an *edge list* format. There are other ways to store this structure into a file, such as Graph Modeling Language (GML) [1] and Pajek [2], among others.

Several definitions that the reader may require in the next sections are:

Degree: Given a node i, its *degree* is defined as the total number of connections it has with other nodes. It is equal to its total number of neighbors. In the particular case of directed networks (graphs where the edges have a direction), the node degree is divided into three categories:

In-degree: Number of connections pointing a node.

Out-degree: Number of connections from a given node to others.

Total-degree: The sum of the in- and out-degrees of a node.

Graphic degree sequence: A sequence of numbers that can be used for creating a graph. Several works state the necessary conditions for using a numerical sequence to create a

[1] https://gephi.org/users/supported-graph-formats/gml-format/

[2] https://gephi.org/users/supported-graph-formats/pajek-net-format/

network Tripathi and Vijay [2003]. According to the *Fulkerson - Ryser theorem* [Kim et al., 2012], one of the conditions for a sequence being graphic in terms of the directed network is:

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$

where x and y belong to the in- and out-degree sequences, respectively. Most of the descriptive tables show this in the following examples, since the average of the in- and out-degree sequences are equal.

Degree sequence: Contains the degrees of all the nodes represented in the network.

Isolated nodes: These types of nodes do not have any connection to any other node or themselves. According to Bak et al. [1987] and McKelvey et al. [2018], they could play an important role in the evolution of the network, since a change in the behavior of isolated nodes could induce considerable changes in the network dynamics. The analysis of isolated nodes plays a fundamental role in communication networks, where having nodes in isolation results in information loss. This phenomenon is common in wireless and sensor networks. Observe that the Zipf-PSS extension is the only one able to deal with zero-degree nodes. In Section 5.3.2, we show an example of an application that needs to deal with these types of nodes.

Loops: These indicate nodes with edges connecting themselves.

Multi-edges: Two or more edges connecting the same pair of nodes.

Mean degree: The mean degree is the arithmetic mean of the degree sequence. It can also be computed as:

$$Mean = \frac{2 * \#E}{\#N}.$$

5.2 Protein-Protein interaction networks

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. The example analyzed in this section is related to the research performed by the author of this thesis in the National Institute of Biology in Slovenia. This 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.

Most of the nodes in the network have more 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 5.1 contains the main statistics for the AtCKN degree sequence. Note that, the data show large variability as well as high skewness value, which allows us to hypothesize the suitability of the proposed models.

Table 5.1 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 5.2 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 at the maximum likelihood parameters estimations, 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, the positive Zipf-PSS and the Zipf-Mandelbrot. These four 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 Zipf-PE, the MOEZipf and the positive Zipf-PSS. The fit obtained with the Zipf-Mandelbrot is slightly better than the one of the DGX, but still worse than our first three models.

Figure 5.1 illustrates the fits obtained for each considered model. Observe the Zipf's model lack of flexibility in adapting the top-concave pattern. On the other hand, the models Zipf-PE, MOEZipf, zt-Zipf-PSS, Zipf-Mandelbrot and DGX seem to provide a quite accurate fit. With respect to the PL methodology, it establishes a cut-off equal to 4, above which the distribution is fitted. By fixing this cut-off point, the method loses approximately 61% of the data.

Applying the LRT to compare the Zipf-PE with the Zipf models, that is 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

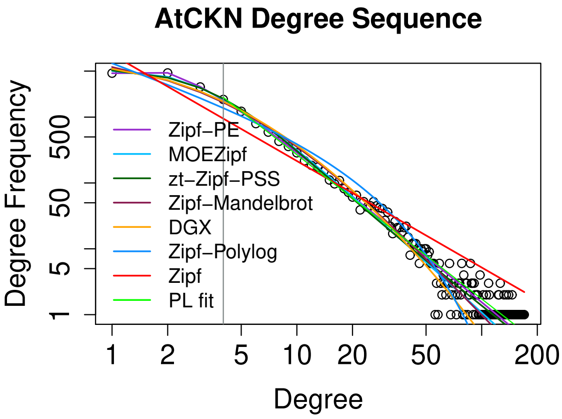

Fig. 5.1 Degree sequence of the PPI network and the fit obtained by each considered model. In addition, the fit obtained using the methodology proposed by Clauset et al. [2009] is also included.

ratio statistic for this degree sequence is equal to $-2\,[-53085.17-(-49429.96)]=7310.42$. By comparing the two 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.

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

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 5 times ($\widehat{E[N]}=4.89$). In general, the maximum expected interaction of all the proteins in the network is equal to 6.87 = $\widehat{E[Y]}$. Moreover, every time a protein is active, we estimate its number of interactions to be with approximately 3 other proteins ($\widehat{E[X]}=2.62$). Conducting the same analysis - but taking into account the parameters of the MOEZipf distribution - results in: the expected number of times that a protein is active until it gets connected with a fixed one is around 9 times ($\widehat{E[N]}=9.31$); and every time that it is active, the expected number of interactions is about 2 ($\widehat{E[X]}=1.82$). Thus, we can expect that, on average, the maximum number of interaction is approximately 5.81 = $\widehat{E[Y]}$.

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.

Table 5.2 Parameter estimates for each analyzed distribution, as well as 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	$\hat{\alpha} = 2.3241$	(2.305, 2.3432)	$\hat{\beta} = 4.8585$	(4.7169, 5.0001)	**-49429.9595**	**98863.9191**
MOEZipf	$\hat{\alpha} = 2.5575$	(2.5313, 2.5837)	$\hat{\beta} = 9.3057$	(8.8438, 9.7676)	-49518.0492	99040.0984
zt-Zipf-PSS	$\hat{\alpha} = 2.1698$	(2.1517, 2.1878)	$\hat{\lambda} = 1.6747$	(1.6377, 1.7116)	-49563.1824	99130.3648
Zipf-Mandelbrot	$\hat{\alpha} = 3.0144$	(2.93, 3.0989)	$\hat{V} = 4.9172$	(4.602, 5.2323)	-49782.0628	99568.1257
DGX	$\hat{\mu} = 0.9308$	(0.9054, 0.9563)	$\hat{\sigma} = 1.1616$	(1.1424, 1.1807)	-49818.0469	99640.0939
Zipf-Polylog	$\hat{\alpha} = 1.0091$	(0.9859, 1.0322)	$\hat{\beta} = 0.9454$	(0.9425, 0.9483)	-50816.0231	101636.0462
Zipf	$\hat{\alpha} = 1.6174$	(1.6086, 1.6262)	-	-	-53085.1701	106172.3402

5.3 Social networks

5.3.1 Facebook 100, the University of California, Santa Cruz network

In the work by Traud et al. [2012] is 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 8991 nodes and 224584 edges. The degree sequence is available through the git-hub repository: https://github.com/adbroido/SFAnalysis, mentioned in the paper by Broido and Clauset [2019]. The main statistics associated with the degree sequence can be observed in Table 5.3. Note that the sequence has a large variability, but, it does not seem to be a skew data set.

Table 5.3 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
8991	224584	453	164.0578	6958.8665	0.5285

Table 5.4 contains the results obtained after fitting all the candidate models. In this case, 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 respectively. In addition, the Zipf-Mandelbrot distribution is not included because it gives place to numerical problems when the MLE is computed.

Figure 5.2 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.

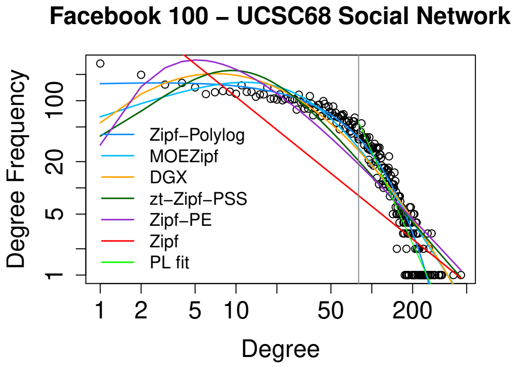

Fig. 5.2 Degree sequence of the Facebook network at UCSC, and the fit obtained by each of the considered models.

By means of the log-likelihood values of the Zipf and Zipf-Polylog models (see Table 5.4), 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 significatively different from one.

Given that $\hat{\alpha} \in (-1,0)$, one can assume, for instance, 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, we can state that a high number of nodes with degree equal to 1 are being fitted by the distribution. Based on Theorem 12, this dataset is not fitted by an MZTP because $\hat{\alpha}$ is negative.

5.3.2 The Math-Overflow platform

Here we analyze the MathOverflow data set, which appears in a recent paper by Paranjape et al. [2017]. This data set is available through the SNAP repository [Leskovec and Krevl, 2014] and contains interactions among the members of the MathOverflow platform, which was developed in 2009 by a group of PhD students and post-docs from Berkeley University. It is an on-line question-and-answer (Q&A) forum for research mathematicians on the Stack Overflow network [Keller, 2010].

Table 5.4 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	$param_1$	CI_{param_1}	$param_2$	CI_{param_2}	**Log-like**	**AIC**
Zipf-Polylog	$\hat{\alpha} = -0.0566$	(-0.0879, -0.0252)	$\hat{\beta} = 0.9789$	(0.9782, 0.9797)	**-44059.6760**	**88123.3521**
MOEZipf	$\hat{\alpha} = 2.5038$	(2.4767, 2.5309)	$\hat{\beta} = 401.8319$	(355.6692, 447.9946)	-44847.7235	89699.4471
DGX	$\hat{\mu} = 3.4082$	(3.383, 3.4335)	$\hat{\sigma} = 1.2087$	(1.19, 1.2274)	-44936.4704	89876.9409
zt-Zipf-PSS	$\hat{\alpha} = 1.7104$	(1.7008, 1.7201)	$\hat{\lambda} = 6.6136$	(6.4763, 6.7509)	-46066.8610	92137.7219
Zipf-PE	$\hat{\alpha} = 1.7408$	(1.7298, 1.7518)	$\hat{\beta} = 11.4596$	(11.0489, 11.8702)	-46666.5545	93337.1091
Zipf	$\hat{\alpha} = 1.2542$	(1.2489, 1.2595)	-	-	-51935.0908	103872.1816

In the repository, the synergy among users is divided into three categories: *answers to questions*, *comments to questions* and *comments to answers*. Here, we only use the category, *answers to questions*, where an edge is created between users u and v when user u answers a question by user v. This gives rise to a directed network, since the origin and final nodes play different roles. The network comprises a total of 21688 nodes and 107581 edges collected from September 9, 2009 to March 6, 2016. As a consequence of including the component time during edge formation, it is possible to have multi-edges between different pairs of nodes. Moreover, due to the fact that a user can answer their own question, loops are also possible in the network.

Each annual network was extracted from the global data set. Because these are directed networks, the degree sequence was split into three sequences: *the in-degree*, *the out-degree* and *the total-degree*. The in-degree sequence corresponds to situations in which the node user's question receives a reply. The out-degree corresponds to situations where the node user answers a question. Finally, the total-degree corresponds to the sum of the in- and out-degrees, and it explains the total activity of a user. Observe that the in-degree as well as the out-degree sequences may have isolated nodes. Thus, it is an appropriate network to test the Zipf-PSS family. For that reason, the in- and out-degree sequences are fitted with the Zipf-PSS as well as the NB and the discrete Weibull distributions that also include the zero value on their support. The total degree sequences are fitted with the same models considered in the previous examples.

Table 5.5 summarizes the main statistics calculated for each annual network. It contains the number of nodes (#N), edges (#E), loops (L) and multi-edges (M), and – after splitting the degree sequence – the number of isolated nodes (I), the range (Range), the mean (Mean), the variance (Var) and the skewness (Skew) of the total-, in- and out-degree sequences. Observe that a greater number of zeros occurs in the out-degree sequences, with the exception of the first year, which is when the community was created and only contains information covering approximately 4 months. One of the reasons for this may be that many people do not answer any questions while a few people answer many questions, which turns out to be an in-degree sequence with fewer nodes that have zero degrees. Observe that the mean values of the in- and out-degree sequences are equal for all years. This is because each edge involves a connection from the in- and the out-degree sequences, respectively. Moreover, this is one of the conditions of a graphical sequence, as stated in the *Fulkerson-Ryser's theorem* [Kim et al., 2012]. The mean values of the first (incomplete) and second year are clearly larger than the other years (the last year is also incomplete). This provides evidence indicating that the community was very active when it was created and the activity stabilized as the time passed. Note the large variance in the sequences, meaning that some users are quite active in answering while others answer only few or no questions. The higher skewness value associated with each sequence leads us to assume that all the degree sequences will show a pronounced right tail.

Fitting these data with the Zipf-PSS and the NB distributions allows insights into the network's data generation mechanism as a direct consequence of the fact that both are PSS. Thus, in both cases, $\widehat{E[N]}$ is interpreted as the average number of times that a user is active on the network. Each time that a user is active, she or he receives (in-degree) or answers (out-degree) an expected $\widehat{E[X_i]}$ number of questions. In contrast, the DW has no straightforward parameter interpretation.

The total number of degree sequences fitted is 24, which is obtained by multiplying 8 (years) by 3 (type of sequences). Initial values to find the MLE of the Zipf-PSS model needs to be found numerically, initial values for each of the parameters are required. These values are computed as follows:

1) The initial value for α is set to be equal to the α estimate obtained for the Zipf distribution proposed by Güney et al. [2017].

2) The initial value for λ is set to be equal to 1, because when the mean is $\lambda = 1$, we have just one term in the summation (3.1.1).

Tables 5.6, 5.7, 5.8 and 5.9 summarize the results obtained. The main conclusions are:

a) In 93.75% of the in- and out-degree sequences (15 out of 16), the Zipf-PSS gives the best fit. In the remaining case, the DW is the best model.

b) The behavior of the NB and the DW is quite similar, independently of the type of sequence, but the DW is always the best of the two.

c) The values of $\hat{\lambda}$ for the out-degree sequences are always smaller than one, while they are slightly larger than one for the in-degree sequences. It is expected that, over the course of one year, less than one person is activated to answer the question of a given node, while the node is active more than once over the same period to answer someone else's question(s).

d) Since there are not zeros in the total-degree sequence we consider only families of distributions with support in the strictly positive numbers and the positive version of the Zipf-PSS. Note that, the DGX and the Zipf-Mandelbrot provide the best fit to 50% of the sequence respectively. Also, the difference between the fits provide for both distributions are practically imperceptible, being always the first and the second distributions for all the total-degree sequences.

Table 5.5 For each degree sequence: Year of the sequence; number of nodes (#N); number of edges (#E); number of loops (L); number of multiple-edges (M); type of the sequence; number of isolated nodes (I) and its percentage; (Range); (Mean); variance (Var) and skewness (Skew) values.

Year	#N	#E	L	M	Type	I (%)	Range	Mean	Var	Skew
2009	1278	7115	1363	188	total	-	438	11.1346	807.5355	7.1768
					in	41.16	323	5.5673	297.6897	9.3863
					out	26.06	209	5.5673	227.2089	6.8121
2010	4648	24799	3120	547	total	-	468	10.6708	825.3136	7.2716
					in	31.58	304	5.3354	224.1519	8.6873
					out	39.57	354	5.3354	335.1896	8.1041
2011	5358	18468	1712	529	total	-	590	6.8936	326.6939	11.1849
					in	29.23	184	3.4468	70.0781	9.0476
					out	46.7	417	3.4468	164.939	11.6483
2012	5687	15954	1485	484	total	-	451	5.6107	243.9244	11.6652
					in	26.9	225	2.8053	46.555	13.979
					out	51.17	372	2.8053	141.5011	12.996
2013	6101	14779	1245	500	total	-	313	4.8448	149.4092	9.2134
					in	23.85	175	2.4224	27.2381	10.639
					out	56.7	194	2.4224	92.1437	9.9783
2014	5556	12574	1074	537	total	-	293	4.5263	124.5039	9.9265
					in	24.77	193	2.2631	25.4512	15.1054
					out	56.17	179	2.2631	75.7392	10.1688
2015	5409	11833	1166	566	total	-	355	4.3753	137.6435	12.7604
					in	24.7	214	2.1877	33.4409	18.9311
					out	55.2	297	2.1877	76.0929	13.3474
2016	1654	2059	83	92	total	-	65	2.4897	13.1357	6.8186
					in	33.13	23	1.2449	3.2727	5.1725
					out	52	46	1.2449	9.0574	6.7078

Table 5.6 For each of the degree sequences, one has: in the first column, the names of the models fitted; from the second to the fourth columns, the maximum likelihood parameter estimates together with their corresponding confidence intervals (CI); and the values of the AIC in the last column.

Year	Type	Distribution	$param_1$	CI_{param_1}	$param_2$	CI_{param_2}	Log-like	AIC
2009	total	**DGX**	$\hat{\mu} = 0.0877$	(-0.3423, 0.5177)	$\hat{\sigma} = 2.0246$	(1.8099, 2.2393)	**-3681.0923**	**7366.1845**
		Zipf-Polylog	$\hat{\alpha} = 1.2754$	(1.2158, 1.335)	$\hat{\beta} = 0.9892$	(0.986, 0.9923)	-3682.8368	7369.6735
		Zipf-Mandelbrot	$\hat{\alpha} = 1.8310$	(1.7124, 1.9496)	$\hat{V} = 1.2674$	(0.8198, 1.715)	-3683.6221	7371.2442
		MOEZipf	$\hat{\alpha} = 1.9255$	(1.8517, 1.9994)	$\hat{\beta} = 3.2173$	(2.6112, 3.8233)	-3694.0488	7392.0975
		Zipf-PE	$\hat{\alpha} = 1.8417$	(1.7832, 1.9001)	$\hat{\beta} = 2.0783$	(1.7163, 2.4404)	-3699.3507	7402.7014
		zt-Zipf-PSS	$\hat{\alpha} = 1.7595$	(1.7095, 1.8095)	$\hat{\lambda} = 0.9858$	(0.8163, 1.1552)	-3702.9714	7409.9429
		Zipf	$\hat{\alpha} = 1.5519$	(1.521, 1.5829)	-	-	-3760.4799	7522.9597
	in	**D. Weibull**	$\hat{q} = 0.5717$	(0.5465, 0.597)	$\hat{\beta} = 0.5239$	(0.4967, 0.5511)	**-3083.3667**	**6170.7334**
		Zipf-PSS	$\hat{\alpha} = 1.7636$	(1.7111, 1.816)	$\hat{\lambda} = 0.9752$	(0.906, 1.0444)	-3106.0120	6216.0239
		Neg. Bin.	$\hat{k} = 0.2734$	(0.2484, 0.2984)	$\hat{q} = 0.9532$	(0.9469, 0.9595)	-3118.6434	6241.2867
	out	**Zipf-PSS**	$\hat{\alpha} = 1.9202$	(1.8633, 1.977)	$\hat{\lambda} = 1.1738$	(1.1009, 1.2467)	**-3121.3763**	**6246.7526**
		D. Weibull	$\hat{q} = 0.6386$	(0.6157, 0.6616)	$\hat{\beta} = 0.6112$	(0.5841, 0.6383)	-3237.7605	6479.5210
		Neg. Bin.	$\hat{k} = 0.3761$	(0.3448, 0.4074)	$\hat{q} = 0.9367$	(0.9294, 0.9441)	-3281.7638	6567.5277
2010	total	**Zipf-Mandelbrot**	$\hat{\alpha} = 1.8500$	(1.7884, 1.9116)	$\hat{V} = 1.1308$	(0.9179, 1.3436)	**-12921.6372**	**25847.2745**
		DGX	$\hat{\mu} = -0.2472$	(-0.5291, 0.0347)	$\hat{\sigma} = 2.1032$	(1.9746, 2.2317)	-12921.7090	25847.4179
		Zipf-Polylog	$\hat{\alpha} = 1.3390$	(1.3081, 1.3699)	$\hat{\beta} = 0.9908$	(0.9892, 0.9923)	-12938.8673	25881.7345
		MOEZipf	$\hat{\alpha} = 1.9325$	(1.8923, 1.9728)	$\hat{\beta} = 2.9440$	(2.6514, 3.2367)	-12953.5134	25911.0267
		Zipf-PE	$\hat{\alpha} = 1.8595$	(1.8271, 1.8918)	$\hat{\beta} = 1.9488$	(1.7575, 2.14)	-12967.1950	25938.3901
		zt-Zipf-PSS	$\hat{\alpha} = 1.7787$	(1.7511, 1.8063)	$\hat{\lambda} = 0.9029$	(0.8156, 0.9903)	-12976.0815	25956.1630
		Zipf	$\hat{\alpha} = 1.5757$	(1.5587, 1.5926)	-	-	-13158.0762	26318.1523
	in	**Zipf-PSS**	$\hat{\alpha} = 1.9048$	(1.8744, 1.9352)	$\hat{\lambda} = 1.2202$	(1.179, 1.2614)	**-11558.4437**	**23120.8875**
		D. Weibull	$\hat{q} = 0.6361$	(0.6238, 0.6485)	$\hat{\beta} = 0.6118$	(0.5968, 0.6267)	-11696.7115	23397.4229
		Neg. Bin.	$\hat{k} = 0.3736$	(0.3564, 0.3908)	$\hat{q} = 0.9346$	(0.9305, 0.9387)	-11851.6053	23707.2106
	out	**Zipf-PSS**	$\hat{\alpha} = 1.8503$	(1.8194, 1.8813)	$\hat{\lambda} = 0.8625$	(0.8311, 0.894)	**-10174.7132**	**20353.4265**
		D. Weibull	$\hat{q} = 0.5340$	(0.521, 0.547)	$\hat{\beta} = 0.5066$	(0.4936, 0.5197)	-10518.2129	21040.4257
		Neg. Bin.	$\hat{k} = 0.2493$	(0.2378, 0.2608)	$\hat{q} = 0.9553$	(0.9522, 0.9585)	-10714.8762	21433.7524

Table 5.7 For each of the degree sequences, one has: in the first column, the names of the models fitted; from the second to the fourth columns, the maximum likelihood parameter estimates together with their corresponding confidence intervals (CI); and the values of the AIC in the last column.

Year	**Type**	**Distribution**	$param_1$	CI_{param_1}	$param_2$	CI_{param_2}	**Log-like**	**AIC**
2011	total	**DGX**	$\hat{\mu} = -0.6337$	(-0.9348, -0.3326)	$\hat{\sigma} = 1.9742$	(1.8517, 2.0966)	**-12933.3840**	**25870.7681**
		Zipf-Mandelbrot	$\hat{\alpha} = 2.0495$	(1.9805, 2.1186)	$\hat{V} = 1.1162$	(0.9248, 1.3075)	-12940.8995	25885.7990
		Zipf-Polylog	$\hat{\alpha} = 1.4333$	(1.3993, 1.4672)	$\hat{\beta} = 0.9859$	(0.9834, 0.9884)	-12951.4498	25906.8996
		MOEZipf	$\hat{\alpha} = 2.0365$	(1.9928, 2.0803)	$\hat{\beta} = 2.6267$	(2.3786, 2.8749)	-12954.9064	25913.8127
		Zipf-PE	$\hat{\alpha} = 1.9771$	(1.9406, 2.0135)	$\hat{\beta} = 1.8101$	(1.6235, 1.9967)	-12961.9901	25927.9802
		zt-Zipf-PSS	$\hat{\alpha} = 1.8805$	(1.8502, 1.9108)	$\hat{\lambda} = 0.7291$	(0.6543, 0.8039)	-12971.3522	25946.7044
		Zipf	$\hat{\alpha} = 1.6710$	(1.6525, 1.6896)	-	-	-13136.9918	26275.9837
	in	**Zipf-PSS**	$\hat{\alpha} = 2.1226$	(2.0864, 2.1588)	$\hat{\lambda} = 1.2067$	(1.17, 1.2434)	**-11849.8717**	**23703.7434**
		D. Weibull	$\hat{q} = 0.6393$	(0.628, 0.6507)	$\hat{\beta} = 0.7229$	(0.7069, 0.7388)	-12159.3827	24322.7655
		Neg. Bin.	$\hat{k} = 0.5101$	(0.4867, 0.5336)	$\hat{q} = 0.8711$	(0.8642, 0.8779)	-12275.7841	24555.5682
	out	**Zipf-PSS**	$\hat{\alpha} = 1.94$	(1.9051, 1.975)	$\hat{\lambda} = 0.7191$	(0.6931, 0.7451)	**-10080.7118**	**20165.4236**
		D. Weibull	$\hat{q} = 0.4753$	(0.463, 0.4876)	$\hat{\beta} = 0.5180$	(0.5048, 0.5313)	-10432.8123	20869.6246
		Neg. Bin.	$\hat{k} = 0.2407$	(0.2294, 0.2521)	$\hat{q} = 0.9347$	(0.9302, 0.9392)	-10636.8105	21277.6210
2012	total	**DGX**	$\hat{\mu} = -0.7479$	(-1.0469, -0.449)	$\hat{\sigma} = 1.8698$	(1.753, 1.9865)	**-12626.5898**	**25257.1795**
		Zipf-Mandelbrot	$\hat{\alpha} = 2.1787$	(2.1001, 2.2573)	$\hat{V} = 1.1331$	(0.942, 1.3242)	-12626.6583	25257.3166
		MOEZipf	$\hat{\alpha} = 2.1220$	(2.0751, 2.169)	$\hat{\beta} = 2.5851$	(2.343, 2.8271)	-12639.1821	25282.3642
		Zipf-PE	$\hat{\alpha} = 2.0612$	(2.0218, 2.1006)	$\hat{\beta} = 1.7954$	(1.6078, 1.9829)	-12645.9667	25295.9334
		zt-Zipf-PSS	$\hat{\alpha} = 1.9591$	(1.9261, 1.992)	$\hat{\lambda} = 0.6796$	(0.61, 0.7493)	-12651.7092	25307.4184
		Zipf-Polylog	$\hat{\alpha} = 1.5048$	(1.469, 1.5406)	$\hat{\beta} = 0.9836$	(0.9805, 0.9867)	-12664.2093	25332.4185
		Zipf	$\hat{\alpha} = 1.7321$	(1.7124, 1.7519)	-	-	-12818.9781	25639.9562
	in	**Zipf-PSS**	$\hat{\alpha} = 2.3127$	(2.2705, 2.355)	$\hat{\lambda} = 1.2328$	(1.1976, 1.2681)	**-11761.9010**	**23527.8020**
		D. Weibull	$\hat{q} = 0.6473$	(0.6366, 0.6581)	$\hat{\beta} = 0.8046$	(0.788, 0.8211)	-12205.7467	24415.4935
		Neg. Bin.	$\hat{k} = 0.6456$	(0.6151, 0.6761)	$\hat{q} = 0.8130$	(0.804, 0.822)	-12299.6374	24603.2748
	out	**Zipf-PSS**	$\hat{\alpha} = 1.9926$	(1.9545, 2.0306)	$\hat{\lambda} = 0.6330$	(0.6098, 0.6563)	**-9704.8249**	**19413.6498**
		D. Weibull	$\hat{q} = 0.4352$	(0.4233, 0.4472)	$\hat{\beta} = 0.5146$	(0.5014, 0.5278)	-10077.7163	20159.4326
		Neg. Bin.	$\hat{k} = 0.2244$	(0.2136, 0.2352)	$\hat{q} = 0.9259$	(0.9208, 0.9311)	-10306.6735	20617.3471

Table 5.8 For each of the degree sequences, one has: in the first column, the names of the models fitted; from the second to the fourth columns, the maximum likelihood parameter estimates together with their corresponding confidence intervals (CI); and the values of the AIC in the last column.

Year	Type	Distribution	$param_1$	CI_{param_1}	$param_2$	CI_{param_2}	Log-like	AIC
2013	total	**Zipf-Mandelbrot**	$\hat{\alpha} = 2.1298$	(2.0561, 2.2035)	$\hat{V} = 0.7376$	(0.5889, 0.8863)	**-12500.4646**	**25004.9291**
		DGX	$\hat{\mu} = -1.6996$	(-2.2043, -1.195)	$\hat{\sigma} = 2.0740$	(1.9143, 2.2337)	-12506.1822	25016.3643
		MOEZipf	$\hat{\alpha} = 2.1210$	(2.0724, 2.1696)	$\hat{\beta} = 2.1057$	(1.9108, 2.3005)	-12513.0853	25030.1706
		Zipf-PE	$\hat{\alpha} = 2.0849$	(2.0425, 2.1273)	$\hat{\beta} = 1.4411$	(1.2562, 1.6259)	-12514.7396	25033.4793
		Zipf-Polylog	$\hat{\alpha} = 1.6127$	(1.5768, 1.6485)	$\hat{\beta} = 0.9853$	(0.982, 0.9885)	-12519.7207	25043.4414
		zt-Zipf-PSS	$\hat{\alpha} = 1.9846$	(1.9506, 2.0186)	$\hat{\lambda} = 0.4952$	(0.4311, 0.5593)	-12520.5797	25045.1593
		Zipf	$\hat{\alpha} = 1.7981$	(1.7772, 1.819)	-	-	-12627.6025	25257.2049
	in	**Zipf-PSS**	$\hat{\alpha} = 2.4882$	(2.4412, 2.5353)	$\hat{\lambda} = 1.2234$	(1.1905, 1.2562)	**-11861.0801**	**23726.1601**
		D. Weibull	$\hat{q} = 0.6520$	(0.6418, 0.6622)	$\hat{\beta} = 0.8721$	(0.8553, 0.8888)	-12503.9177	25011.8353
		Neg. Bin.	$\hat{k} = 0.7861$	(0.7487, 0.8235)	$\hat{q} = 0.7548$	(0.7442, 0.7655)	-12565.2384	25134.4768
	out	**Zipf-PSS**	$\hat{\alpha} = 1.9572$	(1.9196, 1.9949)	$\hat{\lambda} = 0.5484$	(0.5277, 0.5691)	**-9721.2156**	**19446.4311**
		D. Weibull	$\hat{q} = 0.3951$	(0.3835, 0.4066)	$\hat{\beta} = 0.4896$	(0.4765, 0.5028)	-10001.1292	20006.2583
		Neg. Bin.	$\hat{k} = 0.1977$	(0.188, 0.2075)	$\hat{q} = 0.9245$	(0.9192, 0.9299)	-10192.9956	20389.9913
2014	total	**DGX**	$\hat{\mu} = -1.5901$	(-2.086, -1.0942)	$\hat{V} = 2.0093$	(1.8501, 2.1684)	**-11175.2516**	**22354.5032**
		Zipf-Mandelbrot	$\hat{\alpha} = 2.1599$	(2.0785, 2.2414)	$\hat{V} = 0.7490$	(0.5879, 0.91)	-11176.6277	22357.2554
		Zipf-Polylog	$\hat{\alpha} = 1.6098$	(1.5703, 1.6492)	$\hat{\beta} = 0.9823$	(0.9784, 0.9862)	-11184.5143	22373.0286
		MOEZipf	$\hat{\alpha} = 2.1318$	(2.0802, 2.1835)	$\hat{\beta} = 2.0712$	(1.8689, 2.2734)	-11190.1869	22384.3739
		Zipf-PE	$\hat{\alpha} = 2.0927$	(2.0476, 2.1377)	$\hat{\beta} = 1.3869$	(1.1933, 1.5805)	-11193.5554	22391.1109
		zt-Zipf-PSS	$\hat{\alpha} = 1.9950$	(1.9587, 2.0313)	$\hat{\lambda} = 0.4747$	(0.4073, 0.5421)	-11199.8717	22403.7434
		Zipf	$\hat{\alpha} = 1.8130$	(1.7907, 1.8353)	-	-	-11288.8767	22579.7534
	in	**Zipf-PSS**	$\hat{\alpha} = 2.5422$	(2.4898, 2.5947)	$\hat{\lambda} = 1.2115$	(1.1772, 1.2458)	**-10573.2343**	**21150.4685**
		D. Weibull	$\hat{q} = 0.6485$	(0.6378, 0.6592)	$\hat{\beta} = 0.8943$	(0.8763, 0.9123)	-11106.9162	22217.8324
		Neg. Bin.	$\hat{k} = 0.8385$	(0.7949, 0.8821)	$\hat{q} = 0.7297$	(0.7175, 0.7419)	-11150.4633	22304.9266
	out	**Zipf-PSS**	$\hat{\alpha} = 1.9792$	(1.9391, 2.0194)	$\hat{\lambda} = 0.5551$	(0.5333, 0.577)	**-8814.9405**	**17633.8811**
		D. Weibull	$\hat{q} = 0.4000$	(0.3879, 0.4121)	$\hat{\beta} = 0.5064$	(0.4923, 0.5206)	-9061.9975	18127.9951
		Neg. Bin.	$\hat{k} = 0.2108$	(0.1998, 0.2218)	$\hat{q} = 0.9148$	(0.9086, 0.921)	-9223.2682	18450.5365

Table 5.9 For each of the degree sequences, one has: in the first column, the names of the models fitted; from the second to the fourth columns, the maximum likelihood parameter estimates together with their corresponding confidence intervals (CI); and the values of the AIC in the last column.

Year	Type	Distribution	$param_1$	CI_{param_1}	$param_2$	CI_{param_2}	Log-like	AIC
2015	total	**Zipf-Mandelbrot**	$\hat{\alpha} = 2.2064$	(2.1222, 2.2906)	$\hat{V} = 0.7589$	(0.5981, 0.9197)	**-10588.7673**	**21181.5346**
		DGX	$\hat{\mu} = -1.7482$	(-2.2947, -1.2017)	$\hat{\sigma} = 2.0181$	(1.8488, 2.1874)	-10591.5958	21187.1915
		MOEZipf	$\hat{\alpha} = 2.1665$	(2.1121, 2.2209)	$\hat{\beta} = 2.0814$	(1.8741, 2.2886)	-10595.9954	21195.9908
		Zipf-PE	$\hat{\alpha} = 2.1319$	(2.0841, 2.1796)	$\hat{\beta} = 1.4292$	(1.2287, 1.6297)	-10597.1536	21198.3073
		zt-Zipf-PSS	$\hat{\alpha} = 2.0264$	(1.9884, 2.0645)	$\hat{\lambda} = 0.4687$	(0.4025, 0.535)	-10602.0540	21208.1079
		Zipf-Polylog	$\hat{\alpha} = 1.6528$	(1.6129, 1.6928)	$\hat{\beta} = 0.9839$	(0.98, 0.9879)	-10608.1998	21220.3996
		Zipf	$\hat{\alpha} = 1.8354$	(1.8121, 1.8587)	-	-	-10692.2039	21386.4077
	in	**Zipf-PSS**	$\hat{\alpha} = 2.6118$	(2.5554, 2.6683)	$\hat{\lambda} = 1.1825$	(1.1487, 1.2164)	-10001.4149	20006.8297
		D. Weibull	$\hat{q} = 0.6352$	(0.6244, 0.6461)	$\hat{\beta} = 0.8830$	(0.8655, 0.9004)	**-10639.2672**	**21282.5344**
		Neg. Bin.	$\hat{k} = 0.8251$	(0.782, 0.8683)	$\hat{q} = 0.7262$	(0.7137, 0.7386)	-10699.6448	21403.2895
	out	**Zipf-PSS**	$\hat{\alpha} = 2.0214$	(1.9791, 2.0637)	$\hat{\lambda} = 0.5688$	(0.5464, 0.5912)	**-8540.3847**	**17084.7695**
		D. Weibull	$\hat{q} = 0.4058$	(0.3936, 0.4181)	$\hat{\beta} = 0.5222$	(0.5077, 0.5367)	-8814.8100	17633.6200
		Neg. Bin.	$\hat{k} = 0.2230$	(0.2112, 0.2348)	$\hat{q} = 0.9075$	(0.9008, 0.9142)	-8982.6090	17969.2179
2016	total	**Zipf-Mandelbrot**	$\hat{\alpha} = 2.8462$	(2.53, 3.1623)	$\hat{V} = 1.1906$	(0.7347, 1.6466)	**-2543.3546**	**5090.7092**
		DGX	$\hat{\mu} = -0.7904$	(-1.2879, -0.2928)	$\hat{\sigma} = 1.3946$	(1.213, 1.5761)	-2544.2196	5092.4393
		MOEZipf	$\hat{\alpha} = 2.5653$	(2.431, 2.6997)	$\hat{\beta} = 2.5182$	(2.0251, 3.0114)	-2546.9593	5097.9187
		Zipf-PE	$\hat{\alpha} = 2.5053$	(2.3876, 2.623)	$\hat{\beta} = 1.8478$	(1.4194, 2.2762)	-2547.7728	5099.5457
		Zipf-Polylog	$\hat{\alpha} = 1.6217$	(1.4972, 1.7462)	$\hat{\beta} = 0.9218$	(0.8942, 0.9495)	-2547.8006	5099.6011
		zt-Zipf-PSS	$\hat{\alpha} = 2.3520$	(2.2553, 2.4486)	$\hat{\lambda} = 0.4750$	(0.3703, 0.5798)	-2548.8617	5101.7234
		Zipf	$\hat{\alpha} = 2.0433$	(1.9896, 2.097)	-	-	-2586.5090	5175.0179
	in	**Zipf-PSS**	$\hat{\alpha} = 3.3093$	(3.1162, 3.5024)	$\hat{\lambda} = 0.9708$	(0.9178, 1.0238)	**-2446.6728**	**4897.3457**
		Neg. Bin.	$\hat{k} = 1.7500$	(1.4651, 2.0348)	$\hat{q} = 0.4156$	(0.3738, 0.4575)	-2524.5524	5053.1047
		D. Weibull	$\hat{q} = 0.5992$	(0.5782, 0.6203)	$\hat{\beta} = 1.1396$	(1.094, 1.1852)	-2533.0659	5070.1318
	out	**Zipf-PSS**	$\hat{\alpha} = 2.3938$	(2.2856, 2.502)	$\hat{\lambda} = 0.6210$	(0.5787, 0.6633)	**-2378.3336**	**4760.6672**
		D. Weibull	$\hat{q} = 0.4397$	(0.4173, 0.4621)	$\hat{\beta} = 0.7250$	(0.6896, 0.7604)	-2442.1930	4888.3861
		Neg. Bin.	$\hat{k} = 0.4404$	(0.3903, 0.4905)	$\hat{q} = 0.7387$	(0.7113, 0.766)	-2458.7030	4921.4060

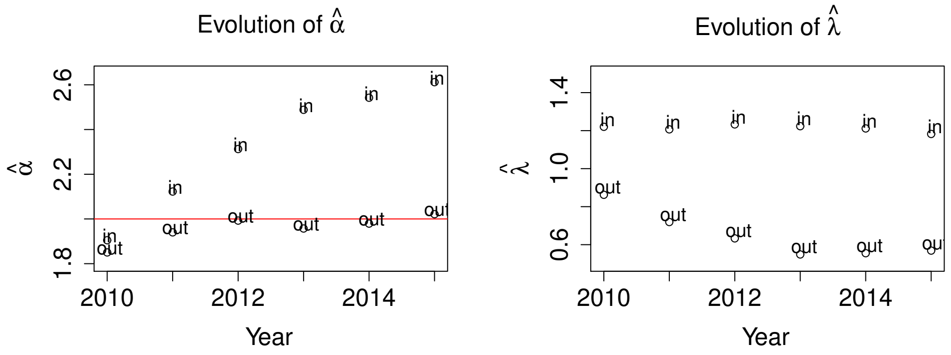

Fig. 5.3 Evolution of the $\hat{\alpha}$ (left-hand side) and the $\hat{\lambda}$ (right-hand side), MLE of the parameters of the Zipf-PSS distribution over the period 2010-2015 for the in-degree (in) and the out-degree (out) sequences.

Figure 5.3 shows the behavior of the parameter estimates over time. Since the data set contains only partial information related to the years 2009 and 2016, these years have been excluded from the figure. On the left-hand side of 5.3, we can observe the evolution of the system in terms of the $\hat{\alpha}$, which increases significantly in the in-degree sequence and slightly in that of the out-degree. The work of McKelvey et al. [2018], states that exponents of the PL that are lower than 2 appear in nascent systems, which agrees with our in-degree results, since a value smaller than two occurs only for the year 2010. The differences between the in- and out-degree sequence of the Q&A network are shown. While the $\hat{\alpha}$ of the in-degree sequences quickly departs from the threshold established in the literature, the $\hat{\alpha}$ for the out-degree stabilized at around 2.0, and thus time does not change the number of questions answered by a member of the community each time that she or he is active. With respect to the $\hat{\lambda}$ behavior (see, the right-hand side of 5.3), it seems to remain constant for the in-degree from nearly the beginning while it takes more time to stabilize for the out-degree. From this, it seems that, in all the years analyzed, the users are active in the networks approximately the same number of times.

It is important to observe that the fact that $\hat{\lambda}$ quickly stabilizes over time has a lot of sense, because all of them correspond to a one year period of time. The $\hat{\alpha}$ takes longer to stabilize because it is reasonable that, at the beginning, the questions receive less replies, and also that people answer less questions, as a consequence of the fact that there is less people in the platform. After observing this, and given that the Zipf-PSS is partially closed under addition, as it has been mentioned in Chapter 3, it would has sense to aggregate the years, and to analyze if the appropriate distribution for fitting such aggregation is the Zipf-PSS with the same α parameter and the λ parameter equal to k times the λ corresponding to one year. This is left as future work and it is mentioned at the end of this thesis.

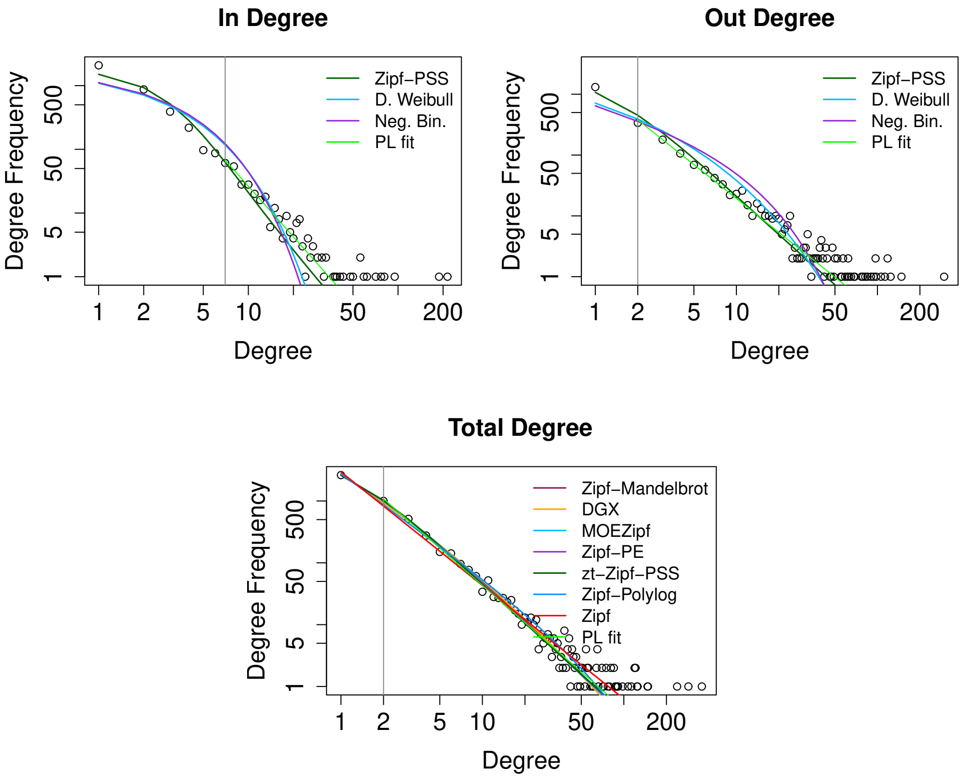

Fig. 5.4 In-degree (left-hand side), out-degree (right-hand side) total-degree (bottom) of the 2015 network, jointly with the fit obtained by each of the considered models. It also includes the estimate of the PL and the cut-off point achieved using the methodology proposed by Clauset et al. [2009].

Figure 5.4 shows the fits obtained by the models studied using the network's degree sequences for the year 2015. The plots also incorporate the fit obtained from the PL distribution by using the methodology proposed by Clauset et al. [2009], which initially determines a value x_{min} (cut-off), from which the PL is fitted. Establishing a cut-off point equal to 6 in the in-degree sequence generates a loss in observations of 92.3 %. In the case of the out-degree sequence, the cut-off is taken to be equal to 2, which implies that those users with less than two replies are not considered in the analysis, which accounts for 79.4% of the nodes in the sequence for the year 2015. Finally, setting a cut-off equal to 2 for the total degree sequence implies to not considering 48.49% of the information. The Zipf-PSS, the NB and the DW, avoid lost information by covering the whole range of the degree sequence. Note that the Zipf-PSS is the only bi-parametric model that can maintain the linearity in the tail. Comparing the fits of the considered models shows that the NB and DW have a more pronounced curvature than the Zipf-PSS; which, together with the earlier decay of their probabilities, demonstrates a clear deviation from the pattern shown by the data.

Since the Zipf-PSS and the NB are PSS distributions, their estimated parameters capture some insights into the behavior of the community members. For instance, for 2015 and the in-degree sequence, the parameter estimates of the Zipf-PSS and the NB are equal to: $\hat{\alpha} = 2.6118, \hat{\lambda} = 1.1825$ and $\hat{k} = 0.8251, \hat{q} = 0.7262$ (see, Table 5.9) respectively. From this, one has that the expected number of active users answering someone's questions is approximately equal to one ($\widehat{E[N]} = \hat{\lambda} = 1.1828$ for the Zipf-PSS and $\widehat{E[N]} = -\hat{k}\,log(1-\hat{q}) = 1.0689$ for the NB). Moreover, given that according to (1.1.7) $\widehat{E[X]} = 1.732$ for the Zipf-PSS, and $\widehat{E[X]} = -\hat{q}/(log(1-\hat{q})(1-\hat{q})) = 2.05$ for the NB, the expected number of answers provided by the active user is around 2.

On the other hand, for 2015 and the out-degree sequence, the parameter estimates of the Zipf-PSS and the NB are equal to: $\hat{\alpha} = 2.0214, \hat{\lambda} = 0.5688$ and $\hat{k} = 0.223, \hat{q} = 0.9075$ respectively. Thus, the expected number of times that a particular user is active for answering questions is approximately 0.5 ($\widehat{E[N]} = 0.5686$ for the Zipf-PSS and $\widehat{E[N]} = 0.5309$ for the NB). Once the user becomes active, the expected number of answered questions is $\widehat{E[X]} = 28.844$ for the Zipf-PSS and $\widehat{E[X]} = 4.12$ for the NB.

Note that the models agree in their estimates for the in-degree activity, but they provide quite different results for the number of answers in the out-degree. However, even though the Zipf-PSS is the model that gives the best fit to the data, the estimated number of a user's answers (out-degree) may be highly influenced by the large variability observed in the sample variance of the out-degree, which is larger than that obtained for the in-degree sequence. In addition, that result might also be a consequence of using the Riemann zeta function with a parameter close to the lower boundary of its parameter space; according to (1.1.7), its numerator is $\zeta(\hat{\alpha}-1) = \zeta(1.0214)$, which may lead to a less robust result.

5.4 Communication networks

5.4.1 University Rovira i Virgili

In this case study we analyze 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 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 5.10 summarizes the main statistics related to the network and its degree sequence. The large variability in the sequence may be a symptom of the existence of some nodes highly connected and others showing only a few number of connections. On the other hand, the skewness value does not allows to assume a long tight tail.

Table 5.10 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
1133	5451	70	9.6222	87.3059	1.7689

Table 5.11 contains the results obtained after fitting the degree sequence of the network by means of the Zipf distribution, the four extensions of this thesis, the Zipf-Mandelbrot and the DGX distributions. 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 5.5 shows the fits obtained by the each one of the considered models. 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.

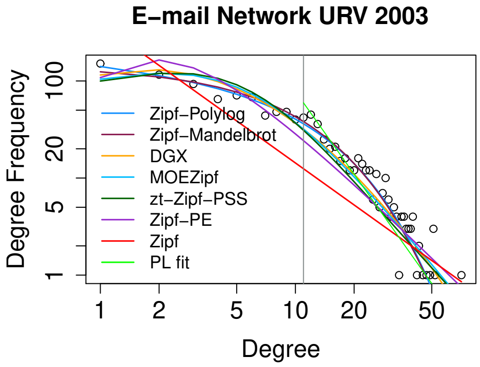

Fig. 5.5 Degree sequence of the URV e-mail network for the year 2003, and the fit obtained by each of the considered models.

LRT is performed to compare the fit of the Zipf-Polylog distribution with that of the Zipf, that is, to test if $H_0 : \beta = 1$ vs. $H_1 : \beta < 1$. 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 LR statistics is equal to $-2\,[-4106.629 - (-3632.26)] = 948.738$. Under the null hypothesis, the LR statistic follows a 50:50 mixture of χ^2_0 and χ^2_1 as in the example of section 5.3.1. Thus, the critical value for $\alpha = 0.05$ is equal to $0.5\,\chi^2_{0.95,1} = 0.5 \cdot 3.84 = 1.92$. 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 4.1 of Chapter 4, 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 12, the 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.

5.4.2 EU institution

In this example we analyze the degree sequence of a communication network that represents the email traffic of a European Research Institution for a period of 18 months (October 2003 - May 2005). It has a total of 265214 nodes, representing email accounts, and 420045 directed edges. In particular, an edge is created between nodes i and j if they have exchanged emails in either directions; that is, if node i has sent at least one message to node j and vice-versa. This network was previously analyzed in the work by Leskovec et al. [2007], in which the authors studied how the densification and average distance of the network behave over time. The data set is publicly available at the Stanford Network Analysis Project (SNAP) repository [Leskovec and Krevl, 2014] (http://snap.stanford.edu/data/email-EuAll.html).

Because it is a directed network, the degree sequence needs to be split into three different sequences: the *in-degree*, containing the information related to the amount of emails received; the *out-degree*, representing the number of emails sent; and the *total-degree*, which corresponds to the total activity of the accounts.

Table 5.11 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 at the University of California in 2003.

Distribution	$param_1$	CI_{param_1}	$param_2$	CI_{param_2}	**Log-like**	**AIC**
Zipf-Polylog	$\hat{\alpha} = 0.1774$	(0.0689, 0.2859)	$\hat{\beta} = 0.9108$	(0.8994, 0.9223)	**-3632.2648**	**7268.5295**
Zipf-Mandelbrot	$\hat{\alpha} = 17.83187$	(-6.367, 42.0306)	$\hat{V} = 145.3025$	(-74.8625, 365.4674)	-3635.6155	7275.2310
DGX	$\hat{\mu} = 1.7524$	(1.6848, 1.82)	$\hat{\sigma} = 1.0924$	(1.0385, 1.1463)	-3673.7135	7351.4271
MOEZipf	$\hat{\alpha} = 2.4980$	(2.4105, 2.5855)	$\hat{\beta} = 28.8413$	(22.1355, 35.547)	-3698.9643	7401.9286
zt-Zipf-PSS	$\hat{\alpha} = 2.0056$	(1.9524, 2.0588)	$\hat{\lambda} = 3.118$	(2.9051, 3.3309)	-3722.6859	7449.3719
Zipf-PE	$\hat{\alpha} = 2.0102$	(1.959, 2.0613)	$\hat{\beta} = 5.9809$	(5.3691, 6.5927)	-3770.9683	7545.9366
Zipf	$\hat{\alpha} = 1.43744$	(1.4116, 1.4632)	-	-	-4106.6291	8215.2582

Table 5.12 summarizes the main characteristics of the network. It can be observed that it contains 1089 loops, meaning that some email accounts have acted as both sender and receiver of the same email. With respect to the basic statistics of the sequences, note that the means of the in- and out-degree sequences are equal, as in the Math-Overflow example, which is one of the conditions required for being a valid degree sequence in a directed network as pointed out in Section 5.1. The large values for the variance is a frequent characteristic in real networks because some nodes have very different behaviors, due to the fact that the degrees of the nodes may differ in orders of magnitude. Finally, the values of the skewness statistic are huge, because the data distribution has a long right tail in all the three sequences.

Table 5.12 Characteristics of each degree sequence: number of nodes (#N); number of edges (#E); number of loops (L); number of multi-edges (M); type of sequence (Type); (Range); (Mean); variance (Var); skewness (Skew).

#N	#E	L	M	Type	Range	Mean	Var	Skew
265214	420045	1089	0	total	7635	3.17	1810.56	73.45
				in	7631	1.58	1334.24	97.55
				out	930	1.58	99.68	38.41

Table 5.13 contains the MLE for each distribution considered, jointly with their confidence intervals. It also contains the log-likelihood at the maximum likelihood estimations, and the AIC goodness-of-fit measure. Observe that in the three cases, the MOEZipf and the Zipf-PE distributions behave quite similar and are clearly better than the Zipf model. However, the Zipf-PE model is the one offering the best fit to the three degree sequences.

Figure 5.6 shows, in log-log scale, the fits obtained for the each one of the sequences. The reason why this example has been chosen is that it shows a top-convex pattern as it can be appreciated in this figure. This is not common in real data sets which usually show either a top-concave pattern or a straight line. Since not all the distributions considered before are able to fit this type of pattern, the plots include only the fits associated with distributions that allow for linearity or top-convexity. Observe that the Zipf-Mandelbrot and the Zipf-Polylog are providing the same fit than Zipf distribution. They parameter values confirm that both distributions families are in the particular case when they become Zipf. Moreover, the fit achieved using the methodology by Clauset et al. [2009] is included only in the plot of the out-degree (right-hand side). Observe that, after setting the cut-off value (vertical line) 97.7% of the observations are left out of the fitted area. The results for the in-degree and the total-degree are not included in the plot because, after fixing the cut-off value, only 0.04% and 0.047% of the observations should be considered, respectively, which has no sense.

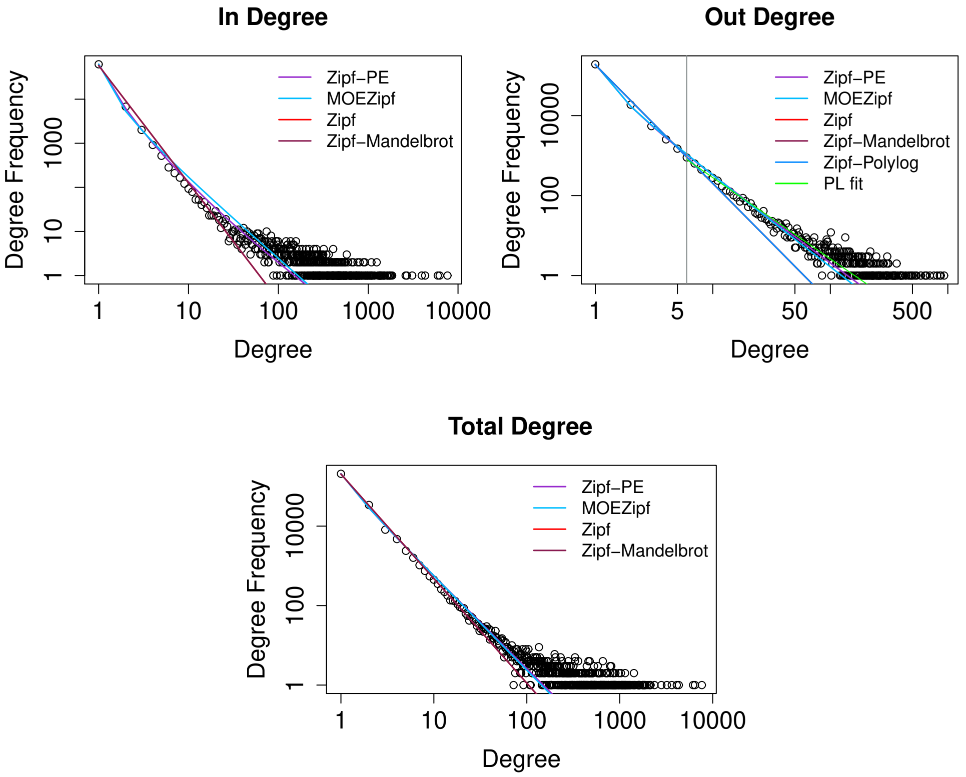

Fig. 5.6 Fit associated with the in-degree (left hand side), the out-degree (right hand side) and the total-degree (centered). It also includes the fit obtained using the methodology proposed by Clauset et al. [2009].

We perform the LRT to compare the Zipf distribution with the Zipf-PE extension, i.e, to test $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$. Assuming a significance level of $\alpha = 0.05$, the critical point is equal to $\chi^2_{0.95,1} = 3.84$. The likelihood ratio statistic for the in- and out-degree sequences is equal, respectively, to $-2\,[-66218.20 - (-63246.21)] = 5943.98$ and $-2\,[-156765.21 - (-154287.55)] = 4955.32$. Since both statistics are larger than the critical point, the null hypothesis can be clearly rejected, which ensures that the Zipf-PE provides a better fit that the Zipf distribution.

The MOEZipf and Zipf-PE models agree, meaning that the data distribution is defined in both cases in terms of minimums. This is because $\hat{\beta} < 0$ for the Zipf-PE and $\hat{\beta} \in (0,1)$ for the MOEZipf.

Considering the parameter estimates for the in-degree sequence, the expected number of times that users are connected for receiving emails, after being notified by the system, is $\widehat{E[N]} = \hat{\beta}/(1 - e^{-\hat{\beta}}) = 4.73$ for the Zipf-PE. For the MOEZipf distribution, $\widehat{E[N]} = 1/\hat{\beta} = 4.55$ may be interpreted as the number of times that a person checks the email until a particular contact sends an email to him/her. We are not able to calculate the expected number of emails received every time that a user is connected, because the $\hat{\alpha}$ for both distributions does not fulfill the condition $\hat{\alpha} > 2$, which is the one required in order that the expected value of the Zipf distribution exists. Since the same requirement is needed for calculating the expected values of the Zipf-PE and MOEZipf distributions, a global estimation of the number of received emails cannot be computed for none of these cases. This situation bring us to interpreted that the number of emails received by the user is huge, close to infinity.

With respect to the out-degree sequence, the expected number of times that the user was connected for sending emails is $\widehat{E[N]} = 2.39$ for the Zipf-PE distribution. For the MOEZipf, $\widehat{E[N]} = 2.56$ may be interpreted as the number of times that a user connects to the email application until he/she reaches a particular contact. Every time that the user is active, the expected number of emails sent is estimated by $\widehat{E[X]} = 4.51$ for the Zipf-PE and $\widehat{E[X]} = 2.89$ for the MOEZipf distribution. The expected number of sent emails in the entire network is, respectively, $\widehat{E[Y]} = 1.68$ for the Zipf-PE and $\widehat{E[Y]} = 1.61$ for the MOEZipf.

5.5 Collaboration networks

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 5.14 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.

Table 5.13 The parameter estimations for each one of the analyzed distributions as well as their confidence interval, the log-likelihood, and the AIC goodness-of-fit measure.

Type	**Distribution**	$param_1$	CI_{param_1}	$param_2$	CI_{param_2}	**Log-like**	**AIC**
total	**Zipf-PE**	$\hat{\alpha} = 2.3321$	(2.3141, 2.3501)	$\hat{\beta} = -0.8444$	(-0.8924, -0.7963)	**-240334.3928**	**480672.7855**
	MOEZipf	$\hat{\alpha} = 2.3703$	(2.3545, 2.386)	$\hat{\beta} = 0.6830$	(0.6682, 0.6978)	-240389.4561	480782.9122
	Zipf	$\hat{\alpha} = 2.6316$	(2.6244, 2.6387)	-	-	-241004.5560	482011.1119
	Zipf-Mandelbrot	$\hat{\alpha} = 2.6316$	(2.6078, 2.6553)	$\hat{V} = 0$	(-0.0185, 0.0185)	-241004.4979	482012.9959
in	**Zipf-PE**	$\hat{\alpha} = 1.4789$	(1.4509, 1.5069)	$\hat{\beta} = -4.6858$	(-4.918, -4.4537)	**-63246.2079**	**126496.4158**
	MOEZipf	$\hat{\alpha} = 1.7755$	(1.7516, 1.7993)	$\hat{\beta} = 0.2182$	(0.208, 0.2284)	-63600.1243	127204.2486
	Zipf	$\hat{\alpha} = 2.6609$	(2.6472, 2.6746)	-	-	-66218.2045	132438.4089
	Zipf-Mandelbrot	$\hat{\alpha} = 2.6609$	(2.551, 2.7709)	$\hat{V} = 0$	(-0.0927, 0.0927)	-66218.1920	132440.3841
out	**Zipf-PE**	$\hat{\alpha} = 2.1643$	(2.1369, 2.1918)	$\hat{\beta} = -2.0901$	(-2.1667, -2.0135)	**-154287.5483**	**308579.0967**
	MOEZipf	$\hat{\alpha} = 2.2839$	(2.2639, 2.3039)	$\hat{\beta} = 0.3897$	(0.3788, 0.4006)	-154399.8201	308803.6402
	Zipf	$\hat{\alpha} = 2.9677$	(2.9579, 2.9774)	-	-	-156765.2091	313532.4182
	Zipf-Mandelbrot	$\hat{\alpha} = 2.9677$	(2.9144, 3.0209)	$\hat{V} = 0$	(-0.035, 0.035)	-156765.0725	313534.1451
	Zipf-Polylog	$\hat{\alpha} = 2.9677$	(2.9579, 2.9774)	$\hat{\beta} = 1$	(0.9997, 1.0003)	-156765.0907	313534.1814

During the analysis of this degree sequence, we considered two approaches. The first one is devoted to analyzing only the nodes that have at least one connection in the network. Consequently, this part of the study considered the distributions with support in the strictly positive numbers. The second approach also considers the 851 isolated nodes, and the degree sequence is fitted with the distributions that include the zero value in its support.

Table 5.14 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 5.15 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 5.7 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, 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.

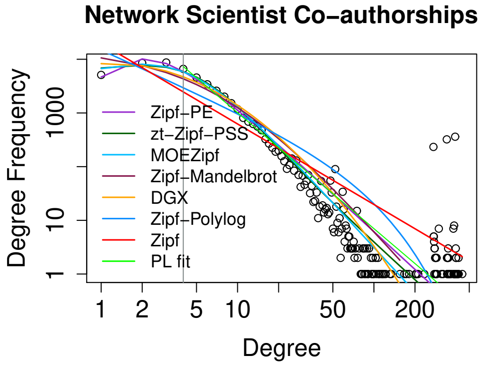

Fig. 5.7 Degree sequence of the co-authorship network and, the fit obtained by each one of the considered models.

Based on the parameter interpretation of the Zipf-PE distribution, we can say that an author published an average of 8.24 ($\widehat{E[N]} = 8.2382$) papers in the period 1998-2019. Also, the average number of co-authors in one author's publication is around 2.49 ($\widehat{E[X]} = 2.48779$). Moreover, it is expected that each author has an average of $\widehat{E[Y]} = 9.52$ co-authors in this period. 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. 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.

In a similar way, the Zipf-PSS is the distribution that provides the best fit when the isolated nodes are included in the sequence. Table 5.16 summarizes the statistics obtained during the fitting process, and Figure 5.7 illustrates the fits obtained by each one of the distributions used as part of this experiment. Note that the fit of the Zipf-PSS is clearly better than those obtained with the NB and DW distributions. The cut-off set by Clauset's methodology remains equal to 4, but the percentage of nodes not included in the analysis increases until reaching 55.23%, which is a consequence of adding the isolated nodes.

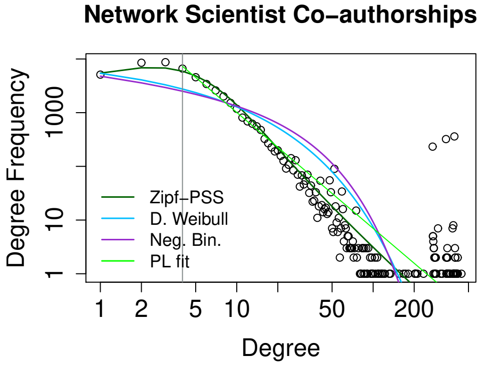

Fig. 5.8 Degree sequence of the co-authorship network (including isolated nodes) and, the fit obtained by each one of the considered models.

An interpretation based on the parameters of the Zipf-PSS suggests that the expected number of papers published by an author over the analyzed period is equal to 3 ($\widehat{E[N]} =$ 3.0084). Observe that this estimate is smaller than the one obtained in the previous scenario. This may be a consequence of the fact that the sequence contains authors who have not published with other members of the network, perhaps because these authors have worked in the field of network science for only a limited amount of time. Also, the average number of co-authors is set to be equal to 2.68 ($\widehat{E[X]} = 2.683409$), which agrees with the number of co-authors estimated above. In general, the results indicate that each author has, on average, about 8.073 ($\widehat{E[Y]} = 8.072767$) co-authors in this period.

Table 5.15 Distributions fitted jointly with their parameter estimates, confidence intervals, log-likelihood and the AIC goodness-of-fit measure for the degree sequence of the co-author network.

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
Zipf-Mandelbrot	$\hat{\alpha} = 3.3498$	(3.2903, 3.4093)	$\hat{V} = 9.5651$	(9.2339, 9.8964)	-148073.9400	296151.8800
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

Table 5.16 Fitted distributions jointly with their parameter estimates, confidence intervals, log-likelihood and the AIC goodness-of-fit measure, for the degree sequence of the co-author network.

Distribution	$param_1$	CI_{param_1}	$param_2$	CI_{param_2}	**Log-like**	**AIC**
Zipf-PSS	2.3095	(2.2992, 2.3198)	3.0084	(2.9872, 3.0296)	**-152392.6602**	**304789.3203**
D. Weibull	0.8029	(0.8004, 0.8054)	0.7037	(0.7, 0.7074)	-175125.0232	350254.0463
Neg. Bin.	0.5695	(0.5633, 0.5756)	0.9566	(0.956, 0.9573)	-181282.1796	362568.3592

Conclusions and future work

Ever since the Zipf distribution [Zipf, 1949] was first introduced, it has been widely associated with the linguistic domain and used extensively to measure the frequency of words appearing in a given text. In recent years, its applicability has also been extended to the area of Network Analysis, where the Zipf distribution is used as a model for the degree distribution of real networks. The use of this distribution is justified mainly because the networks usually contain a large number of nodes with just a few of connections and few nodes that are highly connected.

However, the Zipf distribution is not flexible enough to capture the pattern shown by the degree sequence of many real networks when plotted in log-log scale. In these plots, one usually observes a top-concave or top-convex pattern that deviates from a perfect straight line corresponding to the Zipf distribution. This justifies the need to define alternative models that are more flexible, especially in the first integer values. Several methodologies have been developed to deal with such phenomena, but the most commonly used ones incur a heavy loss of information, because valuable data is left out of the analysis. This is not the case of the four extensions presented in this thesis, which allow to fit the data in its entire range. In what follows we describe the main results of this work.

From the theoretical point of view:

1. We have proved that the Zipf distribution is a continuous mixture of geometric distributions, and that it is also an MZTP but not a ZTMP distribution.
2. In order to overcome the lack of flexibility in the Zipf distribution, we have focused on the Zipf extensions that are defined based on their ability to:
   a) Fit the data in its entire range by maintaining linearity in the tail, which avoids the selection of a cut-off point and, consequently, prevents the loss of information.
   b) Provide information about the underlying data generation mechanism. In other words, the interpretation of the proposed model's parameters can reveal some insights about the data generation mechanism.

3. For each one of the four bi-parametric extensions considered, we have performed the following:

    a) The work related to the first extension, the MOEZipf, is a continuation of the research work begun by A. Casellas during her master's thesis [Casellas, 2013]. Here, we have obtained new properties related to this distribution family, such as the asymptotic relationship between the tail of a MOEZipf and a Zipf distribution with the same α parameter. In addition, we have developed an efficient methodology for generating random numbers from a Zipf random number generator.

    b) The second and third extensions, the Zipf-PE and the Zipf-PSS, are completely novel distributions. Therefore, they have been defined during this research work, with their main properties having been analyzed and included in this thesis. Both of these families of distributions include the Poisson distribution in their definition, which results in a natural interpretation of their parameters. The Zipf-PE allows for top-concavity and top-convexity while the Zipf-PSS only allows for top-concavity. However, the Zipf-PSS contains the zero value in its support, which allows fitting real data that contain this value, such as graphs that contain isolated nodes. Both families of distributions have a linear tail, which is asymptotically related to the Zipf distribution. Also, we have proposed a methodology for generating random numbers from these two distributions.

    c) Finally, we have introduced the Zipf-Polylog family of distributions, which are obtained as a natural generalization of the PGF of the Zipf distribution. However, this family is already present in the literature under the names *PL distribution with an exponential cut-off* and *hybrid geometric/power model*. To the best of our knowledge, we have proved several properties of this distribution that do not appear in the literature. Among them, we have demonstrated that: it belongs to the exponential family; it is unlike the others because it has moments of any order; and it is a continuous mixture of geometric distributions. Apart from that, we have set the conditions for it to be ZTMP and/or MZTP. Furthermore, we have adapted the classical inversion method for efficiently generating random numbers.

From the practical point of view:

1. We have developed and released the *zipefextR* R-package, which contains the implementation of all the Zipf extensions included in this thesis.

2. We have prepared a guide that helps for distinguishing from among all the models the most appropriate for fitting a particular data set.

3. This document provides the fittings obtained for 32 degree sequences associated with real networks. These networks pertain to different research areas such as biology, social networks, collaboration or communication networks. Most of the sequences analyzed show a top-concave pattern and they do not contain isolated nodes. However, others contain the information related to isolated nodes and one exhibits a top-convex pattern. The outcome of the fitting processes allows stating the significance of the proposed models compared to the results obtained from the Zipf or from its other bi-parametric alternatives that are usually used in the literature.

Future Work

In what follows, we describe several ideas that have emerged while developing this thesis, which we propose as new lines of research that we would like to pursuer in our investigations.

Analysis of the degree distribution of temporal networks

Since the Zipf-PSS distribution has the property of being partially closed under addition, as mentioned in Section 3.3 of Chapter 3, we propose to study the degree sequence distribution of temporal networks, also known as time-varying networks. The ability to properly model the evolution of a network over time is very important, and a good starting point for this is to ascertain the evolution of the degree sequence. By way of example, it is necessary to know the evolution of a cloud-hosted social network, in order to predict the resources that it will require in the coming years.

Using the proposed Zipf extensions in the field of linguistics

The Zipf and the Zipf-Mandelbrot distributions are perhaps the ones most commonly used in the field of linguistics. Here, we propose to analyze how the Zipf extensions of this thesis perform when modeling the frequency of words in a given text. In particular, we would like to begin by considering the books included in the Gutenberg[3] Project. The Project Gutenberg was created in 1971 with the goal of creating an electronic library containing all versions of the free books for which their U.S copyright has expired. This library contains more than 60000 freely available books in different languages.

[3] https://www.gutenberg.org/

Our objective is to select different books by the same author and compare the fits obtained by our models with those obtained using other distributions such as the Zipf, Zipf-Mandelbrot, NB or Inverse-Gaussian Poisson distribution. If any of our models turn out to be appropriate for this field of study, then it can help extract features about the authors and identify the otherwise unknown authorship of some texts.

In order to obtain the frequency counts, we propose to use the software developed in the work by Gerlach and Font-Clos [2018].

Using the proposed Zipf extensions in PPI networks

Thanks to the grant Ferran Sunyer i Balaguer, from the Institut d'Estudis Catalans, which the author of this thesis received in 2018, it was possible to reach out to the researchers Živa Ramšak and Kristina Gruden of the National Institute of Biology of Slovenia. During my one-month stay in this institute, it was possible to begin understanding how PPI networks are constructed. Some of these networks are composed of three different layers of information: protein-protein interaction, transcriptional regulation and regulation through microRNA.

The example shown in Section 5.2 of Chapter 5 of this thesis was analyzed during the research stay. However, we would like to go further and study the evolution of the *Arabidopsis thaliana* and the *Solanum tuberosum* networks every time that a new layer of information is included. Also, we would like to add a new layer of information resulting from a link prediction approach. This study requires, on the one hand, developing a methodology for imputing the missing links in a network and, on the other, measuring changes in the degree distribution when extra information is added.

Using the proposed Zipf extensions in regression models

A way of continuing the research started in this thesis could be to consider regression models where the response variable follows one of the Zipf extensions proposed. This idea has an extra difficulty for the MOEZipf, Zipf-PE, and Zipf-PSS distributions because the expected value is not defined in all the parameter space. This is not the case for the Zipf-Polylog distribution which has always moments of any order, and it also belongs to the exponential family. Hence, in this new research line, we propose to start considering regression models with the Zipf-Polylog as the distribution of the dependent variable.

Using the proposed Zipf extensions in generating synthetic graphs

People who work in different areas that require graphs analytics are faced with data privacy limitations. In most of those cases, the use of personal data prevents the release of real data sets for research. Thus, it is necessary to create synthetic data sets based on the characteristics observed in reality. In this future research line, we propose to use random degree sequences generated from the distributions presented in this thesis, as input parameters of well know random graphs generators algorithms as, for instance, the one introduced by Aiello et al. [2000] or DATAGEN [Erling et al., 2015], and to compare the resulting graphs with respect to the ones observed in reality. For this comparison we can use several of the measures of the graphs, such diameter, clustering coefficient or number of communities.

References

Adamic, L. (2011). Complex systems: Unzipping Zipf's law. *Nature*, 474(7350):164.

Adamic, L. A. and Huberman, B. A. (2002). Zipf's law and the Internet. *Glottometrics*, 3(1):143–150.

Aiello, W., Chung, F., and Lu, L. (2000). A random graph model for massive graphs. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 171–180.

Apostol, T. M. (1974). Mathematical analysis.

Arnold, T. B. and Emerson, J. W. (2011). Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions. *R Journal*, 3(2).

Auerbach, F. (1913). Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, 59:74–76.

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

Baek, S. K., Bernhardsson, S., and Minnhagen, P. (2011). Zipf's law unzipped. *New Journal of Physics*, 13(4):043004.

Bak, P., Tang, C., and Wiesenfeld, K. (1987). Self-organized criticality: An explanation of the 1/f noise. *Physical review letters*, 59(4):381.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Barabási, A.-L. and Bonabeau, E. (2003). Scale-free networks. *Scientific american*, 288(5):60–69.

Barabási, A.-L. and Pósfai, M. (2016). *Network science*. Cambridge university press.

Barndorff-Nielsen, O. (2014). *Information and exponential families: in statistical theory*. John Wiley & Sons.

Bhattacharya, A., Chen, B., van der Hofstad, R., and Zwart, B. (2020). Consistency of the plfit estimator for power-law data. *arXiv preprint arXiv:2002.06870*.

Bi, Z., Faloutsos, C., and 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*, pages 17–26. ACM.

Boginski, V., Butenko, S., and Pardalos, P. M. (2005). Statistical analysis of financial networks. *Computational statistics & data analysis*, 48(2):431–443.

Broido, A. D. and Clauset, A. (2019). Scale-free networks are rare. *Nature communications*, 10(1):1–10.

Cancho, V. G., Louzada-Neto, F., and Barriga, G. D. C. (2011). The Poisson-exponential lifetime distribution. *Comput. Statist. Data Anal.*, 55(1):677–686.

Caron, Y., Makris, P., and 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.

Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703.

Conover, W. J. (1972). A kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association*, 67(339):591–596.

Corral, A., Serra, I., and Ferrer-i Cancho, R. (2019). The distinct flavors of Zipf's law in the rank-size and in the size-distribution representations, and its maximum-likelihood fitting. *arXiv preprint arXiv:1908.01398*.

Csermely, P. (2009). *Weak links: the universal key to the stability of networks and complex systems*. Springer Science & Business Media.

Duarte-López, A., Casellas, A., and Pérez-Casany, M. (2017). *moezipfR: Marshall-Olkin Extended Zipf*. R package version 1.0.2.

Duarte-López, A. and Pérez-Casany, M. (2020). *zipfextR: Zipf Extended Distributions*. R package version 1.0.2.

Duarte-López, A., Pérez-Casany, M., and Valero, J. (2020a). Random Stopped Extreme Zipf Extensions. (Submmited, 2020).

Duarte-López, A., Pérez-Casany, M., and Valero, J. (2020b). The Zipf–Poisson-stopped-sum distribution with an application for modeling the degree sequence of social networks. *Comput. Statist. Data Anal.*, 143:106838.

Duarte-López, A., Prat-Pérez, A., and Pérez-Casany, M. (2015). Using the Marshall-Olkin extended Zipf distribution in graph generation. In *European Conference on Parallel Processing*, pages 493–502. Springer.

Ectors, W., Kochan, B., Janssens, D., Bellemans, T., and Wets, G. (2018). Exploratory analysis of Zipf's universal power law in activity schedules. *Transportation*, pages 1–24.

Englehardt, J. D. and Li, R. (2011). The discrete Weibull distribution: an alternative for correlated counts with confirmation for microbial counts in water. *Risk Analysis: An International Journal*, 31(3):370–381.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

Erling, O., Averbuch, A., Larriba-Pey, J., Chafi, H., Gubichev, A., Prat, A., Pham, M.-D., and Boncz, P. (2015). The LDBC social network benchmark: Interactive workload. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 619–630.

Estoup, J. (1916). Gammes Sténographiques. *Paris: Institut Sténographique de France*.

Evert, S. and Baroni, M. (2007). zipfr: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 29–32. Association for Computational Linguistics.

Feller, W. (1971). An introduction to probability theory and its applications. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1971, 3rd ed.*

Fenner, T., Levene, M., and Loizou, G. (2005). A stochastic evolutionary model exhibiting power-law behaviour with an exponential cutoff. *Physica A: Statistical Mechanics and its Applications*, 355(2-4):641–656.

Ferrer-i Cancho, R. and Vitevitch, M. S. (2018). The origins of Zipf's meaning-frequency law. *Journal of the Association for Information Science and Technology*, 69(11):1369–1379.

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of eugenics*, 6(1):13–25.

Floudas, C. (2009). Pardalos, et Panos. Encyclopedia of Optimization.

Gerlach, M. and Font-Clos, F. (2018). A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *arXiv preprint arXiv:1812.08092*.

Ghitany, M., Al-Awadhi, F., and Alkhalfan, L. (2007). Marshall–Olkin extended Lomax distribution and its application to censored data. *Communications in Statistics - Theory and Methods*, 36(10):1855–1866.

Ghitany, M., Al-Hussaini, E., and Al-Jarallah, R. (2005). Marshall–Olkin extended Weibull distribution and its application to censored data. *Journal of Applied Statistics*, 32(10):1025–1034.

Gillespie, C. (2015). Fitting Heavy Tailed Distributions: The poweRlaw Package. *Journal of Statistical Software, Articles*, 64(2):1–16.

Girvan, M. and Newman, M. E. (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.

Gomez-Lievano, A., Youn, H., and Bettencourt, L. M. (2012). The statistics of urban scaling and their connection to Zipf's law. *PloS one*, 7(7):e40393.

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

Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., and Arenas, A. (2003). Self-similar community structure in a network of human interactions. *Physical review E*, 68(6):065103.

Gulisashvili, A. (2012). *Analytically tractable stochastic stock price models*. Springer Finance. Springer, Heidelberg.

Güney, Y., Tuaç, Y., and Arslan, O. (2017). Marshall-Olkin distribution: parameter estimation and application to cancer data. *J. Appl. Stat.*, 44(12):2238–2250.

Günther, R., Levitin, L., Schapiro, B., and Wagner, P. (1996). Zipf's law and the effect of ranking on probability distributions. *International Journal of Theoretical Physics*, 35(2):395–417.

Hill, B. M. and Woodroofe, M. (1975). Stronger forms of Zipf's law. *Journal of the American Statistical Association*, 70(349):212–219.

Jessen, H. A. and Mikosch, T. (2006). Regularly varying functions. *Publications de l'Institut Mathematique*, 80(94):171–192.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate discrete distributions*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition.

Jose, K. (2011). Marshall-Olkin family of distributions and their applications in reliability theory, time series modeling and stress-strength analysis. *Proc. ISI 58th World Statist. Congr Int Stat Inst, 21st-26th August*, pages 3918–3923.

Karlis, D. and Xekalaki, E. (2005). Mixed poisson distributions. *International Statistical Review*, 73(1):35–58.

Keller, J. (2010). Beyond Facebook: How the World's Mathematicians Organize Online. https://www.theatlantic.com/technology/archive/2010/09/beyond-facebook-how-the-worlds-mathematicians-organize-online/63422/. (Accessed on 10/01/2019).

Kemp, A. W. (2010). Families of power series distributions, with particular reference to the lerch family. *Journal of statistical planning and inference*, 140(8):2255–2259.

Kim, H., Del Genio, C. I., Bassler, K. E., and Toroczkai, Z. (2012). Constructing and sampling directed graphs with given degree sequences. *New Journal of Physics*, 14(2):023012.

Koziol, J., Griffin, N., Long, F., Li, Y., Latterich, M., and Schnitzer, J. (2013). On protein abundance distributions in complex mixtures. *Proteome science*, 11(1):5.

Krumme, C., Llorente, A., Cebrian, M., Moro, E., et al. (2013). The predictability of consumer visitation patterns. *Scientific reports*, 3:1645.

Kuş, C. (2007). A new lifetime distribution. *Comput. Statist. Data Anal.*, 51(9):4497–4509.

Kunegis, J. (2013). Konect: the koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1343–1350. ACM.

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

Lee, M. H. (1997). Polylogarithms and Riemann's ζ function. *Phys. Rev. E*, 56:3909–3912.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2.

Leskovec, J. and Krevl, A. (2014). SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.

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

Low, W. J., Wilson, P., and Thelwall, M. (2016). Stopped sum models and proposed variants for citation data. *Scientometrics*, 107(2):369–384.

Mahanti, A., Carlsson, N., Mahanti, A., Arlitt, M., and Williamson, C. (2013). A tale of the tails: Power-laws in internet measurements. *IEEE Network*, 27(1):59–64.

Malone, D. and Maher, K. (2012). Investigating the distribution of password choices. In *Proceedings of the 21st international conference on World Wide Web*, pages 301–310. ACM.

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

Mandelbrot, B. (1965). Information theory and psycholinguistics. *BB Wolman and E.*

Marshall, A. W. and 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.

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.

Meng, S. and Gao, G. (2018). Compound Poisson Claims Reserving Models: Extensions and Inference. *ASTIN Bulletin: The Journal of the IAA*, pages 1–20.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251.

Molontay, R. and Nagy, M. (2019). Two Decades of Network Science as seen through the co-authorship network of network scientists. *arXiv preprint arXiv:1908.08478.*

Nakagawa, T. and Osaki, S. (1975). The discrete Weibull distribution. *IEEE Transactions on Reliability*, 24(5):300–301.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351.

Panjer, H. H. (1981). Recursive evaluation of a family of compound distributions. *Astin Bull.*, 12(1):22–26.

Paranjape, A., Benson, A. R., and Leskovec, J. (2017). Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 601–610. ACM.

Patel, Y. C. (1976). Estimation of the parameters of the triple and quadruple stuttering-Poisson distributions. *Technometrics*, 18(1):67–73.

Patil, G. P. and Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, pages 179–189.

Pérez-Casany, M. and Casellas, A. (2013). Marshall-Olkin Extended Zipf Distribution. *arXiv preprint arXiv:1304.4540.*

Pérez-Casany, M., Valero, J., and 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.

Podur, J. J., Martell, D. L., and Stanford, D. (2010). A compound Poisson model for the annual area burned by forest fires in the province of Ontario. *Environmetrics*, 21(5):457–469.

Puig, P. and Valero, J. (2006). Count data distributions: some characterizations with applications. *J. Amer. Statist. Assoc.*, 101(473):332–340.

Puig, P. and Valero, J. (2007). Characterization of count data distributions involving additivity and binomial subsampling. *Bernoulli*, 13(2):544–555.

Ramos, P. L., Dey, D. K., Louzada, F., and Lachos, V. H. (2019). An extended poisson family of life distribution: A unified approach in competitive and complementary risks. *Journal of Applied Statistics*, pages 1–17.

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

Reed, W. J. and Jorgensen, M. (2004). The double Pareto-lognormal distribution—a new parametric model for size distributions. *Communications in Statistics-Theory and Methods*, 33(8):1733–1753.

Saghir, A., Hamedani, G., Tazeem, S., and Khadim, A. (2017). Weighted Distributions: A Brief Review, Perspective and Characterizations. *International Journal of Statistics and Probability*, 6(3):109.

Sala, A., Zheng, H., Zhao, B. Y., Gaito, S., and Rossi, G. P. (2010). Brief announcement: revisiting the power-law degree distribution for social graph analysis. In *Proceedings of the 29th ACM SIGACT-SIGOPS symposium on Principles of distributed computing*, pages 400–401. ACM.

Satterthwaite, F. (1942). Generalized Poisson distribution. *The Annals of Mathematical Statistics*, 13(4):410–417.

Schapiro, B. (1994). An approach to the physics of complexity. *Chaos, Solitons & Fractals*, 4(1):115–123.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

Seshadri, M., Machiraju, S., Sridharan, A., Bolot, J., Faloutsos, C., and Leskove, J. (2008). Mobile call graphs: beyond power-law and lognormal distributions. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 596–604. ACM.

Smolinsky, L. (2017). Discrete power law with exponential cutoff and Lotka's law. *Journal of the Association for Information Science and Technology*, 68(7):1792–1795.

Sundt, B. r. and Jewell, W. S. (1981). Further results on recursive evaluation of compound distributions. *Astin Bull.*, 12(1):27–39.

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

Teicher, H. (1961). Maximum likelihood characterization of distributions. *The Annals of Mathematical Statistics*, 32(4):1214–1222.

Traud, A. L., Mucha, P. J., and Porter, M. A. (2012). Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165 – 4180.

Tripathi, A. and Vijay, S. (2003). A note on a theorem of erdős & gallai. *Discrete Mathematics*, 265(1-3):417–420.

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

Valero, J., Pérez-Casany, M., and Ginebra, J. (2010). On zero-truncating and mixing Poisson distributions. *Advances in Applied Probability*, 42(4):1013–1027.

Valero, J., Pérez-Casany, M., and Ginebra, J. (2013). On Poisson-stopped-sums that are mixed Poisson. *Statist. Probab. Lett.*, 83(8):1830–1834.

Vilfredo, P. (1896). Cours d'économie politique. *Œuvres complètes, tI-II, Genève, Droz.*

Visser, M. (2013). Zipf's law, power laws and maximum entropy. *New Journal of Physics*, 15(4):043021.

Wang, D., Cheng, H., Wang, P., Huang, X., and Jian, G. (2017). Zipf's law in passwords. *IEEE Transactions on Information Forensics and Security*, 12(11):2776–2791.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684):440.

Wickham, H. (2014). *Advanced R*. Chapman and Hall/CRC.

Yee, T. W. (2019). *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.1-1.

Yeh, H.-C. (2004). The generalized Marshall–Olkin type multivariate Pareto distributions. *Communications in Statistics-Theory and Methods*, 33(5):1053–1068.

Young, D. S. et al. (2010). Tolerance: an R package for estimating tolerance intervals. American Statistical Association.

Zhang, H. and Li, B. (2016). Characterizations of discrete compound Poisson distributions. *Comm. Statist. Theory Methods*, 45(22):6789–6802.

Zipf, G. (1935). K.(1968). the psycho-biology of language: An introduction to dynamic philology.

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

Zörnig, P. (2015). Zipf's law for randomly generated frequencies: explicit tests for the goodness-of-fit. *Journal of Statistical Computation and Simulation*, 85(11):2202–2213.

Zörnig, P. and Altmann, G. (1995). Unified representation of Zipf distributions. *Computational Statistics & Data Analysis*, 19(4):461–473.

Appendix A

Guidalines for choosing the most appropriate extension for a given data set

The aim of this appendix is to provide a guide that helps the practitioner decide which Zipf extension is more suitable for a given data set. At the same time, we present a complete summary of the main properties associated with the distributions families discussed in the previous chapters.

Before choosing the model, it is important to deeply understand the data from a descriptive point of view. Additionally, if possible, it is convenient to decide if any of the mechanisms that generate the proposed extensions fit to their data generation mechanism. Table A.1 explains the mechanism that generates each of the distributions defined in the previous chapters. As before, Y is an r.v. that follows one of the Zipf extensions defined previously, and the r.v. X is Zipf distributed.

Table A.1 Mechanisms that generate the presented Zipf extensions.

Distribution	**Interpretation**
MOEZipf(α,β)	$Min(X_1,X_2,\ldots,X_N)$ or $Max(X_1,X_2,\ldots,X_N)$, where $X_i \sim Zipf(\alpha)$ and $N \sim Geom(\beta)$ or $Geom(1/\beta)$
Zipf-PE(α,β)	$Min(X_1,X_2,\ldots,X_N)$ or $Max(X_1,X_2,\ldots,X_N)$, where $X_i \sim Zipf(\alpha)$ and $N \sim zt-Po(\lvert\beta\rvert)$
Zipf-PSS(α,λ)	$X_1+X_2+\ldots+X_N$, where $X_i \sim Zipf(\alpha)$ and $N \sim Po(\lambda)$
Zipf-Polylog(α,β)	$\begin{cases} \text{weighted } Zipf(\alpha);\ w(x;\beta)=\beta^x>0, & \alpha>1 \\ \text{weighted } Zipf(\alpha+1);\ w(x;\beta)=\beta^x x, & \alpha\in(0,1) \\ \text{weighted } Zipf(\alpha+2);\ w(x;\beta)=\beta^x x^2, & \alpha\in(-1,0) \\ \text{weighted } Zipf(-\alpha);\ w(x;\beta)=\beta^x x^{-2\alpha}, & \alpha<-1 \end{cases}$

In order to facilitate the comparison, Table A.2 contains the summary of the main properties of the models introduced in this thesis. Observe that only the most important properties are included in this table. For example, note that all models allow for top-concavity, and only those defined as RSED also allow for top-convexity. With the exception of the Zipf-PSS that contains the zero value in its support, the rest of them are defined in strictly positive integer numbers. The first three generalizations have a linear tail proportional to the tail of the Zipf distribution with the same α value. The Zipf-Polylog distribution is the only distribution that has moments of any order, when not considering the particular parameter values for which the distribution is equal to the Zipf.

Figure A.1 shows a flow diagram that helps decide which model may provide a reasonable fit to the data. In order to follow the diagram, it is necessary to have either frequency of frequencies data or rank data, plus the log-log plot of such data set. By looking at the tail and the top-concavity (top-convexity) of the plotted data, it is possible to follow the diagram and determine if any of the proposed models are suitable for the data set under analysis.

Table A.2 Summary of the main properties associated with the four Zipf generalizations included in this thesis.

	MOEZipf	**Zipf-PE**	**Zipf-PSS**	**Zipf-Polylog**
Type of family	RSED	RSED	PSS	Exponential Family $\forall \beta \in (0,1)$
Support	$\{1,2,3,\ldots\}$	$\{1,2,3,\ldots\}$	$\{0,1,2,3,\ldots\}$	$\{1,2,3,\ldots\}$
Parameters	$\alpha > 1; \beta > 0$	$\alpha > 1; \beta \in \mathbb{R} \backslash 0$	$\alpha > 1; \lambda > 0$	$\alpha \in \mathbb{R}; \beta \in (0,1)$ or $\alpha > 1; \beta = 1$
Zipf Dist. **as particular case**	$\beta = 1$	$\beta = 0$	$\lambda \to 0$	$\alpha > 1, \beta = 1$
Pattern **in log-log scale**	Top-Concave Top-Convex	Top-Concave Top-Convex	Top-Concave	Top-Concave

Existence of the
***k-th* moment**
	$\alpha > k+1$	$\alpha > k+1$	$\alpha > k+1$	if $\beta \in (0,1),\ \forall \alpha \in \mathbb{R}$
				if $\beta = 1,\ \text{for } \alpha > k+1$

Prob. at one	$P(Y=1) < (>) P(X=1),\, \forall \beta > (<) 1$	$P(Y=1) < (>) P(X=1),\, \forall \beta > (<) 0$	$P(Y_{ZT}=1) < P(X=1),\, \forall \lambda > 1$	$P(Y=1) \geq P(X=1),\, \forall \beta$
Infinite Divisibility	-	-	Yes	-

Mixture of Poissons	-	-	MP	if $\alpha > 0,\ MZTP$
				if $\alpha = 0,\ ZTMP$ and $MZTP$

Ratio $\frac{P(Y=x+1)}{P(Y=x)}$	$\frac{P(X=x+1)}{P(X=x)} \cdot \frac{\zeta(\alpha) - \overline{\beta}\zeta(\alpha,x)}{\zeta(\alpha) - \overline{\beta}\zeta(\alpha,x+2)}$	$\frac{e^{\beta P(X=x+1)} - 1}{1 - e^{-\beta P(X=x)}}$	$\frac{P(X=x+1)}{P(X=x)} \cdot \left(\frac{x}{x+1}\right)^{1-\alpha} \cdot h(x,\alpha,\lambda)$	$\beta \, \frac{P(X=x+1)}{P(X=x)}$
Ratio $\frac{P(Y=x)}{P(X=x)}$	$\frac{\zeta(\alpha)^2}{[\zeta(\alpha) - \overline{\beta}\zeta(\alpha,x)][\zeta(\alpha) - \overline{\beta}\zeta(\alpha,x+1)]}$	$\frac{e^{\beta} e^{\frac{-\beta\zeta(\alpha,\beta)}{\zeta(\alpha)}} (e^{\frac{\beta x^{-\alpha}}{\zeta(\alpha)}} - 1)\zeta(\alpha)}{(e^{\beta} - 1)x^{-\alpha}}$	-	$\frac{\beta^x \zeta(\alpha)}{Li_\alpha(\beta)}$

Tail	$P(Y > x) \sim \beta \, P(X > x)$	if $\beta < 0, P(Y > x) \sim \frac{-\beta \, e^{\beta}}{1 - e^{\beta}} \, P(X > x)$
		if $\beta > 0, P(Y > x) \sim \frac{\beta}{1 - e^{\beta}} \, P(X > x)$
			$P(Y > x) \sim \lambda \, P(X > x)$	-

Random Data **$u \sim N(0,1)$**	Inversion method of the Zipf applied to: $u' = \frac{u\beta}{1 + u(\beta - 1)}$	Inversion method of the Zipf applied to: $u' = \frac{\log(u(e^{\beta} - 1) + 1)}{\beta}$	Inversion method applied directly to the Zipf-PSS	First $x \geq 1$ verifying that: $u \cdot Li_\alpha(\beta) \leq \sum_{i=1}^{x} i^{-\alpha} \cdot \beta^i$

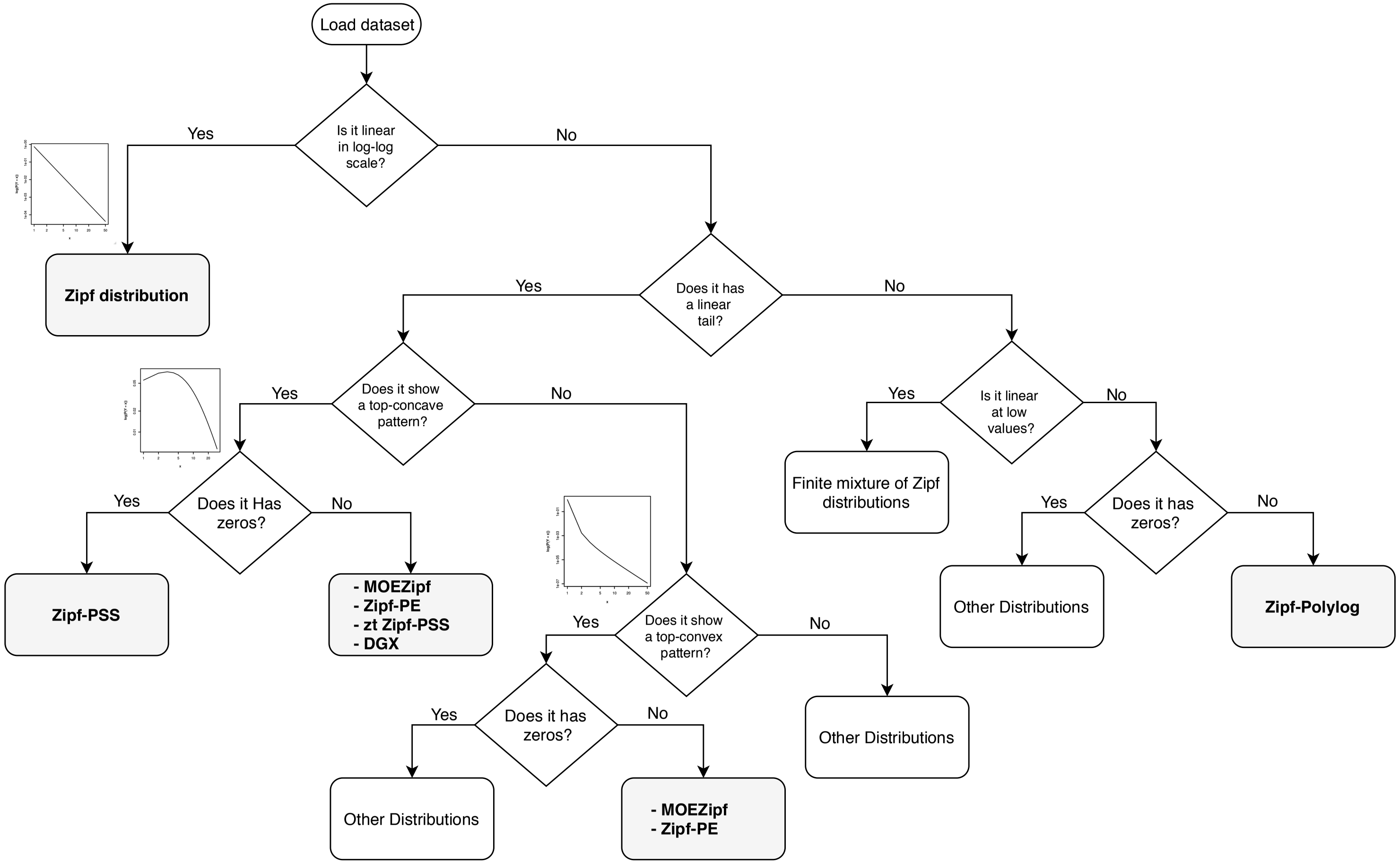

Fig. A.1 Flow diagram for choosing the most appropriate model, based on the characteristics of the data set.

Appendix B

The *zipfextR* Package

The ubiquity of the Zipf distribution has made it necessary to implement this distribution in the most popular statistical software, such as *R*. The readers may find at CRAN several packages that implement the Zipf distribution or other related families of distributions. For example, the package *VGAM* [Yee, 2019] provides a full implementation of the Zipf family, as does the package *tolerance* [Young et al., 2010], in which implementations of Zipf-Mandelbrot and Zeta distributions can also be found. Note that the Zipf random number generator used in our package is the one implemented in the above mentioned package. In the same way, *zipfR* [Evert and Baroni, 2007] implements some Large-Number-of-Rare-Events models for modeling word frequency distributions. The latest version of the *zipfR* package is from October 2019 (see http://zipfr.r-forge.r-project.org/).
Another R-package for dealing with heavy tail distributions is *poweRlaw* [Gillespie, 2015], which provides an interface for using the methodology developed by Clauset et al. [2009]. However, for the methodology used in Chapter 5, we have directly used the scripts available from the author's web page: http://tuvalu.santafe.edu/~aaronc/powerlaws/plfit.r.

In order to facilitate the use of the Zipf generalizations presented in this thesis to practitioners of different areas of research, we first implemented the MOEZipf distribution in the *moezipfR* package [Duarte-López et al., 2017]. Later, we created the R-package *zipfextR* [Duarte-López and Pérez-Casany, 2020], which is currently available at CRAN. This package implements the four extensions presented in this thesis.

Similarly to other distributions implemented in R, the package *zipfextR* implements the PMF, CDF, quantile function and a function for generating random numbers. For each family, a function is included for numerically estimating their parameters via MLE. Appendix D.2 includes the main functions used for estimating the parameter values in the applications that appear in Chapter 5.

The functions are implemented using the R class named *S3* [Wickham, 2014, p–102]. As a consequence, the results from a fitting process can directly access functions such as: *summary, print, coef, AIC, BIC, fitted, residuals*, and others. The manual Duarte-López and Pérez-Casany [2020] contains detailed documentation of each function in the package. Table B.1 summarizes the functions developed for each family of distributions.

Table B.1 Functionalities implemented in the R-package *zipfextR*.

	MOEZipf	**Zipf-PE**	**Zipf-PSS**	**Zipf-Polylog**
PMF	dmoezipf	dzipfpe	dzipfpss	dzipfpolylog
CDF	pmoezipf	pzipfpe	pzipfpss	pzipfpolylog
Quartile Function	qmoezipf	qzipfpe	qzipfpss	qzipfpolylog
Random Number Generator	rmoezipf	rzipfpe	rzipfpss	rzipfpolylog
Moments	moezipfMoments	zipfpeMoments	zipfpssMoments	zipfPolylogMoments
Mean	moezipfMean	zipfpeMean	zipfpssMean	zipfpolylogMean
Variance	moezipfVariance	zipfpeVariance	zipfpssVariance	zipfpoylogVariance
MLE	moezipfFit	zipfpeFit	zipfpssFit	zipfPolylogFit

The package *zipfextR* was released in March, 2018 and most recently updated in July, 2020. According to the CRAN logs, the package was downloaded 7502 times between its publication and up until early May 2020. Figure B.1 shows the number of downloads of the *zipfextR* package by country. Table B.2 ranks the ten top countries with the largest numbers of downloads. For the construction of the figure and the table we have excluded the information on 619 downloads, for which no country data was available.

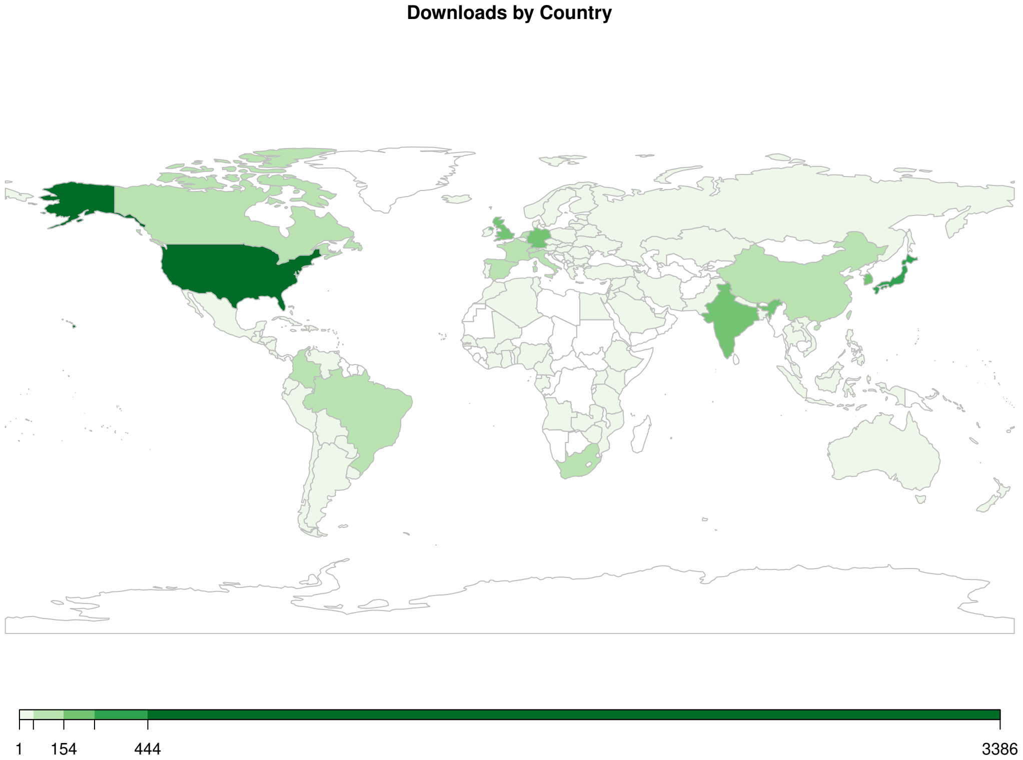

Fig. B.1 Downloads of the R-package *zipfextR* by countries.

Table B.2 Ten top countries with the largest number of downloads of the *zipfextR* package until May 2020.

Country	**Downloads**
United States of America	3386
Japan	444
Republic of Korea	260
Germany	255
United Kingdom	196
India	195
Switzerland	154
China	145
Canada	141
Spain	108

Appendix C

CDF of maximums and minimums of randomly stopped extreme Zipf distributions

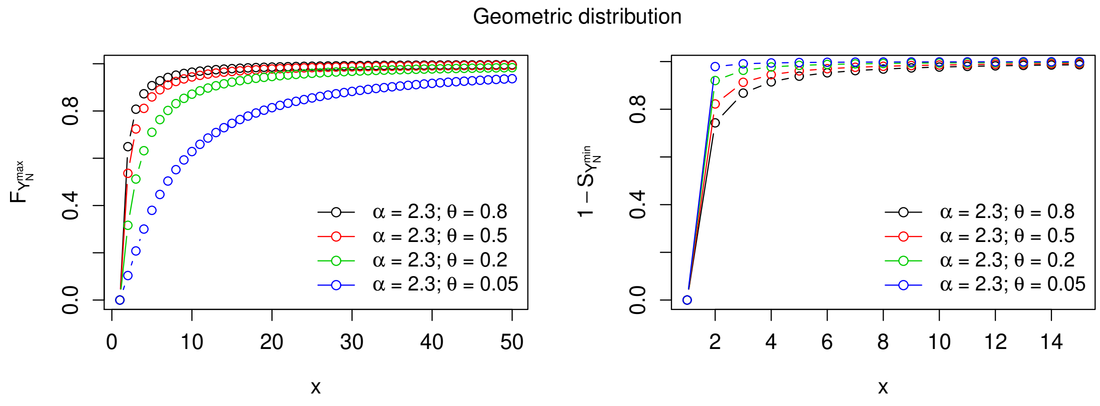

Fig. C.1 CDFs of the RSEZipf distribution with a geometric stopping distribution for $\alpha = 2.3$ and $\theta = 0.05, 0.2, 0.5$ and 0.8. On the left-hand side for maximums, and on the right-hand side for minimums.

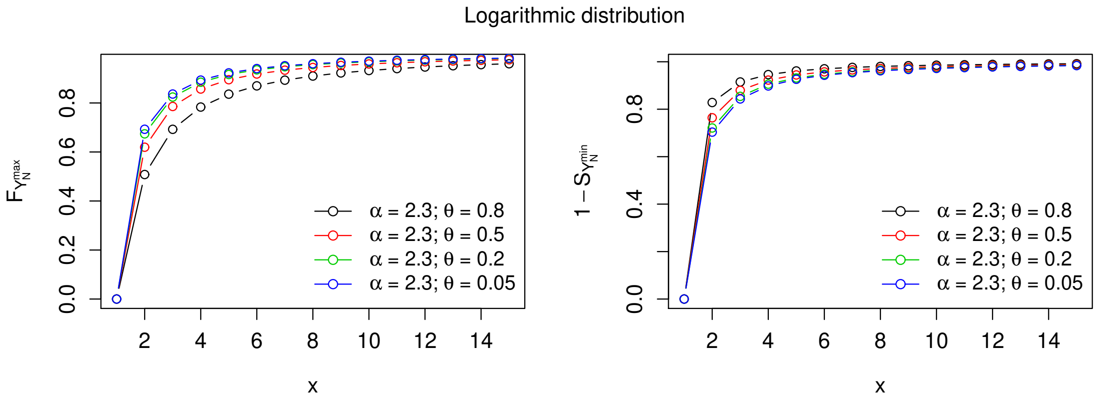

Fig. C.2 CDFs of the RSEZipf distribution with a logarithmic stopping distribution for $\alpha = 2.3$ and $\theta = 0.05, 0.2, 0.5$ and 0.8. On the left-hand side for maximums, and on the right-hand side for minimums.

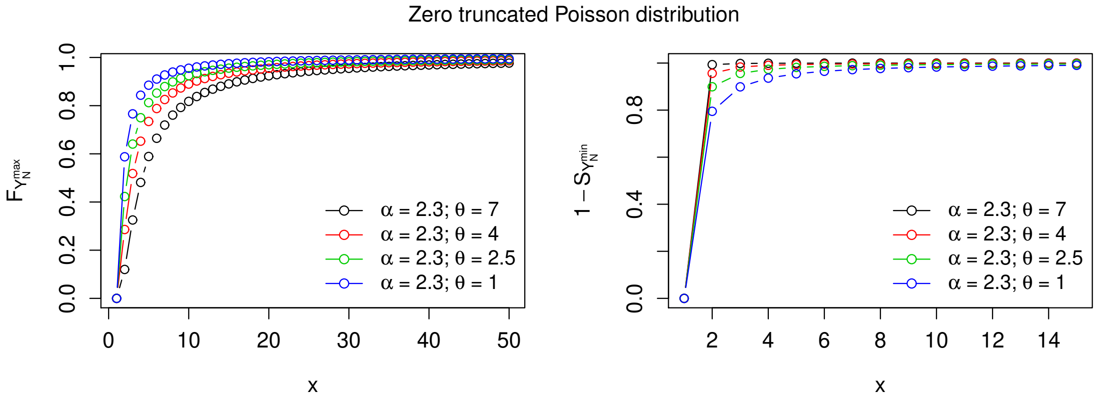

Fig. C.3 CDFs of the RSEZipf distribution with a zero-truncated Poisson stopping distribution for $\alpha = 2.3$ and $\theta = 1, 2.5, 4$ and 7. On the left-hand side for maximums, and on the right-hand side for minimums.

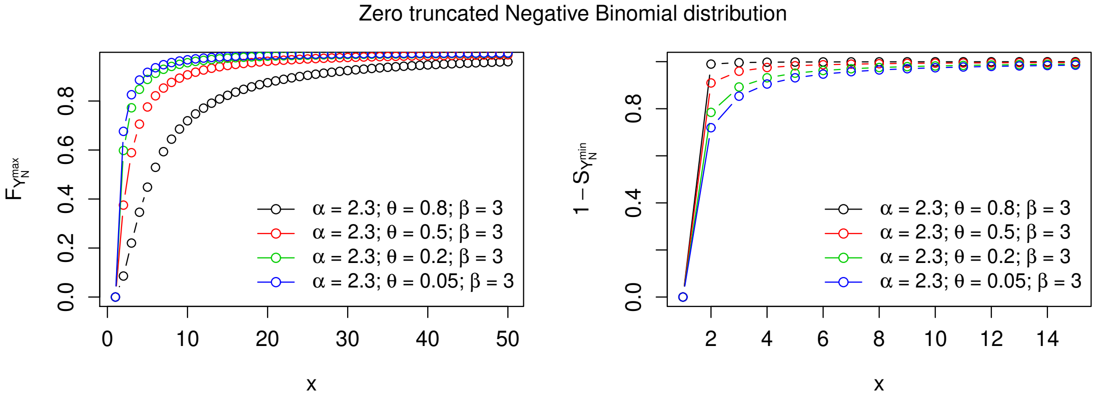

Fig. C.4 CDFs of the RSEZipf distribution with a zero-truncated negative binomial stopping distribution for $\alpha = 2.3$ and $\theta = 0.05, 0.2, 0.5, 0.8$ and $\beta = 3$. On the left-hand side for maximums, and on the right-hand side for minimums.

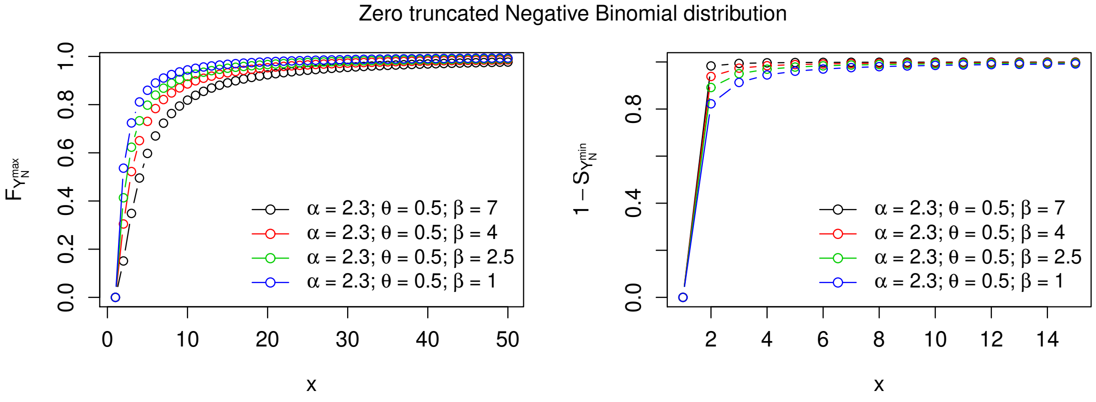

Fig. C.5 CDFs of the RSEZipf distribution with a zero-truncated negative binomial stopping distribution for $\alpha = 2.3$ and $\theta = 0.5$ and $\beta = 1, 2.5, 4$ and 7. On the left-hand side for maximums, and on the right-hand side for minimums.

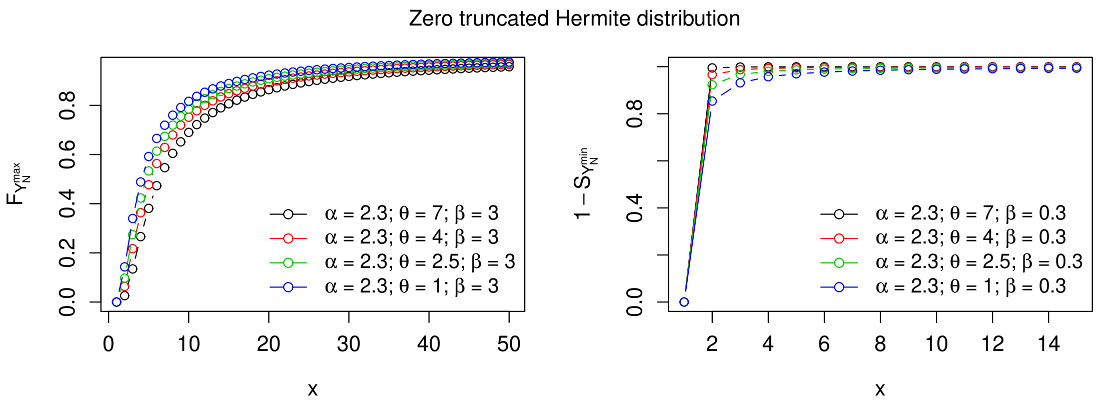

Fig. C.6 CDFs of the RSEZipf distribution with a zero-truncated Hermite stopping distribution for $\alpha = 2.3$ and $\theta = 1, 2.5, 4, 7$. On the left-hand side for $beta = 3$, and on the right-hand side for $\beta = 0.3$. On the left-hand side for maximums, and on the right-hand side for minimums.

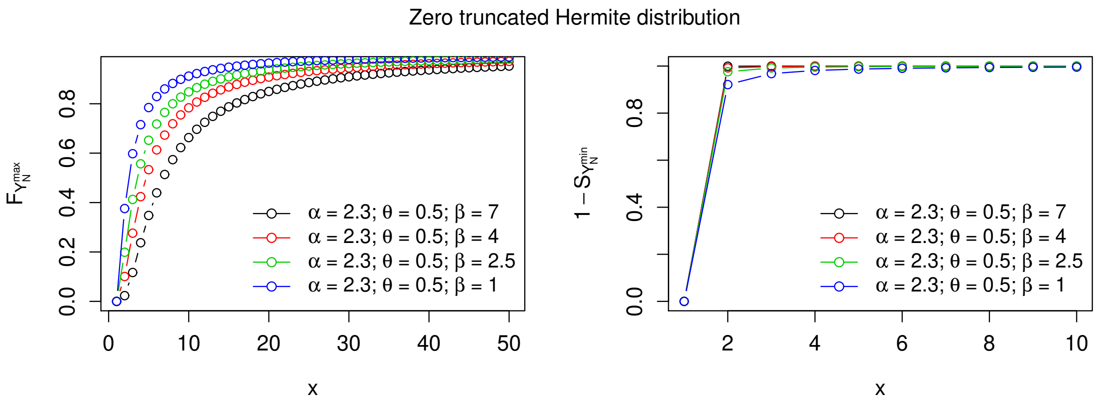

Fig. C.7 CDFs of the RSEZipf distribution with a zero-truncated Hermite stopping distribution for $\alpha = 2.3$ and $\theta = 0.5$ and $\beta = 1, 2.5, 4$ and 7. On the left-hand side for maximums, and on the right-hand side for minimums.

Appendix D

Source code

D.1 Source code for the computation of the Kolmogorov-Smirnov test

D.1.1 Sequence generation

```
library(zipfextR)
library(foreach)
library(parallel)
library(doParallel)

sampleSize <- c(100, 1000)
samples <- 500
alphas <- c(1.25, 2, 3.5, 5)
betas <- c(0.1, 0.25, 0.5, 1, 1.75, 2.25, 3.5, 10)
dirName <-
    '/home/aduarte/work/degree/kolmogorovThesis/sequences/zipfpss/%s/%s'

for (ss in sampleSize) {
  for (a in alphas) {
    for (b in betas) {
      dirN <- sprintf(dirName, ss, sprintf('%s-%s', gsub('.', '_', a, fixed
          = T), gsub('-', 'n', gsub('.', '_', b, fixed = T), fixed = T)))
      dir.create(dirN, showWarnings = F, recursive = T)
      cores <- detectCores(all.tests = FALSE, logical = TRUE)
      cl <- parallel::makeCluster(cores-1)
      doParallel::registerDoParallel(cl)
      parLapply(cl, 1:samples, function(i, alpha, beta, ss, dirN){i
        library(R.utils)
   withTimeout({
       library(zipfextR)
       #data <- rmoezipf(ss, alpha, beta)
       data <- rzipfpss(ss, alpha, beta)
   #data <- rzipfpe(ss, alpha, beta)
       write.table(data, file = sprintf('%s/%s.csv', dirN, i), row.names =
           F, quote = F, col.names = F)
   }, timeout = 1800, onTimeout = "warning")
      }, alpha = a, beta = b, ss = ss, dirN = dirN)
      parallel::stopCluster(cl)
    }
  }
}
```

D.1.2 KS test computation

```
library(parallel)
library(doParallel)

args <- commandArgs(TRUE)
seqFolder <- '<sequencesPath>'
model <- as.character(args[1])
sampleSize <- as.numeric(args[2])
simKS <- as.logical(args[3])
if (is.na(simKS)){
  stop('Parameter has to be T or F.')
}
print(model)
print(sampleSize)

seqDir <- file.path(seqFolder, model, sampleSize)
seqDirFolders <- list.files(seqDir)

sapply(seqDirFolders, function(x, parentFolder){
```

```
tryCatch({
  folderOutput <- unlist(strsplit(x,'-', fixed = T))
  #alpha <- as.numeric(gsub('_', '.', folderOutput[1], fixed = T))
  alpha <- as.numeric(gsub('_', '.', gsub('n', '-', folderOutput[1], fixed
      = T), fixed = T))
  beta <- as.numeric(gsub('_', '.', gsub('n', '-', folderOutput[2], fixed
      = T), fixed = T))
  print(c(alpha, beta))
  seqList <- list.files(file.path(parentFolder, x))

  cores <- detectCores(all.tests = FALSE, logical = TRUE)
  cl <- parallel::makeCluster(cores-1, outfile = sprintf(''))
  doParallel::registerDoParallel(cl)
  parLapply(cl, seqList, function(i, alpha, beta, seqFolder, model,
      sampleSize){
    tryCatch({
      library(zipfextR)
      library(dgof)
      data <- read.csv(file.path(seqFolder, i))
      spectrum <- table(data)
      data_frame <- data.frame(spectrum)
      data_frame[, 1] <- as.numeric(levels(data_frame[, 1]))[data_frame[,
         1]]
      data_frame[, 2] <- as.numeric(data_frame[, 2])

      switch (model,
        moezipf = {
          cdfTheoDist <- pmoezipf(data_frame[,1], alpha, beta)
        },
        zipfpe = {
          cdfTheoDist <- pzipfpe(data_frame[,1], alpha, beta)
        },
        zipfpss = {
          cdfTheoDist <- pzipfpss(data_frame[,1], alpha, beta)
        },
        zipfpolylog = {
          cdfTheoDist <- pzipfpolylog(data_frame[,1], alpha, beta)
        },
        stop('Incorrect model name.')
        )
        theoStepFn <- stepfun(data_frame[,1], c(0, cdfTheoDist))
        ksObj <- dgof::ks.test(data[,1], theoStepFn, simulate.p.value = simKS)
        csvLine <-data.frame(i, model, alpha, beta, sampleSize,
            ksObj$statistic, ksObj$p.value, simKS, ifelse(ksObj$p.value >
            0.05, 1, 0))
        write.table(csvLine, file='ksResults.csv', sep = ',', eol = '\n',
            append = T, row.names = F, col.names = F)
      }, error = function(e){
        message(e)
        print(c(model, sampleSize, alpha, beta, simKS, file.path(seqFolder,
            i)))
      })
      gc()
    }, alpha = alpha, beta = beta, seqFolder = file.path(parentFolder, x),
        model = model, sampleSize = sampleSize)
    parallel::stopCluster(cl)
  }, error = function(e){
    print(e)
    print(file.path(parentFolder, x))
  })
}, parentFolder = seqDir)
```

D.2 Source code for parameter estimation

```
library(igraph)
library(baselineDistR)
library(zipfextR)
library(tolerance)
library(dplyr)
library(xtable)

set.seed(2019)

getDegreeSequence <- function(g, type, loops = TRUE){
  totalDegree <- degree(g, v = V(g), mode = type, loops = loops)
  dataProc <- as.data.frame(totalDegree)
  dataProc <- table(dataProc[,1])
  dataProc <- as.data.frame(dataProc)
  dataProc[,1] <- as.numeric(as.character(dataProc[,1]))
  dataProc[,2] <- as.numeric(as.character(dataProc[,2]))
  return(dataProc)
}

get_AIC <- function(loglike, K) {
  -2*loglike + 2*K
}

fitSequence <- function(numSeq, type = '', includeDeltaAIC = TRUE, ids = '',
    initAlphaPoly=0.5, initBetaPoly=-0.01){
  dfResults <- data.frame(character(), character(), numeric(), character(),
      numeric(), character(), numeric(), character(), numeric(), numeric())
  cNames <- c('IDS','SeqType', 'nVals', 'Distribution', 'p1', 'CI_p1', 'p2',
      'CI_p2', 'loglik', 'AIC')
  colnames(dfResults) <- cNames

  nVals <- nrow(numSeq)
  if(0 %in% numSeq[,1]){

    ## Negative Binomial
    ztNB <- NULL
    tryCatch({
      ztNB <- negbinZTFit(numSeq, init_gamma = 0.5, init_p = 0.5)
      tempDF <- data.frame(ids, type, nVals, 'Neg. Bin.', coef(ztNB)[1,1],
                 sprintf('(%s, %s)', round(ztNB$gammaCI[1], 4),
                     round(ztNB$gammaCI[2], 4)),
                 coef(ztNB)[2,1],
                 sprintf('(%s, %s)', round(ztNB$pCI[1], 4),
                     round(ztNB$pCI[2], 4)),
                 logLik(ztNB), AIC(ztNB))
      colnames(tempDF) <- cNames
      dfResults <-rbind(dfResults, tempDF)
    }, error = function(e){
      print(e)
    })
    ## Discrete Weibull
    ztDW <- NULL
    tryCatch({
      ztDW <- discWeibullZTFit(numSeq,init_p = 0.5, init_v = 0.5)
      tempDF <- data.frame(ids, type, nVals, 'D. Weibull', coef(ztDW)[1,1],
                     sprintf('(%s, %s)', round(ztDW$pCI[1], 4),
                         round(ztDW$pCI[2], 4)),
                     coef(ztDW)[2,1],
                     sprintf('(%s, %s)', round(ztDW$vCI[1], 4),
                         round(ztDW$vCI[2], 4)),
                     logLik(ztDW), AIC(ztDW))
      colnames(tempDF) <- cNames
      dfResults <-rbind(dfResults, tempDF)
    }, error = function(e){
      print(e)
    })

    zipfpssD <- NULL
    tryCatch({
      zipfpssD <- zipfpssFit(numSeq)
      tempDF <- data.frame(ids, type, nVals, 'Zipf-PSS', coef(zipfpssD)[1,1],
                     sprintf('(%s, %s)', round(zipfpssD$alphaCI[1], 4),
                         round(zipfpssD$alphaCI[2], 4)),
                     coef(zipfpssD)[2,1],
                     sprintf('(%s, %s)', round(zipfpssD$lambdaCI[1], 4),
                         round(zipfpssD$lambdaCI[2], 4)),
                     logLik(zipfpssD), AIC(zipfpssD))
      colnames(tempDF) <- cNames
      dfResults <-rbind(dfResults, tempDF)
    }, error = function(e){
      print(e)
    })
  } else {
    ## Zipf Distribution
    zipfD <- NULL
    tryCatch({
      zipfD <- zipfFit(numSeq, init_alpha = 1.5)
      tempDF <- data.frame(ids, type, nVals, 'Zipf', coef(zipfD)[1,1],
                 sprintf('(%s, %s)', round(zipfD$alphaCI[1], 4),
                    round(zipfD$alphaCI[2], 4)),
                 -1, -1, logLik(zipfD), AIC(zipfD))
  colnames(tempDF) <- cNames
  dfResults <-rbind(dfResults, tempDF)

}, error = function(e){
  print(e)
})

## DGX
dgxD <- NULL
tryCatch({
  dgxD <- dgxFit(numSeq, init_mu = 1.2, init_sig = 1.5)
  tempDF <- data.frame(ids, type, nVals, 'DGX', coef(dgxD)[1,1],
                 sprintf('(%s, %s)', round(dgxD$muCI[1], 4),
                    round(dgxD$muCI[2], 4)),
                 coef(dgxD)[2,1],
                 sprintf('(%s, %s)', round(dgxD$sigCI[1], 4),
                    round(dgxD$sigCI[2], 4)),
                 logLik(dgxD), AIC(dgxD))
  colnames(tempDF) <- cNames
  dfResults <-rbind(dfResults, tempDF)
}, error = function(e){
  print(e)
})

## MOEZipf
moezipfD <- NULL
tryCatch({
  moezipfD <- moezipfFit(numSeq)
  tempDF <- data.frame(ids, type, nVals, 'MOEZipf', coef(moezipfD)[1,1],
                 sprintf('(%s, %s)', round(moezipfD$alphaCI[1], 4),
                    round(moezipfD$alphaCI[2], 4)),
                 coef(moezipfD)[2,1],
                 sprintf('(%s, %s)', round(moezipfD$betaCI[1], 4),
                    round(moezipfD$betaCI[2], 4)),
                 logLik(moezipfD), AIC(moezipfD))
  colnames(tempDF) <- cNames
  dfResults <-rbind(dfResults, tempDF)
}, error = function(e){
  print(e)
})

## Zipf-PE
zipfpeD <- NULL
tryCatch({
  zipfpeD <- zipfpeFit(numSeq)
  tempDF <- data.frame(ids, type, nVals, 'Zipf-PE', coef(zipfpeD)[1,1],
                 sprintf('(%s, %s)', round(zipfpeD$alphaCI[1], 4),
                    round(zipfpeD$alphaCI[2], 4)),
                 coef(zipfpeD)[2,1],
                 sprintf('(%s, %s)', round(zipfpeD$betaCI[1], 4),
                    round(zipfpeD$betaCI[2], 4)),
                 logLik(zipfpeD), AIC(zipfpeD))
  colnames(tempDF) <- cNames
  dfResults <-rbind(dfResults, tempDF)
}, error = function(e){
  print(e)
})

## Zipf-PSS (zero-truncated)
zipfpssD <- NULL
tryCatch({
  zipfpssD <- zipfpssFit(numSeq, isTruncated = T)
  tempDF <- data.frame(ids, type, nVals, 'zt-Zipf-PSS',
      coef(zipfpssD)[1,1],
                 sprintf('(%s, %s)', round(zipfpssD$alphaCI[1], 4),
                    round(zipfpssD$alphaCI[2], 4)),
                 coef(zipfpssD)[2,1],
                 sprintf('(%s, %s)', round(zipfpssD$lambdaCI[1], 4),
                    round(zipfpssD$lambdaCI[2], 4)),
                 logLik(zipfpssD), AIC(zipfpssD))
  colnames(tempDF) <- cNames
  dfResults <-rbind(dfResults, tempDF)
}, error = function(e){
  print(e)
})
```

```
    ## Zipf-Mandelbrot
    zipfMand <- NULL
    tryCatch({
      x <- rep(numSeq[,1], numSeq[,2])
      x <- table(factor(x, levels = min(x):max(x)))
      zipfMand <- zm.ll(x, N = max(numSeq[,1]), dist = "Zipf-Man")
      zipfMandSTD <- sqrt(diag(solve(zipfMand@details$hessian)))

      tempDF <- data.frame(ids, type, nVals, 'Zipf-Mandelbrot',
          zipfMand@details$par[1],
       sprintf('(%s, %s)',
round(zipfMand@details$par[1]-stats::qnorm(1-((1-0.95)/2))*zipfMandSTD[1],
    4),
    round(zipfMand@details$par[1]+stats::qnorm(1-((1-0.95)/2))*zipfMandSTD[1],
    4)),
zipfMand@details$par[2],
sprintf('(%s, %s)',
round(zipfMand@details$par[2]-stats::qnorm(1-((1-0.95)/2))*zipfMandSTD[2],
    4),
round(zipfMand@details$par[2]+stats::qnorm(1-((1-0.95)/2))*zipfMandSTD[2],
    4)),
-1*zipfMand@details$value, get_AIC(-1*zipfMand@details$value, 2))
      colnames(tempDF) <- cNames
      dfResults <-rbind(dfResults, tempDF)
    }, error = function(e){
      print(e)
    })

    zipfPoly <- NULL
    tryCatch({
      zipfPoly <- zipfPolylogFit(numSeq, init_alpha = initAlphaPoly,
          init_beta = initBetaPoly, method="L-BFGS-B", lower=c(-Inf,-Inf),
          upper=c(Inf,0))
      # zipfPolySTD <- sqrt(diag(solve(zipfPoly$hessian)))
      tempDF <- data.frame(ids, type, nVals, 'Zipf-Polylog',
          coef(zipfPoly)[1,1],
                        sprintf('(%s, %s)', round(zipfPoly$alphaCI[1], 4),
                            round(zipfPoly$alphaCI[2], 4)),
                        coef(zipfPoly)[2,1],
                        sprintf(’(%s, %s)’, round(zipfPoly$betaCI[1], 4),
                            round(zipfPoly$betaCI[2], 4)),
                        logLik(zipfPoly), AIC(zipfPoly))
      colnames(tempDF) <- cNames
      dfResults <-rbind(dfResults, tempDF)
    }, error = function(e){
      print(e)
      print(ids)
    })
  }

  if(includeDeltaAIC){
    dfResults <- dfResults %>%
      group_by(IDS) %>%
      mutate(DAIC = AIC - min(AIC)) %>%
      arrange(DAIC)
  }
  return(dfResults)
}
```