Beyond Zipf's law: Exploring the discrete generalized beta distribution in open-source repositories

Przemyslaw Nowak, Marc Santolini, Chakresh Singh, Grzegorz Siudem, Liubov Tupikina

Beyond Zipf's Law: Exploring the Discrete Generalized Beta Distribution in Open-Source Repositories

Przemysław Nowak[a], Marc Santolini[b,c], Chakresh Singh[b], Grzegorz Siudem[a], Liubov Tupikina[c,d]

[a] *Warsaw University of Technology, Faculty of Physics, ul. Koszykowa 75, Warsaw, 00-662, Poland*
[b] *Université Paris Cité, Inserm, System Engineering and Evolution Dynamics, F-75004, Paris, France*
[c] *Learning Planet Institute, Research Unit Learning Transitions (UR LT, joint unit with CY Cergy Paris University), F-75004, Paris, France*
[d] *Nokia Bell labs, France, Paris, France*

Abstract

Rank-size distributions, such as Zipf's Law, have been instrumental in providing insights into the emergence of hierarchies across diverse systems, from linguistic corpuses to urban structures. However, the application of Zipf's Law reveals limitations, particularly in its focus on distribution tails, sometimes overlooking a large proportion of the data which might play a pivotal role in system dynamics. Yet, fitting rank-size distributions other than a straight line on the log-log scale requires caution. In this study, we re-evaluate the utility of rank-size distributions by contrasting the traditional Zipf's Law with the Discrete Generalized Beta Distribution (DGBD). We show the need of cautious fitting techniques for rank distributions, including the use of binning to prevent overfitting to data tails. Through both analytical derivation and empirical validation on commit data of open-source repositories, we show that DGBD consistently improves over Zipf distribution for concave rank distributions of large datasets ($N \geq 100$). This approach contributes to the advancement of methodologies for analyzing hierarchical systems.

Keywords: DGBD, Discrete Generalized Beta Distribution, Rank-size distribution, Fitting methods, Open-source data

Preprint submitted to Physica A: Statistical Mechanics and its Applications August 10, 2024

1. Introduction – rank-size distributions

When dealing with numerical data, it is sometimes useful to consider the ordered list of values, or ranking. Indeed, sorting numerical values, e.g. from largest to smallest, reveals the subset of observations achieving the highest results, indicating the top of a given set. Such rankings are present in almost every area of life (e.g., rankings of the largest cities, most popular musicians, wealthiest people, and most famous universities, to name a few) [1, 2]. The first scholarly discussion of the regularities in rankings can be attributed to Felix Auerbach in the context of city sizes (for English translation and reference to original German article, see [3]), later extended by George Zipf to the frequency of words in corpora of natural languages [4, 5]. Zipf showed that the frequency of any word is inversely proportional to its rank in the frequency table, following a power law. This law has been later widely observed in numerous systems [6, 7, 8, 9]. Zipf's Law states that for the ordered sample $x(1) \geqslant x(2) \geqslant \cdots \geqslant x(N)$ of size N

$$x(r) = \frac{C}{r^{\alpha}} \;\Rightarrow\; \log\left(x(r)\right) = -\alpha \log(r) + \log(C), \tag{1}$$

where r is the rank (i.e., 1, 2, ..., N) and $x(r)$ can be any numeric property of the considered objects such as frequency, size, value, etc. The two constants α, $C > 0$ are parameters, and α is usually called the Pareto exponent (for more connection with the Pareto distribution, see below), with $\alpha \simeq 1$. Eq. (1) is one of the most famous rank-size distributions, linking rank r to size $x(r)$ across the whole sample, observed widely in real datasets. However, it is not the only such distribution. Before considering others, let us note the connections between rank-size and probability distributions.

When considering fat-tailed distributions, Newman, Clauset and Shalizi [7, 10] suggest using the Complementary Cumulative Distribution Function (CCDF, sometimes also called the survival function):

$$S(x) = \mathrm{Prob}(X \geq x) = \int_{x}^{\infty} f(t)dt, \tag{2}$$

rather than the probability density function $f(x)$. Such an approach offers a robust perspective on extreme values and rare events within a dataset, effectively reducing uncertainties associated with infrequent occurrences [7, 10]. One can show (see Appendix A) that there is a strict relation between

the expected rank-size distribution $r(x)$ for a sample of size N and the CCDF $S(x)$ of the distribution from which the sample was drawn:

$$x = S^{-1}\left(\frac{r(x)-1}{N-1}\right). \tag{3}$$

From the above formula in Eq. (3), one can see that the rank-size distribution can be expressed in terms of the inverse CCDF, connecting these two different methods for the description of the ranked data. In the case of Zipf Law of Eq. (1), the survival function, which is the Pareto type I distribution [11], is analytically invertible so that one can express both the CCDF and the rank-size distribution in a compact and explicit form. Unfortunately, in general, this property does not hold. One faces the problem that a close analytical formula is given either for the CCDF or the rank-size distribution, while the other is available only implicitly. As a consequence, apart from the elegant (because reversible) Zipf Law, there are in the literature no compact formulas for both CCDF and rank-size domains. When considering the CCDF, the most significant competitors to Zipf are Pareto type II [6], log-normal ([12]), Tsallis-Pareto [13], Tsallis q-exponential distribution [14], Generalized Beta family of distributions [15], other Pareto modifications [9] and Gumbel [16] to name just a few. Alternative rank-size distributions include, among others, Price Model [6, 17] and the Discrete Generalized Beta Distribution (DGBD, sometimes named generalized Lavalette distribution [18]) [19, 20, 21, 22, 23, 24, 25]. Importantly, DGBD focuses on the whole spectrum of ranks, while Zipf's Law describes only the tail of the data, which in many cases covers a small part of the data [7]. While in the context of probability distributions, the tail encompasses largest values (e.g. hubs in degree distributions), for rank distributions the tail corresponds to low values with large ranks. Being able to describe the top ranking elements along with the tail of lower ranking elements is therefore important to get a full picture of the mechanism at play.

Since the Zipf distribution has an explicit form in the rank-size and probability domains, it allows the direct use of fit and testing methods and techniques derived classically in the language of probability distributions to describe the rank-size distribution as well [7]. Unfortunately, many of these methods fail for models with explicit form in the rank-size domain, and their application may lead to erroneous conclusions. While some of the rank-size models (e.g., [6]) are asymptotically reversible between the rank-size and probability domains, this is not always the case. In this work, we focus on

the more complex case of the DGBD rank-size distribution, for which there are solid indications that it is not reversible [20]. The DGBD distribution is a straightforward generalization of Zipf's Law (compare Eqs. (1) and (4)) with a larger applicability domain, provided it is appropriately applied, which we demonstrate later in the work.

The paper outline is as follows: In section 2, we introduce and provide new derivations of the properties of DGBD and of its parameters constraints. Next, in section 3, we review methods from earlier works considering DGBD [19, 20, 21, 22, 23, 24, 25], discuss their weaknesses, and point out some limitations of them. In section 4, we we empirically compare DGBD and Zipf Law using a unique dataset of ranked commit data from open-source GitHub repositories. In particular, to reduce the importance of low ranking data on the fit, we investigate the impact of binning ranks to improve the model fit. Finally, in section 5 we draw conclusions from the theoretical considerations and empirical investigation.

2. Discrete Generalized Beta Distribution

We focus in this study on the Discrete Generalized Beta Distribution (DGBD) [19]. This rank-size distribution serves as a versatile generalization of Zipf's Law (Eq. 1), extending its applicability and enhancing its potential to capture the nuances of real-world data through the formula:

$$x = C\frac{(N - r(x) + 1)^{\beta}}{(r(x))^{\alpha}}, \tag{4}$$

where $r(x) \in [1, N]$ is the rank of the numerical property x, N is the sample size, and C, α, β are parameters.

Despite the undoubted elegance of Zipf's law, not all natural and artificial phenomena generate data from this distribution. This was undoubtedly one motivation behind studying the generalized DGBD distribution and its presence in observational data. Starting with population distribution of urban areas, authors in [20] introduced Zipf's law generalization in the form of DGBD. Other researchers have also explored the use of the DGBD for diverse applications, including fitting city size distributions [21], biomass distribution in weighted food webs [22], character frequency distribution in the Chinese language [23, 24], letter frequency distribution in various languages [25], baby name popularity [26], stochastic resonance [27], and in the dynamics of nonlinear maps [28]. It is worth mentioning that the term "generalized Lavalette

distribution" has also been used instead of DGBD [18]. In all of these works, authors claim that DGBD offers a more accurate data description than Zipf's Law. One of the most significant advantages of the DGBD distribution is that it usually fits the entire data set, not just the tail, in contrast to Zipf's law [7]. In some cases, this means that Zipf's law describes only a tiny portion of the data, and what is not in the tail is overlooked. This is concerning, since in rank distributions, the first few items are the top-ranking elements which can be the key drivers of the dynamics of the system. In Figure 1, the top elements (i.e., those with the highest values) correspond to the first ranks, which are visible on the left side in the Rank-size representation. Conversely, in the probability domain, these same top elements are visible in the tail on the right side, where Zipf's Law is typically fitted. On the other hand, the DGBD model considers the entire range, simultaneously focusing on the largest and the smallest values.

2.1. Selected properties

DGBD is parameterized by two exponents: β primarily influences large rank values r, while α plays a more significant role for lower rank values. This distribution generalizes Zipf's Law, which relies only on a single exponent, α. It is important to note that one of the most common properties of Zipf's Law is its representation as a straight line in a double-logarithmic scale. This relationship can be expressed as $\log(x(r)) = -\alpha \log(r) + \log(C)$ (as shown in Eq. (1)), where the exponent α corresponds to the slope of the line. One might assume that since the DGBD is a generalization of Zipf's Law, their properties would also generalize, resulting in two straight lines when plotted in a double logarithmic scale. However, this assumption is only partially true. To illustrate this feature, we show in Figure 1 two lines at low and high rank values. Nevertheless, the slopes of these lines do not correspond to the exponents α and β, as seen in Zipf's Law. It can be demonstrated that in a double logarithmic scale, the DGBD (4) can be approximated using a Maclaurin series as follows:

$$\log x = \log(CN^{\beta}) + \left(-\alpha - \frac{\beta}{N}\right) z + O(z^2),$$

where $z = \log r$. It means that the slope is equal to $-\alpha - \frac{\beta}{N}$ for low values of rank r. Similarly, the Taylor expansion near the end of DGBD (for $r = N$) takes the form:

$$\log x = \log(CN^{-\alpha}) + (-\alpha - \beta N)\zeta + O(\zeta^2),$$

where $\zeta = \log r - \log N$. In that case, the slope is equal to $-\alpha - \beta N$ for high values of rank r. This observation is depicted in Figure 1. Among different probability distributions mentioned in the Introduction, the Beta Prime distribution yields similar property, as it exhibits two power-law dependencies when $x \to 0$ and $x \to \infty$.

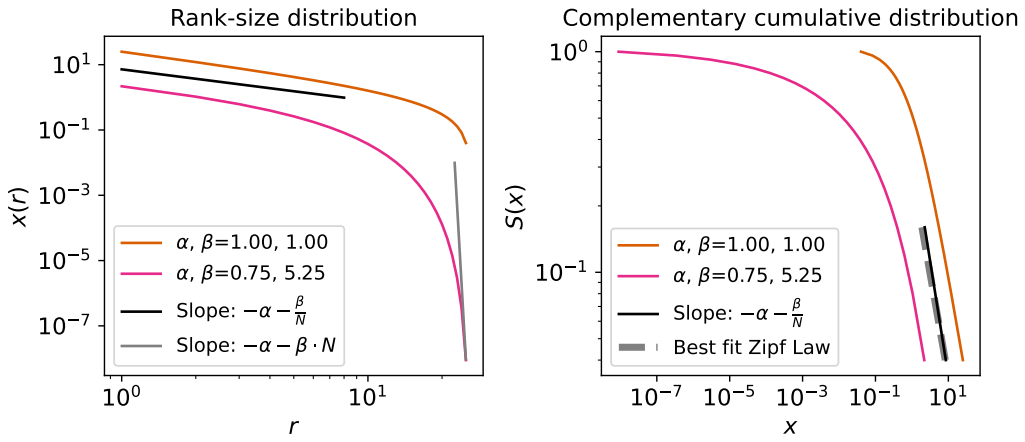

Figure 1: A comparison of two different distributions of DGBD on the double log scales. On the left side, we present rank-size distribution (data in the domain of sizes versus ranks). In this plot, we observe two lines - pink and orange - which, despite having different parameters α and β, exhibit the same slope at the beginning of the graph. This phenomenon can be exploited in the probabilistic domain (illustrated on the right side of the figure, where we plot DGBD's complementary cumulative distribution functions). In such a situation, fitting the Zipf's law to both sets of data, as proposed in [7], would yield a line with identical results (in other words, the same slope).

As we discuss DGBD properties, we need to address the topic of parameter boundaries. Remembering that DGBD is a rank-size function and must exhibit a decreasing trend is crucial. As the rank r increases, the values $x(r)$ should decrease accordingly. Interestingly, this property does not hold for every combination of the parameters C, N, α, and β of the DGBD. This aspect has not been widely recognized in recent studies concerning DGBD but is of significant importance. Careless fitting of data to functions can lead to incorrect distributions. To ensure the validity of such fits, one must satisfy the following inequalities, which we have derived in Appendix B:

$$\begin{cases} N\alpha + \beta \geq 0, \\ \alpha + N\beta \geq 0. \end{cases}$$

The DGBD itself can represent several well-known distributions, as noted in [20]: Zipf's Law is obtained by setting $\beta = 0$, a Uniform distribution can be approximated with $\alpha = 0$ and $\beta = 1$, a Dirac delta Distribution can be modeled with $\alpha = \beta = 0$, and the Lavalette distribution corresponds to $\alpha = \beta$, closely resembling the lognormal distribution [29]. However, it also comes with limitations. DGBD has limited applicability in cases where the data exhibit a convex shape in a double logarithmic scale. As demonstrated in Appendix B, the minimum possible value of parameter β is $\beta = -\frac{\alpha}{N}$ and DGBD is convex when $\beta < 0$. As such, when the size of the data N is big, DGBD will be more prone to fail serving as a suitable model. Examples of this phenomena are shown in Figures 2 and 3. The left panel of Figure 2 depicts concave data on a double logarithmic scale, where DGBD can be a good model (and indeed, the fit is satisfactory). On the other hand, the right panel presents convex data, where the limitation of DGBD can be observed. Despite the fact that negative β is allowed, its minimum value $\beta = -\frac{\alpha}{N}$ is so small that it prevents the model from fitting the data properly. The best fit shown there is convex, but this "convexity" is barely noticeable, only at the very end. Another example of convex data is visible in Figure 3, where we discuss the impact of the rank-size domain on the Kolmogorov-Smirnov test.

In summary, before fitting DGBD to the given data, it is necessary to verify whether it appears concave on a double logarithmic scale. This can be estimated either visually or by attempting to fit a quadratic function $a(\log r)^2 + b(\log r) + c$. The coefficient a determines whether the function is concave ($a < 0$) or convex ($a > 0$) and can serve as a criterion to examine concavity.

3. Fitting and testing rank-size distributions: state-of-the-art review

In order to compare the efficacy of Zipf and DGBD to describe rank-size distributions, we need to address what constitutes a "good fit" when analyzing rank-size distributions. In statistical analysis and data modeling, a "good fit" measures the extent to which a chosen model accurately represents the underlying data. However, the answer is not straightforward, as various methods are commonly employed to assess the goodness of fit, each with its advantages and disadvantages [30]. Importantly, it is important to keep in mind what is the mechanism of interest for which an accurate assessment is most needed. For example, in the context of ranked distributions, the

first few elements are of primordial importance in the system, since they can drive the overall attention dynamics of the system (e.g. the largest cities in a country, or the most frequent words in a language). This importance is usually highlighted by the use of a graphical log-log distributions, giving disproportionate visual weight to the top-ranked elements of the distribution. Such a log-log scale can reveal important discrepancies of a fitting method on a small yet essential subset of the data (the first few elements at the start of the distribution) that a statistical analysis, focusing on the much larger number of observations in the tail, would overlook. Here we survey methods and provide heuristics to ensure an equal treatment of the top and the tail of the rank-size distribution, using the DGBD rank-size distribution as a case study.

We surveyed fitting methods (Table 1) and goodness of fit measures (Table 2) for data described by the DGBD distribution. The approaches employed for the fitting process and the computation of goodness of fit within the literature are diverse and lack consistency: contrary to Zipf's law [7, 10], for the DGBD and other generalized rank-size distributions, there is no gold standard. Therefore, here we propose a thorough analysis of the methods employed, including a detailed examination of their limitations and their impact on the quality of the conclusions.

3.1. Overview of fitting procedures for DGBD

Method	Formula	Publications
Multiple Linear Regression	Eq. (5)	[20, 22, 23, 24, 28, 31, 32, 33]
Nonlinear Least Squares	Eq. (6)	[20, 22, 23, 24, 25, 26, 34]
Methods of Moments	Eq. (7)	[35]
Maximum Likelihood Estimation	Eq. (8)	[21, 36]

Table 1: Review of the fitting methods used previously for DGBD.

The methods listed in Table 1 and used in previous work to fit DGBD encompass Multiple Linear Regression, Nonlinear Least Square, and Maximum Likelihood Estimation (MLE). These methods are equivalent under the assumption of independent variables and Gaussian noise. However, in the case of the DGBD distribution, the variables are not independent, and

as we shall see the likelihood function does not have a closed form (Appendix C).

The first method listed in Table 1 is the Multiple Linear Regression, a widely used statistical technique for analyzing the relationship between a dependent variable and multiple independent variables:

$$\log x = \log C + \beta \log (N - r + 1) - \alpha \log r + \varepsilon, \tag{5}$$

where $\log x$ is the dependent variable, α, β, C are coefficients to be estimated, $\log r$ and $\log (N - r + 1)$ are variables, and ε is the error term, usually Gaussian distributed. In this approach, the dependent variable is a linear combination of the independent variables, and the coefficients quantify the impact of each independent variable on the dependent one. However, the estimation of the parameters is unstable when the variables have some degree of multicollinearity. In the case of DGBD, the variables $\log r$ and $\log (N - r + 1)$ are dependent since they both are functions of the rank r. Therefore, caution should be exercised when using regression as the assumptions of independence are not met.

The second technique in Table 1 is Nonlinear Least Squares. It is an optimization technique that fits nonlinear models, such as DGBD, to observed data. It minimizes a chosen loss function, in our case the sum of the square of errors on the logarithmic scale:

$$\min_{\alpha,\beta,C} \sum_{r=1}^{N} \left(\log x(r) - \log C - \beta \log (N - r + 1) + \alpha \log r\right)^2. \tag{6}$$

Unlike linear regression, this approach accommodates nonlinear relationships between variables by estimating parameters that minimize the sum of squared residuals. It step-by-step refines the parameter estimates until the gradient of Eq. 6 is sufficiently close to zero. This method was applied to fit rank distributions in diverse scientific domains, including social systems [19, 23, 25, 37], bibliometrics [6] and biology [22]. Its flexibility and ability to handle nonlinearities make it useful for the investigation of complex systems. However, the sensitivity to initial parameter estimates remains critical, as improper choices may lead to convergence towards local minima rather than the global minimum. Additionally, the vulnerability to outliers poses a challenge, potentially compromising the accuracy of model fitting.

The third method in Table 1 is the Method of Moments. In this method, moments (such as mean, variance, skewness, etc.) of the observed data are

equated to the corresponding theoretical moments of the probability distribution

$$\mu_i(\alpha, \beta) = \hat{\mu}_i(\hat{\alpha}, \hat{\beta}), \tag{7}$$

where μ_i and $\hat{\mu}_i$ are the i-th moment and sample moment. By solving these equations, estimates of the parameters $\hat{\alpha}, \hat{\beta}$ of the DGBD are obtained. This method is widely used in statistics, including for Beta Rank Functions (BRF) serving as the continuous equivalent of DGBD [35]. However, deriving a closed-form expression for DGBD poses a challenge. Nonetheless, the derivation of the first two moments for the log-BRF family of distributions [35] was successful.

The last method in Table 1 is Maximum Likelihood Estimation [21, 36]. This approach aims to determine the parameter values that maximize the likelihood function within a parameter space denoted as $\boldsymbol{\theta} \in \Theta$. The likelihood function is defined as follows:

$$\max_{\alpha,\beta,C} \mathcal{L}_{\mathcal{R}}(x; \alpha, \beta, C) = \max_{\alpha,\beta,C} \prod_{i=1}^{n} f(x_i; \alpha, \beta, C), \tag{8}$$

Here $f(x_i; \alpha, \beta, C)$ is a probability of obtaining x_i given α, β, C parameters. Maximizing the likelihood, one can identify the most reasonable set of parameter values that explain the observed data. It is important to note that MLE was designed to estimate parameters in the probability domain. As we said earlier, for DGBD one faces the problem of having no compact analytical formula on the probability domain, and one needs to compute numerical values of its inverse $x^{-1}(r)$ (see Appendix A). We derive these formulas analytically for the case of DGBD in Appendix C, resulting in implicit equations for the value of α and β:

$$\frac{\sum_{r=1}^{N} \log(r) r^{-\alpha} (N - r + 1)^{\beta}}{\sum_{r=1}^{N} r^{-\alpha} (N - r + 1)^{\beta}} = \frac{1}{C} \sum_{r=1}^{N} x(r) \log(r), \tag{9}$$

$$\frac{\sum_{r=1}^{N} \log(N - r + 1) r^{-\alpha} (N - r + 1)^{\beta}}{\sum_{r=1}^{N} r^{-\alpha} (N - r + 1)^{\beta}} = \frac{1}{C} \sum_{r=1}^{N} x(r) \log(N - r + 1). \tag{10}$$

There is no analytical solution for the above system of equations for α and β, so we solve them numerically using Powell hybrid method.

3.2. Exploring the idea of a good fit

The concept of a "good fit" holds fundamental importance for emprirical investigations across numerous fields of study, spanning physics, statistics, and mathematics [30, 38]. A "good fit" refers to the degree of agreement between some observed data and a theoretical or expected model. This section will briefly explore the meaning and importance of a good fit for the rank-size distribution domain.

Method	Formula	Publications
Coefficient of determination (R^2)	Eq. (11)	[18, 19, 22, 28, 34, 31, 32, 33, 39, 37]
Mean squared error	$\frac{1}{N}\sum(y_i - y_{\text{pred}})^2$	[23, 25, 26, 33]
KS distance	Eq. (12)	[20, 21, 36]
KS test	% of $D_{\text{sim}} < D_n$	[20]

Table 2: Review of the goodness of fit measures used for DGBD.

One of the most widely used measures to assess the goodness of fit is the coefficient of determination, also known as R^2. This coefficient measures the proportion of the variance in the dependent variable that is predictable from the independent variable when using a linear model [40]:

$$R^2 = 1 - \frac{\sum_r \left(x(r) - C\frac{(N-r+1)^\beta}{r^\alpha}\right)^2}{\sum_r \left(x(r) - \bar{x}\right)^2}, \tag{11}$$

where $x(r)$ is the observed rank data and $\bar{x}$ is the mean of the data.

A high coefficient of determination indicates a linear association between the predicted and observed data, commonly interpreted as a good fit. In the later sections of the article, we will highlight the limitations and potential weaknesses of this coefficient.

The other measures mentioned in Table 2 are closely associated with hypothesis testing. The Kolmogorov-Smirnov test (KS test) [41] is the only non-parametric statistical test that has been used for computing goodness of fit for DGBD, despite the existence of other tests such as the chi-square test. KS test compares the observed data empirical cumulative distribution function $F_{\text{emp}}(x)$ to the theoretical or expected cumulative distribution function $F_\Theta(x)$ by computing the maximum distance. This measure is commonly

referred to as the KS distance, denoted as:

$$D_{\boldsymbol{x}_N} = \sup_{x \in \boldsymbol{x}_N} |F_{\text{emp}}(x) - F_{\Theta}(x)|, \tag{12}$$

where $\sup_x$ is taken over the whole sample $\boldsymbol{x}_N$. Through Monte Carlo Simulation and the bootstrap method, we obtain the probability of getting the observed KS distance by chance alone, referred to as the p-value [7].

3.3. Pitfall and limitations of the considered methods

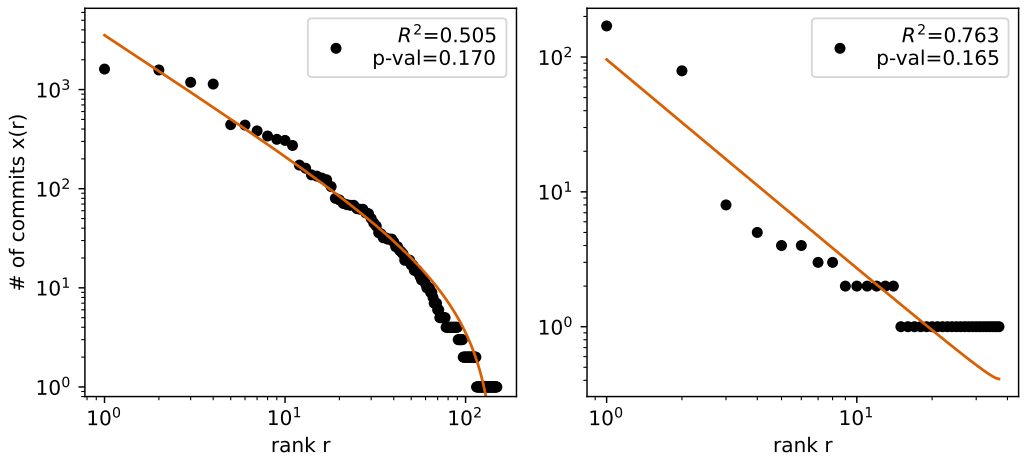

Figure 2: In both plots, the focus is on the potentially misleading R^2 coefficient. On the left, it is evident that the data closely aligns with the orange fit. On the right, the visual fit appears worse, but the R^2 value is higher, which is deceptive. This discrepancy may be attributed to the non-linearity inherent in DGBD. The data used for this comparison is from two different GitHub repositories.

While R-squared is a commonly used measure for assessing the quality of a regression model, it can be misleading in some instances. We emphasize, based on Table 2, that many authors rely on the R-squared coefficient (Eq. 11) and visual assessment to confirm the presence of a DGBD. However, as illustrated in Figure 2, a high R^2 value does not necessarily imply that the model best fits the data. Other factors, such as model complexity and sample size, must also be considered.

Among the methods listed in Table 2, the Kolmogorov-Smirnov test is the only statistical test that compares the result obtained to expectation to

derive a p-value. Therefore, it is the only method that provides the significance level at which one obtains the result [42]. However, the KS test operates at the level of the cumulative distribution $S(x)$, while we are interested in the quality of fit at the level of the rank-size distribution $x(r)$ in double logarithmic scale. Because of the power-law nature of the distribution, this means that the KS test will focus on deviations happening at low x values, corresponding to high rank values. As shown in Fig 3, when the data is inverted, the axes are swapped, meaning that the KS test measures the horizontal distance on the rank-size plot. The green arrow in the right-hand plot represents the KS distance at the commit count equal to 1. However, this effect is not readily apparent in the rank-size plot due to the double logarithmic scale, which makes hundreds of points in the tail appear as minor deviations from the best fit. Therefore, the KS test will generate a strong bias towards fitting the tail of the rank-size distribution, which might not be representative of the broader distribution, especially the top ranked elements. This discrepancy highlights how a fit yielding a relatively good p-value can be rejected based on convexity in double logarithmic scale.

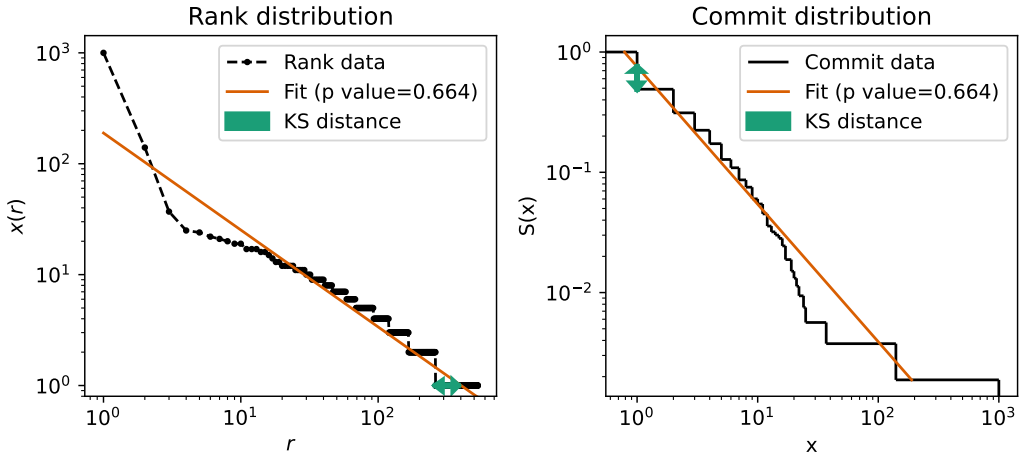

Figure 3: An illustrative example of a potentially misleading p-value stemming from the KS test is presented. Despite the p-value being relatively high at 0.664, suggesting that the data conforms to the DGBD model, a visual inspection reveals that the fit is notably poor. On the right-hand side of the figure, we provide insight into what the KS distance is evaluating, as indicated by the green arrow. This example is based on the repository named *Faker*.

4. Application to commit distributions in open-source software

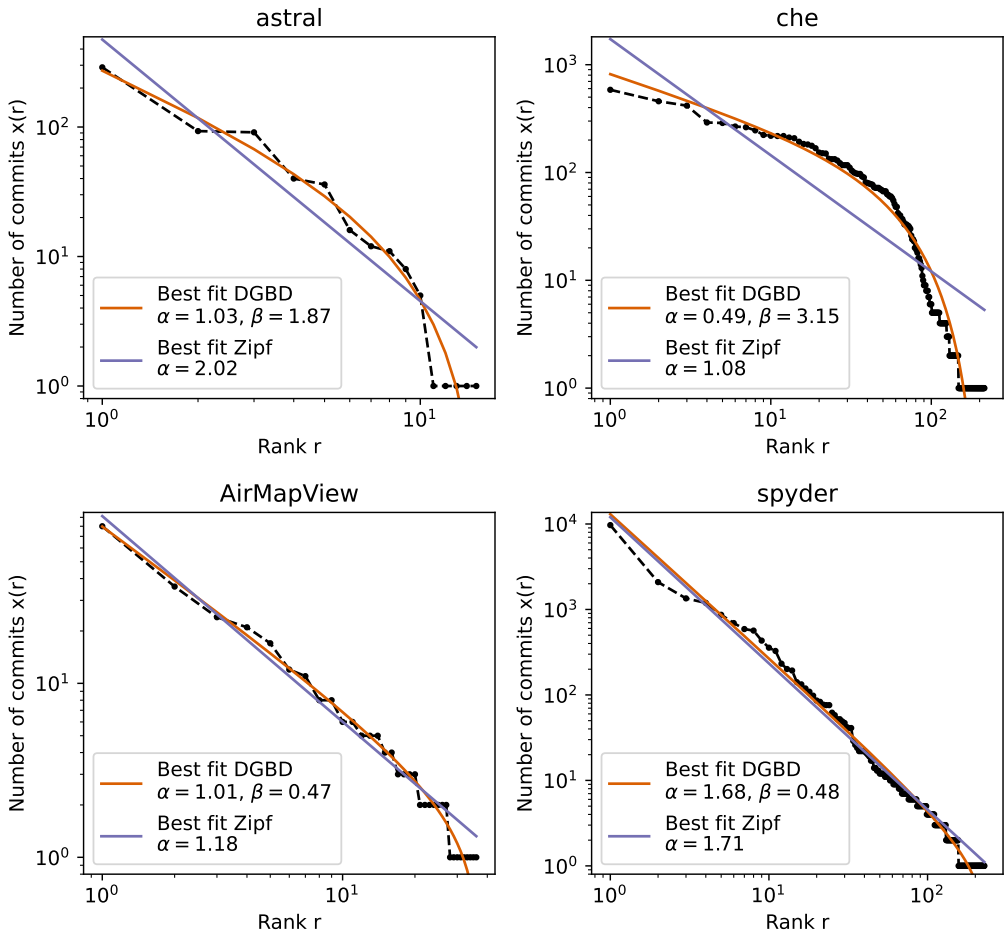

Figure 4: Rank-size distribution showing the number of commits as a function of a user rank across 4 open-source repositories selected based on their concave behavior in a double log scale. We have divided the charts in such a way that the charts on the left show small repositories, while the charts on the right show large repositories. The top two plots show examples where DGBD (orange line) is significantly better than Zipf's law (blue line), while in the bottom both models produce similar results.

In order to empirically explore the fit of the DGBD distribution, we use commit data from the web-based collaborative software development platform GitHub. It allows users to store and manage their code, track changes made to the code (referred to as commits), and collaborate with

others through pull requests, code reviews, and issue tracking. It is the most prominent source code host, with over 83 million developers and more than 200 million repositories [43]. Commit distributions are known to be heavy-tailed and GitHub repositories have been investigated in the context of rank dynamics [1].

As we mentioned in Section 2, fitting the DGBD model is meaningful only when the data exhibits concavity in a double log scale, as we elaborated in Appendix B. Consequently, we selected repositories from GitHub that showed this concavity. In Figure 4, we showcase 4 repositories where we conducted the fitting process and assessed the goodness of fit. The estimators were computed using the nonlinear least square method with the Trust Region Reflective algorithm [44] to minimize residuals, as introduced in Section 3. The outcomes of these analyses are presented in Figure 4. We also explored maximum likelihood estimation, but it yielded similar or worse results compared to nonlinear least square, and sometimes did not converge. MLE does not describe well the heavy tails of CCDF and the same applies to the rank-size distribution.

4.1. Binning data

We first note the importance of binning the data to improve the fitting process. Indeed, in the heavy tail processes, the tail contains most data points, dominating the early part of the rank-size distribution. However, this early part includes the top individuals, meaning that a traditional fitting procedure yields very noisy estimates over a crucial part of the data. To remedy this problem, before the fitting procedure we use bins on commit data to give equal weights to all ranks in double log space. When one fits by minimizing residuals using nonlinear least squares, binning methods provide a *fairer* fit over all the distribution and avoids treating the essential part of top-ranked elements as outliers.

We compared three types of data binning methods. The first is log binning, also referred to as logarithmic binning or log-spaced binning. This technique groups data points into bins or intervals so that each bin covers a range of values that increases exponentially [7]. However, log binning comes with certain drawbacks. Firstly, the quality of the analysis is highly dependent on the choice of the number of bins. Selecting too few or too many bins can result in an oversimplified or overly detailed representation of the data. Furthermore, when employing log binning, comparing fits can be complicated because different numbers of bins can yield different fit results. This

lack of consistency can pose challenges when making comparisons between other datasets or studies, impacting the reproducibility and generalizability of the analysis.

In contrast to log binning, we introduce mean binning as an alternative for integer-type data, such as the number of commits. The approach is straightforward: for each integer value of commits, denoted as x, we compute its rank $\hat{r}_x$ as the mean value of all ranks corresponding to a given number of commits $x(r) = x$:

$$\hat{r}_x = \frac{\sum_{\{r\,:\,x(r)=x\}} r}{\#\{r \;:\; x(r) = x\}}. \tag{13}$$

One of the primary advantages of mean binning over log binning is its parameter-free nature. As mentioned earlier, the analysis, including best-fit computation, when based on log binning, is significantly influenced by the choice of the number of bins. In contrast, mean binning does not require specifying a bin count, eliminating this parameter.

Third, we introduce geometrical binning, where for each value of commits x we calculate the geometrical mean of its corresponding ranks:

$$\hat{r}_x = \left(\prod_{\{r\,:\,x(r)=x\}} r \right)^{1/\#\{r\,:\,x(r)=x\}}. \tag{14}$$

On a double log scale, the midpoint between two points x and y is determined by the equation:

$$\frac{\log x + \log y}{2} = \log(\sqrt{xy}),$$

where $\sqrt{xy}$ is called the geometric mean. Hence, we refer to this binning method as geometric bin.

Figure 5 shows the impact of binning on fit performance. In the left-hand graph, one can see an example repository where the best fit without binning (represented by the orange line) yields poor results. Applying any of the three binning methods visibly improves the best fit. On the right-hand side of the same Figure, we show the distribution of the maximum log-fold change between the best fit and actual data, computed as follow:

$$\log_2 \mathrm{FC} = \log_2 \left(\max \left\{ \max_r \frac{x(r)}{x_{\mathrm{DGBD}}(r)}, \max_r \frac{x_{\mathrm{DGBD}}(r)}{x(r)} \right\} \right), \tag{15}$$

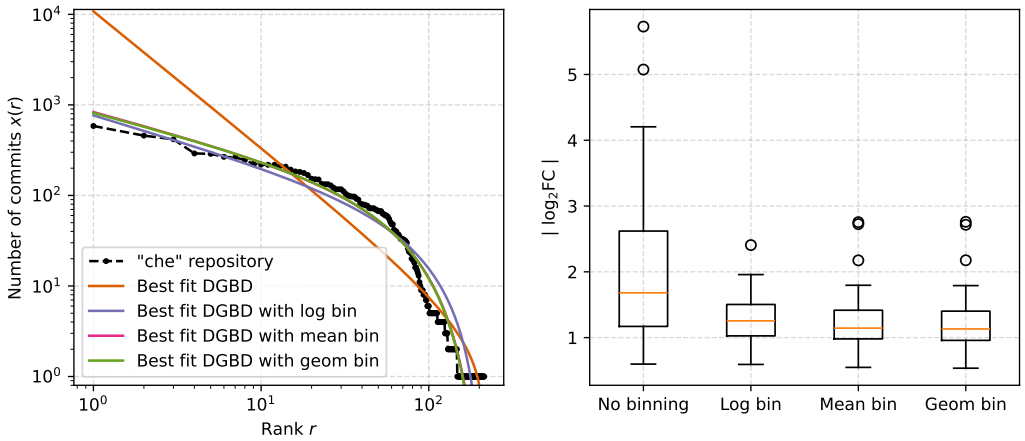

Figure 5: The left-hand plot illustrates the impact of binning data on the best fit. Binning is necessary for the good fit to emphasize the tail of the rank-size distribution, leading to an uneven representation of points in the double log scale. Consequently, the orange line yields visually suboptimal results, especially when compared to the log bin or geometrical bin. On the right-hand plot, we visualize the logarithmized maximum fold change distribution between the data and the best fit. This representation highlights that the maximum distance is several times greater when data is not binned compared to the binned versions.

where $x(r)$ is actual data and $x_{\mathrm{DGBD}}(r)$ is data determined from the best fit using formula (4). The log-fold change describes the difference between the observed and expected distributions in the log scale, and is similar to the KS technique, adapted for rank-size distributions (i.e. vertical deviations rather than horizontal ones). Additionally, when DGBD takes values below 1, for calculations of log-fold change, we assume DGBD=1, as the smallest number of commits one can have is precisely 1 and the comparison is meaningful.

We find that the maximum log-fold change is on average 2.54 times greater when no binning is applied, yielding significantly poorer results compared with any binning methods ($p < 0.01$, Mann-Whitney test). Finally, we find no significant difference between the different binning methods ($p > 0.05$ between each pair of binning methods, Mann-Whitney test). Boxplots of maximum log-fold changes of the different binning methods are shown in Figure 5.

As such, it is not possible to unequivocally determine which type of data binning is best. Among the concave repositories we tested, we can find at least one for which one of the four approaches is best in terms of Fold Change

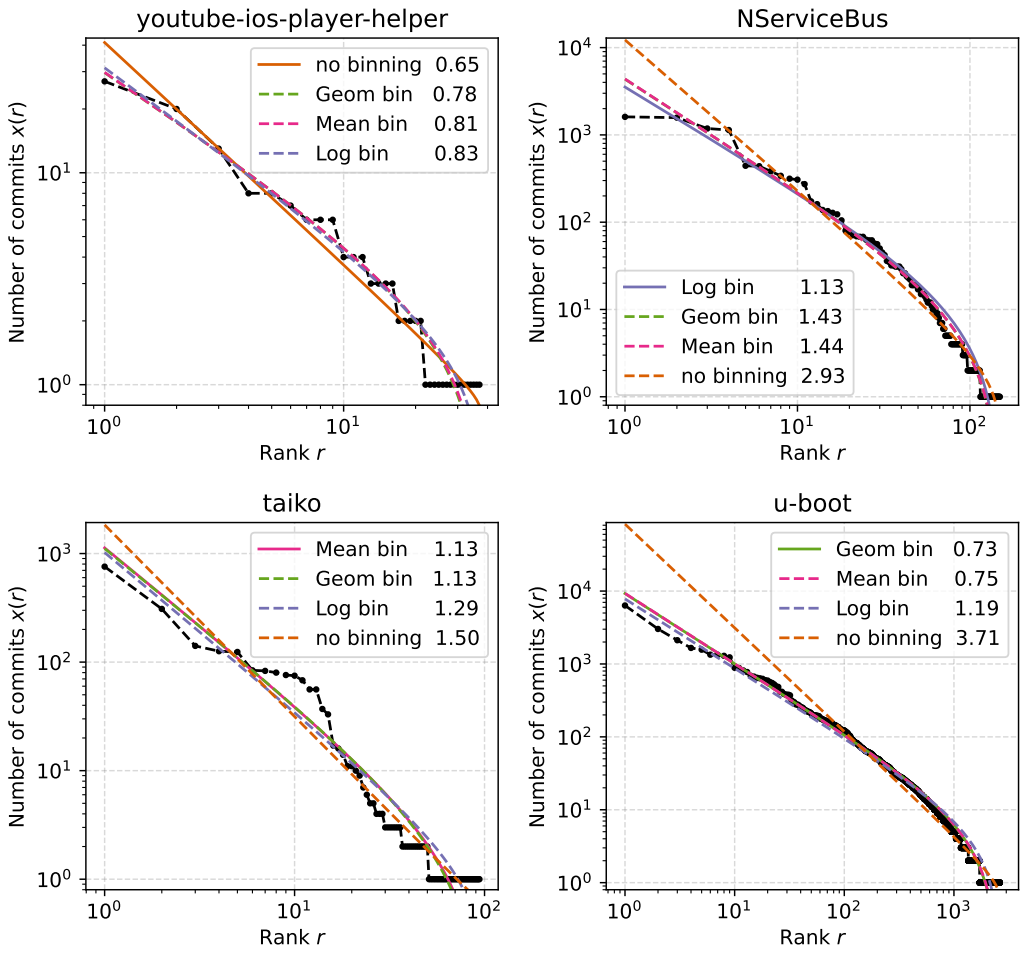

Figure 6: Comparison of four different repositories regarding the use of various binning methods. They were selected in such a way that a different binning method performs better on each of them compared to the others. The numbers in the legend represent the absolute value of maximum log-fold change computed by 15. In the top-left corner, no binning is the best, in the top-right corner, log binning, in the bottom-left corner, mean binning, and in the bottom-right corner, geom binning is the best.

measure. In Figure 6, we highlight four repositories in which we have marked the increasing value of FC calculated by formula (15) for the different types of binning, with the best method having the lowest FC. Therefore, when analyzing individual datasets, caution should be exercised in the choice of binning, as the statistically best option may not necessarily be the best choice

for that specific dataset.

4.2. Performance of DGBD compared with Zipf Law

In the introduction, we mentioned that DGBD serves as a natural generalization of Zipf's Law, with the potential to extend its applicability. To assess this, we compared their goodness of fit to the commit distributions using p-values from the KS test described in Section 3. For the case of DGBD we utilized the same schema as in [7], which is a bootstrap method with Monte Carlo simulations to ensure the validity of p-values. To study the quality of fit, we need to generate M samples, where each sample consists of N pseudorandom values from DGBD distribution with parameters N, α, β and C separately obtained from each repository. Then we estimate the distribution parameters $\hat{\alpha}$, $\hat{\beta}$ and $\hat{C}$ and compare the estimated distribution with the initial values. We compute M KS distances using Eq. (12) and compare them with the corresponding value from the given repository to obtain a p-value. This step is crucial because DGBD lacks an analytical inverse function (as detailed in Appendix A), necessitating Monte Carlo simulations.

We show the results of the statistical analysis in Figure 7. We first show a volcano plot comparing the statistical significance (p-value) against the magnitude of change (maximum log-fold change) for each fit. Inspecting both measures simultaneously allows for a better confirmation of DGBD and Zipf's Law. We set the thresholds to $p = 0.05$ for the p-value and $|\log_2 FC| = 2$ for the maximum log-fold change. For example, we excluded three cases where, despite a high p-value, the maximum fold change was exceedingly large (see two gray triangles and one gray circle on the left side of the plot). As we can see, most of the DGBD fits pass the threshold, while a larger number of Zipf tests fail.

Another comparison of Zipf's Law and DGBD is illustrated in the right-hand plot of Figure 7. It shows the maximum fold change for both models alongside a diagonal line. We find that the points are concentrated above the diagonal line, meaning that the maximum fold change for Zipf's Law is generally larger than that of DGBD. As such, DGBD yields better results than Zipf's Law, except for three repositories, where DGBD has worse FC. Some repositories fall on or close to the diagonal line, meaning that both Zipf's Law and DGBD are good models to fit such repositories. These correspond to the cases where the data is shaped like a straight line across all ranks (see repositories at the bottom of Figure 4).

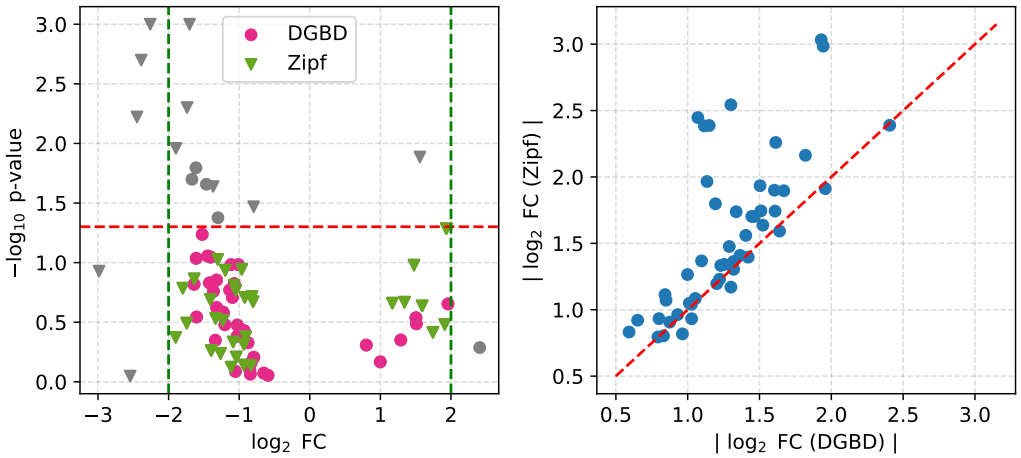

Figure 7: On the left-hand side, the volcano plot compares the statistical significance (p-value from KS test) with maximum fold change of each fit to a commit distribution. Circles and triangles respectively represent DGBD and Zipf fits. Goodness of fit is set with two thresholds at $p > 0.05$ for p-value and $\log_2 \mathrm{FC} < 2$ for the fold change, and indicated by the use of pink (DGBD) and green (Zipf) colors. Comparison between the two model shows that we reject Zipf more often than DGBD. On the right-hand graph we compare the absolute value of the log-fold changes for Zipf and DGBD across repositories. Points generally lie above the diagonal, indicating that DGBD provides a the same or better fit in almost every case, except of three repositories.

To address the concern regarding the improvement of model fit not being solely attributed to an increased number of parameters (DGBD has two exponents α and β, while Zipf's Law has only one exponent α), we employed a k-fold cross-validation. This involved partitioning the data set into 75% for training and 25% for testing, and repeating this process multiple times to compute the average accuracy and its standard error on the test set. The accuracy was measured using a normalized sum of residuals, which is a what we minimized when determining the best fit via nonlinear least squares. Through this method, we aimed to demonstrate that the DGBD provides on average better fit compared to the Zipf's Law, beyond the mere inclusion of additional parameters.

Our findings are supported by a Figure 8, where we plot model accuracy as a function of repository size. For small repositories ($N < 30$), the Zipf's Law performs better due to the tendency of DGBD to overfit the initial few data points. This is particularly evident in k-fold cross-validation, where the probability of randomly selecting the most influential contributors is higher

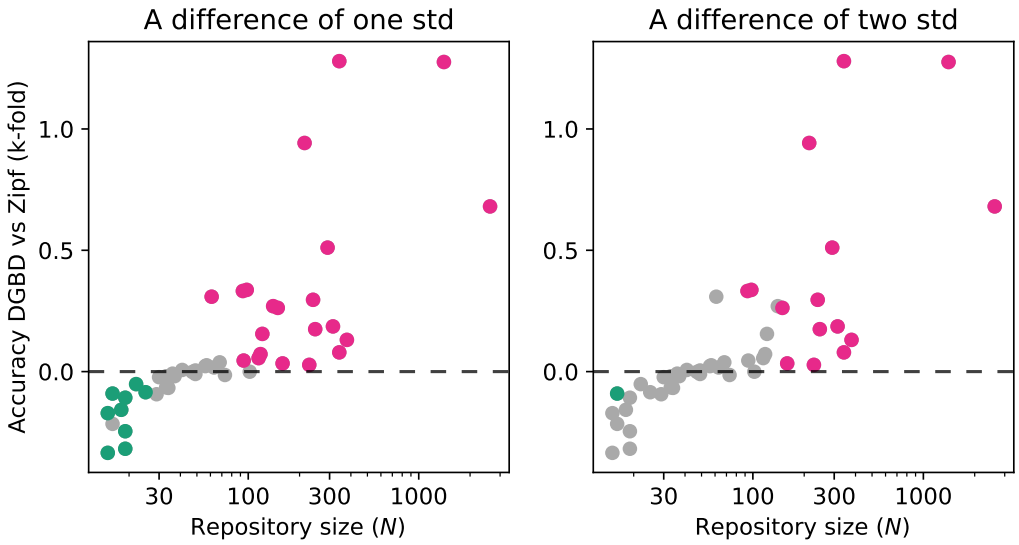

Figure 8: Comparison of accuracy of the DGBD and Zipf's Law across varying repository sizes, assessed using k-fold cross-validation. The accuracy difference is computed as the difference in the normalized sum of residuals between Zipf and DGBD models. Values above zero indicate that the DGBD does not overfit and performs better than the Zipf's Law, whereas values below zero signify that the DGBD overfits the data and has worse accuracy compared to the Zipf's Law. The points on the left-hand plot are color-coded as follows: green points indicate repositories where the Zipf's Law performs better than the DGBD within one standard deviation, pink points denote repositories where the DGBD outperforms the Zipf's Law within one standard deviation, and gray points represent repositories where both models perform equally well. Right-hand plot shows the same accuracy measure but highlights repositories where the performance difference between the two models exceeds two standard deviations.

in smaller repositories, leading to less reliable performance for DGBD. However, as repository size increases ($N > 100$), DGBD consistently outperforms Zipf's Law. This improvement is not only statistically significant but also aligns with theoretical expectations that larger data sets benefit more from the additional flexibility provided by DGBD. Figure 8 clearly shows this transition: green points indicate regions where Zipf performs better, pink points highlight areas where DGBD excels, and gray points show where both models perform equally well. Thus, while Zipf may suffice for small datasets, DGBD proves to be the superior choice for larger datasets, validating DGBD.

5. Conclusions

The methodologies employed in recent works investigating DGBD contain limitations and weaknesses that can give misleading results (see Figure 2). Neglecting these limitations may lead to the incorrect conclusion that DGBD is suitable for the data, even when the visual fit is clearly poor (as exemplified in Figure 3). The solutions proposed in this paper address several of the issues encountered in analyzing rank-size distributions. We found that employing log binning, mean binning, or geometrical binning can significantly improve the quality of best-fit models by reducing the importance of high r values in the tail. Furthermore, we showed that DGBD is primarily limited to datasets exhibiting a concave pattern when plotted on a double log-scale. This limitation was derived and established in Section Appendix B, and its significance was illustrated in Figure 3, where we showed the consequences of attempting KS tests on convex data. We found that in the case of concave rank-size distributions from commit data, DGBD consistently provides better fit than Zipf.

Future directions include evaluating approaches for convex data patterns in a double-log scale. Our examination encompassed a wide range of skewed rank-size distributions and probability distributions, including Zipf, Pareto type II [6], log-normal [12], Tsallis-Pareto [13], Tsallis q-exponential distribution [14], other Pareto modifications [9] and Gumbel [16]. None of these distributions exhibited convex patterns on a double log scale. However, this exploration raises an intriguing challenge concerning modeling datasets characterized by convex patterns, which remains an open area for statistical research.

Conflict of interest

All authors certify that they have no affiliations with or involvement in any organisation or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Acknowledgement

This research was supported by the POB Research Centre Cybersecurity and Data Science of Warsaw University of Technology within the Excellence

Initiative Program – Research University (ID-UB). This work was also supported by the French Agence Nationale de la Recherche (ANR), under grant agreement ANR-21-CE38-0002.

CRediT authorship contribution statement

Przemysław Nowak: Conceptualization, Formal Analysis, Methodology, Software, Visualisation, Writing – original draft **Marc Santolini:** Conceptualization, Data curation, Methodology, Project administration, Resources, Supervision, Writing – review & editing **Chakresh Singh:** Conceptualization, Methodology, Writing - review & editing **Grzegorz Siudem:** Conceptualization, Formal Analysis, Methodology, Writing – review & editing **Liubov Tupikina:** Conceptualization, Methodology, Writing – review & editing

Appendix A. Ranks and cumulative distributions

In the following section, we derive the relation between the rank-size distribution $x(r)$ and the cumulative distribution function $S(x)$. It bridges the classical domain of probability theory with the rank domain described with rank-size distribution. For that purpose, let us consider an ordered sample $x(1) \geqslant x(2) \geqslant \cdots \geqslant x(N)$ of size N drawn from the independent identically distributed ($X_i \sim X,\ i = 1, \ldots, N$) random variables $\mathbf{X}_N = (X_1, X_2, \ldots X_N)$. Let us define the expected rank function $\mathfrak{R}_N$ for the variables $\mathbf{X}_N$ of the variable x from the support of variable X given with the formula

$$\mathfrak{R}_N(x) = \mathbb{E}\left(r(\mathbf{X}_N)\right),$$

i.e. it is expected value of the rank for the observation of value x drawing from the variables $\mathbf{X}_N$. In the following steps we derive the formula for $\mathfrak{R}_N$ and then connect it with the empirical ranks $r(x_i)$.

Let us consider firstly the probability that an item of value x is on the i-th place in the ordered sample drawn from $\mathbf{X}_N$, i.e.:

$$\mathbb{P}\left(X_{(N)} \leq \cdots \leq X_{(i+1)} \leq x \leq X_{(i-1)} \leq \cdots \leq X_{(1)}\right) = (\triangle), \tag{A.1}$$

where $X_{(1)}, \ldots, X_{(N)}$ are order statistics of variables $X_1, \ldots, X_N$. One can further express Eq. (A.1) as follows:

$$\begin{aligned}
(\triangle) =& \mathbb{P}\left(\exists_{k_1,\ldots,k_{i-1}} X_{k_1}, \ldots, X_{k_{i-1}} \geq x \ \wedge \forall_{l \notin \{k_1,\ldots,k_{i-1}\}}\ x \geq X_l\right) = \\
& \sum_{\sigma \in S_{N-1}} \mathbb{P}\left(Y^{(i)}_{\sigma_{(1)}} \geq \ldots Y^{(i)}_{\sigma_{(i-1)}} \geq x \geq Y^{(i)}_{\sigma_{(i)}} \geq \cdots \geq Y^{(i)}_{\sigma_{(N-1)}}\right) = \\
& \binom{N-1}{i-1} \mathbb{P}^{i-1}\left(X \geq x\right) \mathbb{P}^{N-i}\left(X \leq x\right) = \\
& \binom{N-1}{i-1} \left(\underbrace{S(x) + \mathbb{P}(X = x)}_{u(x)}\right)^{i-1} \left(\underbrace{1 - S(x)}_{v(x)}\right)^{N-i} = \\
& \binom{N-1}{i-1} u^{i-1}(x) v^{N-i}(x),
\end{aligned}$$

where we expand the sum over all possible permutations of the considered sample (with the notion $Y^{(i)}_\ell = X_\ell$ for $\ell < i$ and $Y^{(i)}_\ell = X_{\ell+1}$ for $\ell > i$), but,

since the probabilities under the sum are independent, it could be further simplified with the notionn of $u(x) = S(x) + \mathbb{P}(X = x)$ and $v(x) = 1 - S(x)$.

Expected value respect to the above probabilities gives desired expected rank function $\mathfrak{R}_N$ (for brevity we skip argument writing u and v instead of $u(x)$ and $v(x)$):

$$\begin{aligned}
\mathfrak{R}_N(x) = &\sum_{i=1}^{N} i \binom{N-1}{i-1} u^{i-1} v^{N-i} = \sum_{j=0}^{N-1} (j+1) \binom{N-1}{j} u^j v^{N-j-1} = \\
&\sum_{j=1}^{N-1} \frac{(N-1)!}{(j-1)!(N-j)!} u^j v^{N-j-1} + \sum_{j=0}^{N-1} \binom{N-1}{j} u^j v^{N-j-1} = \\
&\sum_{k=0}^{N-2} \frac{(N-1)!}{k!(N-k-2)!} u^{k+1} v^{N-k-2} + (u+v)^{N-1} = \\
&u(N-1)(u+v)^{N-2} + (u+v)^{N-1} = (uN+v)(u+v)^{N-2}.
\end{aligned}$$

For continuous random variables, where $P(X = x) = 0$ holds, one obtains:

$$\mathfrak{R}_N(x) = (N-1)S(x) + 1,$$

which could be further inverted into the formula for x in terms of its expected rank: $\mathfrak{R}_N$

$$x = S^{-1}\left(\frac{\mathfrak{R}_N - 1}{N-1}\right).$$

The question remains: can we replace the expected rank $\mathfrak{R}_N$ with an empirical rank function obtained from the sample we observe $\{x(1), x(2), \ldots, x(N)\}$? The answer to this question is positive if the observed sample is large enough. This asymptotics can be argued in various ways: starting from the Law of Large Numbers (for empirical ranks), Glivenko–Cantelli theorem (note that the rank function is the inverse of the survival function, and, therefore, related to the cumulative distribution function) or, more in the style of physicists, replacing the appropriate integrals their maximum values, just like [45]. So finally we can assume that $r(x) \approx \mathfrak{R}_N(x)$ which gives desired relationship between the empirical rank function and its expected probabilistic version,

$$x = S^{-1}\left(\frac{r-1}{N-1}\right),$$

which also establishes the expected relationship between empirical ranks and the inversion of the survival function.

Appendix B. Selected properties of DGBD

In this section, we derive two selected DGBD properties - limitations of the parameters' values and the convex-concave analysis.

Appendix B.1. Boundaries for the parameters

Every rank-size distribution (in particular DGBD) must, by definition, be non-decreasing. However, not all functional forms of rank-size distributions should be monotonic for every set of parameters. Therefore, before any analysis, it is essential to determine the parameter set for which the mentioned functions are non-decreasing. To ensure this, it suffices to demonstrate when the first derivative is non-positive. So, let us differentiate Eq. (4):

$$\frac{dx(r)}{dr} = \frac{(N-r+1)^{\beta-1}}{r^{\alpha+1}} \left(-\beta r - \alpha \left(N-r+1\right)\right). \tag{B.1}$$

For the monotonicity, one requires that DGBD satisfies $x'(r) \leq 0$ for $r \in [1, N]$. The fraction in (B.1) is always positive, so the expression in brackets needs to be negative and thus:

$$r(\alpha - \beta) - \alpha(N+1) \leq 0. \tag{B.2}$$

The left-hand side of the above inequality is an affine function of r. To ensure the distribution's monotonicity, it is enough to satisfy the inequality for the endpoints of the interval, i.e., for $r = 1$ and $r = N$, which gives us:

$$\begin{cases} N\alpha + \beta \geq 0, \\ \alpha + N\beta \geq 0. \end{cases} \tag{B.3}$$

The conclusion drawn from Eq. (B.3) is crucial, considering that α and β can potentially assume negative values. However, it is essential to emphasize that their negativity is limited due to stringent constraints. The absolute values of α and β (for negative values) must remain small, mainly when dealing with large sample sizes N.

Appendix B.2. The convex-concave analysis

We demonstrate that DGBD can serve as an effective model only when data exhibit concavity (i.e., curved or rounded inward) in a double log scale. To illustrate this, let us express the second derivative of DGBD in logarithmic

terms, using variables $y = \log x(r)$, $z = \log r$. Logarithms of DGBD (4) takes the following form:

$$y(z) = \log C - \alpha z + \beta \log (N - e^z + 1). \tag{B.4}$$

Second derivative is:

$$\frac{d^2y}{dz^2} = -\beta(N+1)\frac{e^z}{(N - e^z + 1)^2}. \tag{B.5}$$

Analysis of Eq. (B.5) concludes that DGBD exhibits convexity when $\beta < 0$ and concavity when $\beta > 0$. However, it is essential to note that the latter case is minimal due to the previously derived boundaries (B.3). The minimum value for β is $\beta = -\frac{\alpha}{N}$, which restricts its use as a suitable choice for fitting data that appears convex on a double log scale.

Appendix C. Maximum likelihood estimation

In this section, we derive maximum likelihood estimators (MLE) for BGBD. Usually, MLE is based on the probability density function, so we must extend the typical approach to the rank-size distributions. Let us assume that the rank distribution is given with the general formula:

$$x(r) = C f_\Theta(r), \tag{C.1}$$

where $C^{-1} = \sum_{k=1}^{N} f_\Theta(r)$ and $f_\Theta(r)$ is a frequency of ranks, possibly depends on some parameters $\Theta = (\theta_1, \theta_2, \ldots, \theta_p)$. We aim to estimate the parameters, knowing the data $\boldsymbol{x}_N$ and assuming the form of the frequency distribution f_Θ. Introduced setup results in the following form of the likelihood function:

$$\mathcal{L}_\mathcal{R}(x_1, x_2, \ldots, x_N; \theta_1, \ldots, \theta_p) = \prod_{r=1}^{N} \left(f_\Theta(r)\right)^{x(r)}, \tag{C.2}$$

which simplifies after taking logarithm:

$$\log \left(\mathcal{L}_\mathcal{R}(x_1, x_2, \ldots, x_N; \theta_1, \ldots, \theta_p)\right) = \sum_{r=1}^{N} x(r) \log \left(f_\Theta(r)\right). \tag{C.3}$$

We aim to look for the parameters Θ, which maximizes $\mathcal{L}_{\mathcal{R}}$ with given ranked data $\boldsymbol{x}_N$. For DGBD there are two parameters (i.e. $\Theta = (\alpha,\, \beta)$) and the formula for the rank-size distribution is given with:

$$f_{\alpha,\,\beta}(r) = \frac{1}{H_N(\alpha,\, \beta)} \frac{(N+1-r)^\beta}{r^\alpha} \tag{C.4}$$

where we define notion of H_N, similar to harmonic number, for the case of two parameters:

$$H_N(\alpha,\, \beta) = \sum_{r=1}^{N} \frac{(N+1-r)^\beta}{r^\alpha}.$$

Thus $\log(\mathcal{L}_{\mathcal{R}})$ takes the following form:

$$\begin{aligned}
&\log\left(\mathcal{L}_{\mathcal{R}}(x_1,\, \ldots,\, x_N;\, \alpha,\, \beta)\right) \\
&= \sum_{r=1}^{N} x(r)\left(\beta \log(N+1-r) - \alpha \log(r) - \log(H_N(\alpha,\, \beta))\right) = \\
&= \underbrace{-\sum_{r=1}^{N} x(r) \log[H_N(\alpha,\, \beta)]}_{=C} - \alpha \sum_{r=1}^{N} x(r) \log(r) + \beta \sum_{r=1}^{N} x(r) \log(N+1-r).
\end{aligned}$$

The derivation respect to the α and β gives:

$$\begin{aligned}
\frac{\partial \log(\mathcal{L}_{\mathcal{R}})}{\partial \alpha} &= \frac{-C}{H_N(\alpha,\, \beta)} \frac{\partial H_N(\alpha,\, \beta)}{\partial \alpha} - \sum_{r=1}^{N} x(r) \log(r), \\
\frac{\partial \log(\mathcal{L}_{\mathcal{R}})}{\partial \beta} &= \frac{-C}{H_N(\alpha,\, \beta)} \frac{\partial H_N(\alpha,\, \beta)}{\partial \beta} + \sum_{r=1}^{N} x(r) \log(N+1-r).
\end{aligned}$$

We set both derivatives to zero, which gives the following implicit equations for α and β:

$$\frac{\sum_{r=1}^{N} \log(r) r^{-\alpha} (N-r+1)^\beta}{\sum_{r=1}^{N} r^{-\alpha} (N-r+1)^\beta} = \frac{1}{C} \sum_{r=1}^{N} x(r) \log(r), \tag{C.5}$$

$$\frac{\sum_{r=1}^{N} \log(N-r+1) r^{-\alpha} (N-r+1)^\beta}{\sum_{r=1}^{N} r^{-\alpha} (N-r+1)^\beta} = \frac{1}{C} \sum_{r=1}^{N} x(r) \log(N-r+1). \tag{C.6}$$

There is no analytical solution for the above system of equations for α and β, so we solve them numerically using Powell hybrid method.

References

[1] G. Iñiguez, C. Pineda, C. Gershenson, A.-L. Barabási, Dynamics of ranking, Nature Communications 13 (2022) 1646.

[2] P. Holme, Universality out of order, Nature Communications 13 (2022) 2355.

[3] F. Auerbach, A. Ciccone, The Law of Population concentration, Environment and Planning B: Urban Analytics and City Science 50 (2023) 290–298.

[4] G. K. Zipf, The Psycho-Biology of Language., Psychology Press, 1936.

[5] G. K. Zipf, Human Behavior and the Principle of Least Effort, Ravenio Books, 1949.

[6] G. Siudem, P. Nowak, M. Gagolewski, Power laws, the Price model, and the Pareto type-2 distribution, Physica A: Statistical Mechanics and its Applications 606 (2022) 128059.

[7] A. Clauset, C. R. Shalizi, M. E. J. Newman, Power-Law Distributions in Empirical Data, SIAM Review 51 (2009) 661–703.

[8] W. Li, Zipf's Law Everywhere, Glottometrics 5 (2002).

[9] S. Arshad, S. Hu, B. N. Ashraf, Zipf's law and city size distribution: A survey of the literature and future research agenda, Physica A: Statistical Mechanics and its Applications 492 (2018) 75–92.

[10] M. E. Newman, Power laws, Pareto distributions and Zipf's law, Contemporary physics 46 (2005) 323–351.

[11] B. C. Arnold, Pareto Distributions, Chapman and Hall/CRC, New York, NY, USA, 2015.

[12] M. Brzezinski, Power laws in citation distributions: evidence from Scopus, Scientometrics 103 (2015) 213–228.

[13] Z. Néda, L. Varga, T. S. Biró, Science and Facebook: The same popularity law!, PLOS ONE 12 (2017) 1–11.

[14] K. Gangopadhyay, B. Basu, City size distributions for india and china, Physica A: Statistical Mechanics and its Applications 388 (2009) 2682–2688.

[15] J. Liu, R. Serota, Rethinking Generalized Beta family of distributions, The European Physical Journal B 96 (2023) 24.

[16] C. K. Singh, E. Barme, R. Ward, L. Tupikina, M. Santolini, Quantifying the rise and fall of scientific fields, PLOS ONE 17 (2022) 1–15.

[17] G. Siudem, B. Żogała Siudem, A. Cena, M. Gagolewski, Three dimensions of scientific impact, Proceedings of the National Academy of Sciences 117 (2020) 13896–13900.

[18] M. Ausloos, Two-exponent Lavalette function: A generalization for the case of adherents to a religious movement, Phys. Rev. E 89 (2014) 062803.

[19] G. Martínez-Mekler, R. A. Martínez, M. B. del Río, R. Mansilla, P. Miramontes, G. Cocho, Universality of Rank-Ordering Distributions in the Arts and Sciences, PLOS ONE 4 (2009) 1–7.

[20] O. Fontanelli, P. Miramontes, G. Cocho, W. Li, Population patterns in World's administrative units, Royal Society Open Science 4 (2017) 170281.

[21] A. Ghosh, B. Basu, Universal City-size distributions through rank ordering, Physica A: Statistical Mechanics and its Applications 528 (2019) 121094.

[22] J. Zhang, Y. Feng, Common patterns of energy flow and biomass distribution on weighted food webs, Physica A: Statistical Mechanics and its Applications 405 (2014) 278–288.

[23] W. Li, Characterizing Ranked Chinese Syllable-to-Character Mapping Spectrum: A Bridge between the Spoken and Written Chinese Language, Journal of Quantitative Linguistics 20 (2013) 153–167.

[24] W. Li, Fitting Chinese syllable-to-character mapping spectrum by the beta rank function, Physica A: Statistical Mechanics and its Applications 391 (2012) 1515–1518.

[25] W. Li, P. Miramontes, Fitting Ranked English and Spanish Letter Frequency Distribution in US and Mexican Presidential Speeches, Journal of Quantitative Linguistics 18 (2011) 359–380.

[26] W. Li, Analyses of baby name popularity distribution in U.S. for the last 131 years, Complexity 18 (2012) 44–50.

[27] E. Lugo, R. Doti, J. Faubert, Characterization of Stochastic Resonance by the Discrete General Beta Distribution, Springer Singapore, 2019, pp. 751–758.

[28] R. Alvarez-Martinez, G. Cocho, G. Martinez-Mekler, Rank ordered beta distributions of nonlinear map symbolic dynamics families with a first-order transition between dynamical regimes, Chaos: An Interdisciplinary Journal of Nonlinear Science 28 (2018) 075515.

[29] O. Fontanelli, P. Miramontes, Y. Yang, G. Cocho, W. Li, Beyond Zipf's Law: The Lavalette Rank Function and Its Properties, PLOS ONE 11 (2016) 1–14.

[30] D. Bates, D. Watts, Nonlinear Regression Analysis and Its Applications, 2008, pp. 32 – 66.

[31] R. Alvarez-Martinez, G. Martinez-Mekler, G. Cocho, Order–disorder transition in conflicting dynamics leading to rank–frequency generalized beta distributions, Physica A: Statistical Mechanics and its Applications 390 (2011) 120–130.

[32] R. Mansilla, E. Köppen, G. Cocho, P. Miramontes, On the Behavior of Journal Impact Factor Rank-Order Distribution, Journal of Informetrics 1 (2007) 155–160.

[33] W. Li, P. Miramontes, G. Cocho, Fitting Ranked Linguistic Data with Two-Parameter Functions, Entropy 12 (2010) 1743–1764.

[34] M. Ausloos, R. Cerqueti, A Universal Rank-Size Law, PLOS ONE 11 (2016) 1–15.

[35] O. Fontanelli, P. Miramontes, R. Mansilla, G. Cocho, W. Li, Beta rank function: A smooth double-pareto-like distribution, Communications in Statistics - Theory and Methods 51 (2022) 3645–3668.

[36] A. Ghosh, P. Shreya, B. Basu, Maximum entropy framework for a universal rank order distribution with socio-economic applications, Physica A: Statistical Mechanics and its Applications 563 (2021) 125433.

[37] M. Beltrán del Río, G. Cocho, G. Naumis, Universality in the tail of musical note rank distribution, Physica A: Statistical Mechanics and its Applications 387 (2008) 5552–5560.

[38] C. Schunn, D. Wallach, Evaluating Goodness-of-Fit in Comparison of Models to Data, Psychologie der Kognition: Reden and Vorträge anlässlich der Emeritierung von Werner Tack (2005).

[39] R. Cerqueti, M. Ausloos, Cross ranking of cities and regions: population versus income, Journal of Statistical Mechanics: Theory and Experiment 2015 (2015) P07002.

[40] M. H. Kutner, C. J. Nachtsheim, J. Neter, W. Li, Applied Linear Statistical Models, 5th ed., McGraw-Hill, 2004.

[41] F. J. Massey, The Kolmogorov-Smirnov Test for Goodness of Fit, Journal of the American Statistical Association 46 (1951) 68–78.

[42] R. Nuzzo, Scientific method: Statistical errors, Nature 506 (2014) 150–2.

[43] E. Kalliamvakou, G. Gousios, K. Blincoe, L. Singer, D. German, D. Damian, The promises and perils of mining github, Empirical Software Engineering (2015).

[44] A. R. Conn, N. I. M. Gould, Ph. L. Toint, Trust-Region Methods, SIAM, Philadelphia, PA, USA, 2000.

[45] T. Mora, W. Bialek, Are biological systems poised at criticality?, Journal of Statistical Physics 144 (2011) 268–302.