J. Appl. Prob. **57,** 1276–1297 (2020)
*doi:*10.1017/jpr.2020.64

ZIPF'S LAW FOR ATLAS MODELS

RICARDO T. FERNHOLZ,* *Claremont McKenna College*
ROBERT FERNHOLZ,** *Intech Investment Management, LLC*

Abstract

A set of data with positive values follows a *Pareto distribution* if the log–log plot of value versus rank is approximately a straight line. A Pareto distribution satisfies *Zipf's law* if the log–log plot has a slope of -1. Since many types of ranked data follow Zipf's law, it is considered a form of universality. We propose a mathematical explanation for this phenomenon based on Atlas models and first-order models, systems of strictly positive continuous semimartingales with parameters that depend only on rank. We show that the stationary distribution of an Atlas model will follow Zipf's law if and only if two natural conditions, conservation and completeness, are satisfied. Since Atlas models and first-order models can be constructed to approximate systems of time-dependent rank-based data, our results can explain the universality of Zipf's law for such systems. However, ranked data generated by other means may follow non-Zipfian Pareto distributions. Hence, our results explain why Zipf's law holds for word frequency, firm size, household wealth, and city size, while it does not hold for earthquake magnitude, cumulative book sales, and the intensity of wars, all of which follow non-Zipfian Pareto distributions.

Keywords: Zipf's law; Pareto distribution; Atlas model; first-order model

2010 Mathematics Subject Classification: Primary 60H30
Secondary 91B70; 91G70; 91D20

1. Introduction

A set of empirical data with positive values follows a *Pareto distribution* if the log–log plot of the values versus rank is approximately a straight line. Pareto distributions are ubiquitous in the social and natural sciences, appearing in a wide range of fields from geology to economics [3, 34, 38]. A Pareto distribution satisfies *Zipf's law* if the log–log plot has a slope of -1, following Zipf [44], who noticed that the frequency of written words in English follows such a distribution. We shall refer to these distributions as *Zipfian*. Zipf's law is considered a form of universality, since Zipfian distributions occur almost as frequently as Pareto distributions. Nevertheless, according to Tao [41], 'mathematicians do not have a fully satisfactory and convincing explanation for how the law comes about and why it is universal'.

We propose a mathematical explanation of Zipf's law based on *Atlas models* and *first-order models*, systems of strictly positive continuous semimartingales with parameters that depend only on rank. Atlas and first-order models were introduced by Fernholz [14] to model the distribution of capital in stock markets, and a mathematical development of these models can be found in [4], [18], and [29]. Atlas and first-order models can be constructed to approximate

Received 9 May 2019; revision received 22 May 2020.
* Postal address: 500 E. Ninth St., Claremont, CA 91711, USA. Email address: rfernholz@cmc.edu
** Postal address: One Palmer Square, Princeton, NJ 08542, USA. Email address: bob@bobfernholz.com

empirical systems of time-dependent rank-based data that exhibit some form of stability, and while the stationary distributions of Atlas models are Pareto, first-order models can be constructed to have any stationary distribution [14].

Many empirical systems of time-dependent rank-based data generate distributions with log–log plots that are not actually straight lines but rather are concave curves with a tangent of slope -1 at some point along the curve. We shall refer to these more general distributions as *quasi-Zipfian*, and we shall use first-order models to approximate the systems that generate them.

The class of empirical systems for which Zipf's law, or its quasi-Zipfian counterpart, is likely to hold comprises large time-dependent systems for which the number of members can vary over time. Frequency of written words in a language, population of cities, and capitalization of US companies all fall into this class. These systems frequently satisfy two natural conditions, *conservation* and *completeness*. Conservation is like conservation of mass in a physical system, and arises, for example, in measuring the frequency of written words. Since it is impossible to count all the written words in a language, a given number of words must be sampled, and conservation is the result of maintaining a constant sample size over time. Hence, conservation is a natural condition that can be expected to hold for many time-dependent rank-based systems of empirical data.

The second condition, completeness, is related to the replacement of members at the bottom of a rank-based empirical system. In a large rank-based system of time-dependent data those members in the lowest ranks will frequently be replaced by new members from outside the system, and completeness ensures that the effect of this replacement is minimal if the system includes enough ranks. As an example, in Section 4 we show that the distribution of capital in the US stock market follows a complete quasi-Zipfian distribution. However, if this distribution is cut off after the top 100 stocks, the resulting incomplete system is no longer quasi-Zipfian. While it is certainly possible to construct incomplete systems, like the top 100 stocks, most such systems seem to be truncated versions of larger complete systems. Accordingly, conservation and completeness are broadly universal properties of large systems of time-dependent rank-based empirical data.

Mathematically, we show that under the assumptions of conservation and completeness, the stationary distribution of an Atlas model will satisfy Zipf's law. However, most time-dependent rank-based systems do not quite satisfy Zipf's law, and also do not quite satisfy the requirements for Atlas models, so in practice we usually must employ more general first-order models. We refer to these more general models as *quasi-Atlas models*, and we show that under conservation and completeness these models will result in quasi-Zipfian distributions as long as the top-ranked process represents less than half the total mass of the system. Quasi-Atlas models can be used to approximate many large rank-based systems, and since conservation and completeness are common characteristics of such systems, this offers an explanation for the universality of quasi-Zipfian distributions in the natural and social sciences.

The dichotomy between the class of Zipfian and quasi-Zipfian distributions versus the class of non-Zipfian Pareto distributions is of interest to us here. We find that Zipfian and quasi-Zipfian distributions are usually generated by systems of time-dependent rank-based data, and it is this class of systems that we can approximate by Atlas models or first-order models. In contrast, data that follow non-Zipfian Pareto distributions are usually generated by other means, often of a cumulative nature. Examples of time-dependent rank-based systems that generate Zipfian or quasi-Zipfian distributions include the market capitalization of companies [14, 37], the population of cities [21], the employees of firms [2], the income and wealth of

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

households [1, 7], and the assets of banks [20]. From the comprehensive survey of Newman [34] we find an assortment of non-Zipfian Pareto distributions: the magnitude of earthquakes, citations of scientific papers, copies of books sold, the diameter of moon craters, the intensity of solar flares, and the intensity of wars, all of which are cumulative systems. Consider, for example, the magnitude of earthquakes: each new earthquake adds a new observation to the data, but once recorded, these observations do not change over time. Such cumulative systems may generate Pareto distributions, but we have no reason to believe that these distributions will be Zipfian.

The mathematical theory of Atlas and first-order models developed in [4] and [29] is based on a number of earlier results. The existence and uniqueness for solutions of these systems comes from [6] and [40]. The behavior of the 'gap processes', the differences between adjacent rank processes, is based on [23, 24, 25, 43]. The long-term behavior of Atlas and first-order models, including the existence of a stationary distribution and a strong law of large numbers, can be found in [31, 32].

The theory of rank-based systems of continuous semimartingales has been extended in several directions, e.g. infinite Atlas systems [9, 10, 35], behavior at triple points [5], existence and nonexistence of triple points [26, 27, 36], convergence to equilibrium [11, 13, 28], behavior of degenerate systems [16, 17], large deviations [12], and second-order stock market models [15].

In the next sections we first review the properties of Atlas and first-order models, and then characterize Zipfian and quasi-Zipfian systems using these models. We apply our results to the capitalization of US companies, with an analysis of the corresponding quasi-Zipfian distribution curve. We also discuss a number of other time-dependent systems, as well as other approaches that have been used to characterize these systems.

2. Atlas and quasi-Atlas models

We use systems of strictly positive continuous semimartingales $\{X_1, \ldots, X_n\}$, with $n > 1$, to approximate systems of time-dependent data. For such a system we define the *rank function* to be the random permutation $r_t \in \Sigma_n$, for $t \geq 0$, such that $r_t(i) < r_t(j)$ if $X_i(t) > X_j(t)$ or if $X_i(t) = X_j(t)$ and $i < j$. Here, Σ_n is the symmetric group on n elements. The *rank processes* $\{X_{(1)} \geq \cdots \geq X_{(n)}\}$ are defined by $X_{(r_t(i))}(t) = X_i(t)$.

For a continuous semimartingale X, we can define the *semimartingale local time at the origin* Λ by the Tanaka–Meyer formula

$$\Lambda(t) \triangleq \frac{1}{2}\Big(|X(t)| - |X(0)| - \int_0^t \operatorname{sgn}(X(s))\, \mathrm{d}X(s)\Big),$$

for $t \geq 0$, where $\operatorname{sgn}(x) = 2\,\mathbf{1}_{\{x>0\}} - 1$, for $x \in \mathbb{R}$ (see [30, (7.7)–(7.9), p. 220]). The local time Λ measures the amount of time that X spends near 0^+. The mapping $t \mapsto \Lambda(t)$ is continuous and nondecreasing, and induces the random measure $\mathrm{d}\Lambda$ with support contained in the set $\{t \geq 0 : X(t) = 0\}$ (see [30, Theorem 7.1(ii), p. 218]).

We have assumed that the semimartingales X_i are strictly positive, so we can consider the logarithmic processes $\log X_1, \ldots, \log X_n$. For $1 \leq k < \ell \leq n$, let $\Lambda^X_{k,\ell}$ denote the local time at the origin for $\log X_{(k)} - \log X_{(\ell)}$, with $\Lambda^X_{0,1} = \Lambda^X_{n,n+1} \equiv 0$. The processes $\log X_1, \ldots, \log X_n$ have a *triple point* at time $t > 0$ if there exist $j < k < \ell$ such that $\log X_j(t) = \log X_k(t) = \log X_\ell(t)$. Multidimensional Brownian motion almost surely has no triple points (see [30, Proposition 3.22, p. 161]), but some of the systems we consider satisfy only the weaker condition that the processes $\log X_1, \ldots, \log X_n$ *accumulate no local time at triple points*, by which

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

we mean that, for all $\ell \geq k+2$, we have $\Lambda_{k,\ell} \equiv 0$, almost surely (a.s.). If the $\log X_i$ accumulate no local time at triple points, then [5, Theorem 2.5] shows that the rank processes $\log X_{(k)}$ satisfy

$$\mathrm{d}\log X_{(k)}(t) = \sum_{i=1}^{n} \mathbf{1}_{\{r_t(i)=k\}}\, \mathrm{d}\log X_i(t) + \frac{1}{2}\mathrm{d}\Lambda^X_{k,k+1}(t) - \frac{1}{2}\mathrm{d}\Lambda^X_{k-1,k}(t), \quad \text{a.s.,} \tag{2.1}$$

for $t \geq 0$ and $k = 1, \ldots, n$.

Let us define the processes $X_{[k]} \triangleq X_{(1)} + \cdots + X_{(k)}$, for $k = 1, \ldots, n$. The following lemma shows that the local time process $\Lambda^X_{k,k+1}$ measures the flow into and out of $X_{[k]}$.

Lemma 2.1. *Let $X_1, \ldots, X_n$ be strictly positive continuous semimartingales that satisfy (2.1). Then*

$$\frac{1}{2}X_{(k)}(t)\mathrm{d}\Lambda^X_{k,k+1}(t) = \mathrm{d}X_{[k]}(t) - \sum_{i=1}^{n} \mathbf{1}_{\{r_t(i)\leq k\}}\mathrm{d}X_i(t), \quad \text{a.s.,} \tag{2.2}$$

for $t \geq 0$ and $k = 1, \ldots, n$.

Proof. Suppose that the rank processes $X_{(k)}$ satisfy (2.1), so we have

$$\mathrm{d}\log X_{(k)}(t) = \sum_{i=1}^{n} \mathbf{1}_{\{r_t(i)=k\}}\mathrm{d}\log X_i(t) + \frac{1}{2}\mathrm{d}\Lambda^X_{k,k+1}(t) - \frac{1}{2}\mathrm{d}\Lambda^X_{k-1,k}(t), \quad \text{a.s.,}$$

for $t \geq 0$ and $k = 1, \ldots, n$. By Itô's rule this is equivalent to

$$\begin{aligned}\frac{\mathrm{d}X_{(k)}(t)}{X_{(k)}(t)} &= \sum_{i=1}^{n} \mathbf{1}_{\{r_t(i)=k\}}\frac{\mathrm{d}X_i(t)}{X_i(t)} + \frac{1}{2}\mathrm{d}\Lambda^X_{k,k+1}(t) - \frac{1}{2}\mathrm{d}\Lambda^X_{k-1,k}(t) \\ &= \sum_{i=1}^{n} \mathbf{1}_{\{r_t(i)=k\}}\frac{\mathrm{d}X_i(t)}{X_{(k)}(t)} + \frac{1}{2}\mathrm{d}\Lambda^X_{k,k+1}(t) - \frac{1}{2}\mathrm{d}\Lambda^X_{k-1,k}(t), \quad \text{a.s.,}\end{aligned}$$

for $t \geq 0$ and $k = 1, \ldots, n$. From this, we have

$$\begin{aligned}\mathrm{d}X_{(k)}(t) &= \sum_{i=1}^{n} \mathbf{1}_{\{r_t(i)=k\}}\mathrm{d}X_i(t) + \frac{1}{2}X_{(k)}(t)\mathrm{d}\Lambda^X_{k,k+1}(t) - \frac{1}{2}X_{(k)}(t)\mathrm{d}\Lambda^X_{k-1,k}(t) \\ &= \sum_{i=1}^{n} \mathbf{1}_{\{r_t(i)=k\}}\mathrm{d}X_i(t) + \frac{1}{2}X_{(k)}(t)\mathrm{d}\Lambda^X_{k,k+1}(t) - \frac{1}{2}X_{(k-1)}(t)\mathrm{d}\Lambda^X_{k-1,k}(t), \quad \text{a.s.,}\end{aligned}$$

for $t \geq 0$ and $k = 1, \ldots, n$, since the support of $\mathrm{d}\Lambda^X_{k-1,k}$ is contained in the set $\{t : \log X_{(k-1)}(t) = \log X_{(k)}(t)\}$. Now we can add up $\mathrm{d}X_{(1)}(t) + \cdots + \mathrm{d}X_{(k)}(t) = \mathrm{d}X_{[k]}(t)$, and we have

$$\mathrm{d}X_{[k]}(t) = \sum_{i=1}^{n} \mathbf{1}_{\{r_t(i)\leq k\}}\mathrm{d}X_i(t) + \frac{1}{2}X_{(k)}(t)\mathrm{d}\Lambda^X_{k,k+1}(t), \quad \text{a.s.,}$$

for $t \geq 0$ and $k = 1, \ldots, n$, and (2.2) follows. □

The local time process $\Lambda^X_{k,k+1}$ compensates for turnover into and out of the top k ranks. Over time, some of the higher-ranked processes will decrease and exit from the top ranks,

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

while some of the lower-ranked processes will increase and enter those top ranks. Equation (2.2) measures the replacement of the top k ranks of the system by the lower ranks.

We are interested in systems that show stability by rank, at least asymptotically. Since we must apply our definition of stability to systems of empirical data as well as to continuous semimartingales, we use asymptotic time averages rather than expectations for our definitions. We shall show below that for the systems of continuous semimartingales we consider, a law of large numbers implies that the asymptotic time averages are equal to the corresponding expectations.

Definition 2.1. (Fernholz [14]) Let $\{X_1, \ldots, X_n\}$ be a system of strictly positive continuous semimartingales that satisfy (2.1). Then this system is *asymptotically stable* if there exist positive constants $\lambda_{k,k+1}$ and $\sigma^2_{k,k+1}$, $k = 1, \ldots, n-1$, such that

1. $\lim\limits_{t\to\infty} \dfrac{1}{t}\big(\log X_{(1)}(t) - \log X_{(n)}(t)\big) = 0, \quad$ a.s. (*coherence*);
2. $\lim\limits_{t\to\infty} \dfrac{1}{t}\Lambda^X_{k,k+1}(t) = \lambda_{k,k+1}, \quad$ a.s., for $k = 1, \ldots, n-1$;
3. $\lim\limits_{t\to\infty} \dfrac{1}{t}\big\langle \log X_{(k)} - \log X_{(k+1)} \big\rangle_t = \sigma^2_{k,k+1}, \quad$ a.s., for $k = 1, \ldots, n-1$;

where $\langle \cdot \rangle$ represents quadratic variation.

The simplest system we consider is an *Atlas model*, a system of strictly positive continuous semimartingales $\{X_1, \ldots, X_n\}$ defined by

$$\mathrm{d}\log X_i(t) = -g\,\mathrm{d}t + ng\mathbf{1}_{\{r_t(i)=n\}}\mathrm{d}t + \sigma\,\mathrm{d}W_i(t), \tag{2.3}$$

for $t \geq 0$ and $i = 1, \ldots, n$, where $g > 0$ and $\sigma > 0$ are constants, and $(W_1, \ldots, W_n)$ is a Brownian motion (see [14, Example 5.3.3, p. 103]). Atlas models are asymptotically stable with parameters

$$\lambda_{k,k+1} = 2kg, \qquad \sigma^2_{k,k+1} = 2\sigma^2, \tag{2.4}$$

for $k = 1, \ldots, n-1$ (see [29, Proposition 2]).

A modest generalization of the Atlas model is the first-order model, introduced in [14, Section 5.5]. A *first-order model* is a system of strictly positive continuous semimartingales $\{X_1, \ldots, X_n\}$ with

$$\mathrm{d}\log X_i(t) = g_{r_t(i)}\,\mathrm{d}t + G_n\mathbf{1}_{\{r_t(i)=n\}}\mathrm{d}t + \sigma_{r_t(i)}\,\mathrm{d}W_i(t), \tag{2.5}$$

for $t \geq 0$ and $i = 1, \ldots, n$, where $\sigma^2_1, \ldots, \sigma^2_n$ are positive constants; $g_1, \ldots, g_n$ are constants that satisfy

$$g_1 + \cdots + g_n \leq 0 \quad \text{and} \quad g_1 + \cdots + g_k < 0 \text{ for } k < n; \tag{2.6}$$

$G_n = -(g_1 + \cdots + g_n)$; and $(W_1, \ldots, W_n)$ is a Brownian motion (see [4, (1.1)–(1.6)]). First-order models are asymptotically stable with parameters

$$\lambda_{k,k+1} = -2\big(g_1 + \cdots + g_k\big), \qquad \sigma^2_{k,k+1} = \sigma^2_k + \sigma^2_{k+1}, \qquad \text{a.s.}, \tag{2.7}$$

for $k = 1, \ldots, n-1$ (see [29, Proposition 2]). Here we use a simple form of first-order model in which the drift parameters g_k are constant and the variance parameters σ^2_k grow linearly with

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

rank. Accordingly, we define a *quasi-Atlas model* to be a first-order model determined by three parameters $g > 0$ and $\sigma_2^2 \geq \sigma_1^2 > 0$, such that

$$g_k = -g, \qquad \sigma_k^2 = \sigma_1^2 + (k-1)(\sigma_2^2 - \sigma_1^2), \tag{2.8}$$

for $k = 1, \ldots, n$. Hence, we see that Atlas models are a subclass of quasi-Atlas models, which in turn are a subclass of first-order models.

By [4, Proposition 2.3], each of the processes X_i in a first-order model asymptotically spends equal time in each rank. Due to this ergodicity, the parameters ng in (2.3) and G_n in (2.5) cause the asymptotic growth rate to be zero for each of the processes $\log X_i$, for $i = 1, \ldots, n$. Equations (2.3) and (2.5) can be generalized by the addition of a term $\gamma\, \mathrm{d}t$ on the right-hand side, where the constant γ represents the common logarithmic growth rate of the system, but in our setting it is convenient to make the simplifying assumption that $\gamma = 0$ (see, e.g., [4, (1.1) and (1.6)]). The condition (2.6), along with $G_n = -(g_1 + \cdots + g_n)$, stabilizes the system and prevents it from separating into smaller subsystems over time. A discussion of this stabilizing effect can be found in the Remark following Theorem 8 of [35].

We see from (2.7) that for a first-order model the parameters $\lambda_{k,k+1}$ and $\sigma_{k,k+1}^2$ depend only on ranks 1 through $k+1$ and not on the number n of processes in the model. On a more intuitive level, the parameter G_n is defined so that whatever the size n of the model, the 'upward force' $g_{k+1} + \cdots + g_n + G_n > 0$ from below adjusts to counteract the 'downward force' $g_1 + \cdots + g_k < 0$ from above, with

$$g_{k+1} + \cdots + g_n + G_n = -(g_1 + \cdots + g_k).$$

The local time $\Lambda_{k,k+1}$ between ranks k and $k+1$ is determined by these upward and downward forces since they push these two ranks together, and the value of $\lambda_{k,k+1}$ depends on this local time.

Lemma 1 in [29] shows that the processes $\log X_1, \ldots, \log X_n$ in a first-order model accumulate no local time at triple points. It is also known that a first-order model for which $k \mapsto \sigma_k^2$ is *concave,* i.e. for which $\sigma_{k+1}^2 - \sigma_k^2 \leq \sigma_k^2 - \sigma_{k-1}^2$, for $k = 2, \ldots, n-1$, almost surely has no triple points, and this condition holds for Atlas and quasi-Atlas models [27, 36]. Hence, (2.1) and Lemma 2.1 are valid for Atlas and quasi-Atlas models.

For a first-order model $\{X_1, \ldots, X_n\}$, let us define the processes $\mathcal{X}_1, \ldots, \mathcal{X}_n$ by

$$\mathcal{X}_i(t) = \log X_i(t) - \frac{1}{n} \sum_{j=1}^{n} \log X_j(t), \qquad t \in [0, \infty),$$

for $i = 1, \ldots, n$, along with the corresponding ranked processes $\mathcal{X}_{(1)} \geq \cdots \geq \mathcal{X}_{(n)}$, with

$$\mathcal{X}_{(k)}(t) = \log X_{(k)}(t) - \frac{1}{n} \sum_{j=1}^{n} \log X_j(t), \qquad t \in [0, \infty),$$

for $k = 1, \ldots, n$. Then it follows from [29, Proposition 1], [31, Theorems 3.1 and 3.2], or [32, Theorem 4.1], that $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$, as a process with values in $\mathbb{R}^n$, has a unique stationary distribution. We define the *gap processes* by $\log X_{(k)} - \log X_{(k+1)}$, for $k = 1, \ldots, n-1$, and the stationary distribution for $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ induces a stationary distribution for each gap process $\log X_{(k)} - \log X_{(k+1)} = \mathcal{X}_{(k)} - \mathcal{X}_{(k+1)}$ (see [29, Corollary 2]).

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

For a first-order model $\{X_1, \ldots, X_n\}$, let ξ_k represent the gap process $\log X_{(k)} - \log X_{(k+1)}$ in its stationary distribution, for $k = 1, \ldots, n-1$. For an Atlas or quasi-Atlas model, the ξ_k will be independent and exponentially distributed, so the stationary joint distribution of $(\xi_1, \ldots, \xi_{n-1})$ will be the product of the exponential marginal distributions (this follows from [24, Theorem 9.2] and is a special case of [29, Theorem 2]). It is also known that in this case ξ_k has density function $\alpha_k e^{-\alpha_k x}$, for $x \in [0, \infty)$, with rate parameter

$$\alpha_k = \frac{2\lambda_{k,k+1}}{\sigma^2_{k,k+1}} \tag{2.9}$$

and expectation

$$\mathbb{E}\big[\xi_k\big] = \frac{1}{\alpha_k} = \frac{\sigma^2_{k,k+1}}{2\lambda_{k,k+1}}$$

(see [29, Theorem 2]). For $k = 1, \ldots, n-1$, if $f : [0, \infty) \to \mathbb{R}$ is a measurable function with

$$\int_0^\infty |f(x)| e^{-\alpha_k x}\, dx < \infty,$$

then the strong law of large numbers,

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T f\big(\log X_{(k)}(t) - \log X_{(k+1)}(t)\big) dt = \mathbb{E}\big[f(\xi_k)\big], \qquad \text{a.s.},$$

holds (see [29, Proposition 1], [31, Theorem 3.1], or [32, Theorem 5.1]). It follows from this that

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \big(\log X_{(k)}(t) - \log X_{(k+1)}(t)\big)\, dt = \mathbb{E}\big[\xi_k\big] = \frac{1}{\alpha_k} = \frac{\sigma^2_{k,k+1}}{2\lambda_{k,k+1}}, \quad \text{a.s.}, \tag{2.10}$$

and

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{X_{(k+1)}(t)}{X_{(k)}(t)} dt = \mathbb{E}\big[e^{-\xi_k}\big] = \frac{\alpha_k}{\alpha_k + 1}, \qquad \text{a.s.}, \tag{2.11}$$

for $k = 1, \ldots, n-1$ (see [29, Theorem 1]).

For a first-order model $\{X_1, \ldots, X_n\}$, the asymptotic slope of the tangent to the log–log plot of the $X_{(k)}$ versus rank will be

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{\log X_{(k)}(t) - \log X_{(k+1)}(t)}{\log(k) - \log(k+1)}\, dt \tag{2.12}$$

at rank k, so if we define the *slope parameters* s_k by

$$s_k \triangleq k \lim_{T\to\infty} \frac{1}{T} \int_0^T \big(\log X_{(k)}(t) - \log X_{(k+1)}(t)\big)\, dt, \tag{2.13}$$

for $k = 1, \ldots, n-1$, then

$$-s_k \left(1 + \frac{1}{2k}\right) < \lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{\log X_{(k)}(t) - \log X_{(k+1)}(t)}{\log(k) - \log(k+1)}\, dt < -s_k, \tag{2.14}$$

for $k = 1, \ldots, n-1$. Accordingly, for large enough k, the slope parameter s_k will be approximately equal to minus the slope given in (2.12). For expositional simplicity, we treat the s_k as if

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

they measured the true log–log slopes between adjacent ranks, but it is important to remember that this equivalence is only as accurate as the range of the inequalities in (2.14).

For an Atlas model, it follows from (2.4), (2.10), and (2.13) that

$$s_k = \frac{\sigma^2}{2g}, \qquad \text{a.s.}, \tag{2.15}$$

for $k = 1, \dots, n-1$, so the stationary distribution of an Atlas model follows a Pareto distribution, at least within the approximation (2.14), and when $\sigma^2 = 2g$ it follows Zipf's law. For a quasi-Atlas model, we see from (2.7) and (2.10) that the slope parameters will be

$$s_k = \frac{k\big(\sigma_k^2 + \sigma_{k+1}^2\big)}{2\lambda_{k,k+1}} = \frac{\sigma_k^2 + \sigma_{k+1}^2}{4g}, \qquad \text{a.s.}, \tag{2.16}$$

for $k = 1, \dots, n-1$, so the stationary distributions of quasi-Atlas models are not confined to the class of Pareto distributions.

It is convenient to consider families of first-order models that share the same parameters, and for this purpose we define a *first-order family* to be a sequence of constants $\{g_k, \sigma_k^2\}_{k\in\mathbb{N}}$ with $g_1 + \cdots + g_k < 0$ and $\sigma_k^2 > 0$, for $k \in \mathbb{N}$. A first-order family generates a class of first-order models $\{X_1, \dots, X_n\}$, each defined as in (2.5) with the common parameters g_k and σ_k^2, for $k \in \mathbb{N}$, with $G_n = -(g_1 + \cdots + g_n)$, for $n \in \mathbb{N}$. An *Atlas family* is a first-order family with $g_k = -g < 0$ and $\sigma_k^2 = \sigma^2 > 0$, for $k \in \mathbb{N}$. A *quasi-Atlas family* is a first-order family with $g_k = -g < 0$ and $\sigma_k^2 = \sigma_1^2 + (k-1)(\sigma_2^2 - \sigma_1^2) > 0$, for $k \in \mathbb{N}$.

For a first-order family $\{g_k, \sigma_k^2\}_{k\in\mathbb{N}}$ we shall use the notation $\mathbb{E}_n$ to denote the expectation with respect to the stationary distribution for the system $\{\log(X_{(1)}/X_{(2)}), \dots, \log(X_{(n-1)}/X_{(n)})\}$ defined by that family. For Atlas and quasi-Atlas models it is useful to measure the expected values of the ranked processes $X_{(k)}$ relative to the value of the top process $X_{(1)}$, so we define the *ranked weight ratios*

$$R_k \triangleq \mathbb{E}_n\left[\frac{X_{(k)}(t)}{X_{(1)}(t)}\right], \tag{2.17}$$

for $k = 1, \dots, n$ and $t \geq 0$. Since $\mathbb{E}_n$ assumes the stationary distribution, and since the definition does not depend on weights below the kth rank, the ranked weight ratios are independent of both t and n. With the system in its stationary distribution, the random variables $\log(X_{(k)}(t)/X_{(k+1)}(t))$ are independent, so

$$R_k = \mathbb{E}_n\left[\frac{X_{(k)}(t)}{X_{(k-1)}(t)}\right] \cdot \mathbb{E}_n\left[\frac{X_{(k-1)}(t)}{X_{(k-2)}(t)}\right] \cdots \mathbb{E}_n\left[\frac{X_{(2)}(t)}{X_{(1)}(t)}\right], \tag{2.18}$$

for $2 \leq k \leq n$ and $t \geq 0$, where the terms on the right-hand side can be calculated in terms of (2.11). We can also define, for $n \in \mathbb{N}$,

$$R_{[n]} \triangleq \mathbb{E}_n\left[\frac{X_{[n]}(t)}{X_{(1)}(t)}\right] = R_1 + \cdots + R_n, \tag{2.19}$$

for $t \geq 0$.

For an Atlas or quasi-Atlas family, the parameters $\sigma_{k,k+1}^2$, $\lambda_{k,k+1}$, s_k, and R_k are defined uniquely for $k \in \mathbb{N}$ by (2.4), (2.8), (2.15), (2.16), and (2.17), as the case may be. Let us note

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

that for a quasi-Atlas family the slope parameters s_k and ranked weight ratios R_k do not depend on the number of processes in the model as long as $n > k$, so a quasi-Atlas family defines a unique asymptotic distribution curve. Accordingly, these families will allow us to derive results about asymptotic distribution curves without repeatedly reciting the characteristics of individual models. Moreover, we only consider values derived from a first-order family when the models in the family are in their stationary distribution. Hence, for the Atlas and quasi-Atlas families we consider, we can calculate the values of the s_k and R_k directly from the parameters g, σ_1^2, and σ_2^2, and we can ignore the models themselves.

3. Zipfian Atlas models as approximations of empirical systems

In this section we first consider how empirical systems of time-dependent data can be approximated by first-order models. In the case that these first-order approximations are in fact Atlas or quasi-Atlas models, we show that it is likely that the empirical systems will follow Zipfian or quasi-Zipfian distributions.

Suppose that $\{Y_1, \dots, Y_n\}$, for $n > 1$, is an asymptotically stable system of strictly positive continuous semimartingales with rank function $\rho_t \in \Sigma_n$, for $t \geq 0$, such that $\rho_t(i) < \rho_t(j)$ if $Y_i(t) > Y_j(t)$ or if $Y_i(t) = Y_j(t)$ and $i < j$. Let $\{Y_{(1)} \geq \cdots \geq Y_{(n)}\}$ be the corresponding rank processes with $Y_{(\rho_t(i))}(t) = Y_i(t)$. As in Definition 2.1, for the processes $Y_1 \dots, Y_n$ we can define the parameters

$$\begin{aligned} \boldsymbol{\lambda}_{k,k+1} &\triangleq \lim_{t\to\infty} \frac{1}{t}\Lambda^Y_{k,k+1}(t) > 0, \qquad \text{a.s.,} \\ \boldsymbol{\sigma}^2_{k,k+1} &\triangleq \lim_{t\to\infty} \frac{1}{t}\big\langle \log Y_{(k)} - \log Y_{(k+1)}\big\rangle_t > 0, \qquad \text{a.s.,} \end{aligned} \tag{3.1}$$

for $k = 1, \dots, n-1$.

Definition 3.1. (Fernholz [14]) Let $\{Y_1, \dots, Y_n\}$ be an asymptotically stable system of strictly positive continuous semimartingales with parameters $\boldsymbol{\lambda}_{k,k+1}$ and $\boldsymbol{\sigma}^2_{k,k+1}$, for $k = 1, \dots, n-1$, defined by (3.1). Then the *first-order approximation* of $\{Y_1, \dots, Y_n\}$ is the first-order model $\{X_1, \dots, X_n\}$ with

$$\mathrm{d}\log X_i(t) = g_{r_t(i)}\,\mathrm{d}t + G_n \mathbf{1}_{\{r_t(i)=n\}}\mathrm{d}t + \sigma_{r_t(i)}\,\mathrm{d}W_i(t), \tag{3.2}$$

for $t \geq 0$ and $i = 1, \dots, n$, where $r_t \in \Sigma_n$ is the rank function for the X_i, the parameters g_k and σ_k are defined by

$$\begin{aligned} g_k &= \frac{1}{2}\boldsymbol{\lambda}_{k-1,k} - \frac{1}{2}\boldsymbol{\lambda}_{k,k+1} \qquad \text{for } k = 2, \dots, n-1; \\ g_1 &= -\frac{1}{2}\boldsymbol{\lambda}_{1,2}, \qquad g_n = g_{n-1} \wedge 0; \\ \sigma_k^2 &= \frac{1}{4}\big(\boldsymbol{\sigma}^2_{k-1,k} + \boldsymbol{\sigma}^2_{k,k+1}\big) \qquad \text{for } k = 2, \dots, n-1; \\ \sigma_1^2 &= \sigma_2^2 + \big(\sigma_2^2 - \sigma_3^2\big)\mathbf{1}_{\{2\sigma_2^2 > \sigma_3^2\}}, \qquad \sigma_n^2 = \sigma_{n-1}^2 + \big(\sigma_{n-1}^2 - \sigma_{n-2}^2\big) \vee 0; \end{aligned} \tag{3.3}$$

where σ_k is the positive square root of σ_k^2, $G_n = -(g_1 + \cdots + g_n)$, and $(W_1, \dots, W_n)$ is a Brownian motion.

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

The parameters g_1, g_n, σ_1^2, and σ_n^2 in (3.3) were chosen to preserve the structure of Atlas and quasi-Atlas models. For the first-order model (3.2) with parameters (3.3), equation (2.7) implies that

$$\lambda_{k,k+1} = -2\big(g_1 + \cdots + g_k\big) = \boldsymbol{\lambda}_{k,k+1}, \qquad \text{a.s.}, \tag{3.4}$$

for $k = 1, \ldots, n-1$, and

$$\sigma_{k,k+1}^2 = \sigma_k^2 + \sigma_{k+1}^2 = \frac{1}{4}\big(\boldsymbol{\sigma}_{k-1,k}^2 + 2\boldsymbol{\sigma}_{k,k+1}^2 + \boldsymbol{\sigma}_{k+1,k+2}^2\big), \qquad \text{a.s.},$$

for $k = 2, \ldots, n-2$, so the $\sigma_{k,k+1}^2$ are a smoothed version of the $\boldsymbol{\sigma}_{k,k+1}^2$. Hence, the parameters for a first-order approximation are similar to those of the asymptotically stable system that it approximates. We would also like to have the stable distributions of the two systems $\{\log(X_{(1)}/X_{(2)}), \ldots, \log(X_{(n-1)}/X_{(n)})\}$ and $\{\log(Y_{(1)}/Y_{(2)}), \ldots, \log(Y_{(n-1)}/Y_{(n)})\}$ be similar, with

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \big(\log X_{(k)}(t) - \log X_{(k+1)}(t)\big)\,\mathrm{d}t \cong \lim_{T\to\infty} \frac{1}{T}\int_0^T \big(\log Y_{(k)}(t) - \log Y_{(k+1)}(t)\big)\,\mathrm{d}t, \qquad \text{a.s.},$$

for $k = 1, \ldots, n-1$. From (3.3) and (3.4) we see that if the system $\{Y_1, \ldots, Y_n\}$ is a quasi-Atlas model with parameters $\boldsymbol{g}_k$ and $\boldsymbol{\sigma}_k^2$, then the first-order approximation $\{X_1, \ldots, X_n\}$ will also be a quasi-Atlas model with the same parameters $g_k = \boldsymbol{g}_k$ and $\sigma_k^2 = \boldsymbol{\sigma}_k^2$, for $k = 1, \ldots, n$. In this case it follows from (2.10) that

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \big(\log X_{(k)}(t) - \log X_{(k+1)}(t)\big)\,\mathrm{d}t = \lim_{T\to\infty} \frac{1}{T}\int_0^T \big(\log Y_{(k)}(t) - \log Y_{(k+1)}(t)\big)\,\mathrm{d}t, \qquad \text{a.s.},$$

for $k = 1, \ldots, n-1$, so the stable distributions of the two systems will be the same.

Lemma 2.1 shows that the parameters $\boldsymbol{\lambda}_{k,k+1}$ can be expressed as

$$\boldsymbol{\lambda}_{k,k+1} = \lim_{T\to\infty} \frac{2}{T}\int_0^T \left(\frac{\mathrm{d}Y_{[k]}(t)}{Y_{(k)}(t)} - \sum_{i=1}^n \mathbf{1}_{\{\rho_t(i)\le k\}}\frac{\mathrm{d}Y_i(t)}{Y_{(k)}(t)}\right), \qquad \text{a.s.}, \tag{3.5}$$

for $k = 1, \ldots, n-1$, in which all the terms on the right-hand side of the equation are observable. In a similar fashion we can write

$$\boldsymbol{\sigma}_{k,k+1}^2 = \lim_{T\to\infty} \frac{1}{T}\int_0^T \mathrm{d}\big\langle \log Y_{(k)} - \log Y_{(k+1)}\big\rangle_t, \qquad \text{a.s.}, \tag{3.6}$$

for $k = 1, \ldots, n-1$. These two equations will allow us to define parameters equivalent to $\boldsymbol{\lambda}_{k,k+1}$ and $\boldsymbol{\sigma}_{k,k+1}^2$ for time-dependent systems of empirical data.

Suppose now that we have a time-dependent system $\{Z_1(\tau), Z_2(\tau), \ldots\}$ of positive-valued data observed at times $\tau \in \{1, 2, \ldots, T\}$, where $T > 1$. Let

$$N_\tau = \#\{Z_1(\tau), Z_2(\tau), \ldots\} \qquad \text{and} \qquad N = N_1 \wedge \cdots \wedge N_T, \tag{3.7}$$

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

where # represents cardinality. Let $\rho_\tau : \mathbb{N} \to \mathbb{N}$ be the rank function for the system $\{Z_1(\tau), Z_2(\tau), \ldots\}$ such that ρ_τ restricted to the subset $\{1, \ldots, N_\tau\}$ is the permutation with $\rho_\tau(i) < \rho_\tau(j)$ if $Z_i(\tau) > Z_j(\tau)$ or if $Z_i(\tau) = Z_j(\tau)$ and $i < j$, and for $i > N_\tau$, $\rho_\tau(i) = i$. We define the ranked values $\{Z_{(1)}(\tau) \geq Z_{(2)}(\tau) \geq \cdots\}$ such that $Z_{(\rho_\tau(i))}(\tau) = Z_i(\tau)$ for $i \leq N_\tau$, and for definiteness we can let $Z_{(k)}(\tau) = 0$ for $k > N_\tau$. With these definitions, we have $Z_{[k]}(\tau) = Z_{(1)}(\tau) + \cdots + Z_{(k)}(\tau)$, for $k = 1, \ldots, N$ and $\tau \in \{1, 2, \ldots, T\}$.

We can mimic the time averages (3.5) and (3.6) to define the parameters

$$\boldsymbol{\lambda}_{k,k+1} \triangleq \frac{2}{T-1} \sum_{\tau=1}^{T-1} \left(\frac{Z_{[k]}(\tau+1) - Z_{[k]}(\tau)}{Z_{(k)}(\tau)} - \sum_{i=1}^{N} \mathbf{1}_{\{\rho_\tau(i) \leq k\}} \frac{Z_i(\tau+1) - Z_i(\tau)}{Z_{(k)}(\tau)} \right) \tag{3.8}$$

and

$$\boldsymbol{\sigma}^2_{k,k+1} \triangleq \frac{1}{T-1} \sum_{\tau=1}^{T-1} \Big(\big(\log Z_{(k)}(\tau+1) - \log Z_{(k+1)}(\tau+1) \big) - \big(\log Z_{(k)}(\tau) - \log Z_{(k+1)}(\tau) \big) \Big)^2, \tag{3.9}$$

for $k = 1, \ldots, N-1$.

Definition 3.2. Suppose that $\{Z_1(\tau), Z_2(\tau), \ldots\}$, for $\tau \in \{1, 2, \ldots, T\}$, with $T > 1$, is a time-dependent system of positive-valued data with N, $\boldsymbol{\lambda}_{k,k+1}$, and $\boldsymbol{\sigma}^2_{k,k+1}$ defined as in (3.7), (3.8), and (3.9). The *first-order approximation* of $\{Z_1(\tau), Z_2(\tau), \ldots\}$ is the first-order family $\{g_k, \sigma_k^2\}_{k \in \mathbb{N}}$ with

$$\begin{aligned} &g_k = \frac{1}{2}\boldsymbol{\lambda}_{k-1,k} - \frac{1}{2}\boldsymbol{\lambda}_{k,k+1} \qquad \text{for } k = 2, \ldots, N-1; \\ &g_1 = -\frac{1}{2}\boldsymbol{\lambda}_{1,2}, \qquad g_k = g_{k-1} \wedge 0 \text{ for } k \geq N; \\ &\sigma_k^2 = \frac{1}{4}\big(\boldsymbol{\sigma}^2_{k-1,k} + \boldsymbol{\sigma}^2_{k,k+1}\big) \qquad \text{for } k = 2, \ldots, N-1; \\ &\sigma_1^2 = \sigma_2^2 + \big(\sigma_2^2 - \sigma_3^2\big)\mathbf{1}_{\{2\sigma_2^2 > \sigma_3^2\}}, \quad \sigma_k^2 = \sigma_{k-1}^2 + \big(\sigma_{k-1}^2 - \sigma_{k-2}^2\big) \vee 0 \text{ for } k \geq N. \end{aligned} \tag{3.10}$$

If the first-order model $\{X_1, \ldots, X_N\}$ defined by (3.2) with parameters (3.11) satisfies

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T \big(\log X_{(k)}(t) - \log X_{(k+1)}(t) \big)\, \mathrm{d}t \cong \frac{1}{T} \sum_{\tau=1}^{T} \big(\log Z_{(k)}(\tau) - \log Z_{(k+1)}(\tau) \big), \tag{3.11}$$

for $k = 1, \ldots, N-1$, then we say that the system $\{Z_1(\tau), Z_2(\tau), \ldots\}$ is *rank-based*. If the system $\{Z_1(\tau), Z_2(\tau), \ldots\}$ is rank-based and the first-order model $\{X_1, \ldots, X_N\}$ defined by (3.2) with parameters (3.10) is a quasi-Atlas model, then it follows from (2.10) and (3.11) that

$$\frac{1}{T} \sum_{\tau=1}^{T} \big(\log Z_{(k)}(\tau) - \log Z_{(k+1)}(\tau) \big) \cong \frac{\boldsymbol{\sigma}^2_{k,k+1}}{2\boldsymbol{\lambda}_{k,k+1}}, \tag{3.12}$$

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

for $k = 1, \ldots, N-1$. In this case, the slope parameters for the first-order approximation apply to the distribution curve for the empirical system $\{Z_1(\tau), Z_2(\tau), \ldots\}$, and this motivates the next two definitions.

Definition 3.3. A first-order family is *Zipfian* if its slope parameters $s_k = 1$, for $k \in \mathbb{N}$. A time-dependent rank-based system is *Zipfian* if its first-order approximation is Zipfian.

We see that, in terms of the parameters g and σ^2, an Atlas family is Zipfian if and only if $\sigma^2 = 2g$, in which case $\alpha_k = k$ in (2.9) and

$$R_k = \frac{k-1}{k} \cdot \frac{k-2}{k-1} \cdots \frac{1}{2} = \frac{1}{k}, \tag{3.13}$$

as in (2.11) and (2.18). Since many empirical distributions are not Zipfian but rather quasi-Zipfian, we need to formalize this concept for first-order families.

Definition 3.4. A first-order family is *quasi-Zipfian* if its slope parameters s_k are nondecreasing with $s_1 \leq 1$ and

$$\lim_{k \to \infty} s_k \geq 1,$$

where this limit includes divergence to infinity. A time-dependent rank-based system is *quasi-Zipfian* if its first-order approximation is quasi-Zipfian.

For a quasi-Atlas family that is not an Atlas family, we see that in terms of the parameters g, σ_1^2, and σ_2^2 of (2.8), the family is quasi-Zipfian if and only if $\sigma_1^2 + \sigma_2^2 \leq 4g$.

By these definitions, a Zipfian system is also quasi-Zipfian. Because the slope parameters s_k are approximately equal to minus the slope of a log–log plot of size versus rank, Definition 3.4 implies that a time-dependent rank-based system will be quasi-Zipfian if this log–log plot of its first-order approximation is concave with slope not steeper than -1 at the highest ranks and not flatter than -1 at the lowest ranks.

Zipf's law originally referred to the frequency of words in a written language [44], with the system $\{Z_1(\tau), Z_2(\tau), \ldots\}$, where $Z_i(\tau)$ represents the number of occurrences of the ith word in a language at time τ. To measure the relative frequency of written words in a language it is not possible to observe all the written words in that language. Instead, the words must be *sampled*, where a random sample is selected (without replacement), and the frequency versus rank of this random sample is studied. For example, in Wikipedia [42] 10 million words in each of 30 languages were sampled and the resulting distribution curves were created. If the sample is large enough, the distribution of the sampled data should not differ materially from the distribution of the entire data set, at least for the higher ranks.

An advantage that arises from using sampled data is that it is possible to keep the total number of data in the sample constant over time. The total number of written words that appear in a language is likely to increase over time, and this increase could bias estimates of some parameters. Sampling the data will remove such a trend from the data, since a constant number of words can be sampled at each time. Accordingly, in all cases we shall assume that global trends have been removed from the data, either by sampling or by some other means of detrending.

Since we have assumed that we have a constant sample size or that the data have been detrended, the total count of our sampled data will remain constant, so

$$Z_1(\tau) + Z_2(\tau) + \cdots = \text{ constant}, \tag{3.14}$$

for $\tau \in \{1, 2, \ldots, T\}$, where in the case of the Wikipedia words the constant would be 10 million.

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

Suppose we have a time-dependent system of positive-valued data $\{Z_1(\tau), Z_2(\tau), \ldots\}$, for $\tau \in \{1, 2, \ldots, T\}$ with $T > 1$, and we observe the top n ranks, for $1 < n < N$, with N from (3.7), along with $Z_{[n]}(\tau) = Z_{(1)}(\tau) + \cdots + Z_{(n)}(\tau)$. Since the total value of the sampled data in (3.14) is constant, for large enough n it is reasonable to expect the relative change of the top n ranks to satisfy

$$\frac{Z_{[n]}(\tau + 1) - Z_{[n]}(\tau)}{Z_{[n]}(\tau)} \cong 0, \tag{3.15}$$

for $\tau \in \{1, 2, \ldots, T - 1\}$ as n becomes large, at least on average over time. This condition is essentially a 'conservation of mass' criterion for $\{Z_1(\tau), Z_2(\tau), \ldots\}$, in which the total 'mass' (3.14) of the system remains constant, at least on average over time. It is useful to normalize the values $Z_{(k)}(\tau)$ and $Z_{[n]}(\tau)$ by measuring them relative to the largest value $Z_{(1)}(\tau)$, in which case (3.15) becomes

$$\frac{1}{\big(Z_{[n]}(\tau)/Z_{(1)}(\tau)\big)} \frac{Z_{[n]}(\tau + 1) - Z_{[n]}(\tau)}{Z_{(1)}(\tau)} \cong 0,$$

for $\tau \in \{1, 2, \ldots, T - 1\}$ as n becomes large, at least on average over time. For the first-order family $\{g_k, \sigma_k^2\}_{k \in \mathbb{N}}$, this expression allows us to use the ranked weight ratios R_k and $R_{[n]}$ of (2.17) and (2.19), and motivates the following definition.

Definition 3.5. The first-order family $\{g_k, \sigma_k^2\}_{k \in \mathbb{N}}$ is *conservative* if, for $T > 0$,

$$\lim_{n \to \infty} \frac{1}{R_{[n]}} \mathbb{E}_n \left[\frac{1}{T} \int_0^T \frac{\mathrm{d}X_{[n]}(t)}{X_{(1)}(t)} \right] = 0.$$

For the system $\{Z_1(\tau), Z_2(\tau), \ldots\}$, for $\tau \in \{1, 2, \ldots, T\}$, the replacement of processes in the top $n < N$ ranks by processes in the lower ranks over the time interval $[\tau, \tau + 1]$ is measured by

$$Z_{[n]}(\tau + 1) - \sum_{i=1}^{N} \mathbf{1}_{\{\rho_\tau(i) \le n\}} Z_i(\tau + 1),$$

or

$$\big(Z_{[n]}(\tau + 1) - Z_{[n]}(\tau)\big) - \left(\sum_{i=1}^{N} \mathbf{1}_{\{\rho_\tau(i) \le n\}} \big(Z_i(\tau + 1) - Z_i(\tau)\big) \right).$$

While some replacement from lower ranks is necessary, it seems reasonable to expect that the system will be 'complete' in the sense that, on average, the relative proportion of the mass that is replaced becomes arbitrarily small for large enough n, i.e. that

$$\frac{1}{Z_{[n]}(\tau)} \left(Z_{[n]}(\tau + 1) - Z_{[n]}(\tau) - \sum_{i=1}^{N} \mathbf{1}_{\{\rho_\tau(i) \le n\}} \big(Z_i(\tau + 1) - Z_i(\tau)\big) \right) \cong 0,$$

for $\tau \in \{1, 2, \ldots, T - 1\}$ and large enough n. As in Definition 3.5, in terms of the first-order approximation of $\{Z_1(\tau), Z_2(\tau), \ldots\}$, this becomes

$$\frac{1}{R_{[n]}} \mathbb{E}_n \left[\frac{1}{T} \int_0^T \frac{\mathrm{d}X_{[n]}(t)}{X_{(1)}(t)} - \frac{1}{T} \int_0^T \left(\sum_{i=1}^{N} \mathbf{1}_{\{r_t(i) \le n\}} \frac{\mathrm{d}X_i(t)}{X_{(1)}(t)} \right) \right] \cong 0, \tag{3.16}$$

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

for $T > 0$ and large enough n, where $N > n$ and $\{X_1, \dots, X_N\}$ is a first-order model defined by $\{g_k, \sigma_k^2\}_{k\in\mathbb{N}}$. By Lemma 2.1, this is equivalent to

$$\frac{1}{R_{[n]}}\mathbb{E}_n\left[\frac{1}{T}\int_0^T \frac{X_{(n)}(t)}{2X_{(1)}(t)}\mathrm{d}\Lambda^X_{n,n+1}(t)\right] \cong 0,$$

for $T > 0$ and large enough n. Since

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathrm{d}\Lambda^X_{n,n+1}(t) = \lambda_{n,n+1} = -2\big(g_1 + \cdots + g_n\big), \qquad \text{a.s.},$$

condition (3.16) corresponds to

$$\frac{1}{R_{[n]}}\mathbb{E}_n\left[\frac{1}{T}\int_0^T -\big(g_1 + \cdots + g_n\big)\frac{X_{(n)}(t)}{X_{(1)}(t)}\mathrm{d}t\right] \cong 0,$$

for $T > 0$ and large enough n. Since $\mathbb{E}_n$ assumes the stationary distribution, this is equivalent to

$$-\big(g_1 + \cdots + g_n\big)\frac{R_n}{R_{[n]}} \cong 0,$$

for large enough n, and with $G_n = -\big(g_1 + \cdots + g_n\big)$, we have the following definition.

Definition 3.6. The first-order family $\{g_k, \sigma_k^2\}_{k\in\mathbb{N}}$ is *complete* if

$$\lim_{n\to\infty}\frac{G_n R_n}{R_{[n]}} = 0.$$

For an Atlas or quasi-Atlas family $G_n = ng$, so for these families completeness is equivalent to

$$\lim_{n\to\infty}\frac{nR_n}{R_{[n]}} = 0.$$

The following two propositions show that conservation and completeness are the basis for the Zipfian nature of the distributions of many systems of time-dependent rank-based data.

Proposition 3.1. *An Atlas family is Zipfian if and only if it is conservative and complete.*

Proof. For an Atlas model $\{X_1, \dots, X_n\}$ with parameters $g > 0$ and $\sigma > 0$, Itô's rule implies that

$$\mathrm{d}X_i(t) = \left(\frac{\sigma^2}{2} - g + ng\mathbf{1}_{\{r_t(i)=n\}}\right)X_i(t)\,\mathrm{d}t + \sigma X_i(t)\,\mathrm{d}W_i(t), \qquad \text{a.s.},$$

for $t \geq 0$ and $i = 1, \dots, n$. Hence,

$$\mathrm{d}X_{[n]}(t) = \left(\frac{\sigma^2}{2} - g\right)X_{[n]}(t)\,\mathrm{d}t + X_{[n]}(t)\,\mathrm{d}M(t) + ngX_{(n)}(t)\,\mathrm{d}t, \qquad \text{a.s.},$$

for $t \geq 0$, where M is a local martingale incorporating all of the terms $\sigma\,\mathrm{d}W_i(t)$. From this we have, for $t \geq 0$,

$$\frac{\mathrm{d}X_{[n]}(t)}{X_{(1)}(t)} = \left(\frac{\sigma^2}{2} - g\right)\frac{X_{[n]}(t)}{X_{(1)}(t)}\,\mathrm{d}t + \frac{X_{[n]}(t)}{X_{(1)}(t)}\,\mathrm{d}M(t) + \frac{ngX_{(n)}(t)}{X_{(1)}(t)}\,\mathrm{d}t, \qquad \text{a.s.},$$

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

so, for $T > 0$,

$$\mathbb{E}_n\left[\frac{1}{T}\int_0^T \frac{\mathrm{d}X_{[n]}(t)}{X_{(1)}(t)}\right] = \left(\frac{\sigma^2}{2} - g\right)R_{[n]} + ngR_n,$$

or

$$\frac{1}{R_{[n]}}\mathbb{E}_n\left[\frac{1}{T}\int_0^T \frac{\mathrm{d}X_{[n]}(t)}{X_{(1)}(t)}\right] = \frac{\sigma^2}{2} - g + \frac{ngR_n}{R_{[n]}}. \tag{3.17}$$

If an Atlas family is conservative and complete, then as n tends to infinity the first and last terms of (3.17) converge to zero, so $\sigma^2/2g = 1$ and the family will be Zipfian.

If the Atlas family is Zipfian then $\sigma^2/2g = 1$, in which case (3.13) holds, so $R_k = \frac{1}{k}$, and

$$R_{[n]} = \sum_{k=1}^{n} \frac{1}{k} = O(\log n).$$

It follows that

$$\frac{ngR_n}{R_{[n]}} = \frac{g}{O(\log n)},$$

so the family is complete, and with $\sigma^2/2 = g$ the right-hand side of (3.17) converges to zero as n tends to infinity. Hence, the left-hand side must also converge to zero, so the family is conservative. □

This proposition has a natural counterpart for quasi-Atlas families.

Proposition 3.2. *If a quasi-Atlas family is conservative and complete with*

$$\lim_{n\to\infty} R_{[n]} \geq 2, \tag{3.18}$$

then it is quasi-Zipfian.

Proof. Let $\{X_1, \ldots, X_n\}$ be a quasi-Atlas model with parameters $g, \sigma_1^2 > 0$ and $\sigma_2^2 \geq \sigma_1^2$, such that $g_k = -g$ and $\sigma_k^2 = \sigma_1^2 + (k-1)(\sigma_2^2 - \sigma_1^2)$, for $k = 1, \ldots, n$. Itô's rule implies that

$$\mathrm{d}X_i(t) = \left(\frac{\sigma_{r_t(i)}^2}{2} - g + ng\mathbf{1}_{\{r_t(i)=n\}}\right)X_i(t)\,\mathrm{d}t + \sigma_{r_t(i)}X_i(t)\,\mathrm{d}W_i(t), \qquad \text{a.s.},$$

for $t \geq 0$ and $i = 1, \ldots, n$, so

$$\mathrm{d}X_{[n]}(t) = \sum_{k=1}^{n} X_{(k)}(t)\left(\frac{\sigma_k^2}{2} - g\right)\mathrm{d}t + \mathrm{d}M(t) + ngX_{(n)}(t)\,\mathrm{d}t, \qquad \text{a.s.},$$

for $t \geq 0$, where M is a local martingale incorporating all of the terms $\sigma_{r_t(i)}X_i(t)\,\mathrm{d}W_i(t)$. As with (3.17) above, for $T > 0$,

$$\frac{1}{R_{[n]}}E_n\left[\frac{1}{T}\int_0^T \frac{\mathrm{d}X_{[n]}(t)}{X_{(1)}(t)}\right] = \frac{1}{R_{[n]}}\sum_{k=1}^{n} R_k\left(\frac{\sigma_k^2}{2} - g\right) + \frac{ngR_n}{R_{[n]}}.$$

Since the family is conservative and complete, the first and last terms of this equation converge to zero as n tends to infinity, so

$$\lim_{n\to\infty}\left(\frac{1}{R_{[n]}}\sum_{k=1}^{n} R_k\frac{\sigma_k^2}{2g}\right) = 1. \tag{3.19}$$

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

Let us now show that (3.18) implies that $s_1 \le 1$. Since $0 < \sigma_1^2 \le \cdots \le \sigma_n^2$, (3.19) implies that

$$\begin{aligned} 1 &\ge \lim_{n\to\infty} \frac{1}{R_{[n]}} \frac{\sigma_1^2}{2g} + \lim_{n\to\infty} \left(\frac{1}{R_{[n]}} \sum_{k=2}^{n} R_k \frac{\sigma_2^2}{2g} \right) \\ &= \lim_{n\to\infty} \frac{1}{R_{[n]}} \frac{\sigma_1^2}{2g} + \left(1 - \lim_{n\to\infty} \frac{1}{R_{[n]}} \right) \frac{\sigma_2^2}{2g} \\ &\ge \frac{1}{2} \frac{\sigma_1^2}{2g} + \frac{1}{2} \frac{\sigma_2^2}{2g} = s_1, \end{aligned}$$

where the last inequality follows from (3.18).

We must now show that either $\lim_{k\to\infty} s_k \ge 1$ or the s_k diverge to infinity. Since the σ_k^2 are nondecreasing, as k tends to infinity they must either converge to a finite value $\sigma^2 > 0$ or diverge to infinity. We see from (2.16) that if the σ_k^2 diverge to infinity, the same will be true for the s_k. If $\lim_{k\to\infty} \sigma_k^2 = \sigma^2$ then $\lim_{k\to\infty} s_k = \sigma^2/2g$, and since the σ_k^2 are nondecreasing,

$$1 = \lim_{n\to\infty} \left(\frac{1}{R_{[n]}} \sum_{k=1}^{n} R_k \frac{\sigma_k^2}{2g} \right) \le \frac{\sigma^2}{2g}.$$

It follows that $\lim_{k\to\infty} s_k \ge 1$. □

These two propositions seem remarkably simple. Many empirical systems can be at least roughly approximated by quasi-Atlas models, and conservation and completeness are properties that are almost universal in large time-dependent rank-based systems of empirical data. If these conditions are satisfied, then these two propositions show that Zipf's law, or at least its quasi-Zipfian counterpart, will pertain. Perhaps it is this simplicity that leads to the universality of Zipf's law for these systems.

4. Examples and discussion

Empirical time-dependent systems often behave like quasi-Atlas families, and in Example 4.1 below we consider one such system, the capitalizations of US companies (see Figures 1 and 2). The condition that the variance rates increase with rank seems natural; even in the original observation of [8] it would seem likely that the water molecules would have buffeted the smaller particles more vigorously than the larger ones. Below the top few ranks, the members of empirical time-dependent systems constantly drift among nearby ranks, and this could result in linearity of the σ_k^2, at least throughout the middle ranks. Whether the $g_k = -g$ for all k may be more problematic, but this appears to hold at least in Example 4.1, where we analyze actual data. Since we are usually observing the top part of a larger distribution, there is 'leakage' out of the system, characterized by the last term in (2.2), so the constant $-g$ may represent the universal draw toward extinction in time-dependent rank-based systems.

Example 4.1. (*Market capitalization of companies*.) The market capitalization of US companies was studied early on in [37], and here we follow the methodology of [14]. The capitalization of a company is defined as the price of the company's stock multiplied by the number of shares outstanding. Ample data are available for stock prices, and this allows us to estimate the first-order parameters we introduced in the previous sections.

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

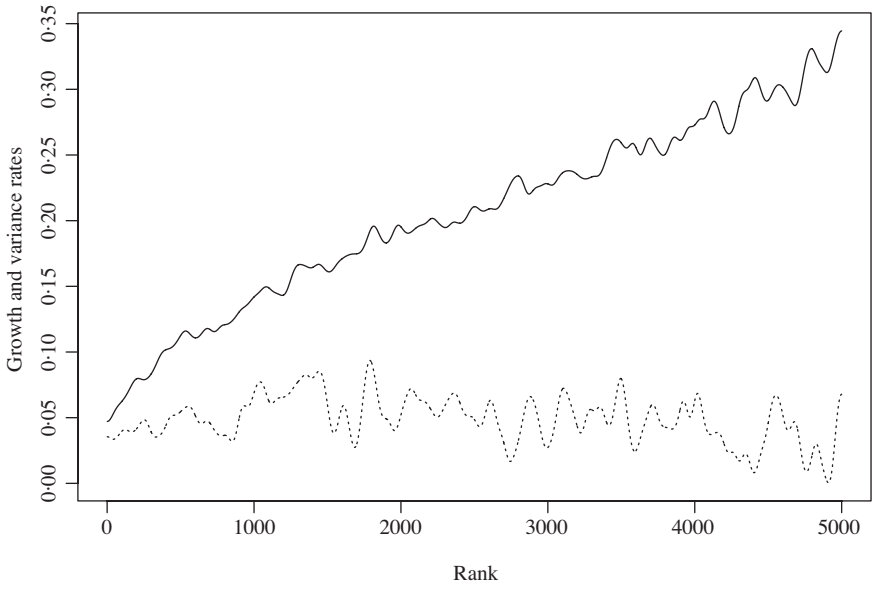

FIGURE 1: US capital distribution first-order parameters (smoothed): σ_k^2 (solid), $-g_k$ (dashed).

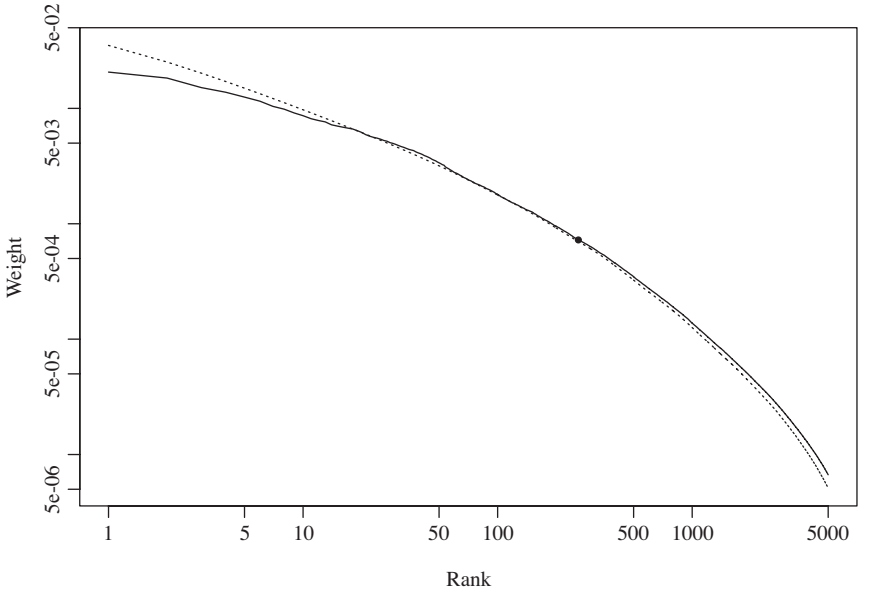

FIGURE 2: US capital distribution, 1990–1999 (solid). First-order approximation (dashed). The dot is the point at which the slope of the tangent is -1.

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

Figure 1 shows the smoothed first-order parameters σ_k^2 and $-g_k$ for the US capital distribution for the ten-year period from January 1990 to December 1999. The capitalization data we used were from the monthly stock database of the Center for Research in Securities Prices at the University of Chicago. The market we consider consists of the stocks traded on the New York Stock Exchange, the American Stock Exchange, and the NASDAQ Stock Market, after the removal of all Real Estate Investment Trusts, all closed-end funds, and those American Depositary Receipts not included in the S&P 500 Index. The parameters in Figure 1 correspond to the 5000 stocks with the highest capitalizations each month. The first-order parameters g_k and σ_k^2 were calculated as in (3.10) from the parameters $\boldsymbol{\lambda}_{k,k+1}$ and $\boldsymbol{\sigma}_{k,k+1}^2$ of (3.8) and (3.9), and then smoothed by convolution with a Gaussian kernel with ± 3.16 standard deviations spanning 100 months on the horizontal axis, with reflection at the ends of the data.

We see in Figure 1 that the values of the parameters $-g_k$ are relatively constant compared to the parameters σ_k^2, which increase almost linearly with rank. The near-constant $-g_k$ and near-linearly increasing σ_k^2 suggest that the first-order approximation can be represented by a quasi-Atlas family. In Figure 2, the distribution curve for the capitalizations is represented by the solid curve, which represents the average of the year-end capital distributions for the ten years spanned by the data. The dashed curve is the first-order approximation of the distribution following (3.12). The two curves are quite close, and this indicates that the time-dependent system of company capitalizations seems to be rank-based. The dot on the curve between ranks 100 and 500 is the point at which the log–log slope of the tangent to the curve is -1, so this is a quasi-Zipfian distribution, consistent with Proposition 3.2. Note that if we had considered only the top 100 companies, the completeness condition, Definition 3.6, would have failed, as we would expect for an incomplete distribution.

Example 4.2. (*Frequency of written words*.) Word frequency is the origin of Zipf's law [44], but testing our methodology with word frequency could be difficult. Ideally, we would like to construct a first-order approximation for the data and compare the first-order distribution to that of the original data. However, the parameters $\boldsymbol{\lambda}_{k,k+1}$ and $\boldsymbol{\sigma}_{k,k+1}^2$ for the top-ranked words in a language are likely to be difficult to estimate over any reasonable time frame, since the top-ranked words probably seldom change ranks. Nevertheless, while the top ranks may require centuries of data for accurate estimates, the lower ranks could be amenable to analysis similar to that which we carried out for company capitalizations. Moreover, it might be possible to combine, for example, all the Indo-European languages and generate accurate estimates of the $\boldsymbol{\lambda}_{k,k+1}$ and $\boldsymbol{\sigma}_{k,k+1}^2$ even for the top ranks of the combined data.

We can see from the remarkable chart in Wikipedia [42] that the log–log plots for 30 different languages are (almost) straight. Actually, these plots seem to be slightly concave, or quasi-Zipfian in nature. It is possible that this slight curvature is due to sampling error at the lower ranks, which would raise the variances and steepen the slope, but this would have to be determined by studying the actual data.

Example 4.3. (*Random growth processes*.) Economists have traditionally used random growth processes to model time-dependent systems with quasi-Zipfian distributions. For example, these processes were used in [21] to model the distribution of city populations and in [7] to construct a piecewise approximation to the distribution curves for the income and wealth of US households. A *random growth process* is an Itô process of the form

$$\frac{\mathrm{d}X(t)}{X(t)} = \mu(X(t))\,\mathrm{d}t + \sigma(X(t))\,\mathrm{d}W(t), \tag{4.1}$$

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

for $t \geq 0$, where W is Brownian motion and μ and σ are well-behaved real-valued functions. We can convert this into logarithmic form by Itô's rule, in which case

$$\mathrm{d}\log X(t) = \left(\mu(X(t)) - \frac{\sigma^2(X(t))}{2}\right)\mathrm{d}t + \sigma(X(t))\,\mathrm{d}W(t), \qquad \text{a.s.,} \tag{4.2}$$

for $t \geq 0$. We shall assume that this equation has at least a weak solution with $X(t) > 0$, a.s., and that the solution has a stationary distribution.

Let us construct n independent and identically distributed copies $X_1, \ldots, X_n$ of X, all defined by (4.1) or, equivalently, by (4.2), and assume that the X_i are all in their common stationary distribution. Let us assume that the $\log X_i$ accumulate no local time at triple points, so we can define the rank processes, and (2.1) and (2.2) will be valid. If the system is asymptotically stable we can calculate the corresponding rank-based growth rates g_k, but if we know the stationary distribution of the original process (4.1), then there is a simpler way to proceed.

If we know the common stationary distribution of the X_i, then we can calculate expectations under this stationary distribution and let

$$g_k = \mathbb{E}\left[\mu(X_{(k)}(t)) - \frac{\sigma^2(X_{(k)}(t))}{2}\right], \qquad \sigma_k^2 = \mathbb{E}\big[\sigma^2(X_{(k)}(t))\big],$$

for $t \geq 0$ and $k = 1, \ldots, n$. Under appropriate regularity conditions on the μ and σ, the expectations here will be equal to the asymptotic time averages of the functions. Since the X_i are in their stationary distribution, the geometric mean $\big(X_1 X_2 \ldots X_n\big)^{1/n} = \big(X_{(1)} X_{(2)} \ldots X_{(n)}\big)^{1/n}$ will also be in its stationary distribution, so for $t \geq 0$,

$$\big(g_1 + \cdots + g_n\big)t = \mathbb{E}\big[\log\big(X_{(1)}(t) \cdots X_{(n)}(t)\big) - \log\big(X_{(1)}(0) \cdots X_{(n)}(0)\big)\big] = 0.$$

Hence, $g_1 + \cdots + g_n = 0$, with $g_1 + \cdots + g_k < 0$, for $k < n$, so the g_k and σ_k^2 define the first-order model

$$\mathrm{d}\log Y_i(t) = g_{r_t(i)}\mathrm{d}t + \sigma_{r_t(i)}\mathrm{d}W_i(t), \tag{4.3}$$

for $t \geq 0$ and $i = 1, \ldots, n$, where $W_1, \ldots, W_n$ is n-dimensional Brownian motion. In this case, $G_n = 0$.

If the functions μ and σ in (4.1) are smooth enough, then the system is likely to be rank based, with the stationary distribution of the first-order model (4.3) close to that of the original system (4.1). More conditions are required for this stationary distribution to be quasi-Zipfian, and to achieve a true Zipfian distribution, a lower reflecting barrier or other equivalent device must be included in the model [22].

Example 4.4. (*Population of cities*.) The distribution of city populations is a prominent example of Zipf's law in social science. However, as the comprehensive cross-country investigation of [39] shows, city size distributions in most countries are not Zipfian but rather quasi-Zipfian. Gabaix [21] hypothesized that the quasi-Zipfian distribution of US city size was caused by higher population variances at the lower ranks, consistent with Proposition 3.2. Which of the deviations from Zipf's law uncovered in [39] are due to population variances that increase with decreasing city size remains an open question.

There is another phenomenon that occurs with city size distributions. Suppose that rather than studying a large country like the US, we consider instead the populations of the cities

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

in New York State. According to the 2010 US census, the largest city, New York City, had a population of 8 175 133, while the second largest, Buffalo, had only 261 310, so this distribution is non-Zipfian. The corresponding population of New York State was 19 378 102, so hypothesis (3.18) of Proposition 3.2 is satisfied, but nevertheless the proposition fails. This calls for an explanation, and we conjecture that while the population of the cities of New York State comprise a time-dependent system, this system is not rank based. The population of New York City is not determined merely by its rank among New York State cities, but is highly city specific in nature. Hence, we cannot expect the stationary distribution for the gap process between New York City and second-ranked Buffalo to be exponential, and we cannot expect the distribution of the system to be quasi-Zipfian.

Example 4.5. (*Assets of banks.*) Fernholz and Koch [19] showed that the distribution of assets held by US bank holding companies, commercial banks, and savings and loan associations are all quasi-Zipfian. This is true despite the fact that these distributions have undergone significant changes over the past few decades. However, as [20] showed, the first-order approximations of these time-dependent rank-based systems generally do not satisfy the hypotheses of Proposition 3.2, since the parameters $\boldsymbol{\sigma}^2_{k,k+1}$ are, in most cases, lower for higher values of k. Nonetheless, the parameters $\boldsymbol{\lambda}_{k,k+1}$ vary with k in such a way as to generate quasi-Zipfian distributions.

Example 4.6. (*Employees of firms.*) Axtell [2] shows that the distribution of employees of US firms is close to Zipfian, with only slight concavity. A number of empirical analyses have shown that for all but the tiniest firms, employment growth in US firms does not vary with firm size [33]. This observation, together with the slight concavity demonstrated in [2], suggests that the first-order approximation of US firm employees might be a quasi-Atlas family, which would explain its quasi-Zipfian nature.

5. Conclusion

We have shown that the stationary distribution of an Atlas family will follow Zipf's law if and only if the family is conservative and complete. We have also shown that a quasi-Atlas family will have a quasi-Zipfian stationary distribution if the family is conservative and complete, provided that the largest member does not represent more than one half of the total weight of the family. Since conservation and completeness are natural conditions for systems of time-dependent rank-based empirical data, and since many such systems can be approximated by Atlas or quasi-Atlas families, our results offer an explanation for the universality of Zipf's law for these systems.

Acknowledgements

We thank Xavier Gabaix, Ioannis Karatzas, members of the Intech SPT seminar, and participants of the 2017 Thera Stochastics Conference for their invaluable comments and suggestions. We are also grateful to an anonymous referee for pointing out a significant error in the original manuscript that led to a major revision of the paper.

References

[1] ATKINSON, A. B., PIKETTY, T. AND SAEZ, E. (2011). Top incomes in the long run of history. *J. Econom. Lit.* **49**, 3–71.
[2] AXTELL, R. (2001). Zipf distribution of U.S. firm sizes. *Science* **293**, 1818–1820.
[3] BAK, P. (1996). *How Nature Works*. Springer, New York.

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

[4] BANNER, A., FERNHOLZ, R. AND KARATZAS, I. (2005). On Atlas models of equity markets. *Ann. Appl. Prob.* **15**, 2296–2330.
[5] BANNER, A. AND GHOMRASNI, R. (2008). Local times of ranked continuous semimartingales. *Stoch. Process Appl.* **118**, 1244–1253.
[6] BASS, R. AND PARDOUX, E. (1987). Uniqueness for diffusions with piecewise constant coefficients. *Prob. Theory Relat. Fields* **76**, 557–572.
[7] BLANCHET, T., FOURNIER, J. AND PIKETTY, T. (2017). Generalized Pareto curves: theory and applications. Technical report. World Wealth & Income Database.
[8] BROWN, R. (1827). Brownian motion. Unpublished experiment.
[9] BRUGGEMAN, C. (2016). Dynamics of large rank-based systems of interacting diffusions. PhD thesis, Columbia University.
[10] CHATTERJEE, S. AND PAL, S. (2010). A phase transition behavior for Brownian motions interacting through their ranks. *Prob. Theory Relat. Fields* **147**, 123–159.
[11] DEMBO, A., JARA, M. AND OLLA, S. (2017). The infinite Atlas process: convergence to equilibrium. *Ann. Inst. H. Poincaré Prob. Statist.* **55**, 607–619.
[12] DEMBO, A., SHKOLNIKOV, M., VARADHAN, S. R. S. AND ZEITOUNI, O. (2016). Large deviations for diffusions interacting through their ranks. *Commun. Pure Appl. Math.* **69**, 1259–1313.
[13] DEMBO, A. AND TSAI, L.-C. (2017). Equilibrium fluctuation of the Atlas model. *Ann. Prob.* **45**, 4529–4560.
[14] FERNHOLZ, E. R. (2002). *Stochastic Portfolio Theory*. Springer, New York.
[15] FERNHOLZ, R., ICHIBA, T. AND KARATZAS, I. (2013). A second-order stock market model. *Ann. Finance* **9**, 1–16.
[16] FERNHOLZ, R., ICHIBA, T. AND KARATZAS, I. (2013). Two Brownian particles with rank-based characteristics and skew-elastic collisions. *Stoch. Process. Appl.* **123**, 2999–3026.
[17] FERNHOLZ, R., ICHIBA, T., KARATZAS, I. AND PROKAJ, V. (2013). A planar diffusion with rank-based characteristics and perturbed Tanaka equations. *Prob. Theory Relat. Fields* **156**, 343–374.
[18] FERNHOLZ, R. AND KARATZAS, I. (2009). Stochastic portfolio theory: an overview. In *Mathematical Modelling and Numerical Methods in Finance: Special Volume, Handbook of Numerical Analysis*, eds A. Bensoussan and Q. Zhang, Vol. XV. North-Holland, Amsterdam, pp. 89–168.
[19] FERNHOLZ, R. T. AND KOCH, C. (2016). Why are big banks getting bigger? Working Paper 1604. Federal Reserve Bank of Dallas.
[20] FERNHOLZ, R. T. AND KOCH, C. (2017). Big banks, idiosyncratic volatility, and systemic risk. *Amer. Econom. Rev.* **107**, 603–607.
[21] GABAIX, X. (1999). Zipf's law for cities: an explanation. *Quart. J. Econom.* **114**, 739–767.
[22] GABAIX, X. (2009). Power laws in economics and finance. *Ann. Rev. Econom.* **1**, 255–294.
[23] HARRISON, J. AND REIMAN, M. (1981). Reflected Brownian motion on an orthant. *Ann. Prob.* **9**, 302–308.
[24] HARRISON, J. M. AND WILLIAMS, R. J. (1987). Brownian models of open queueing networks with homogeneous customer populations. *Stochastics* **22**, 77–115.
[25] HARRISON, J. M. AND WILLIAMS, R. J. (1987). Multidimensional reflected Brownian motions having exponential stationary distributions. *Ann. Prob.* **15**, 115–137.
[26] ICHIBA, T. AND KARATZAS, I. (2010). On collisions of Brownian particles. *Ann. Appl. Prob.* **20**, 951–977.
[27] ICHIBA, T., KARATZAS, I. AND SHKOLNIKOV, M. (2013). Strong solutions of stochastic equations with rank-based coefficients. *Prob. Theory Relat. Fields* **156**, 229–248.
[28] ICHIBA, T., PAL, S. AND SHKOLNIKOV, M. (2013). Convergence rates for rank-based models with applications to portfolio theory. *Prob. Theory Relat. Fields* **156**, 415–448.
[29] ICHIBA, T., PAPATHANAKOS, V., BANNER, A., KARATZAS, I. AND FERNHOLZ, R. (2011). Hybrid Atlas models. *Ann. Appl. Prob.* **21**, 609–644.
[30] KARATZAS, I. AND SHREVE, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer, New York.
[31] KHAS'MINSKII, R. Z. (1960). Ergodic properties of recurrent diffusion processes, and stabilization of the solution to the Cauchy problem for parabolic equations. *Theory Prob. Appl.* **5**, 179–196.
[32] KHAS'MINSKII, R. Z. (1980). *Stochastic Stability of Differential Equations*. Sijthoff and Noordhoff, Amsterdam.
[33] NEUMARK, D., WALL, B. AND ZHANG, J. (2011). Do small businesses create more jobs? New evidence for the United States from the National Establishment Time Series. *Rev. Econom. Statist.* **93**, 16–29.
[34] NEWMAN, M. E. J. (2005). Power laws, Pareto distributions, and Zipf's law. *Contemp. Phys.* **46**, 323–351.
[35] PAL, S. AND PITMAN, J. (2008). One-dimensional Brownian particle systems with rank-dependent drifts. *Ann. Appl. Prob.* **18**, 2179–2207.
[36] SARANTSEV, A. (2015). Triple and simultaneous collisions of competing Brownian particles. *Electron. J. Prob.* **20**, 1–28.
[37] SIMON, H. AND BONINI, C. (1958). The size distribution of business firms. *Amer. Econom. Rev.* **48**, 607–617.
[38] SIMON, H. A. (1955). On a class of skew distribution functions. *Biometrika* **42**, 425–440.

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press

[39] SOO, K. T. (2005). Zipf's law for cities: a cross-country investigation. *Regional Sci. Urban Econom.* **35**, 239–263.
[40] STROOCK, D. W. AND VARADHAN, S. R. S. (2006). *Multidimensional Diffusion Processes*. Springer, Berlin.
[41] TAO, T. (2012). E pluribus unum: from complexity, universality. *Daedalus* **141**, 23–34.
[42] WIKIPEDIA (2020). Zipf's law. https://en.wikipedia.org/wiki/Zipf%27s_law.
[43] WILLIAMS, R. J. (1987). Reflected Brownian motion with skew symmetric data in a polyhedral domain. *Prob. Theory Relat. Fields* **75**, 459–485.
[44] ZIPF, G.(1935). *The Psychology of Language: An Introduction to Dynamic Philology*. MIT Press, Cambridge, MA.

https://doi.org/10.1017/jpr.2020.64 Published online by Cambridge University Press