RIETI Discussion Paper Series 23-E-085

Zipf's Law without the Stationarity Assumption (Revised)

Yoshiyuki ARATA
RIETI

The Research Institute of Economy, Trade and Industry
https://www.rieti.go.jp/en/

RIETI Discussion Paper Series 23-E-085
First Draft: December 2023
Revised: May 2024

Zipf's Law without the Stationarity Assumption*

Yoshiyuki Arata
Research Institute of Economy, Trade and Industry

Abstract

This paper analyzes one of the classic empirical regularities in the literature on firm growth: Zipf's law of the firm size distribution. Firstly, using firm-level data and decomposing the sample by the age of the firms, I found that the Pareto tail is observed within each age cohort. In particular, Zipf's law is observed only among younger firms (e.g., firms under 50 years of age). This empirical finding contradicts previous research which assumes that Zipf's law is observed only when the size distribution of firms from each age cohort is aggregated. To address this empirical inconsistency, this paper provides another explanation for Zipf's law. Specifically, Zipf's law is explained by two assumptions: the random walk assumption (i.e., the log of a firm's sales follows a random walk) and the heavy-tailed assumption that the growth rate distribution has a heavier tail than an exponential. In my analysis, the stationarity assumption (i.e., the firm size distribution is at the stationary state) is not needed. This new explanation resolves the empirical inconsistency and implies that a large firm arises from a few large jumps in size within a short period. Zipf's law reflects this property of firm growth dynamics.

Keywords: Zipf's law; Firm size distribution; Subexponential distributions
JEL classification: D21; D22; L10

*Corresponding author: y.arata0325@gmail.com
This study is a part of the project "Study group on corporate finance and firm dynamics" conducted at the Research Institute of Economy, Trade and Industry (RIETI). Arata appreciates the financial support by the Japan Society for the Promotion of Science (JSPS) (KAKENHI Grant Numbers: 21K13265). Conflict of interest: none.

1 Introduction

Zipf's law is one of the most important stylized fact in economics. Interestingly, this statistical regularity has bee found in different fields, such as the firm size distribution, income and wealth distributions, and the distribution of city sizes. Many economists have examined the mechanism generating this regularity.

However, previous explanations proposed so far have several drawbacks which are inconsistent with empirical data. First, while it is assumed that the observed Zipf's law is the result of aggregation of different age cohorts (i.e., a group of firms born at the same year), empirical data shows that even when focused on firms with the same age, the tail of the firm size distribution exhibits Zipf's law. Specifically, as shown later, Zipf's law is clearer when young firms (e.g., a firm's age is smaller than 50). Since the previous models predict that Zipf's law emerges only when different age cohorts are aggregated, this empirical fact cannot be accounted for. Second, it is known that in realistic time scale, the convergence of the firm size distribution to the stationary distribution is very slow. Especially in the tail region, where Zipf's law holds, the existing models cannot explain changes in the distribution tail or its tail exponent over time. These inconsistencies with empirical data suggest that we need another new explanation about Zipf's law.

This paper aims to provide another new explanation about Zipf's law, which resolve the inconsistencies mentioned above. My explanation is based only on the following two assumptions: the random walk assumption about the underlying growth process and the growth rate is heavy-tailed (or subexponential, whose definition is given later). For the former, I assume that the logarithm of a firm's sales follows a random walk, where a growth rate is an independently and identically distributed (iid) random variable. This assumption has been widely used in the literature and provides a good approximation for the empirical process. For the latter, as confirmed by my empirical data, the growth rate distribution is assumed to be heavy-tailed. More precisely, the tail of the growth rate distribution is not exponentially bounded; that is, so-called Cramer's condition is not met. With these two assumptions, I show that the firm size distribution for each age cohort has a tail following Zipf's law, and that the firm size distribution at the aggregate level (i.e., the firm size distribution for all firms) also satisfies Zipf's law.

A notable feature that differentiate my explanation from the existing models is that the stationarity assumption is not required in my analysis. I show that Zipf's law has nothing to do with the stationary assumption, and that the tail of the firm size distribution, where Zipf's law holds is directly related to the tail of the growth rate distribution. As a result, when the underlying growth rate distribution changes and has a heavier tail than before, it directly affects the tail of the firm size distribution. For this reason, the difficulty about slow convergence in the existing models is fully resolved in my explanation.

My theoretical analysis focuses on the classification of zones where the sum of n iid random variables (i.e., the growth rates) takes. Intuitively speaking, under the subexponential assumption, there are three different zones, where the probability of the sum can be characterized by different mechanisms (see **Figure 1**). The first one is so-called Cramer's zone, which is a zone around the center. The probability of the sum

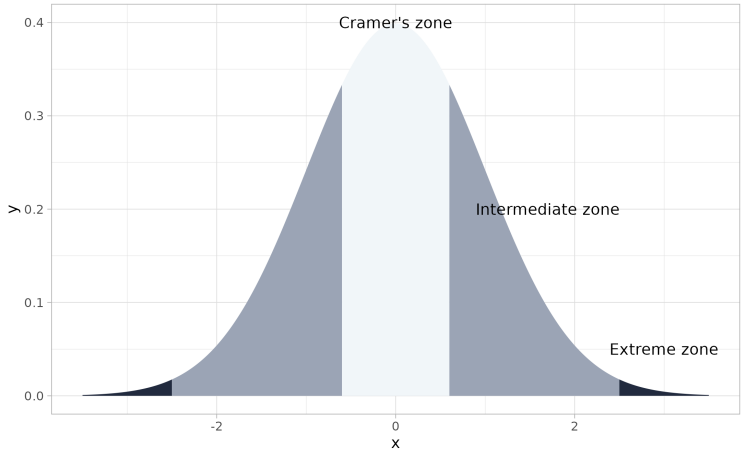


Figure 1: Three different zones for the probability of the sum of n iid random variables.

taking a value in this zone can be characterized by the central limit theorem, i.e., the probability can be well approximated by the Gaussian distribution. This zone expands as n increases at the rate of $n^{1/2}$. The second zone is an extreme zone, which the probability of the sum taking an extremely large value. One can show that when the growth rate follows a subexponential distribution, the probability of the sum is asymptotically equal to the probability of the maximum among n variables. This property is called the principle of a single big jump; that is, the maximum dominates the sum.

The third one is a zone between the two zones above, which is the most relevant to our analysis. In this zone, although neither the normal convergence nor the principle of a single big jump hold, the probability of the sum is still determined by the tail probability of growth rates. I show that widely observed Zipf's law corresponds to the probability of the sum taking a value in this zone, and therefore, it reflects the tail probability of growth rates. Since my explanation requires only the simple two assumptions, it also demonstrates the universality of Zipf's law.

I test whether the theoretical predictions are supported by empirical data, covering a large part of an economy. I use data complied by Orbis for European countries and data complied by Tokyo Shoko Research (TSR) for Japan. These datasets have more than millions firms in each country, and especially for larger firms, the coverage is almost comprehensive. Since the interest of my analysis is in the tail of the firm size distribution (i.e., larger firms), the shape of the tail of the firm size distribution can be properly estimated using these data. Variables of interest are annual sales revenues (as a measure of firm size) and firms' age, which is calculated using the date of incorporation. These two variables are available in both datasets.

Using these data, I first check the two assumptions. For the random walk assumption, I examine the autocorrelation of growth rates and find that the autocorrelation is quite weak. I also try the random walk test. For the subexponentiality of the growth rate distribution, I employ the density estimation, the mean excess function, and the estimation of the Weibull tail coefficient. All of them support that the the growth

rate distribution has a heavier tail than an exponential, and therefore, that it is subexponential.

Empirical findings that strongly support my explanation about Zipf's law and differentiate it from existing models is (1) that the firm size distribution for each age group exhibits Zipf's law especially for young groups and (2) that the tail exponent of the firm size distribution is equal to that of the growth rate distribution. When an age cohort is considered (i.e., n is fixed), we confirm that the firm size distribution for this age group has three zones, and in particular, the distribution exhibits Zipf's law outside Cramer's zone. This is what my theoretical analysis predicts. Especially for (2), I focus on the distribution of growth rates in a longer term (e.g., ten-years growth rates). Note that the long-term growth rate distribution has two aspects. On the one hand, the long-term growth rate is the sum of annual growth rates, and thus, it can be seen as the sum of iid random variables. On the other hand, the firm size distribution is generated by the long-term growth rates, and thus, the long-term growth rates comprise the sum (i.e., firm size). I show that the long-term growth rate distribution has these two properties, which are consistent with my theoretical predictions.

Because of the robustness and universality of Zipf's law, many models and explanations have been proposed in the existing literature. Compared to them, the novelty of my explanation is that it is parsimonious. What is required is only two assumptions, both of which are testable by empirical data. Another feature worth emphasizing is the absence of the stationarity assumption. To the best of my knowledge, the existing models have assumed that Zipf's law is the consequence of the stationary distribution, which is compatible to the idea that an economy is in an equilibrium. As a result, the limiting situation (i.e., $n \to \infty$) is the main interest in such models, and they turn out to struggle with the slow convergence to the stationary distribution. In contrast, my explanation reveals that first of all, Zipf's law has nothing to do with the stationarity assumption, and therefore, is free from the slow convergence issues.

1.1 Related literature

This paper is based on an empirical finding about the shape of the growth rate distribution (for a survey, see Coad (2009) and Dosi et al. (2017)). Since Stanley et al. (1996), it has been widely recognized that the growth rate distribution deviates from a Gaussian. Specifically, compared to a Gaussian, the growth rate distribution has a high kurtosis and a heavier tail, which is well approximated by a Laplace distribution. Furthermore, recent empirical papers have pointed out the possibility that the tail of the growth rate distribution is strictly heavier than an exponential. Bottazzi et al. (2011) and Dosi et al. (2020) introduce the Subbotin family of distributions, which includes Gaussian and Laplace distributions as special cases, and rejects the null hypothesis that the growth rate distribution is exponential. In Arata et al. (2023), the authors consider a distribution family called subexponential distributions, and provide an empirical evidence that the tail of the growth rate distribution is heavier than an exponential. In the present paper, following the line of these empirical papers, it is assumed that the growth rate distribution has a heavier tail than an exponential.

I show that this assumption is crucial to the shape of the firm size distribution, especially to the mechanism of Zipf's law.

This paper contributes to the firm growth literature especially regarding to Zipf's law. The universality of Zipf's law of the firm size distribution has been widely mentioned in the literature; see, e.g., (Axtell (2001); Gabaix (2009); Luttmer (2010)).[1] Theoretical explanation about the mechanism behind Zipf's law has also been proposed. A series of papers by Luttmer (Luttmer (2007); Luttmer (2011)) consider "blueprints" as a source of firm business and model firm growth as the accumulation of blueprints. In particular, Luttmer (2011) emphasizes that it takes a unrealistically longer time to become a large firm, and to address this issue, he introduces the type of blueprint, that is, high- and low-quality blueprints. More recently, Beare and Toda (2022) propose an unified framework incorporating models in previous works and investigate how the exponent of the Pareto tail is determined.

The points that differentiate this paper from the previous literature above are two-folds. First, because of large jumps implied by the heaviness of the growth rate distribution, a small firm can be a giant by a few jumps within a short period. Roughly speaking, in the previous models, a growth process is the accumulation of small shocks (or successes), which explains why it takes long time until it becomes a large firm. In contrast, with a heavier tail of the growth rate distribution, the firm growth process is characterized by a few large jumps. For this reason, there is no need to introduce multiple types into a model. Second, in my analysis, the stationarity assumption that the firm size distribution is the stationary distribution is not imposed. Almost all of the models proposed so far has assumed that the distribution converges to a stationary distribution as $n \to \infty$, which exhibits the observed Zipf's law. In my analysis, I find that Pareto tail is observed only for young firms, suggesting that it is necessary to consider a finite value of n. I show that Zipf's law has nothing to do with the stationarity assumption, and therefore, the slow convergence problem mentioned above does not occur here.

1.2 Outline

The remainder of this paper is organized as follows. Section 2 summarizes existing models for Zipf's law and explains why they are inconsistent with empirical data. Section 3 provides a new explanation about Zipf's law. Section 4 provides its empirical support. Section 5 concludes.

[1] Some recent papers have further analyzed Zipf's law using comprehensive firm-level data. Kondo et al. (2023) use firm level data in the U.S. and find that the firm size distribution is heavy-tailed but not follows a Pareto tail. They argue that the tail of the firm size distribution is better approximated by a log-normal tail.

2 Motivation and previous models

This section reviews previous models for Zipf's law and provides motivating examples. In Section 2.1, I review main mechanisms generating Zipf's law using Beare and Toda (2022). In Section 2.2, I provides empirical findings that are inconsistent with the main mechanism in the previous literature.

2.1 Review of previous models

A stylized empirical regularity regarding firms' sizes is Zipf's law, which suggests that the tail of the firm size distribution follows a Pareto tail. Specifically, letting $\log(\text{sales})$ be S, Zipf's law can be represented as a straight line with a slope of α in the tail of the distribution:

$$\log \mathbb{P}(S > x) = -\alpha x + \text{const.}$$

According to previous literature, it is known that α is close to 1.

As discussed in the Introduction, many theories have been proposed so far to explain Zipf's law. Among them, the most fundamental model assumes that the firm size distribution is a superimposition of the firm size distributions produced by cohorts of firms born in different years. That is, letting S_0 represent the logarithm of a firm's initial size (a random variable) and X_t represent the growth rate in period t, if we consider $S_{i,t}$ as the logarithm of the sales in period t of a firm born in period i, then, by definition, it can be written as follows:

$$\begin{aligned}
S_{i=0,t=n} &= S_0 + X_{t=1} + \quad X_{t=2} + \quad X_{t=3} + \ldots + X_n \\
S_{i=1,t=n} &= S_0 + \quad\quad\quad X_{t=2} + \quad X_{t=3} + \ldots + X_n \\
S_{i=2,t=n} &= S_0 + \quad\quad\quad\quad\quad X_{t=3} + \ldots + X_n \\
&\vdots
\end{aligned}$$

Additionally, firms can exit the market stochastically, leading to a decline in the number of firms established in period i as time goes by. In particular, when a firm's exit follows a Poisson process, the distribution of the number of firms would be described by an exponential distribution.

Letting $\mathbb{P}(S_{i,n} > x)$ be the distribution of firm sizes at time n for firms born at period i, the firm size distribution at the aggregate is given by:

$$\mathbb{P}(S_n > x) = \sum_{i=0}^{n} n_{i,n} \mathbb{P}(S_{i,n} > x)$$

Here, $n_{i,n}$ represents the proportion of firms born in period i at time n. In almost all of the previous models, the limiting distribution of S_n as $n \to \infty$ is considered, i.e., the steady-state distribution of $\mathbb{P}(S_n > x)$, which is used to explain the Pareto tail of Zipf's law. Note that in this model, the size distribution of firms born in period i, represented by $\mathbb{P}(S_{i,n} > x)$, does not necessarily follow Zipf's law. For instance, if X_k follows a Gaussian distribution, the distribution for each age cohort would be also Gaussian. Thus, Zipf's law is observed only at the aggregate level, and for this reason, the proportions of firms born in period i,

denoted by $n_{i,n}$, and its determinant, the "birth rate" of the firms, play a crucial role in explaining Zipf's law in previous literature (cf. Beare and Toda (2022)).

One of the important predictions of this model is that the tail part, which represents Zipf's law, is dominated by firms with a higher age. As mentioned above, when the birth rate of firms is constant, and the exit rate does not depend on the firm's size but exits at a constant rate over time, the number of firms decreases geometrically as their age increases. On the other hand, if the distribution $\mathbb{P}(S_{i,n} > x)$ is Gaussian, its tail decreases rapidly compared to the geometric decrease, specifically at a rate of $\exp(-x^2)$. Therefore, firms that contribute to the tail part of the firm size distribution at the aggregate are those where the variance of $\mathbb{P}(S_{i,n} > x)$ becomes large; in other words, older firms constitute the tail. This characteristic can be seen as a consequence of considering the stationary distribution for the explanation of Zipf's law, and therefore, it is a common feature in most theoretical models proposed in previous studies.

2.2 Firm size distribution

Here, we examine whether Zipf's Law holds true using enterprise-level data.[2] Specifically, we examine whether Zipf's Law holds true as a result of aggregating the samples as previous research has stated, or if it holds true within each age cohort by dividing the samples according to firms's ages.

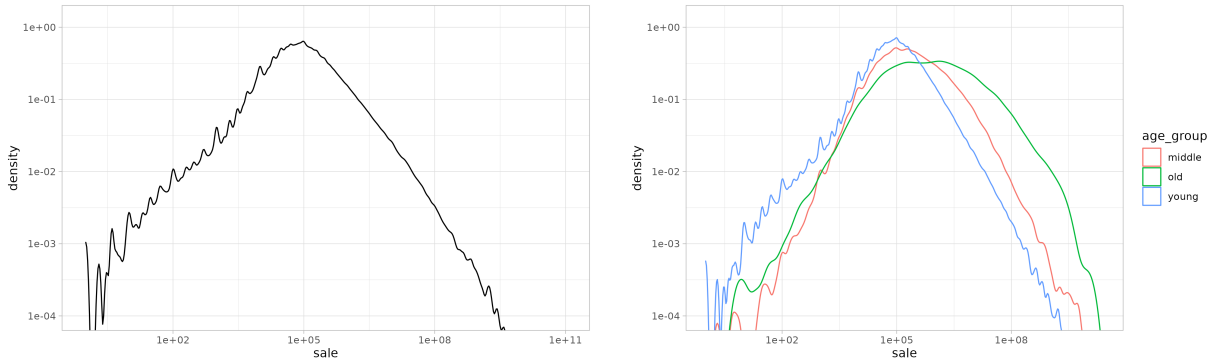
Figure 2 illustrates the size distribution (density) of firms' size on a logarithmic scale. In other words, if Zipf's law holds, it would be represented as a straight line. As shown in **Figure 2**, in the right region, the distribution is well approximated by a straight line, and the slope is also found to be close to -1. Therefore, consistent with the results of previous research, this suggests that Zipf's law holds in my data.

Which age group's sample is responsible for this Pareto tail? Here, to define a firm's age, I utilize the incorporation date of a firm as the firm's birth year. That is, a firm's age is defined as the corresponding year (in this case, 2018) minus the year of incorporation. **Figure 2** (b) compares samples divided into three age groups: those with an age less than 50, those with an age between 50 and 70, and those with an age older than 70. From this figure, it is evident that the distribution of firm sizes for younger-aged firms exhibits a Pareto tail in its tail, whereas for the older-aged firms, it deviates from the Pareto tail and is closer to a Bell-shaped curve. Essentially, the observed Pareto tail that could be observed across the entire sample is actually brought about by samples of younger-aged firms. This result contradicts the theoretical prediction of previous models, which posited that older-aged firms form the Pareto tail.

To further investigate the Pareto tail, the samples are divided based on 5-year intervals of a firm's age, and the distribution of firm sizes for each age group is compared. For instance, here "age 10" refers to firms aged between 5 and 10 years old. **Figure 3** (a) focuses on the age cohorts ranging from 5 to 50 years old, while (b) targets the age cohorts of 50 years and older.

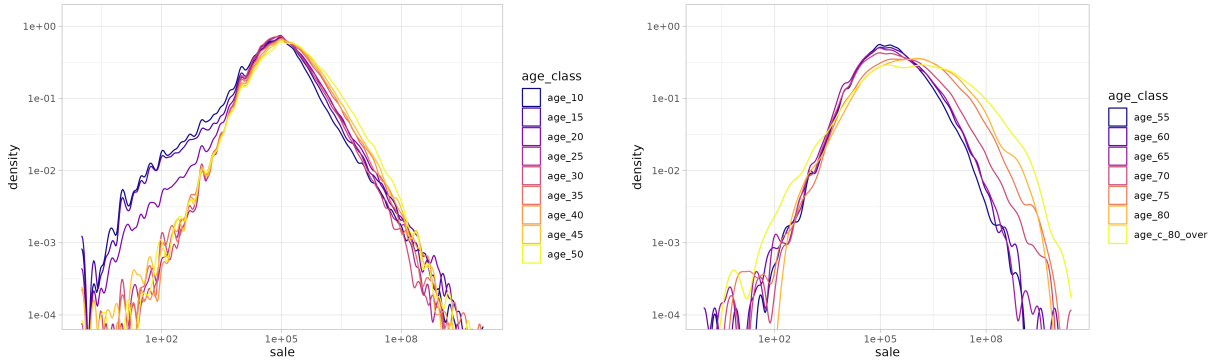
These figures illustrate how the firm size distribution changes as the age of the firms increases. When

[2]Details about the data used are provided in Section 4.1.



(a) Aggregate **(b)** Three age groups

Figure 2: Firm size distribution



(a) Firms with ages below 50 **(b)** Firms with ages 50 and above

Figure 3: Firm size distribution by age cohort.

firms are young, they exhibit Pareto tail in a broad range, particularly in the region of firms' sales exceeding 100 million yen. Moreover, the Pareto tails of the firm size distributions in each sample share a common slope. As the age of firms further increases, the firm size distribution deviates from the straight line of the Pareto tail and shows a curved shape. This difference of the distribution shape according to firms' ages reveals the mechanism behind Zipf's law. In order to comprehend Zipf's law, an explanation consistent with the nature of these age-specific firm size distributions is necessary.

3 New explanation

This section provides a new explanation of Zipf's law without the stationarity assumption. In Section 3.1, I introduce a general setting for a random walk and consider normal convergence. In Section 3.2, I explain three zones of the value of the sum of independent random variables and show that the Pareto tail corresponds to the distribution in intermediate and extreme zones. In Section 3.3, I consider the effect of a firm's initial size on the firm size distribution.

3.1 Setup

As in previous section, let us start with the random walk assumption: Letting X_k be the log growth rate at period k, the log of a firm's sales (denoted by S_n) is given by

$$S_n := S_0 + X_1 + ... + X_n$$

We assume that S_n follows a random walk.

Assumption 3.1. S_n follows a random walk with an initial condition S_0.

In other words, we assume that the growth rates $X_1, ..., X_n$ are iid random variables. For a moment, ignore the contribution of the initial state S_0 to the sum (i.e., assume that $S_0 = 0$). In that case, S_n can be viewed as the sum of n iid random variables. Our main question to be addressed here is as follows: what is the distribution of S_n?

The fundamental theorem of probability theory (i.e., the central limit theorem) tells us that the sum of n iid random variables, when properly normalized, converges to the standard Gaussian distribution as n goes to infinity (see, e.g., Petrov (1995)). More formally, letting

$$Z_n = \sigma^{-1} n^{-1/2} S_n, \quad F_n(x) = \mathbb{P}(Z_n < x)$$

where σ is the standard deviation of X_k, the normal convergence means that for an arbitrary (fixed) x,

$$\frac{1 - F_n(x)}{1 - \Phi(x)} \to 1, \quad \frac{F_n(-x)}{\Phi(-x)} \to 1 \tag{1}$$

Here, Φ is the standard Gaussian distribution. From this theorem, given that n is sufficiently large, one might think that the distribution of S_n can be well approximated by a Gaussian, including a tail zone. However, this idea turns out to be wrong in general, and this point becomes crucial to the analysis of Zipf's law, as we discuss later.

Before discussing general settings, we consider the case where X_k follows a Laplace distribution. In this case, we can obtain the analytical expression of the probability distribution of the sum and shows how the distribution changes as n increases.

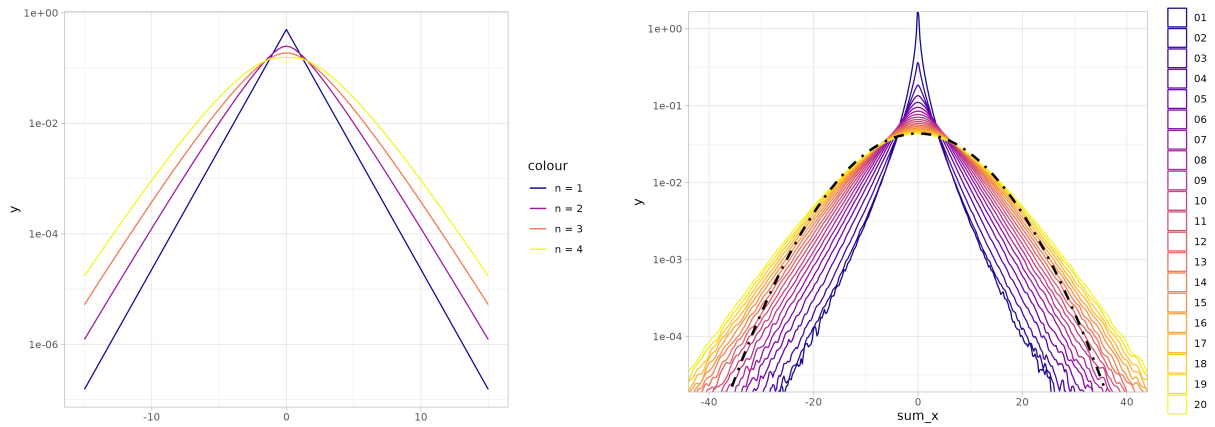
Suppose that the probability density of X_k is given as follows:

$$\mathbb{P}(dx) = \frac{1}{2} \exp(-|x|) dx$$

Then, consider the distribution of the sum $\sum_{k=1}^n X_k$. The densities for $n = 2, 3$ and 4 are given as follows (see Kotz et al. (2001)):

$$\begin{aligned}
\mathbb{P}_{\sum_{k=1}^2 X_k}(dx) &= \frac{1}{2} \cdot \frac{1}{2}(1 + |x|) \exp(-|x|), \\
\mathbb{P}_{\sum_{k=1}^3 X_k}(dx) &= \frac{1}{3} \cdot \frac{9}{16}\left(1 + |x| + \frac{1}{3}|x|^2\right) \exp(-|x|), \\
\mathbb{P}_{\sum_{k=1}^4 X_k}(dx) &= \frac{1}{4} \cdot \frac{1}{24}\left(15 + 15|x| + 8|x|^2 + |x|^3\right) \exp(-|x|)
\end{aligned}$$

Figure 4 presents these densities in the log-scale. When x is small (i.e., x is in the central zone), the densities looks like a bell shape (normal convergence), which expands as n increases. In contrast, when x is large,



(a) Laplace **(b)** Weibull

Figure 4: Density of the sum $\sum_{k=1}^{n} X_k$ for different n.

the density is determined by the term $\exp(-|x|)$, i.e., the same exponential decay as the component X_k. Especially when the log scale is considered (i.e., $\log \mathbb{P}_{\sum_{k=1}^{n} X_k}(dx)$), its density is close to a straight line in the tail.

We can see a similar behavior when a distribution with a Weibull tail is considered, though an exact formula for the density of the sum is not available. We generate pseudo-samples using the following distribution:

$$\mathbb{P}(X > x) = \frac{1}{2}\exp(-|x|^{\alpha})$$

where $\alpha < 1.0$ (in the figure, α is set to 0.7). **Figure 4** (b) describes the evolution of the densities of the sum $\sum_{k=1}^{n} X_k$ as n increases from $n = 1$ to 20. As in the Laplace case, the peak of the density diminishes and gets closer to the Gaussian shape around the center zone as n increases. In contrast, in the tail zone, the densities deviate from a Gaussian and exhibit approximately straight lines, which are parallel.

3.2 Three zone

Let us return to our general settings, i.e., the random walk assumption. I consider the two distribution classes for the growth rate distribution F. We say that the distribution of X_k is light-tailed if it satisfies Cramer's condition: for some $\lambda > 0$,

$$Ee^{\lambda X_k} < \infty$$

Roughly speaking, the distribution is light-tailed if its tail is exponentially bounded. Gaussian and exponential distributions are examples of light-tailed distributions. We say that the distribution is heavy-tailed if the distribution is not light-tailed.

Requiring slight regularity conditions on the distribution tails, we can consider a subclass of the heavy-tailed distributions called subexponential distributions. We say that a distribution on the positive real half-line

$\mathbb{R}^+$ is subexponential if it satisfies the following condition:

$$\lim_{x\to\infty} \frac{\overline{F * F}(x)}{\overline{F}(x)} \tag{2}$$

exists, where $\overline{F}(x) := F[x, \infty)$ and $F * F(x)$ is the convolution of F with itself. We say that a distribution on the whole real line $\mathbb{R}$ is subexponential if the distribution of $X_k^+ := \max\{0, X_k\}$ is subexponential.

It should be noted that the subexponential distributions is a broad subclass of heavy-tailed distributions, and indeed, distributions widely used in empirical applications (e.g, Weibull and Pareto distributions) are subexponential. In particular, when F is a heavy-tailed distribution on $\mathbb{R}^+$, one can show that $\liminf_{x\to\infty} \frac{\overline{F*F}(x)}{\overline{F}(x)} = 2$. That is, the regularity condition only requires the existence of the limit and, if it exists, it is equal to 2. Recall that the probability that the maximum of the elements is larger than x is given by $n\overline{F}(x)$. Thus, the property $\lim_{x\to\infty} \frac{\overline{F*F}(x)}{\overline{F}(x)} = 2$ (or $\lim_{x\to\infty} \frac{\overline{F^{*n}}(x)}{\overline{F}(x)} = n$ in general), means that the probability of the sum is asymptotically equivalent to the probability of the maximum of the elements as $x \to \infty$. In other words, a large deviation of the sum is generated by a large deviation of a single element. For this reason, this property is called the principle of a single big jump and will be discussed below.

In my analysis, I assume that growth rates $X_1, X_2, ..., X_n$ are iid random variables with a common subexponential distribution F. Specially, I consider as F a distribution with a finite variance and a Weibull tail. Its empirical validity is checked in Section 4.3.

Assumption 3.2. The distribution F is subexponential.

Given the two assumptions, what is the distribution of the sum S_n? The discussion and examples in Section 3.1 suggest that the distribution of S_n can be approximated by a Gaussian distribution in the central zone (i.e., for the small deviation zone) due to the central limit theorem. On the other hand, as suggested by the principle of a single big jump, the tail of the distribution of S_n can be approximated by the tail of the distribution of X_k multiplied by a factor n. Thus, the distribution of S_n has three zones of x:

- Cramer's deviation zone: the normal convergence holds.
- Extreme deviation zone: the principle of a single big jump holds.
- Intermediate deviation zone: a zone between the two zones above.

This statement can be made rigorous as follows:

Theorem 3.1 (See Chapter 5 in Borovkov and Borovkov (2008))**.** *For* $x \leq \sigma_1(n)$

$$\mathbb{P}\left(S_n \geqslant x\right) = \left[1 - \Phi\left(\frac{x}{\sqrt{n}}\right)\right] e^{-n\Lambda_\kappa^0(x/n)}(1 + o(1))$$

For $x \gg \sigma_1(n)$

$$\mathbb{P}\left(S_n \geqslant x\right) = ne^{-M}(1 + \varepsilon(x, n))$$

In particular, for $x \gg \sigma_2(n)$

$$\mathbb{P}(S_n \geq x) = nV(x)(1 + o(1))$$

First, the approximation in the case of $x \leq \sigma_1(n)$ corresponds to the so-called Cramer's approximation,

and the factor $e^{-n\Lambda_\kappa^0(x/n)}$ is called Cramer's correction. Especially when $x \ll n^{1/6}$, this correction term converges to 1, and therefore, $\mathbb{P}(S_n \geq x)$ is approximated by a Gaussian distribution. Outside of the zone $x \leq \sigma_1(n)$, the distribution of S_n is controlled mainly by the term M and the tail of the distribution of the element has a more impact. Especially in the log-scale (i.e., when $\log \mathbb{P}(S_n \geq x)$ is considered), M can be simplified: for $x \gg \sigma_1(x)$, we have

$$\log \mathbb{P}(S_n \geq x) = (1 + o(1)) \log nV(x)$$

Obviously, this approximation holds also for $x \gg \sigma_2(n)$.

For later purpose, let us obtain the concrete expression of $\sigma_1(n), \sigma_2(n)$ when F has a Weibull distribution with exponent $\alpha < 1$.

Suppose that the firm size distribution at the aggregate is the superposition of $\mathbb{P}(S_n > x)$ with different n and that each $\mathbb{P}(S_n > x)$ is described by the theorem above. Furthermore, the tail of the distribution $V(x)$ is close to an exponential (i.e., a Weibull tail with α close to 1). Then, in the zone $x \gg \sigma_1(n)$, the distributions $\log \mathbb{P}(S_n > x)$ with different n has a common slope with different intercept. Therefore, when aggregated, the firm size distribution in the log scale has the same slope as that of $\log \mathbb{P}(S_n > x)$. Indeed, suppose

$$\log \mathbb{P}(S_n > x) = b_n - ax,$$

i.e., a straight line with slope a and intercept b_n. If n_n is the fraction of firms with index n, the distribution at the aggregate in the log-scale becomes

$$\log \big(\sum_n n_n \mathbb{P}(S_n > x) \big) = \log \Big(\big(\sum_n n_n B_n \big) \exp(-ax) \Big) = b - ax$$

where $B_n := e^{b_n}$ and $b := \log(\sum_n c_n B_n)$. This is our main mechanism generating Zipf's law at the aggregate.

3.3 Initial growth

So far, the contribution of the initial state S_0 to S_n has been ignored in the analysis. Here, I consider how the distribution of S_n is affected by the distribution of S_0.

Let us consider the random walk with an initial size S_0, whose distribution is denoted by F_0. Thus, S_n is the combination of two random variables: one is S_0 and the other is the sum of $X_1, ..., X_n$, each of which is drawn from a common distribution F. I assume that F_0 is also subexponential distribution and belong to semi-exponential distributions.

Assumption 3.3. The distribution F_0 is subexponential (and belongs to semi-exponential distributions).

This assumption will be empirically checked later. Then, one can get the extension of Theorem 3.1 as follows:

Theorem 3.2 (Theorem 11.3.1 in Borovkov and Borovkov (2008))**.** *Suppose that the conditions* $[\cdot, =]_\tau, [\cdot, =]_\xi$

are met and that functions l_τ, l_ξ satisfy condition [**D**]. *Then, if* $s_1^{(\tau)} \to \infty, s_1^{(\xi)} \to \infty$,

$$\mathbb{P}(S_n \geq x) \sim e^{-M^{(\tau,\xi)}(x,n)} + ne^{-M^{(\xi,\xi)}(x,n)}$$

In particular, if $s_2^{(\tau)} \to \infty, s_2^{(\xi)} \to \infty$,

$$\mathbb{P}(S_n \geq x) \sim V_\tau(x) + nV_\xi(x)$$

First, consider the extreme deviation zone, where $s_2^{(\tau)} \to \infty, s_2^{(\xi)} \to \infty$. This theorem says that as n increases, the second term dominates the distribution of S_n. This is obvious because as a firm's age increases, the contribution of a firm's growth rates $X_1, ..., X_n$ dominates the firm size, and the effect of the initial size becomes less important. Furthermore, this theorem shows that the tail of distribution of S_n is determined by the tails of F_0 and F. In particular, when l_τ and l_ξ have a common slope (but with different intercept), the tail of the distribution of S_0 becomes

$$\mathbb{P}(S_n \geq x) \sim (c_\tau + c_\xi n)e^{-ax}$$

Thus, when $\log \mathbb{P}(S_n \geq x)$ is considered, we would observe the same slope at the aggregate. Suppose that l_τ and l_ξ have different slopes. When the log scale is considered, by taking the derivative of the right-hand side, we have

$$\frac{d}{dx} \log \mathbb{P}(S_n \geq x) = \frac{-l'_\tau(x)V_\tau - nl'_\xi(x)V_\xi}{V_\tau + nV_\xi}$$

That is, the slope of $\log \mathbb{P}(S_n \geq x)$ is given by the weighted average of the slopes of F_0 and F.

For the intermediate deviation zone, we can obtain a similar implication about the tail of the distribution of S_n, though the expression is more complicated. As in Section 3.2, the function $M^{(\tau,\xi)}$ can be approximated by $M^{(\tau,\xi)} = l_\tau(x)(1 + o(1))$. Thus, the slope of $\log \mathbb{P}(S_n \geq x)$ is given by

$$\frac{d}{dx} \log \mathbb{P}(S_n \geq x) = \frac{-(1 + o(1))l'_\tau(x)V_\tau - n(1 + o(1))l'_\xi(x)V_\xi}{V_\tau + nV_\xi}$$

As in the case of the extreme deviation zone, the slope is determined by the weighted average of the slopes of F_0 and F.

4 Empirical results

This section provides empirical evidences for my new explanation of Zipf's law. Section 4.1 provides the summary statistics of firms' sizes and growth rates in my samples. Section 4.2 shows that the random walk assumption provides good approximation for the empirical growth process. Section 4.3 shows that the growth rate distribution is subexponential. Finally, Section 4.5 provides additional empirical support for the idea that the firm growth process is determined by a few large jumps.[3]

[3]Using firm-level data from the Orbis, I have conducted a similar analysis for other countries. I have obtained results similar to those found in this section. These results are available upon request from the author.

Summary statistics of firm size

Data: Tokyo Shoko Research from 2009 to 2019

year	count	mean	sd	q1	median	q3
2009	1222314	11.427	1.779	10.309	11.350	12.429
2010	1260683	11.351	1.807	10.275	11.290	12.372
2011	1289487	11.317	1.824	10.240	11.264	12.346
2012	1281791	11.317	1.841	10.253	11.275	12.346
2013	1279646	11.312	1.870	10.245	11.280	12.352
2014	1279219	11.323	1.893	10.258	11.290	12.388
2015	1302137	11.297	1.917	10.238	11.290	12.376
2016	1309421	11.294	1.918	10.236	11.290	12.369
2017	1310287	11.302	1.929	10.240	11.290	12.387
2018	1300207	11.310	1.947	10.240	11.290	12.390
2019	1266616	11.350	1.923	10.275	11.316	12.429

Table 1: Summary statistics of firms' sizes. The period is from 2009 to 2019. The summary statistics are calculated using the log of annual sales (i.e., $\log(\text{sale})$).

4.1 Summary statistics

The data used in the following is firm-level data complied by Tokyo Shoko Research. It includes both listed and non-listed firms, covering more than one million firms for each year. This data is based on a survey conducted by TSR. Since TSR is a rating agency, the firms surveyed by TSR are determined by requests from TSR's clients. Therefore, almost all large firms are expected to be included in this survey, and their information is anticipated to be updated frequently. Consequently, for the tail part of the distribution of firm sizes, which is of interest in this analysis, it is considered that nearly all firms are thoroughly covered.

In the following analysis, several conditions are imposed on my main sample. First, I use the sales revenue of non-consolidated firms as the definition of firm size. Firms for which sales revenue is not available are excluded from the sample. Within the TSR data, there are firms that report earnings multiple times within a year (i.e., the duration of its accounting period is less than 12 months). In this analysis, only firms with an accounting period of 12 months are considered. In addition, we analyze data from 2009 to 2019 (11 years) as my sample period. This period is chosen to avoid the effects of the global financial crisis and the COVID-19 pandemic. As a result of this procedure, for example, the sample size for the 2018 data is 1,305,878. The summary statistics on firm sizes, including other years, are provided in **Table 1**.

Another important variable in my analysis is the firm's growth rate. For analyzing firm growth rates, I use samples derived from the samples for firm size mentioned above, with two additional conditions imposed. The first condition is that the firm's initial size (sales revenue in 2009) exceeds 100 million yen. The reason for this consideration is that if a firm's sales are too small, the fluctuations in the firm's growth rate become too

large, which deviates significantly from the assumption of Gibrat's Law in theoretical analysis. Additionally, this condition also implies that only firms already in existence in 2009 are included in the analysis, and firms established after 2009 are not considered in the analysis of growth rates. The second condition is about the firm's age. When a firm is young, its growth rate tends to fluctuate more than others, which again deviates from the assumptions of Gibrat's Law. Here, I am analyzing firms that were established before 2005 (i.e., firms that are at least 5 years old as of 2009). The firm's size at its inception and its growth rate immediately after being established are considered separately as the initial growth rate (i.e., S_0). This point will be discussed in more detail in Section 4.5.

The summary statistics for the firm growth rates of these samples are provided in **Table 2**. This table shows the summary statistics of one-year growth rates for different years, as well as the growth rate statistics for longer periods with 2009 as the base year. As is evident from these tables, the fluctuations in one-year firm growth rates are very stable from 2009 to 2019. Another point is that as we consider growth rates over longer periods, the range of growth rate fluctuations increases. While this latter point is obvious, it will be analyzed in more detail in the following sections.

4.2 Random walk assumption

Here, I empirical check whether the random walk assumption actually holds. Specifically, I consider whether the growth rates in each year are independent random variables.

Summary statistics of firm growth rates

Data: Tokyo Shoko Research from 2009 to 2019

year	count	mean	sd	q1	median	q3
g_10_09	865032	−0.042	0.317	−0.113	−0.006	0.049
g_11_09	823346	−0.055	0.394	−0.167	−0.020	0.085
g_12_09	781359	−0.052	0.447	−0.193	−0.019	0.123
g_13_09	752629	−0.042	0.492	−0.213	−0.013	0.166
g_14_09	730618	−0.024	0.534	−0.221	0.000	0.213
g_15_09	716049	−0.035	0.571	−0.251	−0.008	0.223
g_16_09	699410	−0.045	0.602	−0.280	−0.015	0.235
g_17_09	680659	−0.039	0.631	−0.290	−0.010	0.262
g_18_09	659954	−0.031	0.658	−0.301	−0.003	0.288
g_19_09	637852	−0.029	0.684	−0.315	0.000	0.310
g_11_10	806454	−0.016	0.310	−0.087	0.000	0.069
g_12_11	759192	−0.003	0.300	−0.073	0.000	0.074
g_13_12	725604	0.003	0.294	−0.065	0.000	0.080
g_14_13	702144	0.011	0.289	−0.057	0.000	0.088
g_15_14	686128	−0.018	0.288	−0.084	0.000	0.058
g_16_15	673702	−0.018	0.285	−0.079	0.000	0.053
g_17_16	657085	−0.005	0.285	−0.061	0.000	0.065
g_18_17	637109	−0.002	0.282	−0.057	0.000	0.065
g_19_18	616613	−0.008	0.278	−0.067	0.000	0.054

Table 2: Summary statistics of firm growth rates. Here, I present the summary statistics for the one-year firm growth rates for different years, as well as the growth rate statistics for longer periods with 2009 as the base year. For example, g_19_09 represents the summary statistics for the growth rate from 2009 to 2019.

In this analysis, we use two rank correlation coefficients: Spearman's ρ and Kendall's τ.[4] These are defined as follows:

$$\rho := \mathrm{Cor}\left[F_1\left(X_1\right), F_2\left(X_2\right)\right],$$
$$\tau := \mathbb{P}\left[\left(X_1 - X_1'\right)\left(X_2 - X_2'\right) > 0\right] - \mathbb{P}\left[\left(X_1 - X_1'\right)\left(X_2 - X_2'\right) < 0\right]$$

Here, F_1, F_2 are the marginal distributions of X_1, X_2 respectively, and $X' := (X_1', X_2')$ is an independent copy of $X := (X_1, X_2)$. As is clear from the definition, Spearman's ρ is the correlation coefficient of the ranks $F_1(X_1), F_2(X_2)$, rather than the random variables themselves. Kendall's τ measures the degree of concordance. Both coefficients take values from -1 to 1, and in the case of independent random variables, they become 0.

The matrices of correlation coefficients for different years are given in **Table 3** and **Table 4**. As the duration between the two growth rate periods increases, the correlation coefficients become closer to 0. In other words, it is unlikely that shocks pushing up the growth rate over the long term are at work, and shocks from the distant past have little impact on the current growth rate. Also, as shown in these tables, the absolute values of the coefficients of growth rates over two consecutive periods are less than 0.1. This suggests that growth rates for different years can be considered as independent random variables, meaning that the random walk assumption provides a good approximation.

[4]The most widely used coefficient for quantifying the dependence between two variables is probably Pearson's correlation coefficient, which is defined as follows:

$$\rho_{X_1,X_2} = \frac{\mathbb{E}\left[\left(X_1 - \mu_{X_1}\right)\left(X_2 - \mu_{X_2}\right)\right]}{\sigma_{X_1}\sigma_{X_2}}$$

Here, μ_X and σ_X are the expectation and the standard deviation of X, respectively.

However, there are several statistical issues with using Pearson's correlation coefficient, especially in our analysis. First, Pearson's correlation coefficient is an appropriate measure of dependence if the distribution of random variables under consideration are multivariate Gaussian distributions because, in that case, the dependence between the random variables is fully captured by the Pearson's correlation coefficient alone. But, as in our case, when the marginal distributions significantly deviate from a Gaussian distribution, we cannot use this interpretation. This is because Pearson's correlation coefficient captures not only correlation relationships but also has properties dependent on marginal probabilities. To put it more precisely, consider a bivariate distribution, which can be decomposed as follows:

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$$

Here, F_1, F_2 are the marginal distributions of each variable, and the function C (called copula function) determines all the dependence between variables. Pearson's correlation coefficient depends on not only the copula C but also the marginal distributions as well. For example, even if C remains unchanged, Pearson's correlation coefficient can change simply by changing the marginal distributions, making it unsuitable for comparison based on the value of Pearson's correlation coefficient (for copula theory, see Joe (2014)).

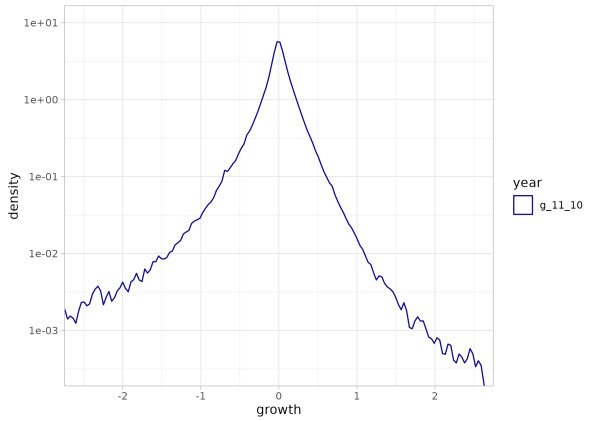
Another more practical problem with Pearson's correlation coefficient is its vulnerability to extremely large values. By its definition, Pearson's correlation coefficient captures the linear relationship between random variables, and therefore, in cases like ours where extremely large values occur, the coefficient is heavily dependent on those samples. In particular, since these extremely large values (i.e., high growth) are of our main interest, it is not appropriate to exclude them as outliers from the analysis. When the tails of the marginal distributions are thick, a more robust dependence measure against extreme values is needed.

Matrix of correlation coefficients Spearman's rank correlation coefficient										
term	g_19_18	g_18_17	g_17_16	g_16_15	g_15_14	g_14_13	g_13_12	g_12_11	g_11_10	g_10_09
g_19_18	NA	−0.122	0.038	0.048	0.046	0.050	0.041	0.047	0.033	0.001
g_18_17	−0.122	NA	−0.108	0.034	0.043	0.063	0.040	0.037	0.050	0.041
g_17_16	0.038	−0.108	NA	−0.130	0.033	0.058	0.055	0.035	0.045	0.058
g_16_15	0.048	0.034	−0.130	NA	−0.112	0.036	0.038	0.041	0.034	0.029
g_15_14	0.046	0.043	0.033	−0.112	NA	−0.106	0.036	0.044	0.041	0.025
g_14_13	0.050	0.063	0.058	0.036	−0.106	NA	−0.101	0.053	0.058	0.034
g_13_12	0.041	0.040	0.055	0.038	0.036	−0.101	NA	−0.087	0.043	0.050
g_12_11	0.047	0.037	0.035	0.041	0.044	0.053	−0.087	NA	−0.088	0.015
g_11_10	0.033	0.050	0.045	0.034	0.041	0.058	0.043	−0.088	NA	−0.089
g_10_09	0.001	0.041	0.058	0.029	0.025	0.034	0.050	0.015	−0.089	NA

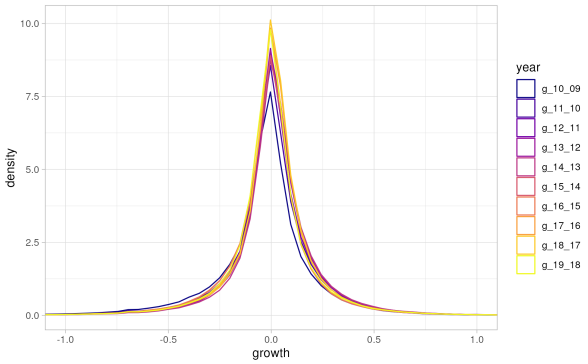
Table 3: Matrix of Spearman's ρ. Samples are the same as in **Table 2**.

Matrix of correlation coefficients Kendall's rank correlation coefficient										
term	g_19_18	g_18_17	g_17_16	g_16_15	g_15_14	g_14_13	g_13_12	g_12_11	g_11_10	g_10_09
g_19_18	NA	−0.082	0.030	0.031	0.030	0.038	0.034	0.027	0.028	0.010
g_18_17	−0.082	NA	−0.089	0.029	0.013	0.040	0.033	0.033	0.024	0.033
g_17_16	0.030	−0.089	NA	−0.114	0.002	0.036	0.072	0.052	0.037	0.040
g_16_15	0.031	0.029	−0.114	NA	−0.104	0.011	0.033	0.034	0.022	0.000
g_15_14	0.030	0.013	0.002	−0.104	NA	−0.087	0.007	0.006	0.020	0.019
g_14_13	0.038	0.040	0.036	0.011	−0.087	NA	−0.068	0.043	0.019	0.024
g_13_12	0.034	0.033	0.072	0.033	0.007	−0.068	NA	−0.040	0.052	0.033
g_12_11	0.027	0.033	0.052	0.034	0.006	0.043	−0.040	NA	−0.072	0.035
g_11_10	0.028	0.024	0.037	0.022	0.020	0.019	0.052	−0.072	NA	−0.083
g_10_09	0.010	0.033	0.040	0.000	0.019	0.024	0.033	0.035	−0.083	NA

Table 4: Matrix of Kendall's τ. Samples are the same as in **Table 2**.



(a) For 2019-2018



(b) For different years

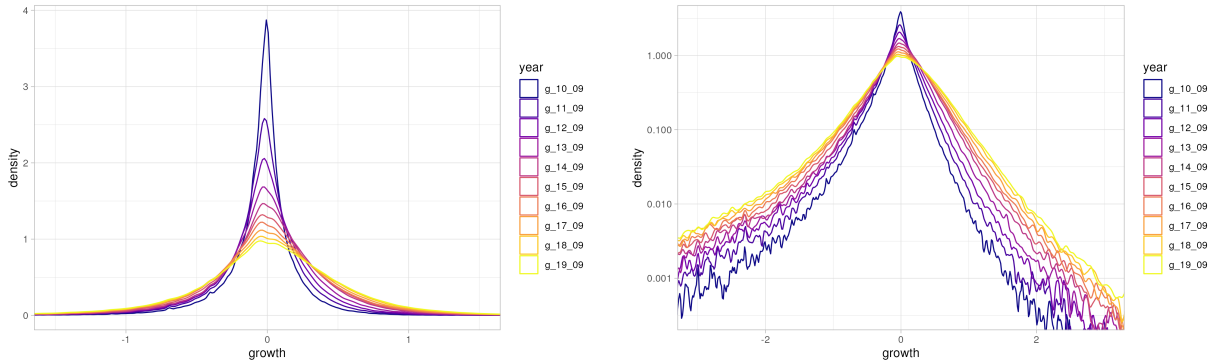
Figure 5: Growth rate distributions.

4.3 Growth rate distribution

In this subsection, I examine the second assumption in my analysis, i.e., whether the distribution of firm growth rates follows a subexponential distribution, specifically a distribution with a Weibull tail. First, I provide the density estimates for one-year growth rates. The results are shown in **Figure 5**. As can be seen in **Figure 5** (a), consistent with previous studies, the distribution deviates from the Gaussian distribution and has a peak at the center and heavy tails. Although previous studies often used the Laplace distribution for the approximation of the growth rate distribution, this figure shows that the tails curve rather than being straight lines, indicating that the tails are heavier than an exponential. This suggests that the distribution is subexponential. Furthermore, **Figure 5** (b) compare the density estimates for different years. As seen in this figure, the shape of the distribution is stable across years.

What is more relevant to my analysis is the distribution of long-term growth rates. I investigate how the growth rate distribution changes as longer periods are considered. **Figure 6** shows the density estimates of the growth rate distributions over k periods ($k = 1, 2, ..., 10$). Both graphs shows the same density estimates, but it has a log-scaled y-axis in the right panel (b). As can be clearly seen from **Figure 6** (a), as k increases, the density around 0 approaches a bell-shaped curve. This is expected by the central limit theorem, and it indicates that the density converges to a Gaussian as $k \to \infty$. However, this bell-shaped curve occurs only around the central region, and outside this region, a different shape emerges. In particular, **Figure 6**(b) shows that in the log scale, the density in the right tail zone has a shape close to a straight line. Additionally, as k increases, the straight line appears to shift upward in parallel.

This feature is consistent with the shape described in Section 3.2: that is, as k increases, the Cramer approximation holds in the central part. Moving further outward from that region, the slope of the density function (in the log scale) is determined by the tail probability of one-year growth rate (i.e., $\mathbb{P}(X_1 > u)$ for a large u), and an increase in k only shifts the intercept in parallel, while keeping the slope unchanged.



(a) Normal scale **(b)** Log scale

Figure 6: Growth rate distributions.

The remainder of this subsection uses statistical methods to verify that the growth rate distribution is subexponential, particularly that it has a tail close to a Weibull tail. First, I use the mean excess function over threshold u (denoted as $e(u)$), which is defined as follows (for details, see Embrechts et al. (1997)):

$$e(u) := E[X_k - u \mid X_k > u] \quad \text{for } u > 0.$$

That is, $e(u)$ represents the conditional expectation of overshoot $X_k - u$ given that X_k exceeds u.

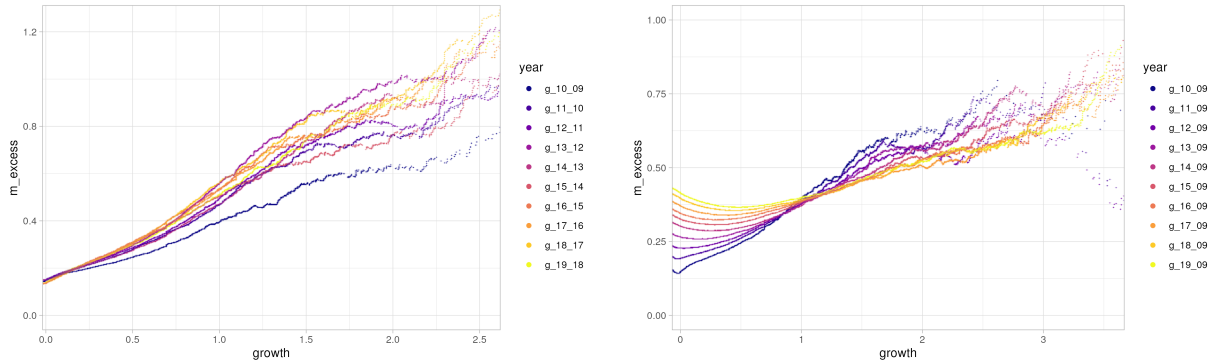
The reason for using the mean excess function $e(u)$ is that it can reveal the tail-heaviness of the distribution of X_k based on whether $e(u)$ is an increasing or decreasing function of u. In particular, if the distribution of X_k is an exponential distribution with parameter λ, then $e(u) = \lambda^{-1}$, i.e., $e(u)$ is constant. Therefore, if $e(u)$ is an increasing function of u, it indicates that the distribution of X_k has a heavier tail than an exponential distribution, which can be used as an evidence that the growth rate distribution is subexponential.

There are two examples for the functional form of the mean excess function relevant to our analysis. The first example is the case where the tail of the growth rate distribution follows a Pareto tail, and in this case, this mean excess function becomes a linear function of u. The second example is when the firm growth rate distribution has a Weibull tail. As u becomes larger, the mean excess function is known to take the following functional form:

$$e(u) = \frac{u^{1-\alpha}}{c\alpha}(1 + o(1))$$

as $u \to \infty$ (cf. Table 3.4.7 in Embrechts et al. (1997)). Depending on which functional form it is closer to, we can identify which class of distributions among subexponential distributions is more appropriate to approximate the tail of the growth rate distribution.

The results of the empirical mean excess function are given in **Figure 7**. In **Figure 7** (a), the mean excess function is calculated using the one-year firm growth rates for each year. As is evident from this figure, the mean excess function is an increasing function of u for each year. Moreover, while the mean excess function is an increasing function of u, its slope is decreasing as the value of u increases. This indicates that



(a) One-year growth rates

(b) k-year growth rates

Figure 7: Mean excess function over threshold u.

the growth rate distribution has a heavier tail than an exponential distribution, and is closer to a Weibull tail rather than a Pareto tail.

In **Figure 7** (b), the mean excess function for growth rates over $k(k = 1, 2, ..., 10)$ years is calculated, using 2009 as the base year. For every k, the mean excess function is an increasing function of u (especially in the tail part), and roughly speaking, these mean excess functions are close to each other (i.e., they do not depend on the value of k). This is because, as a property of subexponential distributions, the tail part of the distribution for k periods is determined by the tail probability of the one-year growth rate.

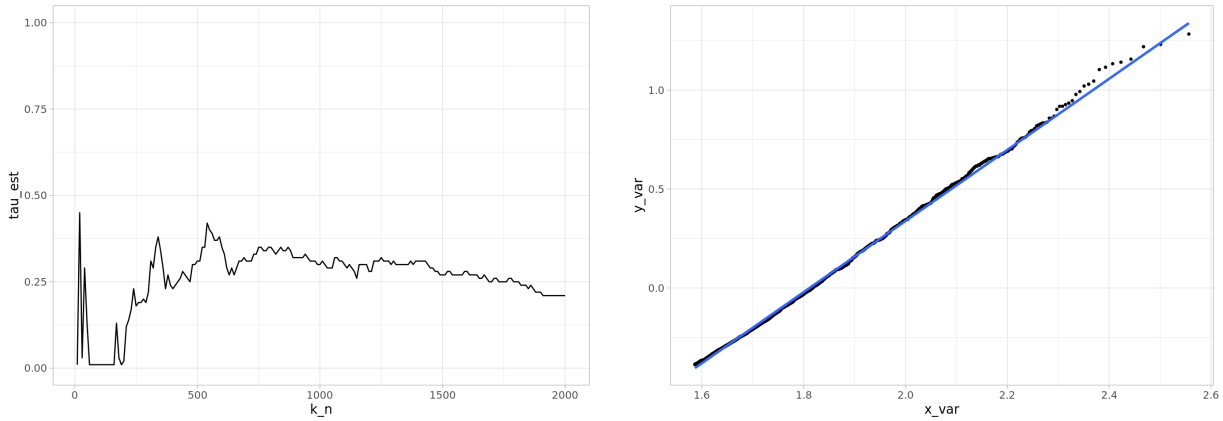
The second statistical method for analyzing the tail of the growth rate distributions is the one proposed by Gardes et al. (2011); El Methni et al. (2012). As discussed above, within the family of subexponential distributions, there exist two groups: distributions with a Pareto tail and ones with a Weibull tail. Our particular interest is to examine which of these two tails better approximate the tail of the growth rate distribution. Their method allows us to statistically verify which of the two is better approximation.

More precisely, letting $K_x(y) = \int_1^y u^{x-1}du$ for $x \in \mathbb{R}$, consider a family of survival distributions with two parameters $\tau \in [0, 1]$ and $\theta > 0$ defined by

$$\overline{F}(x) = \exp(-K_\tau^{\leftarrow}(\log H(x))) \quad \text{for } x \geq x_* > 0, \text{ with } \tau \in [0, 1]$$

where H is an increasing function such that $H^{\leftarrow} \in \mathcal{R}_\theta$ and $\theta > 0$. Here, $\mathcal{R}_\theta$ stands for the regularly varying functions with parameter θ. Parameter τ represents the distribution classes ranging from Weibull-type tail ($\tau = 0$) to Pareto tail ($\tau = 1$). Intuitively, a larger value of τ means a heavier tail of the distribution. Parameter θ is the shape parameter for each distribution tail corresponding to τ; for example, when $\tau = 0$, θ coincides with Weibull-tail coefficient β.

This estimation method consists of two parts. First, I estimate the parameter τ, and based on this estimated value, then I estimate the shape parameter θ of the tail part. Using this method, I can check whether the tail of the growth rate distribution follows a Weibull tail or a Pareto tail, i.e., the two important groups of subexponential distributions. Furthermore, if it belongs to a Weibull tail, I examine whether the



(a) Estimate of τ **(b)** Linear relationship

Figure 8: Estimate of τ

shape parameter θ is less than 1 (i.e., the case of the exponential case) to confirm the assumption made in my theoretical analysis.

The results of the estimation for τ are presented in **Figure 8**. As shown in the figure, most of the estimated values are close to 0, suggesting that the Weibull tail provides a good approximation. These results are consistent with the shape of the empirical mean excess function discussed above.

Additionally, in the case of $\tau = 0$, which corresponds to a Weibull tail, one should expect to see the following linear relationship:

$$\begin{aligned} \log x_u - \log x_v &\simeq \theta \left(K_\tau(-\log u) - K_\tau(-\log v) \right) \\ &= \theta \left(\log(-\log u) - \log(-\log v) \right) \end{aligned}$$

Here, x_u and x_v are the u and v-quantile values of growth rates, respectively. The empirical verification of this relationship can be seen in **Figure 8**(b). As evident from the figure, the relationship aligns with a straight line, supporting the assumption that the tail is approximated by a Weibull tail.

Finally, under the assumption that the tail part follows a Weibull tail (i.e. $\tau = 0$), I estimate the shape parameter θ and calculate the tail probability from it. Then, I compare the tail probability of the actual growth rates to see how well the Weibull tail assumption approximates it. The results are presented in **Figure 9**. It can be said that the estimated tail probability reasonably well approximates the tail probability of the firm growth rates. Therefore, these results support the assumption in my analysis that the growth rate distribution follows a Weibull tail.

4.4 Initial size

Up to this point, I have considered only the growth rates $X_1, X_2, ..., X_n$, assuming the initial size S_0 to be $S_0 = 0$. Also in previous studies, the initial size S_0 is often assumed to be constant and small, and it is often assumed that the initial size does not affect the shape of the limiting distribution of firm sizes as

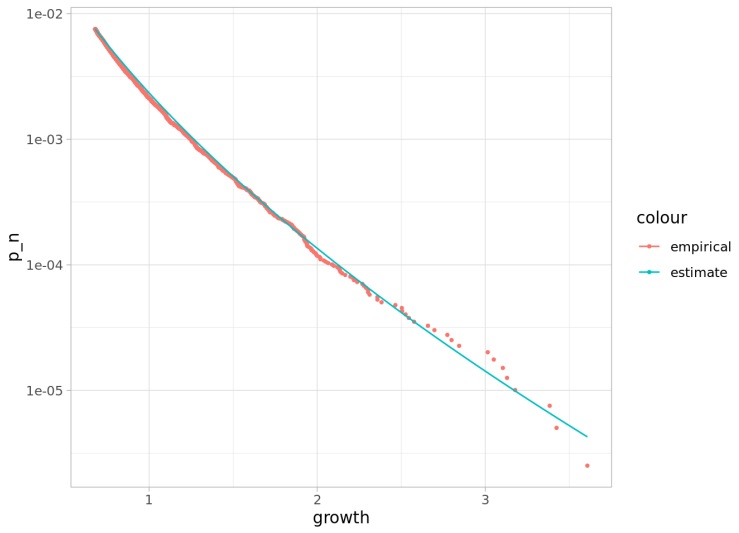


Figure 9: Estimate of the tail probability of firm growth rates. It compares the estimated tail probability based on Gardes et al. (2011) with counter cumulative distribution functions of growth rates.

$n \to \infty$. However, as shown in this subsection, there are firms that are positioned in the tail region of the firm size distribution even immediately after their establishment. Since younger firms generate Zipf's law, the effect of the distribution of S_0 on the overall firm size distribution cannot be ignored.

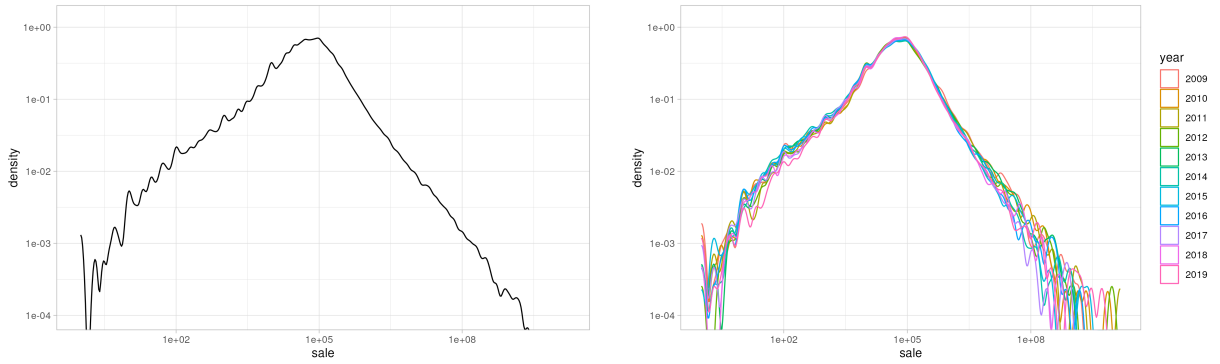
As a premise for the discussion, I explain what "initial size" (i.e., S_0) means in my analysis. By definition, the "initial size" is thought of as the scale (sales) of a firm when it was established. However, since the TSR data used in this analysis is based on surveys, there is a possibility that firms newly established might not be included in the database. In my analysis, to increase its coverage, instead of considering the size of a firm immediately after its establishment, I decided to include the firm's growth over the first five years as part of the initial size, denoted as S_0. For instance, if a firm is established in 2005, the size of the firm in the nth period, S_n, would be as follows.

$$S_n = \underbrace{\log(\text{sales revenue in 2005}) + g_{06} + g_{07} + g_{08} + g_{09}}_{S_0} + \underbrace{g_{10}}_{X_1} + \ldots + \underbrace{g_{19}}_{X_{10}}$$

Here, g_t represents the growth rate in the tth period, and both $\log(\text{sales revenue in 2005})$ and $g_{06} + g_{07} + g_{08} + g_{09}$ are included in S_0. The distribution of S_0 will be analyzed below.[5]

The density estimate for the initial size S_0 is provided in **Figure 10**. As can be seen from this figure, the heterogeneity of the initial size S_0 is substantial. Despite being less than five years old, some firms

[5] Another issue to consider when thinking about firms initial sizes is the case where a firm's establishment results from a reorganization of a corporate group. For instance, there might be situations where a new firm is established by transferring the businesses of multiple firms belonging to a certain corporate group, and this transfer of business operations could span several years. Treating the apparent growth of a firm due to such business transfers the same way as the growth of a firm at other times is not appropriate for the analysis here. To mitigate the impact of these data issues, a firm's growth from the first five years has been included in S_0.



(a) For 2009 **(b)** From 2009 to 2019

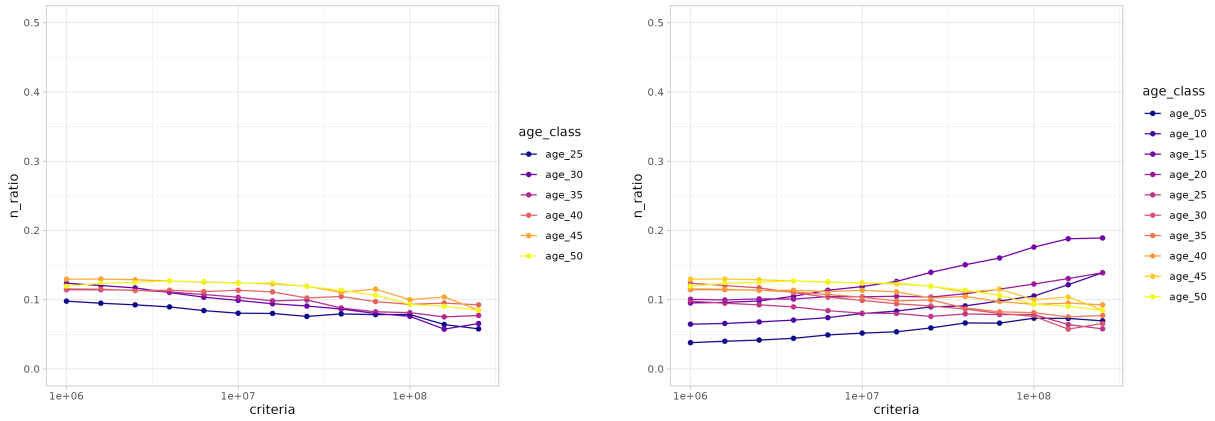
Figure 10: In Panel (b), the density estimates of S_0 for different years are given.

are situated in the tail region of the overall firm size distribution, comprising the Zipf's law. Such high heterogeneity indicates that when analyzing the firm size distribution, one cannot assume the initial size S_0 to be constant. In addition, the tail part of the distribution of S_0 is close to a straight line, confirming that it follows a Pareto tail. When calculated using Hill's method for its slope, the slope is found to be 0.872, indicating that it has a thicker tail than the overall firm size distribution. This also suggests that the distribution (or heterogeneity) of S_0 is an element that cannot be ignored when analyzing the distribution of firm sizes.

4.5 Further empirics

Building on the analysis from Section 4.2 to Section 4.4, we can analyze how the shape of the firm size distribution changes with firm age. Indeed, as explained below, this matches the change in the shape of the distribution observed in **Figure 3**. When n is small, as discussed in Section 3.3, the tail of the firm size distribution is determined by the tail of the distribution of the initial size S_0 and the distribution of growth rates in the intermediate zone and the extreme deviation zone. In particular, when viewed on a log-scale, the slope of the firm size distribution is determined by the weighted average of these two slopes. As n increases, the firm size distribution in the log-scale changes in such a way that its slope remains unchanged, but its intercept rises, i.e., essentially a parallel upward shift. Furthermore, as n becomes larger, the domain of the Cramer approximation expands, and the impact of the initial size diminishes, causing the firm size distribution closer to a Gaussian distribution.

This is the underlying mechanism of forming the firm size distribution and explains how the shape of the firm size distribution, as observed in **Figure 3**, depends on firm age. Notably, this demonstrates that the Pareto tail of the firm size distribution, known as Zipf's law, is explained by groups of younger firms. In the remainder of this section, I will provide further empirical evidence to show that the aforementioned explanation is consistent with the data.



(a) Age cohort from 25 to 50

(b) Age cohort from 10 to 50

Figure 11: Composition of age cohorts in the tail zone.

According to the explanation above, the slope of the distribution of firm sizes by age, especially in the range where the influence of the distribution of S_0 is weak, is determined by the tail probabilities of the growth rate distribution. Specifically, since the differences in the distribution by firm age, when measured in log-scale, involve only a vertical shift upwards in the y-intercept, the proportion in the tail zone explained by an age cohort should not depend on the firm size (see Section 2.1). This property is verified in **Figure 11**. As is clear from this figure, the age composition of firms in the tail zone is independent of firm size, and the ratio remains stable. This result is consistent with my explanation, contrasting with the traditional explanation which suggests that older firms dominate more as we consider the tail part.

Specifically, **Figure 11** (b) takes into account even younger age groups of firms. As can be seen from this figure, when considering larger sizes in the tail zone, the proportion of younger firms increases. This is because, as observed in Section 4.4, the slope of the tail of the distribution of S_0 is steeper than the slope of the growth rate distribution, implying that when considering a larger size range, their proportion rises. This result also serves as evidence that the younger age cohort of firms plays a crucial role in explaining Zipf's law.

Lastly, to verify that a few large jumps determine the tail of the growth rate distribution, we compare the growth rate caused by these rare jumps with the actual growth rate distribution. That is, as seen in **Figure ??**, we consider only growth rates above the 0.97-quantile value (setting all others to 0) and compare whether this growth rate approximates the tail of the actual growth rate well. In other words, we verify whether the growth rate due to jumps, represented by

$$\mathbf{1}_{\{X_1>F_1^{-1}(0.97)\}}X_1 + \mathbf{1}_{\{X_2>F_2^{-1}(0.97)\}}X_2 + \ldots + \mathbf{1}_{\{X_n>F_n^{-1}(0.97)\}}X_n$$

can approximate the distribution of $\sum_{k=1}^{n} X_k$.

The result is given in **Figure 12**. Here, we consider the growth rates from 2009 to 2019 (i.e., $n = 10$) as reference. As indicated by this figure, the tail of the growth rate distribution for $n = 10$ is well approximated

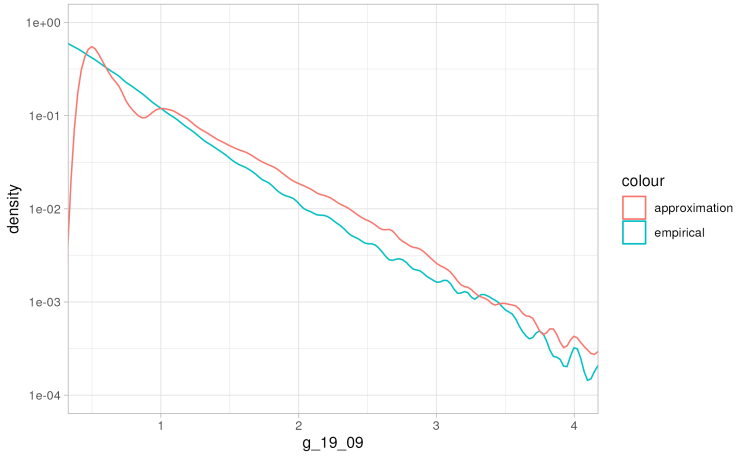


Figure 12: Approximation by a few large jumps. Growth rates from 2009 to 2019 are considered.

by the distribution tail caused by jumps. In other words, rather than achieving high growth over 10 years by continuously realizing moderately large growth rates multiple times, the growth rate for $n = 10$ is determined by a few extremely large jumps. This aligns with the assumptions considered in our previous analysis and explains why even young firms can exist in the tail of the firm size distribution.

5 Conclusion

What was carried out in this analysis is explained by two demonstrable hypotheses related to Zipf's Law. One is the hypothesis of a random walk, and the other is the hypothesis that the probability distribution of corporate growth rates follows a heavy-tailed distribution. Especially the latter implies that the process of corporate growth is determined by jumps or by a process where a few jumps can lead to a significant increase. Using company-level data that covers the entirety of Japan's companies, these two hypotheses are shown to hold true, and it is demonstrated that Zipf's Law can be explained by these two hypotheses.

A significant contribution of this study lies in its explanation of Zipf's Law without assuming a stationary distribution. In most previous research cases, Zipf's Law has been presumed to result from a stationary distribution, implying that larger, more mature companies explain the phenomenon. However, this assumption presented challenges, as older companies with higher ages were expected to be the tail-end large companies that would converge towards a stationary distribution over time. Contrarily, with real data, it becomes apparent that younger companies are more central to Zipf's Law, and older companies actually deviate from the Zipf's Law pattern. What this analysis reveals is that Zipf's Law is not a consequence of a slowly converging stationary distribution over a long time. Instead, it emerges as a result of younger companies experiencing rapid growth within a short span.

References

Arata, Y., Miyakawa, D., and Mori, K. (2023). The U-shaped law of high-growth firms. *NTC Discussion Paper*, 230201-01HJ(2).

Axtell, R. L. (2001). Zipf distribution of us firm sizes. *science*, 293(5536):1818–1820.

Beare, B. K. and Toda, A. A. (2022). Determination of pareto exponents in economic models driven by markov multiplicative processes. *Econometrica*, 90(4):1811–1833.

Borovkov, A. A. and Borovkov, K. (2008). *Asymptotic analysis of random walks*, volume 118. Cambridge University Press.

Bottazzi, G., Coad, A., Jacoby, N., and Secchi, A. (2011). Corporate growth and industrial dynamics: Evidence from french manufacturing. *Applied Economics*, 43(1):103–116.

Coad, A. (2009). *The growth of firms: A survey of theories and empirical evidence*. Edward Elgar Publishing.

Dosi, G., Grazzi, M., Moschella, D., Pisano, G., and Tamagni, F. (2020). Long-term firm growth: an empirical analysis of us manufacturers 1959–2015. *Industrial and Corporate Change*, 29(2):309–332.

Dosi, G., Pereira, M. C., and Virgillito, M. E. (2017). The footprint of evolutionary processes of learning and selection upon the statistical properties of industrial dynamics. *Industrial and Corporate Change*, 26(2):187–210.

El Methni, J., Gardes, L., Girard, S., and Guillou, A. (2012). Estimation of extreme quantiles from heavy and light tailed distributions. *Journal of Statistical Planning and Inference*, 142(10):2735–2747.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media.

Gabaix, X. (2009). Power laws in economics and finance. *Annu. Rev. Econ.*, 1(1):255–294.

Gardes, L., Girard, S., and Guillou, A. (2011). Weibull tail-distributions revisited: a new look at some tail estimators. *Journal of Statistical Planning and Inference*, 141(1):429–444.

Joe, H. (2014). *Dependence modeling with copulas*. CRC press.

Kondo, I. O., Lewis, L. T., and Stella, A. (2023). Heavy tailed but not zipf: Firm and establishment size in the united states. *Journal of Applied Econometrics*.

Kotz, S., Kozubowski, T., and Podgórski, K. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Number 183. Springer Science & Business Media.

Luttmer, E. G. (2007). Selection, growth, and the size distribution of firms. *The Quarterly Journal of Economics*, 122(3):1103–1144.

Luttmer, E. G. (2010). Models of growth and firm heterogeneity. *Annu. Rev. Econ.*, 2(1):547–576.

Luttmer, E. G. (2011). On the mechanics of firm growth. *The Review of Economic Studies*, 78(3):1042–1068.

Petrov, V. V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Clarendon Press.

Stanley, M. H., Amaral, L. A., Buldyrev, S. V., Havlin, S., Leschhorn, H., Maass, P., Salinger, M. A., and Stanley, H. E. (1996). Scaling behaviour in the growth of companies. *Nature*, 379(6568):804–806.