The Universality of Zipf's Law for Random Growth Processes

Ricardo T. Fernholz[†]
Claremont McKenna College

Robert Fernholz[‡]
Intech

January 20, 2018

Abstract

We provide necessary and sufficient conditions for systems of random growth processes to generate an asymptotic size distribution that satisfies Zipf's law. This characterization rigorously establishes an important conjecture in the literature—for a system of random growth processes that follows the strong form of Gibrat's law, Zipf's law obtains in the limit as the friction needed to ensure stationarity of the system vanishes. Our result is achieved using general rank-based methods, and is valid for essentially all random growth models regardless of the specific frictions that are present. We generalize Zipf's law to a less restrictive *quasi-Zipfian* form in which a log-log plot of size versus rank is not necessarily a straight line with slope -1, but rather is concave with a tangent line of slope -1 at some point. Under certain regularity conditions, we show that systems of random growth processes that deviate from Gibrat's law in a specific but realistic manner generate an asymptotic size distribution that is quasi-Zipfian. Because many real-world systems that follow the strong form of Gibrat's law satisfy Zipf's law, and even more systems that do not follow the strong form of Gibrat's law are quasi-Zipfian, our results explain the universality of Zipf's law for random growth processes.

JEL Codes: C14, C46, C60

Keywords: Zipf's law, power law, Pareto distribution, Gibrat's law, nonparametric methods

We thank Ioannis Karatzas, members of the Intech SPT seminar, and participants of the 2017 Thera Stochastics Conference for their comments and suggestions regarding this research. Any errors or omissions are our own.

[†]Claremont McKenna College, 500 E. Ninth St., Claremont, CA 91711, rfernholz@cmc.edu.

[‡]Intech, One Palmer Square, Princeton, NJ 08542, bob@bobfernholz.com.

1 Introduction

A system of random variables follows a power law, or Pareto distribution, if the log-log plot of their values versus rank forms a straight line. Power laws are ubiquitous in the social and natural sciences (Bak, 1996). They are found across many different phenomena, ranging from the frequency of words in a language (Zipf, 1935) to the income and wealth of households (Atkinson et al., 2011; Piketty, 2014) to the market capitalization of stocks (Simon, 1955; Fernholz, 2002).

A system of random variables satisfies Zipf's law, a special case of a power law, if the log-log plot of their values versus rank forms a straight line with slope -1. Many power laws in the social and natural sciences closely conform to Zipf's law, with two of the best-known examples being the distributions of city populations (Gabaix, 1999) and firm employees (Axtell, 2001). Furthermore, many power laws that do not appear to satisfy Zipf's law, such as income and wealth distributions, feature log-log plots of value versus rank that are concave and have a tangent with slope -1 somewhere. The striking empirical regularity of Zipf's law has prompted many researchers to attempt to explain it (Simon, 1955; Steindl, 1965; Gabaix, 1999).

In this paper, we use rank-based methods to characterize necessary and sufficient conditions for general systems of random growth processes to generate a size distribution that satisfies Zipf's law. Our approach is unconventional in that we neither commit to a specific economic model nor do we impose a specific friction on the systems of random growth processes that we analyze. Such a general, model-free approach, however, is essential to any explanation of Zipf's law. Indeed, the observed universality of Zipf's law, which pertains to many different real-world phenomena including the original Zipf's law for the frequency of words in English (Zipf, 1935), means that any explanation should not rely on model-specific details that are only appropriate for certain applications, a point that is emphasized by Gabaix (1999). The results we derive using our rank-based methods satisfy this requirement.

We refer to the objects that make up the systems of random growth processes that we analyze as families, and to the objects that are contained within these families as members. These members within families may take many different forms—people in cities, employees, assets, or capital in firms, income or wealth in households, or occurrences in words. If a system of random growth processes follows the strong form of Gibrat's law (Gibrat, 1931), with growth rates and volatilities that do not vary across different sized families, we show that Zipf's law is equivalent to the system being both *conservative* and *complete*. The first condition, conservation, requires that the expected total size of all families (the total number of family members) is constant over time—the system conserves the "mass" inside it. The second condition, completeness, requires that the expected total size of all families not be dependent on an influx of families from outside the system. In other words, the system must be complete on its own and not depend on some unmodeled phenomena for its growth over time. The first condition must, by construction, hold for a fixed random sample of family members. Since a large random sample should have about the same distribution as the full population, any system that can be sampled should be conservative. The second condition is likely to hold provided that a system includes a sufficient number of families. Thus, our results imply that any economic model of random growth that includes a sufficient number of ranks and follows the strong form of Gibrat's law should be *Zipfian*—the model generates a size distribution that satisfies Zipf's law.

Up to now, the literature on random growth processes and Zipf's law has assumed the existence of a specific friction in order to ensure stationarity. Gabaix (1999, 2009) and Toda (2017), for example, show that

Zipf's law emerges from a geometric Brownian motion that is either reflected near zero, or for which there is a Poisson birth-death process with a rate approximately equal to zero. A similar result has been described in discrete time using "Kesten processes" (Kesten, 1973). These results, however, assume the existence of some specific friction whose disappearance gives rise to Zipf's law.

In contrast to this literature, our result on the universality of Zipf's law for large systems that follow the strong form of Gibrat's law does not commit to or require any specific frictions. We adopt a general econometric approach and model the random growth processes that describe family member dynamics as general Itô processes that are well-approximated by rank-based stochastic processes. These rank-based processes represent the higher ranks of a larger system of random growth processes, so that any friction that may be needed to stabilize the system as a whole occurs outside the model. This means that the specific friction that ensures nondegeneracy and stationarity of these systems is unimportant for our results. In this way, our results confirm the hypothesis of Gabaix (1999, 2009) that for random systems that follow Gibrat's law, Zipf's law emerges in the limit as all frictions acting upon that system vanish. To our knowledge, this is the first paper to rigorously establish this result about the friction-invariant universality of Zipf's law.

While Zipf's law describes the shape of many power laws in the social and natural sciences, there are a number of power laws that do not conform to its strictest form. In particular, many phenomena in the natural and social sciences feature log-log plots of size versus rank that are not straight lines but rather concave curves. Furthermore, practically all of these concave curves have a tangent with slope -1 somewhere. A clear example of this can be seen in Figure 1, which plots the total market capitalization of the largest U.S. stocks relative to all stocks averaged over different ten-year intervals from 1929-1999. The figure shows concave curves that are flatter than -1 at the highest ranks of the distribution over each of these ten-year subperiods. Although these concave curves do not satisfy Zipf's law in any subperiod, it is apparent that each of the curves has a tangent at -1 somewhere.

The same basic pattern emerges for the distributions of firm employees and city populations. Indeed, Figure 1 from Axtell (2001) and Figure 10 from Eeckhout (2004) show concave log-log plots of, respectively, frequency versus U.S. firm sizes and rank versus U.S. city sizes.[1] Although the concavity of these two plots is less pronounced than in Figure 1, it is apparent enough that both authors acknowledge the deviations from a straight line. In this paper, we show that the concave plots shown in our Figure 1, Figure 1 from Axtell (2001), and Figure 10 from Eeckhout (2004) in fact represent an important generalization of Zipf's law that is usually driven by a specific deviation from the strong form of Gibrat's law.

To this end, we generalize the notion of Zipf's law to the less restrictive *quasi-Zipf's law*, in which a plot of the log-values of a system of random variables versus their log-rank need not be a straight line of slope -1, but rather may be concave with a tangent line of slope -1 at some point. We then show that the same two conditions that ensure that Gibrat's law implies Zipf's law—conservation and completeness—also ensure that a specific but realistic deviation from Gibrat's law implies quasi-Zipf's law. In particular, we use results from Banner et al. (2005) to show that a system of random growth processes with growth rates that do not vary across different ranks and volatilities that are monotonically decreasing in family size must, in almost all cases, be *quasi-Zipfian*, provided that the system is conservative and complete. Of course, as discussed earlier, the conditions of conservation and completeness should hold for any real-world system that can be sampled and that includes a sufficient number of ranked observations. Therefore, in the same way

[1]Even though frequency is not the same as rank, the same basic straight-line pattern still represents Zipf's law. In particular, a log-log plot of frequency versus size with slope -2 corresponds to Zipf's law. See Axtell (2001) or Gabaix (2016).

that our results show that essentially all systems of random growth processes that follow the strong form of Gibrat's law should be Zipfian, our results also show that many empirically realistic systems that deviate from Gibrat's law should be quasi-Zipfian.

Both Zipf's law and quasi-Zipf's law represent a form of universality, since systems across both the natural and social sciences frequently follow these distributions. This universality has led many researchers to try to explain it, especially in the case of Zipf's law. Our results are the first to characterize the conditions under which systems of random growth processes will be Zipfian and quasi-Zipfian without restricting ourselves to a specific economic model or imposing a specific friction. The general rank-based stochastic processes that we consider in this paper can reproduce the dynamics and the size distribution of any economic model of random growth with a power law in the upper tail. In fact, these rank-based processes can reproduce the empirical distributions of any stationary systems of random growth processes that can be ranked, and thus our results are applicable to such general systems as well. Because the conditions under which Zipf's law and quasi-Zipf's law obtain are satisfied by many real-world systems, our results offer an explanation for the universality of both forms of Zipf's law.

This paper is organized as follows. Section 2 defines conservation and completeness for systems of random growth processes and then discusses the meaning of these conditions. Section 3 presents results that characterize the stationary size distributions of general systems of random growth processes, and then uses these results to provide necessary and sufficient conditions for Zipf's law. Section 4 uses rank-based systems of random growth processes to show that the strong form of Gibrat's law is equivalent to Zipf's law and that systems that deviate from Gibrat's law in a specific but realistic manner are quasi-Zipfian. Section 5 concludes. All proofs appear in the Appendix.

2 A Rank-Based Approach

Consider a system of positive-valued time-dependent data $\{Z_1(t), Z_2(t), \ldots\}$ of indefinite but finite and potentially time-varying size. These data represent two classes of objects, *members* and *families*. The members are contained within the families, with $Z_i(t)$ indicating the number of members belonging to the i-th family at time t. These members within families can take the form of people belonging to cities, income or wealth belonging to households, employees, assets, or capital belonging to firms, or occurrences belonging to words, as well as many other potential applications.

2.1 Setup

In order to model the top ranks of these positive-valued time-dependent data, we consider systems of random growth processes $X_1, \ldots, X_N$ that are represented by positive Itô processes. Thus, for all $i = 1, \ldots, N$, the random growth process $X_i > 0$ can be written

$$d \log X_i(t) = \gamma_i(t)\, dt + \sum_{z=1}^{M} \delta_{iz}(t)\, dB_z(t), \tag{2.1}$$

where γ_i and δ_{iz}, $z = 1, \ldots, M \geq N$, are measurable processes that are adapted to the filtration generated by the Brownian motion $\mathbf{B}(t) = (B_1(t), \ldots, B_M(t))$. Any plausible economic model of random growth that is continuous can be represented in the form (2.1), since Itô processes are a broad class of stochastic processes that encompass such models. Indeed, the $M \geq N$ sources of randomness allow for a rich structure of time-varying idiosyncratic, correlated, and aggregate shocks that need not conform to any particular distribution. Furthermore, (2.1) also allows for growth rates and volatilities that vary across individual families based on any characteristics.

Our approach is unconventional in that we do not restrict ourselves to a specific economic model. Instead, we derive general results that rely on few assumptions, without imposing a particular model. Such a general, model-free approach is essential to any explanation of Zipf's law since it is a form of universality that is found across many different and unrelated phenomena, ranging from the populations of cities (Gabaix, 1999) to the frequency of words in English (Zipf, 1935) to the employees of firms (Axtell, 2001). In this way, our approach is similar to Gabaix (1999) and Toda (2017), both of whom emphasize that any explanation of Zipf's law should not depend on the details of a specific model.

Up to now, it has been shown that stationary size distributions that satisfy Zipf's law emerge in economic models in which family size dynamics all follow geometric Brownian motions with the same parameters. All of these results, however, assume the existence of a specific "friction"—usually either a lower reflecting barrier or a Poisson birth-death process—and then demonstrate that Zipf's law obtains in the limit as that specific friction disappears (Gabaix, 2009; Toda, 2017). Of course, a geometric Brownian motion with a lower reflecting barrier, a Poisson birth-death process, or any other specific friction is just a special case of the more general random growth framework (2.1). It is this generality that allows us to go further than the previous literature by providing necessary and sufficient conditions for the existence of Zipf's law in any economic model of random growth that can be reasonably approximated by Itô processes.

2.2 Rank-Based Dynamics

In order to characterize the conditions under which the stationary distribution of the system of random growth processes $X_1, \dots, X_N$ satisfies Zipf's law, it is necessary to consider family size dynamics by rank. The rank function arises naturally in the context of Zipf's law, since this law is often stated in terms of a restriction on the shape of a log-size versus log-rank plot. In particular, if the slope of a log-log plot of family size versus rank is a straight line, then the distribution of the processes $X_1, \dots, X_N$ is Pareto, and if the slope of this log-log plot is a straight line with slope -1, then this distribution satisfies Zipf's law (Newman, 2005; Gabaix, 1999).

For the system of random growth processes $X_1, \dots, X_N$, we define the rank function r_t such that $r_t(i) < r_t(j)$ if $X_i(t) > X_j(t)$ or if $X_i(t) = X_j(t)$ and $i < j$. The rank processes $X_{(1)} \geq \cdots \geq X_{(N)}$ are defined by $X_{(r_t(i))}(t) = X_i(t)$. If the processes $\log X_i$ satisfy certain regularity conditions, then, for all $k = 1, \dots, N$, the rank processes satisfy,

$$d \log X_{(k)}(t) = \sum_{i=1}^{n} \mathbb{1}_{\{r_t(i)=k\}} \, d \log X_i(t) + \frac{1}{2} d\Lambda_{k,k+1}(t) - \frac{1}{2} d\Lambda_{k-1,k}(t), \quad \text{a.s.,} \tag{2.2}$$

where $\Lambda_{k,k+1}$ is the local time at the origin for $\log X_{(k)} - \log X_{(k+1)}$, with $\Lambda_{0,1} = \Lambda_{N,N+1} \equiv 0$.[2] For each $k = 1, \dots, N-1$, the local time process $\Lambda_{k,k+1}$ measures the amount of time the process $\log X_{(k)} - \log X_{(k+1)}$ spends near zero (Karatzas and Shreve, 1991). According to (2.2), the dynamics of family size for the k-th

[2] See Fernholz (2001), Banner and Ghomrasni (2008), and Ichiba et al. (2011). The necessary regularity conditions include the requirement that the $\log X_i$ spend no local time at triple points.

largest family are the same as those for the family that is k-th largest at time t, plus two local time terms that measure changes in rank over time as some families grow larger than others.

Lemma 2.1. *For $N > 1$, let $X_1, \ldots, X_N$ be positive random growth processes that satisfy* (2.2)*. For $n < N$, define*

$$X_{[n]} \triangleq X_{(1)} + \cdots + X_{(n)}. \tag{2.3}$$

Then

$$\frac{dX_{[n]}(t)}{X_{[n]}(t)} = \sum_{i=1}^{N} \mathbb{1}_{\{r_t(i) \leq n\}} \frac{dX_i(t)}{X_{[n]}(t)} + \frac{X_{(n)}(t)}{2\, X_{[n]}(t)} d\Lambda_{n,n+1}(t), \quad \text{a.s.} \tag{2.4}$$

Lemma 2.1 describes the dynamic relationship between the combined size of the n largest families, $X_{[n]}$, and the local time process, $\Lambda_{n,n+1}$. This local time process compensates for turnover among the top n ranks. Over time, some of the largest families will decrease in size and exit from the top n ranks, while some smaller families will increase in size and consequently enter into the top n ranks. The intensity of this entry and exit among the top n ranks is measured by the local time process in (2.4).

In our general, model-free framework, specific frictions such as a lower reflecting barrier or a Poisson birth-death process, which stabilize an otherwise nonstationary geometric Brownian motion (Gabaix, 2009), are replaced by the natural entry and exit into and out of the top n ranks of a larger system of random growth processes. This entry and exit, which occurs for any system of continuous random growth processes that can be ranked, captures all of the frictions that affect the system. Because the intensity of this entry and exit is measured by the local time process $\Lambda_{n,n+1}$, it follows that this process measures the intensity of all frictions that act upon the system. In this way, the local time process from (2.4) obviates the need to commit to a specific friction.

In a setting in which families sometimes exit and are replaced by newly entering families, the top $n < N$ ranks of the system (2.1) represent the families that are currently in the system. Under this representation, those families below the top n ranks represent families that are currently outside the system and potentially unobservable. A new family enters—and, correspondingly, an old family exits—when some previously lower-ranked family enters into the top n ranks and replaces a higher-ranked family. In this way, our rank-based framework replaces the exogenous Poisson birth-death processes that are common in many different economic models of random growth (Gabaix, 2009; Benhabib et al., 2016; Toda, 2017) with the natural entry and exit that occurs among the top ranks of any system of random growth processes. The characterization of this natural entry and exit (2.4) is rigorous and does not depend on the underlying mechanism that is driving the entry and exit. Furthermore, this characterization applies both to models in which all families that enter and exit are observed at all times as well as models in which families that exit disappear and are then unobserved.

In addition to its generality and rigor, another advantage of our continuous, rank-based framework relates to the mechanics of Poisson birth-death processes in economic models of random growth. In these models, all families, regardless of how large or small, are equally likely to die and be replaced by a new family. While this assumption may be appropriate for some applications, it is unrealistic for applications such as city and firm size, where Zipf's law is most relevant. In these cases, it is more reasonable to assume that cities and firms first decrease in size, and thus rank, before eventually exiting from the top ranks, and that new cities and firms first increase in size, and thus rank, before eventually entering into the top ranks. Of course, both the process of decline and exit from the top ranks as well as the process of growth and entry into the top ranks are exactly what is described by Lemma 2.1.

Finally, it is worth noting that it is common in empirical applications to encounter data $\{Z_1(t), Z_2(t), \ldots\}$ that are not continuous processes in the manner of (2.1). In this case, the local time process $\Lambda_{n,n+1}$ is defined implicitly by equation (2.4), since all the other terms in that equation are observable. Figure 2 plots local time processes for the distribution of total market capitalizations of U.S. stocks from 1990–1999 using such an implicit construction via (2.4).[3]

2.3 Conservation and Completeness

We wish to examine the behavior of a fixed-size random sample of changing family members over time. We denote this fixed sample by $\{\varXi_1(t), \varXi_2(t), \ldots\}$, with $\varXi_i(t)$ indicating the number of family members from the sample that are contained in the i-th family at time t. As long as this sample is sufficiently large, the distribution of family members for this random sample will be approximately the same as the distribution of family members for the full data $\{Z_1(t), Z_2(t), \ldots\}$. Furthermore, because the number of family members in the random sample is, by construction, constant over time, it follows that for large enough n the size of the largest n families, $\varXi_{[n]}(t)$, should be approximately constant over time as well.

We model the top $n > 1$ ranks of the fixed-size sample $\{\varXi_1(t), \varXi_2(t), \ldots\}$ by the top n ranks of a system of random growth processes $X_1, \ldots, X_N$, where $N > n$. We shall consider the behavior of such systems in the limit as both n and $N > n$ tend to infinity. Furthermore, we restrict our attention in this paper to systems of random growth processes that generate stationary relative size distributions. In other words, we consider systems in which the ranked relative size processes, $\log X_{(k)} - \log X_{(k+1)}$, are assumed to be stationary for all $k = 1, \ldots, N-1$.

Definition 2.2. The system of positive random growth processes $X_1, \ldots, X_N$, is *conservative* if

$$\lim_{N>n\to\infty} \mathbb{E}\left[\frac{dX_{[n]}(t)}{X_{[n]}(t)}\right] = 0, \tag{2.5}$$

where the expectation is for the stationary size distribution.

Definition 2.2 formalizes the notion that, for large n, the expected combined size of the n largest families should be approximately constant. In other words, the system $X_1, \ldots, X_N$ conserves the "mass" inside of the top n ranks. As we shall demonstrate below, the conservation condition of (2.5) is one of the two conditions needed to establish Zipf's law under the strong form of Gibrat's law. It is also one of the conditions needed to establish quasi-Zipf's law when the data only partially follow Gibrat's law.

The conservation condition of Definition 2.2 is equivalent to the restriction that the expected value of a system of random growth processes is constant. This is a key assumption behind the existing explanations of Zipf's law in economics. Gabaix (1999), for example, normalizes the size of cities so that their expected sum is equal to one at all times.[4] This normalization is, of course, equivalent to imposing the conservation condition (2.5) on the random system $X_1, \ldots, X_N$. Similarly, Toda (2017) considers a model with a fixed aggregate size, X, that requires $dX = 0$ at all times.[5] He then argues that a growing aggregate size can

[3]This figure is reproduced from Fernholz (2002).
[4]See footnote 14 of Gabaix (1999).
[5]See Section 3.1 and the proof of Theorem 3.5 in Appendix B of Toda (2017).

be normalized, as in Gabaix (1999), so that $dX = 0$ still holds. As before, this argument is equivalent to imposing (2.5) on a system of random growth processes.

Our motivation for Definition 2.2 differs from the existing literature in an important way, however. As we argued earlier, the conservation condition is a necessary condition for any system of random growth processes used to model a real-world system that could be sampled over time. The reason for this is that if the real-world system, $\{Z_1(t), Z_2(t), \ldots\}$, can be sampled, then a large fixed sample of data from this system, $\{\varXi_1(t), \varXi_2(t), \ldots\}$, can be constructed and the size distribution of this fixed sample will not differ materially from the size distribution of the full data. In this case, any system of random growth processes used to model the full data should maintain a fixed size on average in the same way that the large fixed sample does.

Recall that the local time process $\Lambda_{n,n+1}$ from (2.4) represents the process of entry and exit among the top n ranks of a random system, and that this process of entry and exit is equivalent to any frictions that act upon this system. In order that the system not depend on replacement from infinitely far below, the expected contribution of the local time term in (2.4) should vanish for sufficiently large n. This condition is equivalent to the requirement that any frictions that act upon a system vanish in the limit, even though we do not commit to any specific frictions.

Definition 2.3. The system of positive random growth processes $X_1, \ldots, X_N$, is *complete* if

$$\lim_{N>n\to\infty} \mathbb{E}\left[\frac{X_{(n)}(t)}{X_{[n]}(t)} d\Lambda_{n,n+1}(t)\right] = 0, \tag{2.6}$$

where the expectation is for the stationary size distribution.

In the existing economic models of random growth and Zipf's law, the completeness condition (2.6) is imposed either by assuming that a lower reflecting barrier tends to zero (Gabaix, 1999) or that the Poisson birth-death rate tends to zero (Toda, 2017). Both of these vanishing frictions are subsumed by Definition 2.3, which imposes the more general restriction that the influence on the top n ranks of a system of random growth processes by processes outside of this top subset vanishes in the limit as n grows large. In other words, the top n ranks of the system are complete on their own and are not influenced by some unmodeled, exogenous factor.

The conservation condition (2.5) from Definition 2.2 must, by construction, hold for a fixed random sample of family members, and thus it should be imposed on random growth models of real-world systems for which members can be sampled over time. The completeness condition (2.6) from Definition 2.3 should hold for large random systems of data that do not depend on outside phenomena for their growth, and thus it should be imposed on random growth models that include a sufficiently large number of ranked families. In Section 4, we show that these two conditions are together sufficient to establish Zipf's law under the strong form of Gibrat's law. Furthermore, even if a system does not follow the strong form of Gibrat's law, we show that in many cases conservation and completeness are still sufficient to establish quasi-Zipf's law.

3 Stationary Size Distributions

We restrict our attention in this paper to systems of random growth processes $X_1, \ldots, X_N$ that generate stationary relative size distributions. To this end, we assume that the ranked relative size processes, $\log X_{(k)} -$

$\log X_{(k+1)}$, are stationary for all $k = 1, \dots, N-1$. With this assumption, Fernholz (2002), Banner et al. (2005), and Fernholz (2017b) have shown that the asymptotic size distribution of a system of random growth processes will satisfy

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \big(\log X_{(k)}(t) - \log X_{(k+1)}(t) \big)\, dt = \frac{\sigma^2_{k,k+1}}{2\lambda_{k,k+1}}, \quad \text{a.s.,} \tag{3.1}$$

for $k = 1, \dots, N-1$, where

$$\lim_{t\to\infty} t^{-1}\Lambda_{k,k+1}(t) = \lambda_{k,k+1} > 0, \quad \text{a.s.} \tag{3.2}$$

$$\lim_{t\to\infty} t^{-1}\langle \log X_{(k)} - \log X_{(k+1)}\rangle_t = \sigma^2_{k,k+1} > 0, \quad \text{a.s.} \tag{3.3}$$

Note that (3.1) is valid only if the limits (3.2)-(3.3) both exist and are positive. Even though the characterization (3.1) uses a time-averaged limit, this is equivalent to an expected value if this expected value is taken with respect to the stationary size distribution of the processes $X_1, \dots, X_N$ as in Definitions 2.2 and 2.3.

3.1 Discussion

The characterization of the relative size distribution (3.1) does not require us to assume a specific economic model or impose a specific friction. Instead, this characterization will hold if the stationary distributions of the ranked relative size processes, $\log X_{(k)} - \log X_{(k+1)}$, are exponential with potentially different parameters for all $k = 1, \dots, N-1$. In the literature on Zipf's law in economics, it is common to construct a model in which the stationary distributions of the relative size processes $\log X_{(k)} - \log X_{(k+1)}$, $k = 1, \dots, N-1$, all follow the same exponential distribution. In this way, the characterization (3.1) nests standard economic models of random growth as special cases of a more general framework.

A growing literature in mathematical finance, statistics, and, most recently, economics examines the theoretical conditions that give rise to the characterization (3.1) as well as the empirical size distributions that are well-approximated by this characterization. Banner et al. (2005), for example, show that Atlas models and first-order models, which are systems of exchangeable random growth processes with growth rates and volatilities that vary only by rank, satisfy (3.1). Ichiba et al. (2011) go one step further and show that certain systems of non-exchangeable processes with growth rates and volatilities that vary by both rank and name (denoted by index i) also satisfy (3.1). Later in this section, we show that standard random growth models in economics using either a lower reflecting barrier or a Poisson birth-death process generate stationary size distributions that satisfy (3.1) as well.

Many real-world rank-based random systems have been shown to satisfy (3.1). For example, the distribution of total market capitalizations of U.S. stocks from 1990–1999 is shown in Figure 3 along with the predicted distribution curve generated using the local time estimates reported in Figure 2 together with (2.4). The closeness of the actual and predicted distribution curves in Figure 3 attests to the accuracy of (3.1) for these stock market data. Following the same basic estimation procedure, Fernholz and Koch (2016) and Fernholz (2017b) show that (3.1) provides an accurate description of the empirical distributions of, respectively, the assets of U.S. bank-holding companies, commercial banks, and thrifts and the relative normalized prices of commodities.[6] Documenting classes of empirical data with size distributions that satisfy (3.1) remains an active area of current research.

[6] In addition to these applications, Fernholz (2017b) shows that implicit estimates of the local time processes based on (2.4) can also be used to forecast future relative commodity prices using this same equation.

The mechanism by which the characterization (3.1) arises derives from the central limit theorem. For the smaller relative size processes $\log X_{(k)} - \log X_{(k+1)}$, which correspond to the lower ranks, the growth and variance parameters of the processes $\log X_i$ from (2.1) will be nearly constant while occupying a particular rank. For constant parameters, the central limit theorem implies that the random fluctuations of the relative size processes will be approximately Gaussian, in which case these processes will be (approximately) reflected Brownian motion. Since the stationary distribution of reflected Brownian motion is exponential, the characterization (3.1) follows. For the top ranks, with wider gaps in size between ranks, condition (3.1) may be more tenuous, and in some cases it may be necessary either to observe the system over an extended period of time, or else to augment the system to increase the density of the ranked observations.

3.2 Necessary and Sufficient Conditions for Zipf's Law

We define the asymptotic slope parameters s_k by

$$s_k \triangleq k \lim_{T\to\infty} \frac{1}{T} \int_0^T \big(\log X_{(k)}(t) - \log X_{(k+1)}(t) \big)\, dt, \tag{3.4}$$

for all $k = 1, \dots, N-1$. We can see from (3.4) that, for all $k = 1, \dots, N-1$,

$$-s_k \left(1 + \frac{1}{2k}\right) < \lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{\log X_{(k)}(t) - \log X_{(k+1)}(t)}{\log(k) - \log(k+1)}\, dt < -s_k, \tag{3.5}$$

which implies that for large k,

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{\log X_{(k)}(t) - \log X_{(k+1)}(t)}{\log(k) - \log(k+1)}\, dt \approx -s_k. \tag{3.6}$$

According to (3.5) and (3.6), then, the slope parameters s_k provide an accurate approximation of the slope of a log-size versus log-rank plot at low ranks, but these parameters can be off by a factor of as much as one-half for the top two ranks.[7] For expositional simplicity, we shall treat the slope parameters s_k as the true log-log slopes between adjacent ranks. Nonetheless, it is important to remember that this equivalence is only as accurate as the inequality (3.5).

The distribution of family sizes $X_1, \dots, X_N$ satisfies Zipf's law if the log-log plot of family size versus rank is a straight line with slope -1 (Gabaix, 1999). This is equivalent to the slope parameters s_k being equal to 1, for all $k = 1, \dots, N-1$. We refer to a size distribution that satisfies Zipf's law as a *Zipfian* distribution, and to a system of random growth processes that generates such a distribution as a *Zipfian* system.

Theorem 3.1. *A system of random growth processes $X_1, \dots, X_N$ satisfying* (3.1) *is Zipfian if and only if*

$$\frac{k\sigma^2_{k,k+1}}{2\lambda_{k,k+1}} = 1, \tag{3.7}$$

for all $k = 1, \dots, N-1$.

To our knowledge, Theorem 3.1 provides the first characterization of necessary and sufficient conditions for economic models of random growth to generate Zipf's law. The theorem is a simple consequence of (3.1) and (3.6), the former having been discussed in detail in Section 3.1. One implication of Theorem 3.1 is that all previous papers describing specific Zipfian random systems must necessarily satisfy (3.7).

[7] Interestingly, the correction by a factor of one-half for s_1 shown in (3.5) matches the correction by one-half proposed by Gabaix and Ibragimov (2011) when using OLS to estimate the slope of a straight-line log-log plot of rank versus size.

3.3 The Related Literature Satisfies Theorem 3.1

In order to see how previous characterizations of Zipf's law in economics satisfy (3.7) from Theorem 3.1, let us consider the important special cases of a geometric Brownian motion with a lower reflecting barrier (Gabaix, 1999, 2009) and a geometric Brownian motion with a Poisson birth-death process (Gabaix, 2009; Toda, 2017). For a system of random growth processes of the form (2.1), the restriction that these general Itô processes be geometric Brownian motions with equal growth rates and volatilities—thus satisfying the strong form of Gibrat's law—is equivalent to assumption that, at least for the highest ranks of the system,

$$d \log X_i(t) = g\,dt + \sigma\,dB_i(t), \tag{3.8}$$

or equivalently,

$$\frac{dX_i(t)}{X_i(t)} = \left(g + \frac{\sigma^2}{2}\right) dt + \sigma\,dB_i(t), \tag{3.9}$$

where g and $\sigma > 0$ are constants.

For a system of the form (3.8), Banner et al. (2005) show that, in the presence of some friction that ensures the existence of a stationary relative size distribution, the limits (3.2) and (3.3) exist, with

$$\lambda_{k,k+1} = -2kg, \tag{3.10}$$

and

$$\sigma^2_{k,k+1} = 2\sigma^2. \tag{3.11}$$

It follows from (3.7), then, that Zipf's law obtains in this case if and only if

$$\frac{2k\sigma^2}{-4kg} = -\frac{\sigma^2}{2g} = 1, \tag{3.12}$$

which is equivalent to $\sigma^2 = -2g$. According to Definition 2.2, the condition (3.12) obtains if and only if the system (3.8) is conservative, since

$$E\left[\frac{dX_i(t)}{X_i(t)}\right] = \left(g + \frac{\sigma^2}{2}\right) dt, \tag{3.13}$$

and this expected value equals zero if and only if $\sigma^2 = -2g$.

The previous discussion shows that standard economic models of random growth based on a geometric Brownian motion with a lower reflecting barrier (Gabaix, 1999, 2009), a Poisson birth-death process (Gabaix, 2009; Toda, 2017), or any other friction will generate Zipf's law if the model satisfies the conservation condition (2.5). In most cases, conservation is imposed on the model via either size normalization, as in footnote 14 of Gabaix (1999), or the imposition of a fixed aggregate size, as in Section 3.1 of Toda (2017).

So far, our discussion has left out any mention of the frictions typically used to stabilize the geometric Brownian motions (3.8) and thus it is not fully rigorous. However, such frictions can be rigorously incorporated into our rank-based framework in an easy way using local time processes, as discussed in Section 2. We return to the important special case of geometric Brownian motion in more detail and rigor in the next section.

4 Gibrat's Law, Zipf's Law, and Quasi-Zipfian Distributions

In Section 2, we argued that any random growth model of an economic system that can be sampled over time and that includes a large number of ranked observations should be both conservative and complete (Definitions 2.2 and 2.3). In Section 3, we characterized the size distribution of any stationary system of random growth processes. In this section, we use these results together to show that any system of random growth processes that follows the strong form of Gibrat's law must satisfy Zipf's law, thus rigorously establishing an important hypothesis of Gabaix (1999). We also use these results to show that any system of random growth processes that deviates from Gibrat's law in a specific but realistic manner must satisfy quasi-Zipf's law, which is more general than Zipf's law.

4.1 Rank-Based Systems of Random Growth Processes

According to the characterization (3.1), a system of random growth processes in which the growth rates and volatilities in (2.1) depend only on the rank of each family can replicate the stationary size distribution of any system of random growth processes. Indeed, as long as the rank-based and non-rank-based systems both generate the same limits (3.2) and (3.3), their size distributions will necessarily be identical.

This observation motivates our consideration of the rank-based system of random growth processes $X_1, \ldots, X_N$ defined by

$$d \log X_i(t) = \big(G(t) + g_{r_t(i)}\big)\, dt + \sigma_{r_t(i)}\, dB_i(t), \tag{4.1}$$

where G is the potentially time-varying growth rate of the full system, and g_k and $\sigma_k > 0$, $k = 1, \ldots, N$, are constants satisfying $g_1 + \cdots + g_N = 0$. The constants g_k and σ_k determine the relative growth rates and volatilities of different ranked families in this system. The common growth rate G does not affect the relative size processes, $\log X_{(k)} - \log X_{(k+1)}$, and hence it does not affect the stationary size distribution according to (3.1). Therefore, we can assume without loss of generality that $G = 0$.

We wish to examine the size distribution of the top $n < N$ ranks of the rank-based system of random growth processes (4.1). According to Banner et al. (2005), this top-ranked subset has a stationary size distribution if and only if the relative growth rates $g_1, \ldots, g_n$ satisfy $g_1 + \cdots + g_k < 0$, for all $k \leq n$. In this case, for the top n ranks of the system (4.1) the limits (3.2) and (3.3) exist, with

$$\lambda_{k,k+1} = -2\big(g_1 + \cdots + g_k\big), \quad \text{a.s.}, \tag{4.2}$$

and

$$\sigma^2_{k,k+1} = \sigma^2_k + \sigma^2_{k+1}, \quad \text{a.s.}, \tag{4.3}$$

for $k = 1, \ldots, n-1$. These authors also show that these systems satisfy (3.1), with

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \big(\log X_{(k)}(t) - \log X_{(k+1)}(t)\big)\, dt = \frac{\sigma^2_k + \sigma^2_{k+1}}{2\lambda_{k,k+1}}, \quad \text{a.s.}, \tag{4.4}$$

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

As discussed in Section 2, our approach is unconventional in that we do not restrict ourselves to a specific economic model. Instead, we derive general results that rely on few assumptions, motivated by the view that such generality is essential to any credible explanation of Zipf's law (Gabaix, 1999). This general approach is reflected in the rank-based system (4.1), which, according to (4.4), can generate any power law size distribution by varying the parameters g_k and σ_k across different ranks. For any economic model with a power law size distribution, then, there exists a rank-based system of the form (4.1) that will generate an identical size distribution. In this way, we are able to derive results that are applicable to almost any economic model.

In empirical applications, it is common to estimate the parameters g_k and σ_k using time-series panel data and then to construct a rank-based system of the form (4.1) using these parameters. This estimation procedure yields a system of random growth processes with rank-based parameters $\lambda_{k,k+1}$ and $\sigma^2_{k,k+1}$ that generate a slightly smoothed version of the relative size distribution in the data (Fernholz, 2002).

4.2 Gibrat's Law and Zipf's Law

The strong form of Gibrat's law states that all random growth processes X_i must have equal growth rates and equal volatilities. In terms of the rank-based system of random growth processes (4.1), this implies that for some top-ranked subset $n < N$, the parameters g_k and σ_k must satisfy

$$g_1 = \cdots = g_n = -g \quad \text{and} \quad \sigma_1 = \cdots = \sigma_n = \sigma, \tag{4.5}$$

where g and σ are positive constants. As in Gabaix (1999), the restriction (4.5) implies that some top subset of the largest families have growth rates and volatilities that follow the strong form of Gibrat's law. Unlike Gabaix (1999), however, this restriction allows for any growth rates and volatilities for smaller families which are outside of this top ranked subset, meaning that our results are consistent with any friction that acts upon the lower ranks of the system to ensure the existence of a stationary relative size distribution, including the important special case of a lower reflecting barrier.

A system of random growth processes of the form (4.1) satisfying the restriction (4.5) is also consistent with models that include entry and exit in the form of a Poisson birth-death process. As discussed in Section 2, in this setting the top $n < N$ ranks of the system (4.1) represent families that are currently in the system, while those families below the top n ranks represent families that are currently outside the system and potentially unobservable. For such systems, entry and exit occurs naturally as lower-ranked families overtake higher-ranked families and thus the composition of the top-ranked subset changes over time.

By analyzing systems of random growth processes of the form (4.1) satisfying the restriction (4.5), then, we are able to rigorously derive results that are consistent with models in which some friction affects only the smallest families (Gabaix, 1999, 2009), as well as models in which there is entry and exit that affects all families (Toda, 2017). In other words, we can prove results about the universality of Zipf's law without imposing a specific friction.

For rank-based systems of random growth processes that satisfy the strong form of Gibrat's law (4.5), the result (4.4) implies that

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \big(\log X_{(k)}(t) - \log X_{(k+1)}(t) \big)\, dt = \frac{\sigma^2_{k,k+1}}{2\lambda_{k,k+1}} = \frac{\sigma^2}{2kg}, \quad \text{a.s.,} \tag{4.6}$$

for $k = 1, \ldots, n-1$. According to (4.6), the slope parameters s_k defined by (3.4) are in this case given by

$$s_k = \frac{k\sigma^2}{2kg} = \frac{\sigma^2}{2g}, \tag{4.7}$$

for all $k = 1, \ldots, n-1$, and hence (3.6) implies that for large k, the slope of a plot of log-size versus log-rank satisfies

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{\log X_{(k)}(t) - \log X_{(k+1)}(t)}{\log(k) - \log(k+1)}\, dt \approx -\frac{\sigma^2}{2g}, \quad \text{a.s.} \tag{4.8}$$

Recall that the stationary distribution of a system of random growth processes is Pareto if the log-size versus log-rank plot with slope characterized in (4.8) is a straight line, and that the distribution satisfies Zipf's law if the slope of this log-log plot is equal to -1. According to (4.8), then, systems of rank-based random growth processes satisfying the restriction (4.5) generate a stationary Pareto distribution, and this distribution is Zipfian if and only if $\sigma^2 = 2g$.[8] Given the equivalence between the restriction (4.5) and the strong form of Gibrat's law discussed above, (4.8) implies that the strong form of Gibrat's law generates a Pareto distribution, consistent with Gabaix (1999, 2009).

When applied to systems of random growth processes of the form (4.1), Itô's lemma and Lemma 2.1 imply that the total number of family members in the top n ranks, $X_{[n]} = X_{(1)} + \cdots + X_{(n)}$, satisfies

$$dX_{[n]}(t) = \left(\frac{\sigma^2}{2} - g\right) X_{[n]}(t)\, dt + X_{[n]}(t)\, dM(t) + \frac{1}{2} X_{(n)}(t)\, d\Lambda_{n,n+1}(t), \quad \text{a.s.}, \tag{4.9}$$

where M is a martingale incorporating all of the terms σB_i. It follows, then, that

$$\frac{dX_{[n]}(t)}{X_{[n]}(t)} = \left(\frac{\sigma^2}{2} - g\right) dt + dM(t) + \frac{X_{(n)}(t)}{2X_{[n]}(t)}\, d\Lambda_{n,n+1}(t), \quad \text{a.s.}, \tag{4.10}$$

and hence that

$$\mathbb{E}\left[\frac{dX_{[n]}(t)}{X_{[n]}(t)}\right] = \left(\frac{\sigma^2}{2} - g\right) dt + \mathbb{E}\left[\frac{X_{(n)}(t)}{2X_{[n]}(t)}\, d\Lambda_{n,n+1}(t)\right], \quad \text{a.s.}, \tag{4.11}$$

The conditions of conservation and completeness from Definitions 2.2 and 2.3 require that the expected values of the first and third terms approach zero as the size of the top subset n approaches infinity. In this case, (4.11) implies that $\sigma^2 = 2g$ and hence this rank-based system must satisfy Zipf's law.

Theorem 4.1. *For a system of rank-based random growth processes of the form* (4.1) *satisfying the strong form of Gibrat's law* (4.5)*, the conditions of conservation,* (2.5)*, and completeness,* (2.6)*, are equivalent to Zipf's law.*

Theorem 4.1 shows that a system of random growth processes that follows the strong form of Gibrat's law satisfies Zipf's law if and only if it is conservative and complete.[9] The wide and diverse range of distributions satisfying Zipf's law means that any attempt to explain it should be general and not rely on details that are specific to a particular application or model. The result of Theorem 4.1 and our model-free approach using general rank-based systems of random growth processes are both consistent with this requirement. As we detailed in Section 2, if a system of positive-valued, time-dependent data can be approximated by Itô processes, then a representative model of that data that includes a large number of ranks should be both conservative and complete, provided that a large fixed sample of family members could exist. If these

[8] For a system in which with $\sigma^2/2g \le 1$, we have that $\mathbb{E}\big[X_{[n]}(t)\big]$ diverges to infinity as n increases, so it would appear that the total number of family members becomes infinite. However, this limit as n tends to infinity is used as a mathematical device to characterize the behavior of the system for large n, and there is no intention here of considering infinite systems, even though methods have been developed to accomplish this (see, for example, Pal and Pitman (2008)).

[9] Fernholz and Fernholz (2017) demonstrate the necessity of assuming the strong form of Gibrat's law in Theorem 4.1 by constructing a system that does not follow the strong form of Gibrat's law and generates a Pareto size distribution that violates Zipf's law.

data follow the strong form of Gibrat's law, then this representative model will satisfy the restriction (4.5) and hence it will generate a Zipfian size distribution. In this manner, we provide an explanation for the universality of Zipf's law.

The result in Theorem 4.1 is consistent with the well-known result in the literature that the strong form of Gibrat's law plus a vanishing friction generates Zipf's law (Gabaix, 1999, 2009; Toda, 2017). In our general rank-based framework, all frictions that affect the system $X_1, \ldots, X_N$ are captured by the local time process in (4.10), which, as discussed in Section 2, measures the natural entry and exit in and out of the top n ranks of this system. As a consequence, the result in Theorem 4.1 using rank-based processes does not rely on the presence of any specific frictions. Instead, this result shows that Zipf's law emerges for the top ranks of a large system of random growth processes that follow Gibrat's law regardless of what the underlying frictions that ensure stationarity of that system are. In this manner, our result in Theorem 4.1 confirms a hypothesis of Gabaix (1999) that Zipf's law emerges in the limit as all frictions acting upon a system vanish.

4.3 Quasi-Zipfian Distributions

It is important to note that few real-world systems appear to follow Zipf's law once a large number of ranked observations are considered. In fact, two of the most prominent examples of Zipf's law in economics—the distributions of firm employees and city populations—have log-log plots of size versus rank that are concave rather than straight lines with slope -1. These deviations from Zipf's law can be seen in Figure 1 of Axtell (2001) and Figure 10 of Eeckhout (2004). While both of these log-log plots are close to being straight lines with slope -1, once enough ranked observations are considered, it is clear that a concave curve provides a better fit of the data. In the case of the city size distribution, Gabaix (1999) hypothesized that log-log plots that deviate from Zipf's law in this way are potentially caused by higher volatilities at lower ranks.

In order to rigorously explore this hypothesis, we first define a weaker version of Zipf's law that matches the deviations from the standard Zipf's law often observed in the real world.

Definition 4.2. A system of random growth processes is *quasi-Zipfian* if the slope parameters s_k defined by (3.4) are nondecreasing with $s_1 \leq 1$ and

$$\sup\{s_k, k \in \mathbb{N}\} \geq 1. \tag{4.12}$$

We refer to a system that is quasi-Zipfian as satisfying *quasi-Zipf's law*. Because the slope parameters s_k are approximately equal to minus the slope of a log-log plot of size versus rank, Definition 4.2 states that a system is quasi-Zipfian if this log-log plot is concave with slope not steeper than -1 at the highest ranks and not flatter than -1 at the lowest ranks.

A Zipfian system of random growth processes is also quasi-Zipfian, since the former requires that the log-size versus log-rank plot is a straight line with a slope of -1. Indeed, the tangent of slope -1 required for a quasi-Zipfian system becomes the entire log-log plot if the strong form of Gibrat's law obtains, as shown by Theorem 4.1. In this sense, Definition 4.2 generalizes the notion of Zipf's law to systems of random growth processes that deviate from Gibrat's law and do not asymptotically follow uniform Pareto size distributions. We show that quasi-Zipf's law is universal for systems that deviate from Gibrat's law in a specific but realistic manner, in much the same way that Theorem 4.1 shows that Zipf's law is universal for systems that do follow Gibrat's law.

Theorem 4.3. *Consider a rank-based system of random growth processes of the form* (4.1) *such that*

$$\begin{aligned} g_1 = g_2 = \cdots = g_n = -g < 0, \\ 0 < \sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_n, \end{aligned} \tag{4.13}$$

and

$$\lim_{N>n\to\infty} \mathbb{E}\left[\frac{X_{(1)}(t)}{X_{[n]}(t)}\right] \leq \frac{1}{2}. \tag{4.14}$$

Then, the conditions of conservation, (2.5)*, and completeness,* (2.6)*, imply that the system is quasi-Zipfian.*

A rank-based system of random growth processes that satisfies condition (4.13) from Theorem 4.3 follows Gibrat's law for growth rates but not necessarily for volatilities. Condition (4.13) also implies, however, that the volatilities σ_k weakly increase as family size decreases. Because this condition only imposes equal growth rates for the top n ranks and allows for different volatilities for the families occupying these top ranks, a rank-based system satisfying this condition is more general than a system satisfying the strong form of Gibrat's law condition (4.5).

Theorem 4.3 demonstrates that a rank-based system of random growth processes satisfying conditions (4.13) and (4.14) is quasi-Zipfian if it is conservative and complete. As discussed in Section 2, any realistic economic model of random growth should be conservative and complete. An implication of Theorem 4.3, then, is that all rank-based systems of random growth processes that follow Gibrat's law for growth rates and have volatilities that decrease with family size should be quasi-Zipfian, provided that the largest family does not include more than half of the total population.[10]

According to (3.4), (4.2), and (4.4), a rank-based system that satisfies condition (4.13) from Theorem 4.3 will have slope parameters given by

$$s_k = \frac{k(\sigma_k^2 + \sigma_{k+1}^2)}{2\lambda_{k,k+1}} = \frac{\sigma_k^2 + \sigma_{k+1}^2}{4g}, \tag{4.15}$$

for $k = 1, \ldots, n-1$. Because (4.13) also ensures that the volatilities σ_k are weakly increasing in k, it follows from (4.15) that any system that satisfies (4.13) will have a log-log plot of family size versus rank that is concave. Of course, if the system is also conservative and complete and satisfies condition (4.14), then Theorem 4.3 implies that this concave curve will also have a tangent of -1 somewhere along it. Many real-world size distributions follow this quasi-Zipfian pattern, including the assets of U.S. banks (Fernholz and Koch, 2016), the wealth of U.S. households (Fernholz, 2017a), and the capitalization of U.S. stocks (Figure 3).

The distributions of firm employees and city populations are two of the most prominent examples of Zipf's law in economics. However, as discussed earlier, these examples do not quite conform to the strict requirements of Zipf's law, especially if a large number of ranked observations are considered. Instead, both the U.S. firm and city size distributions have log-log plots of size versus rank that are concave curves with a tangent line of slope -1 somewhere. This concave shape can be seen for firm employees in Figure 1 of Axtell (2001), and for city populations in Figure 4 of Ioannides and Skouras (2013). Definition 4.2 and Theorem 4.3 give meaning to these deviations, and show that the city and firm size distributions in fact satisfy quasi-Zipf's law rather than Zipf's law.

[10]Fernholz and Fernholz (2017) demonstrate the necessity of a condition of the form (4.14) by constructing a conservative and complete system that satisfies (4.13), but that violates (4.14) and is not quasi-Zipfian.

Gabaix (1999) postulates that the concave shape of the log-log plot of U.S. city populations is due to higher volatility parameters σ_k at lower ranks, but it remains an open question to establish empirically just how common this and the first condition of (4.13) from Theorem 4.3 actually are. One system that roughly satisfies both of these conditions is the total market capitalizations of U.S. stocks. Figures 4 and 5 plot, respectively, estimates of the rank-based parameters g_k and σ_k using monthly data on the capitalization of U.S. stocks from 1990–1999. These parameters are estimated using (2.4) and (4.2) for the g_k, and the quadratic variation of the time series for the σ_k. The contrast between the relatively level g_k parameters in Figure 4 and the rising σ_k parameters in Figure 5 is stark, and is consistent with both parts of condition (4.13). Furthermore, Figure 3 shows that the distribution curve implied by the estimated parameters g_k and σ_k closely matches the empirical distribution curve from 1990–1999, so the data-driven rank-based random system (4.1) in this case is accurately describing the data. Not surprisingly, then, Theorem 4.3 provides a valid prediction for this system. Indeed, both the empirical and predicted distribution curves in Figure 3 are flatter than -1 at the highest ranks and steeper than -1 at the lowest ranks, with the point at which a line of slope -1 is tangent to the curve highlighted for emphasis. In other words, the capitalization of U.S. stocks is quasi-Zipfian, as predicted by our results.

Just as with Zipf's law, quasi-Zipf's law is also a form of universality since many different and unrelated random systems satisfy quasi-Zipf's law. As with our results about Zipfian distributions and the strong form of Gibrat's law, our results about quasi-Zipfian distributions do not rely on any details or frictions that are specific to a particular model or application. Because a system of positive-valued, time-dependent data that includes a large number of ranked observations can be modeled by a rank-based system of random growth processes that is both conservative and complete, Theorem 4.3 implies that all such systems that satisfy the conditions (4.13) and (4.14) should satisfy quasi-Zipf's law. In other words, this theorem provides an explanation for the universality of quasi-Zipf's law.

5 Conclusion

We characterize the necessary and sufficient conditions under which systems of random growth processes generate size distributions that satisfy Zipf's law. This characterization is accomplished without committing to an economic model and without imposing a specific friction on the systems that we analyze. The literature on Zipf's law in economics emphasizes that because of its universality, any explanation of Zipf's law should not depend on details that are specific to a particular model or application (Gabaix, 1999). Our model-free, friction-invariant approach satisfies this requirement.

In the paper, we show that a rank-based system that follows the strong form of Gibrat's law, with growth rates and volatilities that do not vary across different ranks, will satisfy Zipf's law if and only if the system is conservative and complete. We also show that a rank-based system of random growth processes that deviates from Gibrat's law in a specific but realistic manner will be quasi-Zipfian—a generalization of Zipf's law in which a log-log plot of size versus rank must only have a tangent at -1 somewhere—if the system is again conservative and complete. Because these conditions of conservation and completeness should be satisfied by almost all systems of random growth processes that cover a large number of ranked observations, our results offer an explanation for the universality of both Zipf's law and quasi-Zipf's law.

Appendix: Proofs

This appendix contains the proofs of Lemma 2.1 and Theorems 4.1 and 4.3.

Proof of Lemma 2.1. Suppose that the rank processes $X_{(k)}$ satisfy (2.2), so we have

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

for all k. By Itô's rule, for all k,

$$\frac{dX_{(k)}(t)}{X_{(k)}(t)} - \frac{1}{2}d\langle \log X_{(k)}\rangle_t = \sum_{i=1}^{N} \mathbb{1}_{\{r_t(i)=k\}} \frac{dX_i(t)}{X_i(t)} - \frac{1}{2}\sum_{i=1}^{N} \mathbb{1}_{\{r_t(i)=k\}} d\langle \log X_i\rangle_t + \frac{1}{2}d\Lambda_{k,k+1}(t) - \frac{1}{2}d\Lambda_{k-1,k}(t),$$

a.s., so

$$\begin{aligned}
\frac{dX_{(k)}(t)}{X_{(k)}(t)} &= \sum_{i=1}^{N} \mathbb{1}_{\{r_t(i)=k\}} \frac{dX_i(t)}{X_i(t)} + \frac{1}{2}d\Lambda_{k,k+1}(t) - \frac{1}{2}d\Lambda_{k-1,k}(t) \\
&= \sum_{i=1}^{N} \mathbb{1}_{\{r_t(i)=k\}} \frac{dX_i(t)}{X_{(k)}(t)} + \frac{1}{2}d\Lambda_{k,k+1}(t) - \frac{1}{2}d\Lambda_{k-1,k}(t), \quad \text{a.s.}
\end{aligned}$$

From this we have

$$\begin{aligned}
dX_{(k)}(t) &= \sum_{i=1}^{N} \mathbb{1}_{\{r_t(i)=k\}} dX_i(t) + \frac{1}{2}X_{(k)}(t)d\Lambda_{k,k+1}(t) - \frac{1}{2}X_{(k)}(t)d\Lambda_{k-1,k}(t) \\
&= \sum_{i=1}^{N} \mathbb{1}_{\{r_t(i)=k\}} dX_i(t) + \frac{1}{2}X_{(k)}(t)d\Lambda_{k,k+1}(t) - \frac{1}{2}X_{(k-1)}(t)d\Lambda_{k-1,k}(t), \quad \text{a.s.,}
\end{aligned}$$

since the support of $d\Lambda_{k-1,k}$ is contained in the set $\{t : \log X_{(k-1)}(t) = \log X_{(k)}(t)\}$. Now we can add up the $dX_{(k)}(t)$ to obtain

$$dX_{[n]}(t) = \sum_{i=1}^{N} \mathbb{1}_{\{r_t(i)\le n\}} dX_i(t) + \frac{1}{2}X_{(n)}(t)d\Lambda_{n,n+1}(t), \quad \text{a.s.,}$$

which is equivalent to (2.4). □

Proof of Theorem 4.1. According to (4.2), equation (4.11) can be replaced by

$$\mathbb{E}\left[\frac{dX_{[n]}(t)}{X_{[n]}(t)}\right] = \left(\frac{\sigma^2}{2} - g\right)dt + \mathbb{E}\left[\frac{ngX_{(n)}(t)}{X_{[n]}(t)}\right]dt. \tag{A.1}$$

Since a rank-based system of random growth processes of the form (4.1) satisfying the strong form of Gibrat's law (4.5) generates a Pareto distribution with log-log plot of size versus rank with slope $-\sigma^2/2g$, we can calculate

$$\mathbb{E}\left[\frac{ngX_{(n)}(t)}{X_{[n]}(t)}\right] = \begin{cases} O(1) & \text{for } \sigma^2/2 < g, \\ O(1/\log n) & \text{for } \sigma^2/2 = g, \\ O\big(n^{(1-\sigma^2/2g)}\big) & \text{for } \sigma^2/2 > g. \end{cases} \tag{A.2}$$

Suppose that $\sigma^2 = 2g$. By (A.2), the right-hand side of (A.1) is $O\big(1/\log n\big)dt$, so the left-hand side must also be, and hence both (2.5) and (2.6) hold. If conservation (2.5) and completeness (2.6) hold, then (4.11) implies that $\sigma^2 = 2g$ must hold. Therefore, (2.5) combined with (2.6) is equivalent to $\sigma^2 = 2g$, which is equivalent to Zipf's law by (4.8). □

Proof of Theorem 4.3. Consider the rank-based system of random growth processes $X_1, \dots, X_N$ with

$$d\log X_i(t) = -g\mathbb{1}_{\{r_t(i)\le n\}}\,dt + g_{r_t(i)}\mathbb{1}_{\{r_t(i)>n\}}\,dt + \sigma_{r_t(i)}\,dB_i(t), \tag{A.3}$$

where $g > 0$, $0 < \sigma_1 \le \sigma_2 \le \cdots \le \sigma_n$, g_k, and $\sigma_k > 0$ are constants. From (4.15), we have that

$$s_k = \frac{\sigma_k^2 + \sigma_{k+1}^2}{4g},$$

for $k = 1, \dots, n-1$, so $s_1 \le \cdots \le s_{n-1}$. We must show that $s_1 \le 1$ and $\sup\{s_k, k \in \mathbb{N}\} \ge 1$.

For a general rank-based random system of the form (4.1), Itô's rule implies that

$$dX_i(t) = \left(\frac{\sigma_{r_t(i)}^2}{2} - g_{r_t(i)}\right)X_i(t)\,dt + \sigma_{r_t(i)}X_i(t)\,dB_i(t), \quad \text{a.s.},$$

for all $i = 1, \dots, N$. Given the restrictions of the system (A.3), it follows from Lemma 2.1 that

$$dX_{[n]}(t) = \sum_{k=1}^n X_{(k)}(t)\left(\frac{\sigma_k^2}{2} - g\right)dt + dM(t) + \frac{1}{2}X_{(n)}(t)d\Lambda_{n,n+1}(t), \quad \text{a.s.},$$

where M is a martingale incorporating all of the terms $\sigma_{r_t(i)}X_i(t)dB_i(t)$. It follows that

$$\mathbb{E}\left[\frac{dX_{[n]}(t)}{X_{[n]}(t)}\right] = \left(\sum_{k=1}^n \mathbb{E}\left[\frac{X_{(k)}(t)}{X_{[n]}(t)}\right]\frac{\sigma_k^2}{2} - g\right)dt + \mathbb{E}\left[\frac{X_{(n)}(t)}{X_{[n]}(t)}d\Lambda_{n,n+1}(t)\right]. \tag{A.4}$$

If the system is conservative and complete and hence satisfies (2.5) and (2.6), then as n tends to infinity the first and third terms of (A.4) vanish and we have

$$\lim_{N>n\to\infty}\sum_{k=1}^n \mathbb{E}\left[\frac{X_{(k)}(t)}{X_{[n]}(t)}\right]\frac{\sigma_k^2}{2g} = 1. \tag{A.5}$$

Let us now show that (4.13) and (4.14) imply that $s_1 \le 1$. By (4.13) the σ_k^2 are nondecreasing, so (A.5) implies that

$$\begin{aligned} 1 &\ge \lim_{N>n\to\infty}\mathbb{E}\left[\frac{X_{(1)}(t)}{X_{[n]}(t)}\right]\frac{\sigma_1^2}{2g} + \left(1 - \lim_{N>n\to\infty}\mathbb{E}\left[\frac{X_{(1)}(t)}{X_{[n]}(t)}\right]\right)\frac{\sigma_2^2}{2g} \\ &\ge \frac{1}{2}\frac{\sigma_1^2}{2g} + \frac{1}{2}\frac{\sigma_2^2}{2g} = s_1, \end{aligned} \tag{A.6}$$

where the second inequality follows from (4.14).

Since the σ_k^2 are nondecreasing, as k tends to infinity they must either converge to a finite value or diverge to infinity. If the σ_k^2 diverge to infinity, the same will be true for the s_k, so they will assume values greater than 1. 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,

$$1 = \lim_{N>n\to\infty}\sum_{k=1}^n \mathbb{E}\left[\frac{X_{(k)}(t)}{X_{[n]}(t)}\right]\frac{\sigma_k^2}{2g} \le \frac{\sigma^2}{2g}.$$

It follows that $\sup\{s_k, k \in \mathbb{N}\} \ge 1$. $\square$

References

Atkinson, A. B., T. Piketty, and E. Saez (2011, March). Top incomes in the long run of history. *Journal of Economic Literature 49*(1), 3–71.

Axtell, R. (2001, September). Zipf distribution of U.S. firm sizes. *Science 293*(5536), 1818–1820.

Bak, P. (1996). *How Nature Works.* New York: Springer-Verlag.

Banner, A., R. Fernholz, and I. Karatzas (2005). Atlas models of equity markets. *Annals of Applied Probability 15*(4), 2296–2330.

Banner, A. and R. Ghomrasni (2008, July). Local times of ranked continuous semimartingales. *Stochastic Processes and their Applications 118*(7), 1244–1253.

Benhabib, J., A. Bisin, and S. Zhu (2016, March). The distribution of wealth in the blanchard-yaari model. *Macroeconomic Dynamics 20*(2), 466–481.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review 94*(5), 1429–1451.

Fernholz, E. R. (2002). *Stochastic Portfolio Theory.* New York, NY: Springer-Verlag.

Fernholz, R. (2001). Equity portfolios generated by functions of ranked market weights. *Finance and Stochastics 5*, 469–486.

Fernholz, R. T. (2017a). The distributional effects of progressive capital taxes. *Journal of Economic Policy Reform 20*(2), 99–112.

Fernholz, R. T. (2017b). Nonparametric methods and local-time-based estimation for dynamic power law distributions. *Journal of Applied Econometrics 32*(7), 1244–1260.

Fernholz, R. T. and R. Fernholz (2017, September). The universality of zipf's law for time-dependent rank-based random systems. *arXiv:1707.04285v2.*

Fernholz, R. T. and C. Koch (2016, February). Why are big banks getting bigger? *Federal Reserve Bank of Dallas Working Paper 1604.*

Gabaix, X. (1999, August). Zipf's law for cities: An explanation. *Quarterly Journal of Economics 114*(3), 739–767.

Gabaix, X. (2009, 05). Power laws in economics and finance. *Annual Review of Economics 1*(1), 255–294.

Gabaix, X. (2016, Winter). Power laws in economics: An introduction. *Journal of Economic Perspectives 30*(1), 185–206.

Gabaix, X. and R. Ibragimov (2011, January). Rank - 1/2: A simple way to improve the OLS estimation of tail exponents. *Journal of Business and Economic Statistics 29*(1), 24–39.

Gibrat, R. (1931). *Les inégalités économiques.* Librairie du Recueil Sirey, Paris.

Ichiba, T., V. Papathanakos, A. Banner, I. Karatzas, and R. Fernholz (2011). Hybrid atlas models. *Annals of Applied Probability 21*(2), 609–644.

Ioannides, Y. and S. Skouras (2013, January). US city size distribution: Robustly Pareto, but only in the tail. *Journal of Urban Economics 73*(1), 18–29.

Karatzas, I. and S. E. Shreve (1991). *Brownian Motion and Stochastic Calculus.* New York, NY: Springer-Verlag.

Kesten, H. (1973). Random difference equations and renewal theory for products of random matrices. *Acta Mathematica 131*(1), 207–248.

Newman, M. E. J. (2005, September-October). Power laws, Pareto distributions, and Zipf's law. *Contemporary Physics 46*(5), 323–351.

Pal, S. and J. Pitman (2008). One-dimensional brownian particle systems with rank-dependent drifts. *The Annals of Applied Probability 18*(6), 2179–2207.

Piketty, T. (2014). *Capital in the Twenty-First Century.* Cambridge, MA: Harvard University Press.

Simon, H. A. (1955, December). On a class of skew distribution functions. *Biometrika 42*(3/4), 425–440.

Steindl, J. (1965). *Random Processes and the Growth of Firms.* New York, NY: Hafner.

Toda, A. A. (2017, May). Zipf's law: A microfoundation. *mimeo, UC San Diego.*

Zipf, G. (1935). *The Psychology of Language: An Introduction to Dynamic Philology.* Cambridge, MA: M.I.T. Press.

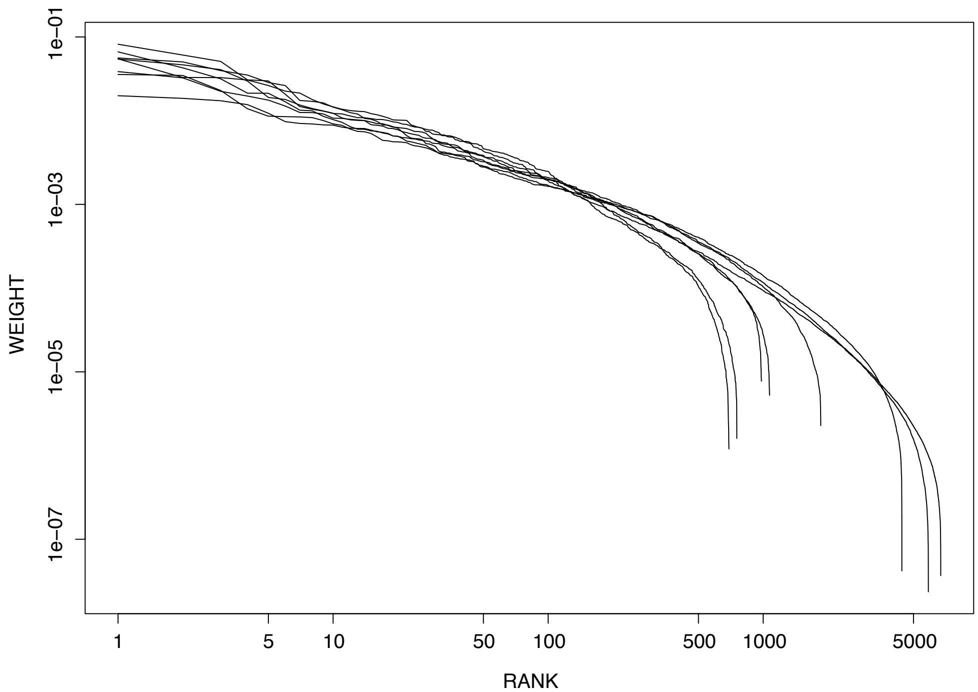

Figure 1: The U.S. capital distribution, 1929–1999. The lines show the distribution at ten-year intervals from 1929–1999.

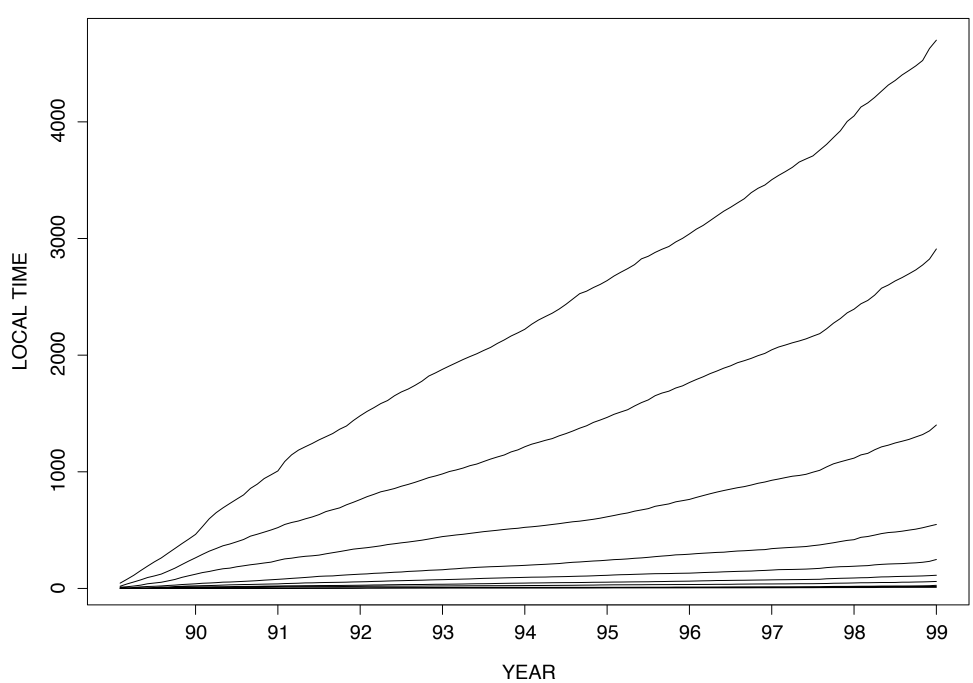

Figure 2: Local time processes $\Lambda_{k,k+1}(t)$ for the U.S. capital distribution from 1990–1999, for $k = 10, 20, 40, \ldots, 5120$.

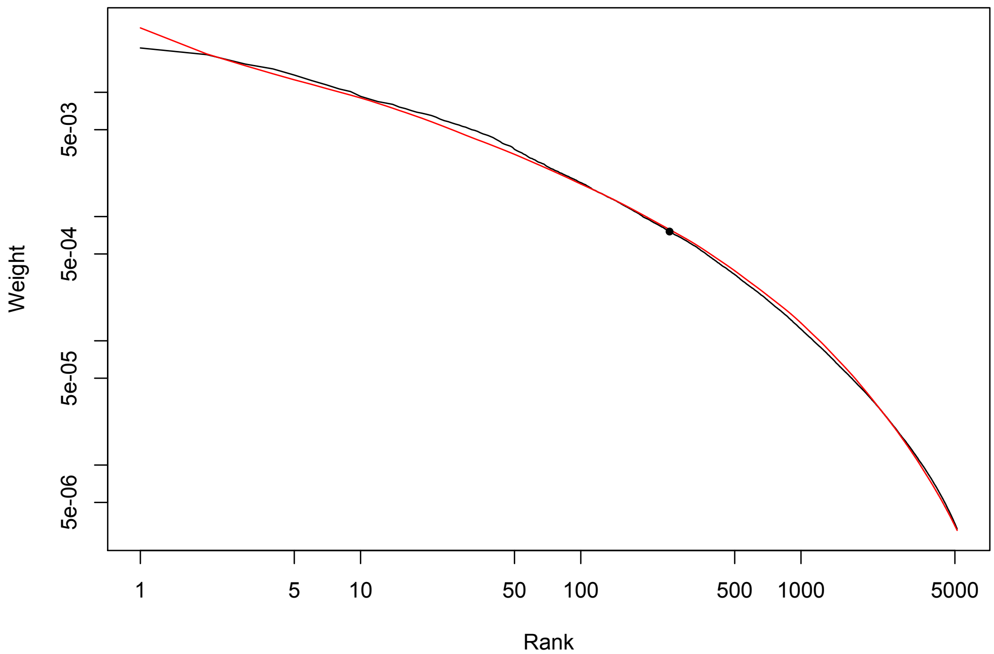

Figure 3: The observed U.S. capital distribution (black line) and the predicted U.S. capital distribution (red line), 1990–1999. The black dot corresponds to the point at which a line of slope -1 is tangent.

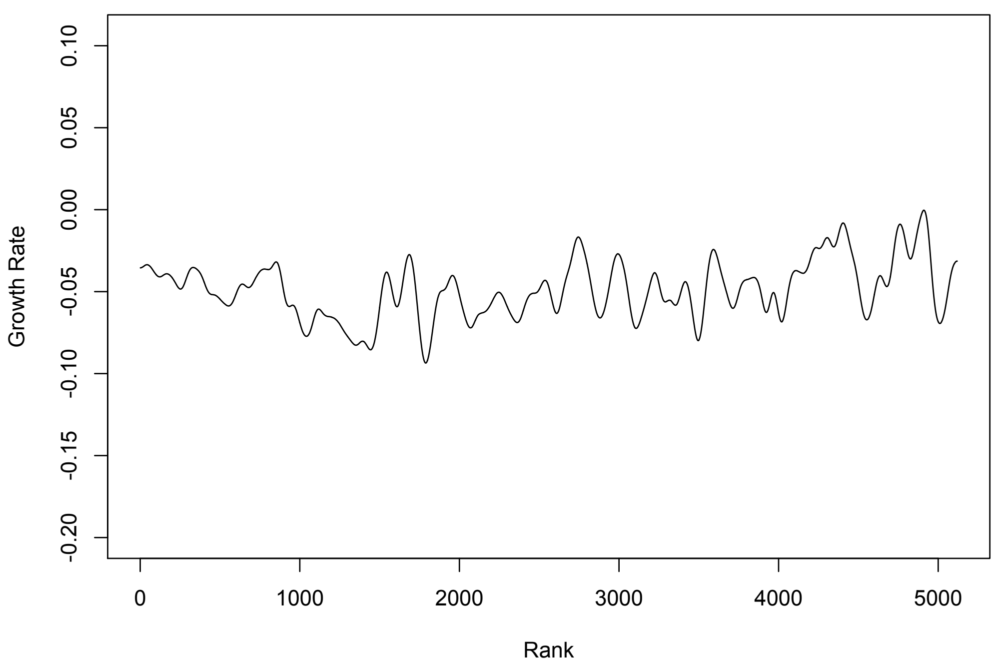

Figure 4: Estimated parameters g_k for the U.S. capital distribution from 1990–1999.

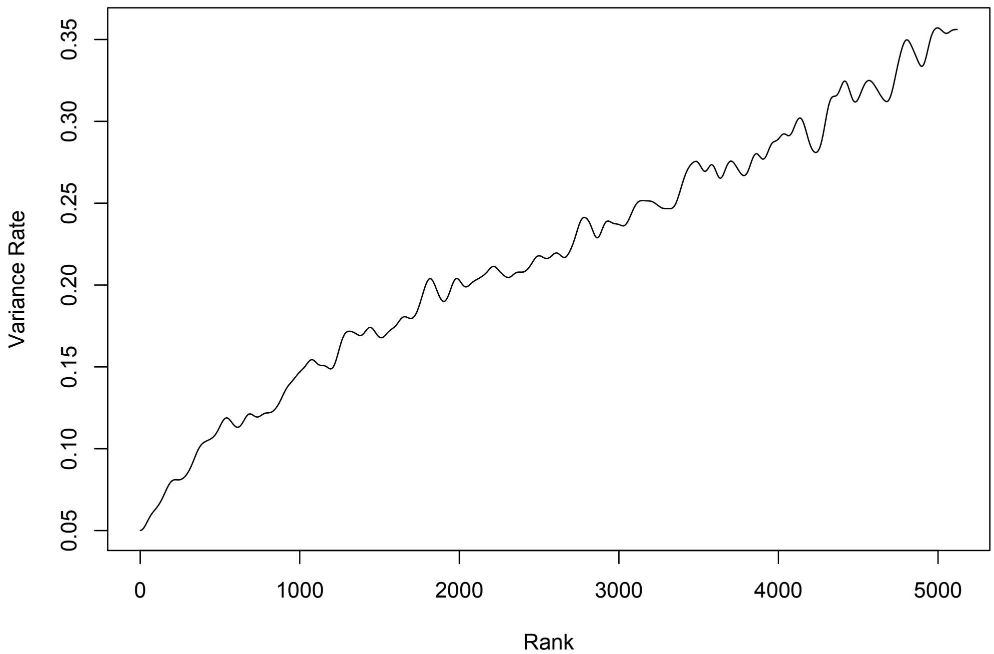

Figure 5: Estimated parameters σ_k^2 for the U.S. capital distribution from 1990–1999.