The Universality of Zipf's Law

CDAR Risk Seminar, UC Berkeley

Ricardo T. Fernholz

Claremont McKenna College

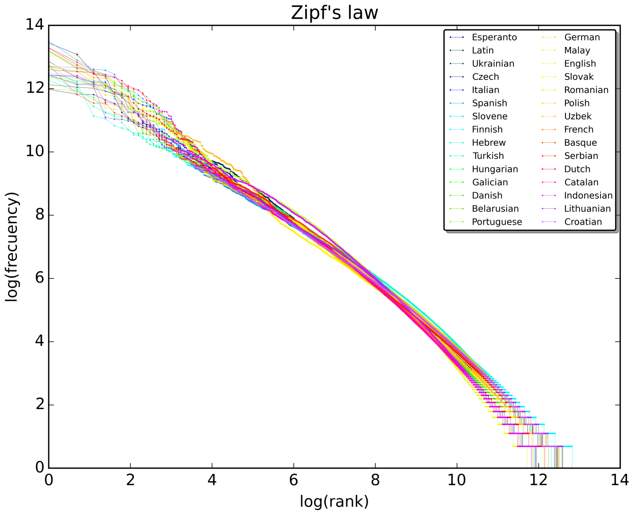
February 1, 2022

Zipf's Law

The original formulation of Zipf's law states that given some collection of natural language text, the *frequency of any word is inversely proportional to its rank* in the frequency table.

The law is named after the American linguist George Kingsley Zipf (1902–1950), who popularized and sought to explain it (Zipf, 1935, 1949).

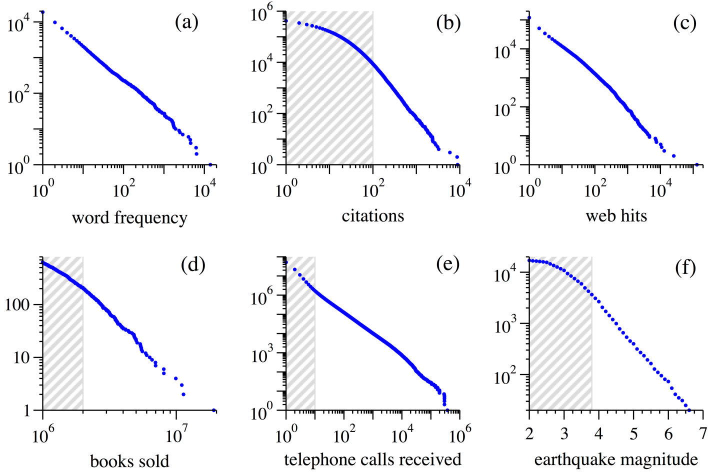
Word Count from Wikipedia



Pareto Distributions and Zipf's Law

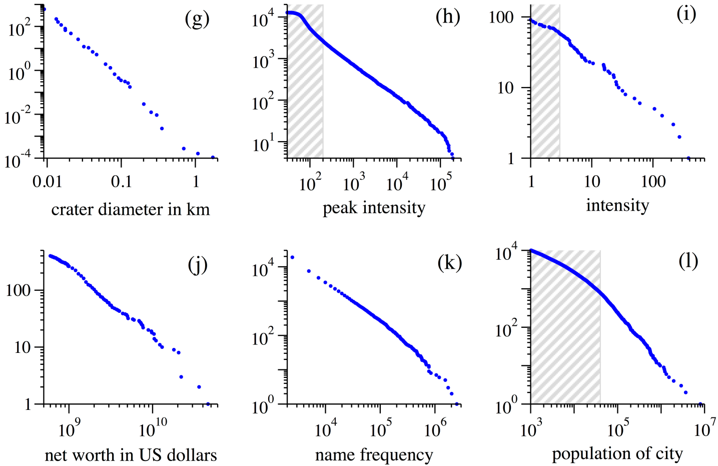
- Pareto distribution, or power law
  - Log-log plot of the data versus rank is approximately a straight line
- Zipf's Law
  - Log-log plot of the data versus rank is approximately a straight line with slope -1
  - Weaker form of Zipf's law requires that log-log plot of the data versus rank is concave with a tangent line of slope -1 at some point

Examples of Pareto Distributions



From Newman (2006)

Examples of Pareto Distributions



From Newman (2006)

Zipf's Law and Universality

According to Tao (2012), "mathematicians do not have a fully satisfactory and convincing explanation for how the law comes about and why it is universal."

Zipf's law appears in many different fields and many different applications. As a consequence, any explanation should appeal to statistics and mathematics rather than field-specific phenomena.

Zipfian and Non-Zipfian Pareto Distributions

- The universality of Zipf's law
  - Firm size, city size, word frequency, income and wealth of households
  - Data generated by time-dependent rank-based systems follow Zipf's law (Fernholz & Fernholz, 2020)
  - Any explanation should not depend on the specific details of a model (Gabaix, 1999; Toda, 2017)
- Non-Zipfian Pareto distributions
  - Earthquake magnitude, cumulative book sales, intensity of wars
  - Data generated by other means, usually of a cumulative nature, do not necessarily follow Zipf's law

Applications

- The distribution of U.S. stock market capitalizations
  - Followed standard quasi-Zipfian distribution for most of U.S. history
  - Empirical estimates confirm market cap dynamics satisfy conditions that yield quasi-Zipfian distribution (Fernholz, 2017)

- Rank- and name-based time-dependent systems (Ichiba et al., 2011)
  - U.S. stock market capitalizations post-2020?
  - City size distributions (Davis & Weinstein, 2002; Soo, 2005)
  - Wealth distribution (Benhabib, Bisin, & Fernholz, 2022)

Ranked Continuous Semimartingales

We use systems of positive continuous semimartingales $\{X_1, \dots, X_n\}$ to approximate systems of time-dependent empirical data. If the X_i satisfy certain regularity conditions, then

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

- *Rank function* is defined such that $r_t(i) < r_t(j)$ if $X_i(t) > X_j(t)$
- *Rank processes* $X_{(1)} \geq \cdots \geq X_{(n)}$ are defined by $X_{(r_t(i))}(t) = X_i(t)$
- $\Lambda^X_{k,k+1}$ is the local time at the origin for $\log(X_{(k)}/X_{(k+1)})$, which measures the effect of crossovers between ranks k and $k+1$

Atlas Models

An *Atlas model* is a system of positive continuous semimartingales $\{X_1, \ldots, X_n\}$ defined by

$$d \log X_i(t) = -g\,dt + ng\,\mathbb{1}_{\{r_t(i)=n\}}dt + \sigma\,dW_i(t),$$

where g and σ are positive constants and $(W_1, \ldots, W_n)$ is a Brownian motion.

- Stationary model when geometric mean of processes X_i is subtracted
- System follows Gibrat's law, with equal growth rates and variances across all ranks

First-Order Models

A *first-order model* is a system of positive continuous semimartingales $\{X_1, \ldots, X_n\}$ defined by

$$d \log X_i(t) = g_{r_t(i)}\, dt + G_n \mathbb{1}_{\{r_t(i)=n\}} dt + \sigma_{r_t(i)}\, dW_i(t),$$

where $\sigma_1^2, \ldots, \sigma_n^2$ are positive constants, $g_1, \ldots, g_n$ are constants satisfying

$$g_1 + \cdots + g_k < 0, \text{ for } k \leq n,$$

$G_n = -(g_1 + \cdots + g_n)$, and $(W_1, \ldots, W_n)$ is a Brownian motion.

- Stationary model when geometric mean of processes X_i is subtracted
- More general than Atlas models

Asymptotic Distribution of Atlas Models

Atlas models satisfy

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

for $k = 1, \dots, n-1$, with the asymptotic parameters

$$\lim_{T\to\infty} T^{-1}\Lambda^X_{k,k+1}(T) = \lambda_{k,k+1} = 2kg, \qquad \text{a.s.},$$

$$\lim_{T\to\infty} T^{-1}\langle \log X_{(k)} - \log X_{(k+1)}\rangle_T = \sigma^2_{k,k+1} = 2\sigma^2, \qquad \text{a.s.}$$

Hence, for large enough k, Atlas models satisfy

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

which implies a Pareto distribution that follows Zipf's law if $\sigma^2 = 2g$.

Atlas Families

An *Atlas family* is a class of Atlas models $\{X_1, \ldots, X_n\}$, for $n \in \mathbb{N}$, with the common parameters $g > 0$ and $\sigma^2 > 0$ defined as in

$$d \log X_i(t) = -g\, dt + ng\, \mathbb{1}_{\{r_t(i)=n\}} dt + \sigma\, dW_i(t).$$

Let $X_{[n]} = X_{(1)} + \cdots + X_{(n)}$, and

$$R_n = \mathbb{E}_n\left[\frac{X_{(n)}(t)}{X_{(1)}(t)}\right] \qquad \text{and} \qquad R_{[n]} = \mathbb{E}_n\left[\frac{X_{[n]}(t)}{X_{(1)}(t)}\right].$$

Fernholz & Fernholz (2020) show that an Atlas family is Zipfian if and only if

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

Conservation

By sampling or detrending, we can ensure that the "total mass" of a system of time-dependent rank-based data $\{Z_1(\tau), Z_2(\tau), \ldots\}$ is constant.

In this case, for large enough n, the mass of the top n ranks,

$$Z_{[n]}(\tau) = Z_{(1)}(\tau) + \cdots + Z_{(n)}(\tau),$$

should also be approximately constant.

Hence, for large enough n, it is reasonable to expect:

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

Conservation

By sampling or detrending, we can ensure that the "total mass" of a system of time-dependent rank-based data $\{Z_1(\tau), Z_2(\tau), \ldots\}$ is constant.

In this case, for large enough n, the mass of the top n ranks,

$$Z_{[n]}(\tau) = Z_{(1)}(\tau) + \cdots + Z_{(n)}(\tau),$$

should also be approximately constant.

Hence, we require that an Atlas family $\{X_1, \ldots, X_n\}$ be *conservative*:

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

Completeness

For a system of data $\{Z_1(\tau), Z_2(\tau), \dots\}$, the replacement of processes in the top n ranks by processes in the lower ranks over the time interval $[\tau, \tau+1]$ is

$$Z_{[n]}(\tau+1) - \sum_{i=1}^{N} 1_{\{r_\tau(i) \leq n\}} Z_i(\tau+1).$$

For large enough n, it is reasonable to expect this replacement to become arbitrarily small, i.e. that the system will be *complete*:

$$\frac{1}{Z_{[n]}(\tau)} \left(\left(Z_{[n]}(\tau+1) - Z_{[n]}(\tau)\right) - \sum_{i=1}^{N} 1_{\{r_\tau(i) \leq n\}} \left(Z_i(\tau+1) - Z_i(\tau)\right) \right) \cong 0.$$

Completeness

In terms of a first-order model, completeness is

$$\frac{1}{R_{[n]}}\mathbb{E}_n\left[\frac{1}{T}\int_0^T \frac{dX_{[n]}(t)}{X_{(1)}(t)} - \frac{1}{T}\int_0^T\left(\sum_{i=1}^N 1_{\{r_t(i)\le n\}}\frac{dX_i(t)}{X_{(1)}(t)}\right)\right] \cong 0.$$

It is not hard to show that this is equivalent to

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

where the last term is the local time at zero of $\log(X_{(n)}/X_{(n+1)})$.

For an Atlas family $\{X_1,\dots,X_n\}$, $d\Lambda^X_{n,n+1}(t)$ is on average equal to $2ng$, and so we require that an Atlas family be *complete*:

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

Zipf's Law: Proof Sketch

An Atlas family is Zipfian (so that $\sigma^2 = 2g$) if and only if it is conservative and complete:

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

For an Atlas model, Itô's rule implies that, a.s.,

$$dX_i(t) = \Big(\frac{\sigma^2}{2} - g + ng\,\mathbb{1}_{\{r_t(i)=n\}}\Big)X_i(t)\,dt + \sigma X_i(t)\,dW_i(t),$$

which, for the total mass $X_{[n]} = X_1 + \cdots + X_n$, implies that, a.s.,

$$dX_{[n]}(t) = \Big(\frac{\sigma^2}{2} - g\Big)X_{[n]}(t)\,dt + X_{[n]}(t)dM(t) + ngX_{(n)}(t)\,dt,$$

where M is a martingale incorporating all the σW_i.

Zipf's Law: Proof Sketch

An Atlas family is Zipfian (so that $\sigma^2 = 2g$) if and only if it is conservative and complete:

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

For the total mass $X_{[n]} = X_1 + \cdots + X_n$, we have, a.s.,

$$dX_{[n]}(t) = \left(\frac{\sigma^2}{2} - g\right) X_{[n]}(t)\, dt + X_{[n]}(t) dM(t) + ngX_{(n)}(t)\, dt,$$

which implies that, a.s.,

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

Zipf's Law: Proof Sketch

An Atlas family is Zipfian (so that $\sigma^2 = 2g$) if and only if it is conservative and complete:

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

For the total mass $X_{[n]} = X_1 + \cdots + X_n$, we have, a.s.,

$$\frac{dX_{[n]}(t)}{X_{(1)}(t)} = \Big(\frac{\sigma^2}{2} - g\Big)\frac{X_{[n]}(t)}{X_{(1)}(t)}\,dt + \frac{X_{[n]}(t)}{X_{(1)}(t)}\,dM(t) + \frac{ngX_{(n)}(t)}{X_{(1)}(t)}dt,$$

which implies that

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

Zipf's Law: Proof Sketch

An Atlas family is Zipfian (so that $\sigma^2 = 2g$) if and only if it is conservative and complete:

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

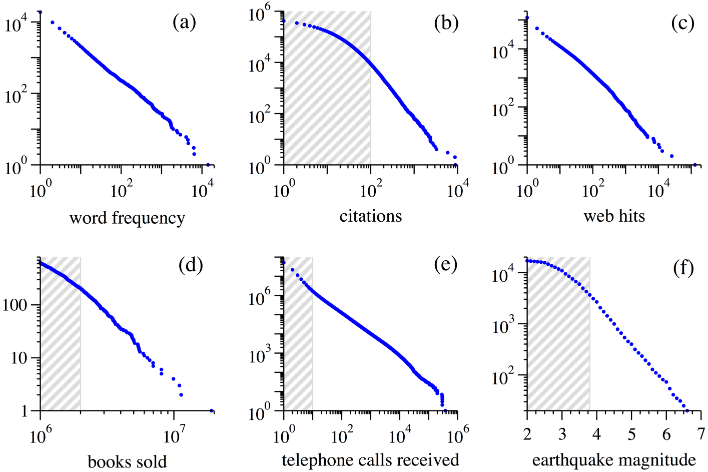
For the total mass $X_{[n]} = X_1 + \cdots + X_n$, we have, a.s.,

$$\frac{dX_{[n]}(t)}{X_{(1)}(t)} = \left(\frac{\sigma^2}{2} - g\right)\frac{X_{[n]}(t)}{X_{(1)}(t)}\,dt + \frac{X_{[n]}(t)}{X_{(1)}(t)}\,dM(t) + \frac{ngX_{(n)}(t)}{X_{(1)}(t)}dt,$$

which implies that

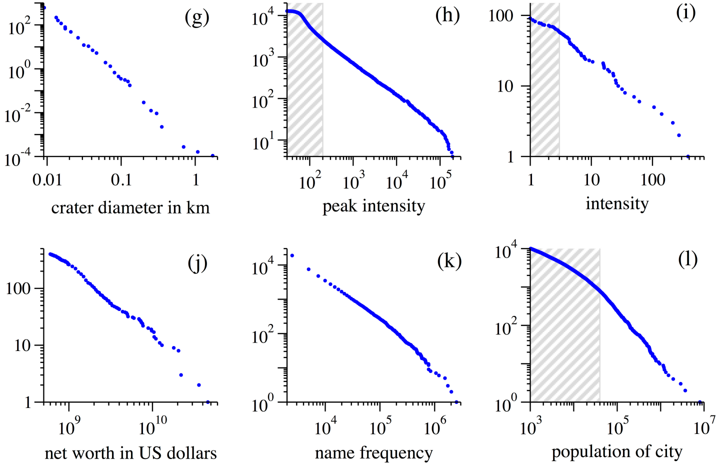
$$\frac{1}{R_{[n]}} \mathbb{E}_n\left[\frac{1}{T}\int_0^T \frac{dX_{[n]}(t)}{X_{(1)}(t)}\right] = \frac{\sigma^2}{2} - g + ng\frac{R_n}{R_{[n]}}.$$

Examples of Pareto Distributions



From Newman (2006)

Examples of Pareto Distributions



From Newman (2006)

Quasi-Atlas Models

A *first-order model* is a system of positive continuous semimartingales $\{X_1, \ldots, X_n\}$ defined by

$$d \log X_i(t) = g_{r_t(i)}\, dt + G_n \mathbb{1}_{\{r_t(i)=n\}} dt + \sigma_{r_t(i)}\, dW_i(t),$$

where $\sigma_1^2, \ldots, \sigma_n^2$ are positive constants, $g_1, \ldots, g_n$ are constants satisfying

$$g_1 + \cdots + g_k < 0, \text{ for } k \leq n,$$

and $G_n = -(g_1 + \cdots + g_n)$. A *quasi-Atlas model* is a first-order model with $g > 0$ and $\sigma_2^2 \geq \sigma_1^2 > 0$, such that

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

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

Asymptotic Distribution of Quasi-Atlas Models

Quasi-Atlas models satisfy

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

a.s., for $k = 1, \dots, n-1$, with the asymptotic parameters

$$\lambda_{k,k+1} = 2kg \qquad \text{and} \qquad \sigma^2_{k,k+1} = \sigma^2_k + \sigma^2_{k+1}, \qquad \text{a.s.}$$

Hence, for large enough k,

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{\log X_{(k)}(t) - \log X_{(k+1)}(t)}{\log(k) - \log(k+1)}\,dt \cong -\frac{\sigma^2_k + \sigma^2_{k+1}}{4g}, \qquad \text{a.s.},$$

so quasi-Atlas models may have non-Pareto stationary distributions.

Quasi-Atlas Families

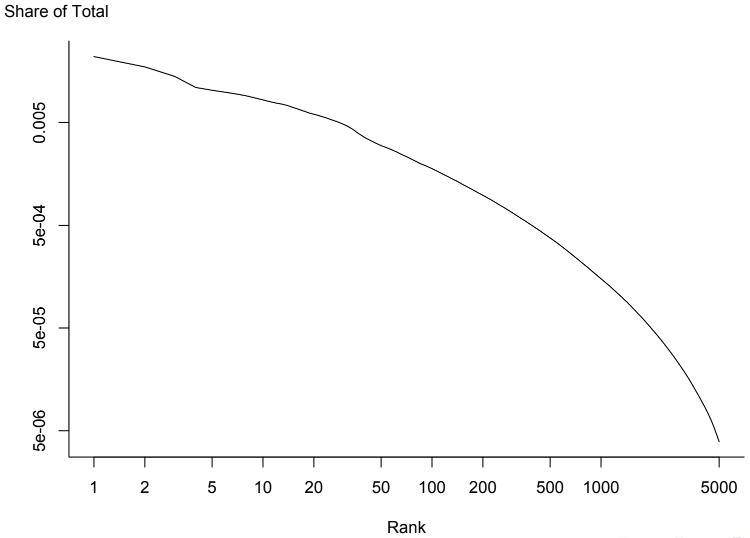
A *quasi-Atlas family* is a class of quasi-Atlas models $\{X_1, \ldots, X_n\}$, for $n \in \mathbb{N}$, with the common parameters $\sigma_k^2 = \sigma_1^2 + (k-1)(\sigma_2^2 - \sigma_1^2) > 0$ for $k \in \mathbb{N}$ and $g > 0$, defined as in

$$d \log X_i(t) = -g\,dt + ng\,\mathbb{1}_{\{r_t(i)=n\}}dt + \sigma_{r_t(i)}\,dW_i(t).$$

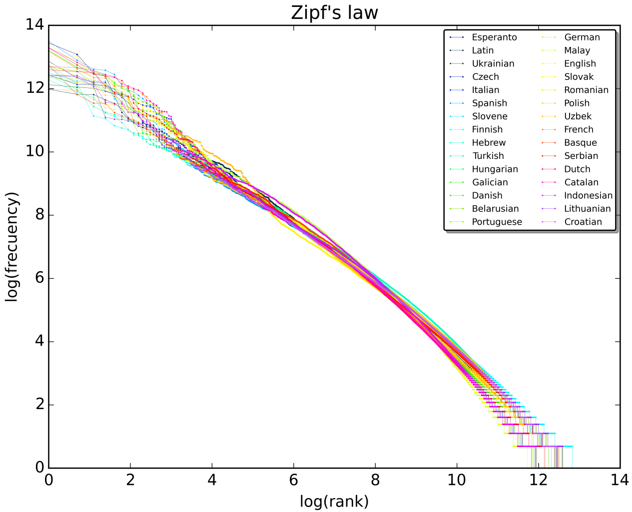
A family is *quasi-Zipfian* if the log-log plot is concave with a tangent of -1 somewhere on the curve. Fernholz & Fernholz (2020) show that a quasi-Atlas family is quasi-Zipfian if it is conservative and complete with

$$\lim_{n\to\infty} \mathbb{E}_n\left[\frac{X_{[n]}(t)}{X_{(1)}(t)}\right] \geq 2.$$

U.S. Stock Market Capitalization Distribution, 2010-2019



Word Count from Wikipedia



Zipfian and Non-Zipfian Pareto Distributions

- Zipfian Pareto distributions
  - Firm size, city size, word frequency, income and wealth of households
  - Data generated by time-dependent rank-based systems will often be Zipfian or quasi-Zipfian
  - Conservation and completeness should always hold in the limit
- Non-Zipfian Pareto distributions
  - Earthquake magnitude, cumulative book sales, intensity of wars
  - Data generated by other means, usually of a cumulative nature, do not necessarily follow Zipf's law

The Universality of Zipf's Law

- Zipfian and quasi-Zipfian distributions appear in many different fields and many different applications
  - Economics, demography, linguistics, etc.
- As a consequence, any explanation should appeal to statistics and mathematics rather than field-specific phenomena
  - Field-specific explanations of the Central Limit Theorem?
- Any time-dependent rank-based system that follows Gibrat's law will, provided enough ranks are sampled, be Zipfian
  - Conservation and completeness should always hold in the limit

First-Order Approximation

Let $\{Z_1(\tau), Z_2(\tau), \ldots\}$, $\tau \in \{1, 2, \ldots, T\}$, be a system of time-dependent data of indefinite size.

The *first-order approximation* for the top n ranks of this system is the first-order model $X_1, \ldots, X_n$ with

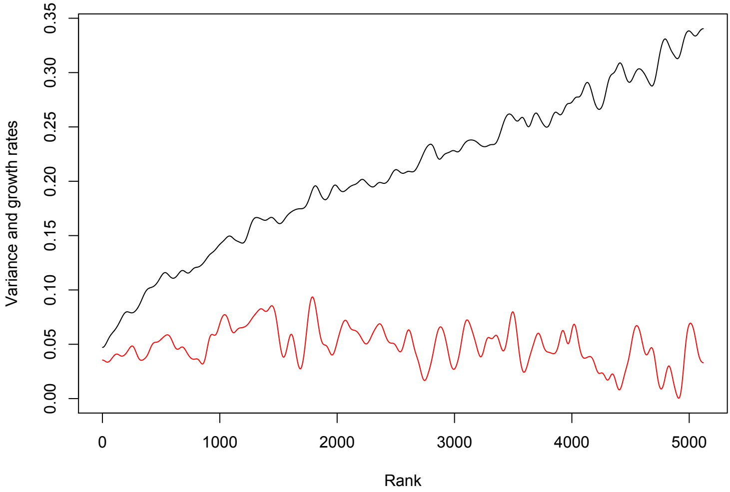
$$d \log X_i(t) = g_{r_t(i)}\, dt + G_n \mathbb{1}_{\{r_t(i)=n\}} dt + \sigma_{r_t(i)}\, dW_i(t),$$

where the parameters $g_1, \ldots, g_n$ and $\sigma_1^2, \ldots, \sigma_n^2$ are estimated using the time-series of $\{Z_1(\tau), Z_2(\tau), \ldots\}$.

First-Order Approximation of U.S. Capital Distribution

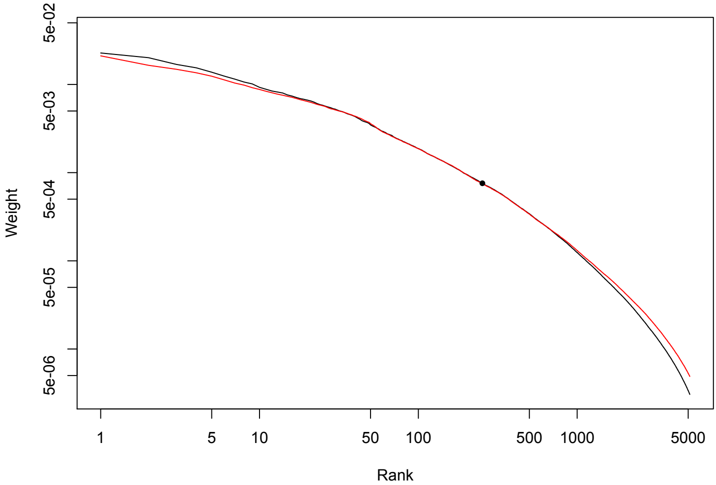
- Construct a first-order approximation of stock market capitalizations of U.S. companies from 1990-99
  - Estimate parameters g_k and σ_k following procedure of Fernholz (2017)
  - Changes in market capitalization also affect returns, so there is a link between capital distribution and stock returns
- First-order approximation is close to a quasi-Atlas model
  - Parameters satisfy $g_1 = g_2 = \cdots = g_n$ and $\sigma_k^2 = \sigma_1^2 + (k-1)(\sigma_2^2 - \sigma_1^2)$
  - If a sufficient number of ranks are considered, then the distribution should be quasi-Zipfian
  - Concave distribution curve with a tangent of -1 somewhere

First-Order Approximation of U.S. Capital Distribution



σ_k^2 (black), $-g_k$ (red)

First-Order Approximation of U.S. Capital Distribution



Actual (black), first-order approximation (red)

Second-Order Models

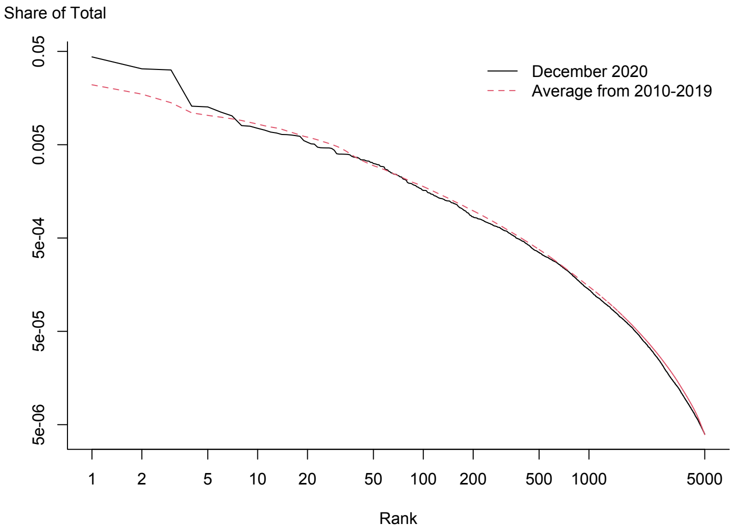
In some cases, the behavior of different entities within a time-dependent system may depend on both rank and name, so that

$$d \log X_i(t) = g_{r_t(i)}\, dt + \gamma_i\, dt + \sigma_{r_t(i)}\, dW_i(t),$$

where $\sigma_1^2, \ldots, \sigma_n^2$ are positive constants, $g_1, \ldots, g_n, \gamma_1, \ldots, \gamma_n$ are constants satisfying $g_1 + \cdots + g_n + \gamma_1 + \cdots + \gamma_n = 0$, as well as a stability condition (Ichiba et al., 2011), and $(W_1, \ldots, W_n)$ is a Brownian motion.

- In such *second-order models*, the processes X_i are not exchangeable
  - Different X_i will spend different amounts of time in each rank
  - No guarantee of Zipfian or quasi-Zipfian stationary distribution, even if parameters g_k and σ_k satisfy conditions from before

U.S. Capital Distribution Pre-2020 vs. End-2020



City Size Distributions

- According to Soo (2005), some city size distributions are neither Zipfian nor quasi-Zipfian
    - France, Argentina, Russia, Mexico, New York State, etc.
    - These systems are not rank-based only, but also name-based (largest cities are fundamentally, persistently different from the rest)
- Davis & Weinstein (2002) show that after the destruction of WWII, the cities that grew to be largest were the same as those from before
    - Japanese city growth is not rank-based only, but also name-based
- Both of these observations can be explained by a second-order model

Wealth Distribution and Long-Run Mobility

- Surprising findings for long-run mobility that are impossible to match using standard random growth models of wealth distribution
  - Wealth-rank coefficient after 585 years is 0.1: Barone & Mocetti (2021)
  - Both parent and grandparent wealth-rank have predictive power for child wealth-rank: Boserup, Kopczuk, & Kreiner (2014)
- Second-order models of intergenerational wealth dynamics can match all of these observations
  - Benhabib, Bisin, & Fernholz (2022)

The End

Thank You