The Universality of Zipf's Law

Ricardo T. Fernholz [1] Robert Fernholz [2]

[1]Claremont McKenna College

[2]Intech Investments

June 29, 2018

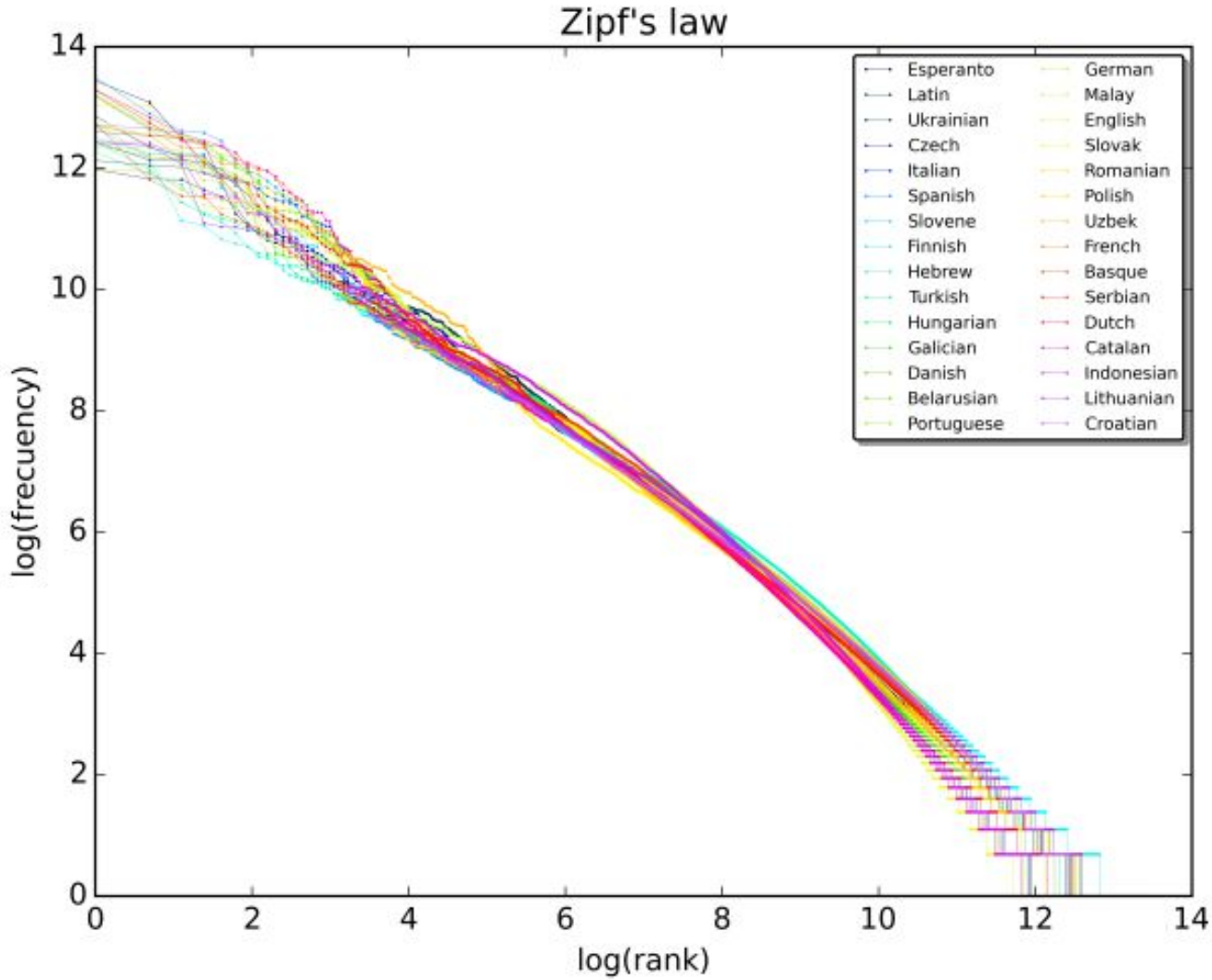
Zipf's Law

"*Zipf's law* states that given some corpus of natural language utterances, 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 it and sought to explain it (Zipf (1935, 1949)), though he did not claim to have originated it." (Wikipedia, 2017)

Zipf's Law

"*Zipf's law* states that given some corpus of natural language utterances, 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 it and sought to explain it (Zipf (1935, 1949)), though he did not claim to have originated it." (Wikipedia, 2017)

Word Count from Wikipedia



Pareto Distributions and Zipf's Law

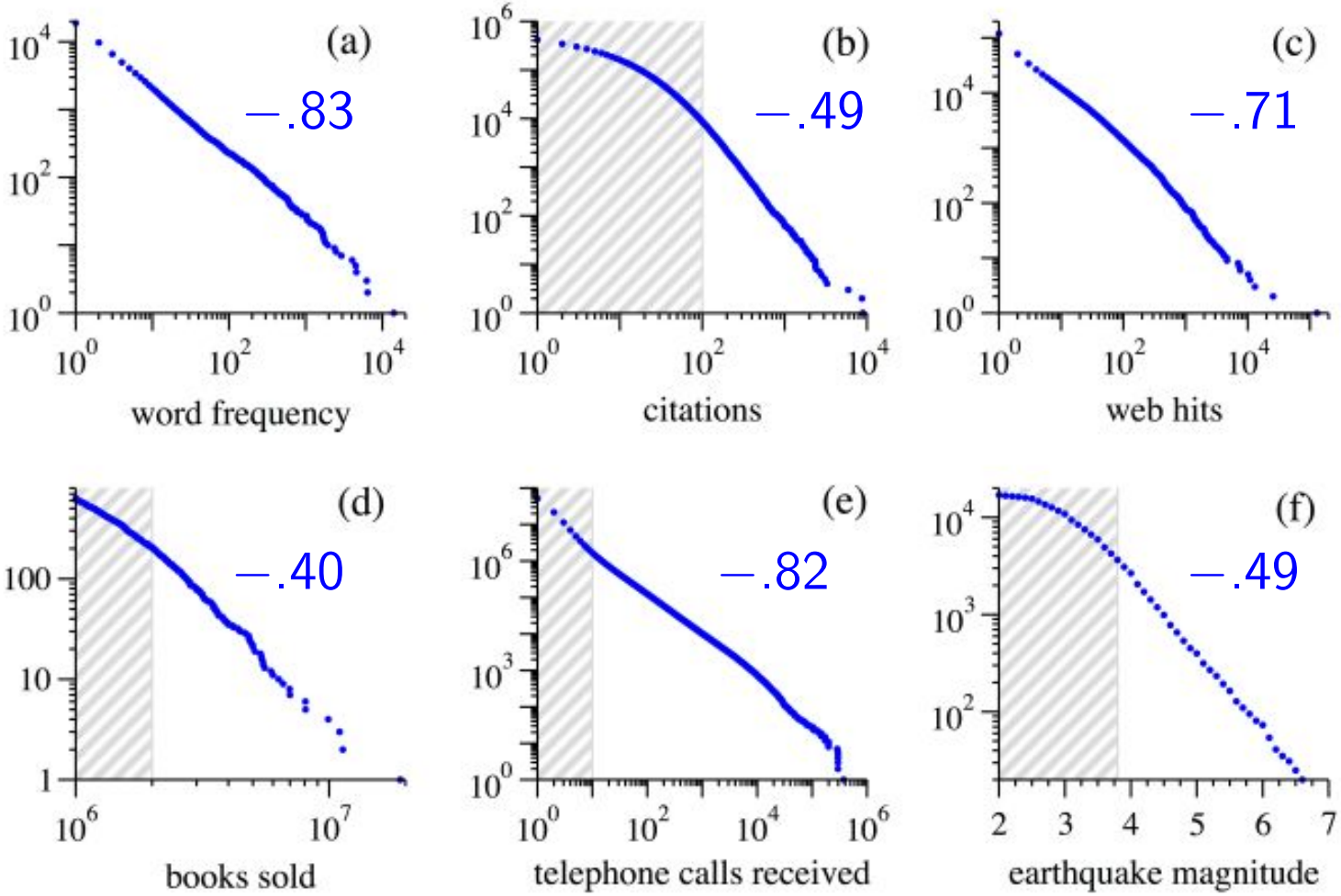
Pareto distribution, or power law

- Log-log plot of the data versus rank is approximately a straight line

Zipf's Law

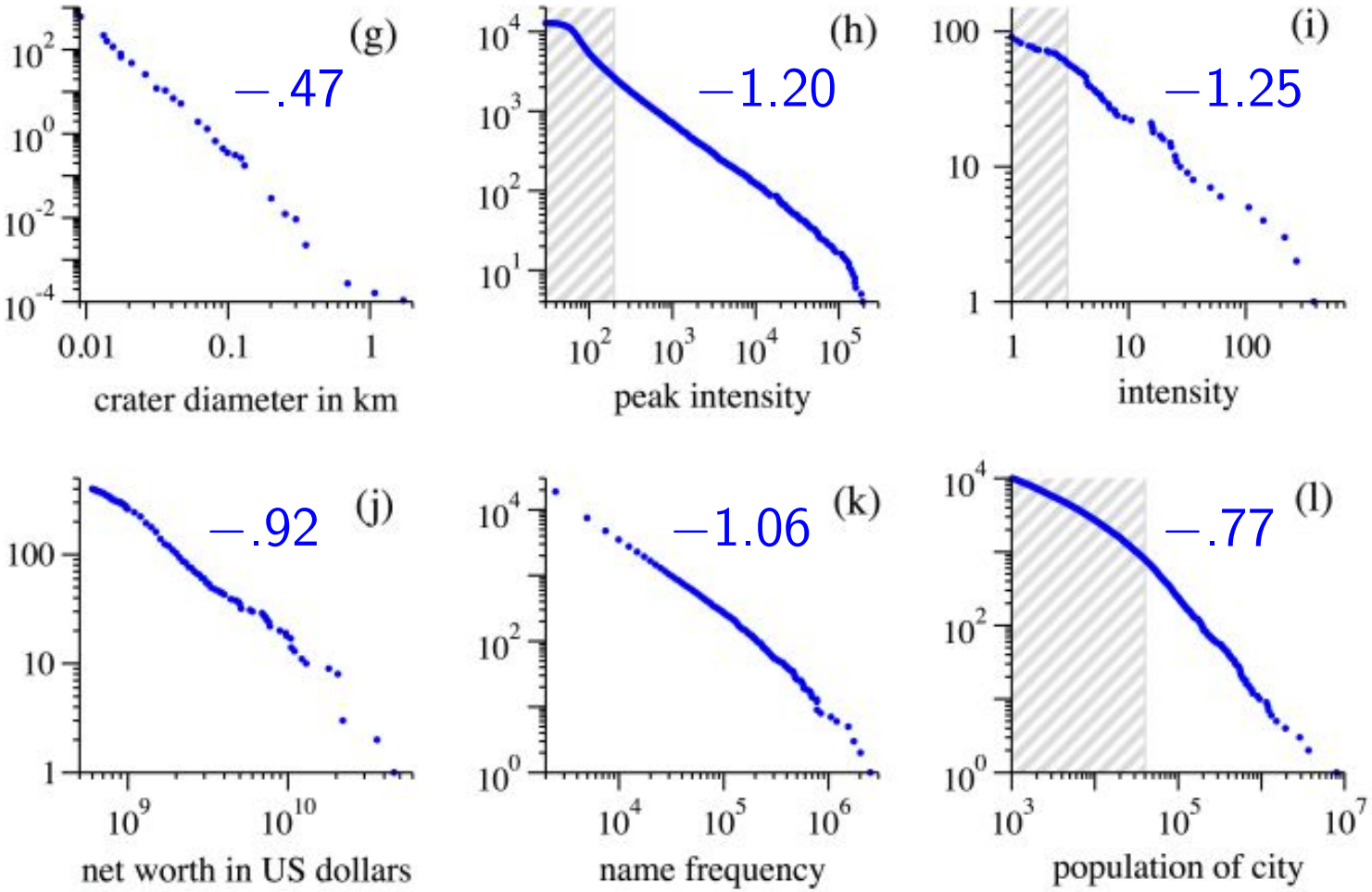
- Log-log plot of the data versus rank is approximately a straight line with slope -1
- Weaker form of Zipf's law requires that log-log plot of the data versus rank is concave with a tangent line of slope -1 at some point

Examples of Pareto distributions



Log-log slopes in blue (From Newman (2006))

Examples of Pareto distributions



Log-log slopes in blue (From Newman (2006))

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

Non-Zipfian Pareto distributions

- Earthquake magnitude, cumulative book sales, intensity of wars
- Data generated by other means, usually of a cumulative nature, do not seem to follow Zipf's law

Ranked Continuous Semimartingales

We use systems of positive continuous semimartingales $\{X_1, \ldots, X_n\}$ to approximate systems of time-dependent empirical data. Fernholz (2002) shows that 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 random permutation $r_t \in \Sigma_n$ such that $r_t(i) < r_t(j)$ if $X_i(t) > X_j(t)$ or if $X_i(t) = X_j(t)$ and $i < j$

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 $\log(X_{(k)}/X_{(k+1)})$, with $\Lambda^X_{0,1} = \Lambda^X_{n,n+1} = 0$

Asymptotic Stability

A system of positive continuous semimartingales $\{X_1, \ldots, X_n\}$ is *asymptotically stable* if

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

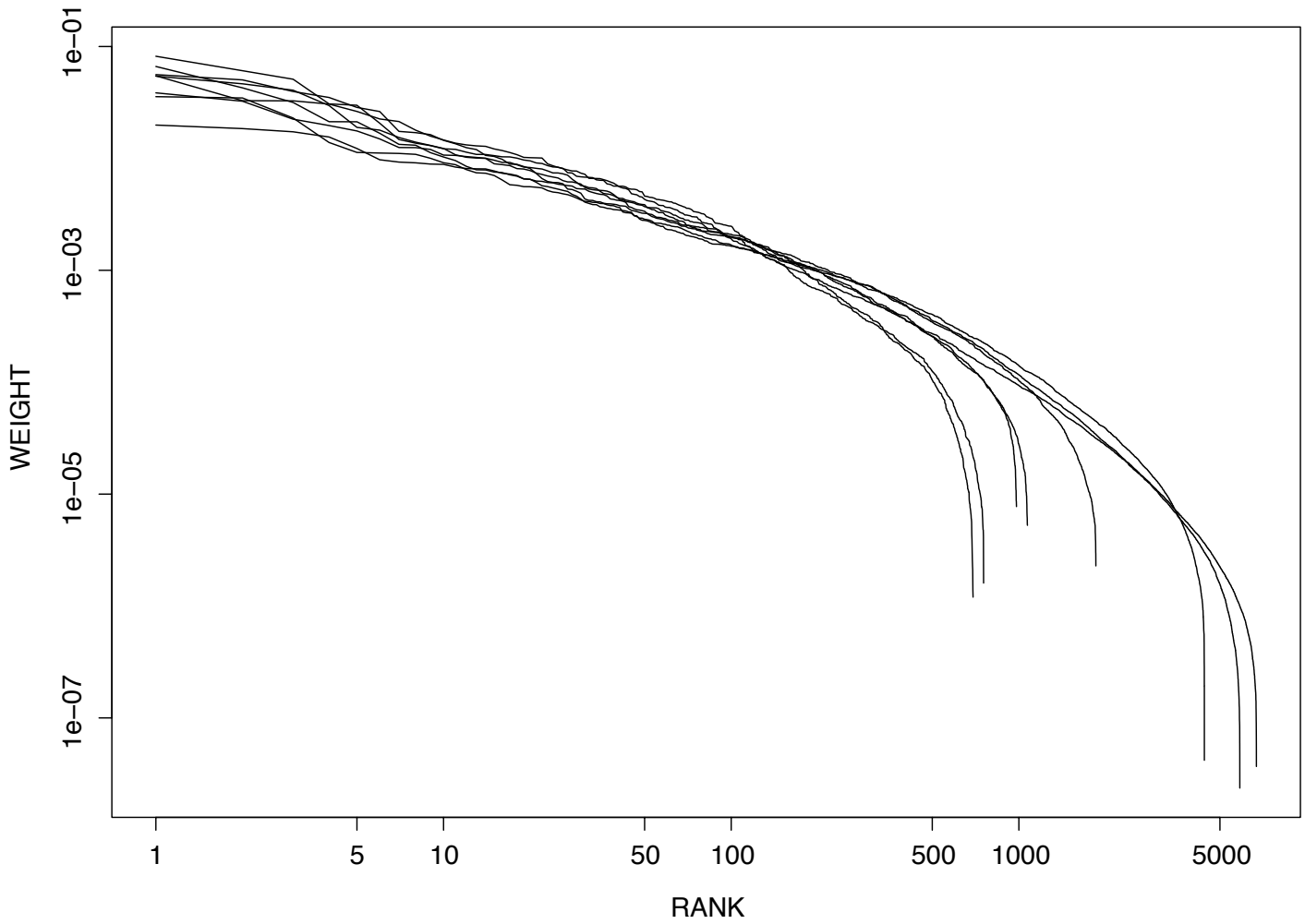
for $k = 1, \ldots, n-1$, where $\lambda_{k,k+1}$ and $\sigma^2_{k,k+1}$ are constants.

The systems of continuous semimartingales we consider will be asymptotically stable and will also satisfy

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

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

U.S. Capital Distribution, 1929 to 1999



Market weight curves (Fernholz, 2002)

Behavior of Ranked Systems

For $k = 1, \ldots, n$, let $X_{[k]} \triangleq X_{(1)} + \cdots + X_{(k)}$, and suppose that $X_1, \ldots, X_n$ satisfy

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

Then,

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

This describes the dynamic relationship between the combined value, or size, $X_{[k]}$, of the k top ranks and the local time process $\Lambda^X_{k,k+1}$. It also allows us to extend the definition of local time to systems of data.

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.

Asymptotically stable model

Processes X_i are exchangeable, so they spend equal time in each rank

Each X_i has zero asymptotic log-drift, so the entire system has zero asymptotic log-drift (Fernholz, 2002; Banner et al., 2005)

Asymptotic Distribution of Atlas Models

Atlas models satisfy

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

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

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

Hence, for large enough k,

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

so Atlas models follow Pareto distributions, and Zipf's law is equivalent to $\sigma^2/2 = g$.

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

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

We wish to show that an Atlas family is Zipfian if and only if

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

and

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{ngX_{(n)}(t)}{X_{[n]}(t)}\right] = 0. \tag{B}$$

Conservation

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

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

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

should also be approximately constant.

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

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{dX_{[n]}(t)}{X_{[n]}(t)}\right] = 0. \tag{A}$$

Completeness

For a system of data $\{Z_1(t), Z_2(t), \ldots\}$, we have the dynamic relationship

$$dZ_{[k]}(t) = \sum_{i=1}^{n} \mathbb{1}_{\{r_t(i) \leq k\}} dZ_i(t) + \frac{1}{2} Z_{(k)}(t) d\Lambda^Z_{k,k+1}(t), \quad \text{a.s.,} \qquad (**)$$

where the last term compensates for entry/exit from $Z_{[k]}$. In order that the system not depend on mass that enters from outside, we require that

$$\lim_{k \to \infty} \mathbb{E}\left[\frac{Z_{(k)}(t)}{Z_{[k]}(t)} d\Lambda^Z_{k,k+1}(t) \right] = 0.$$

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

$$\lim_{n \to \infty} \mathbb{E}\left[\frac{ngX_{(n)}(t)}{X_{[n]}(t)} \right] = 0. \qquad (B)$$

Zipf's Law

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

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

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

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

where M is a martingale incorporating all the σW_i, so

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

where the last term plays the role of entry/exit in the system Z. Hence,

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

Zipf's Law

For an Atlas model we have

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

and we can calculate

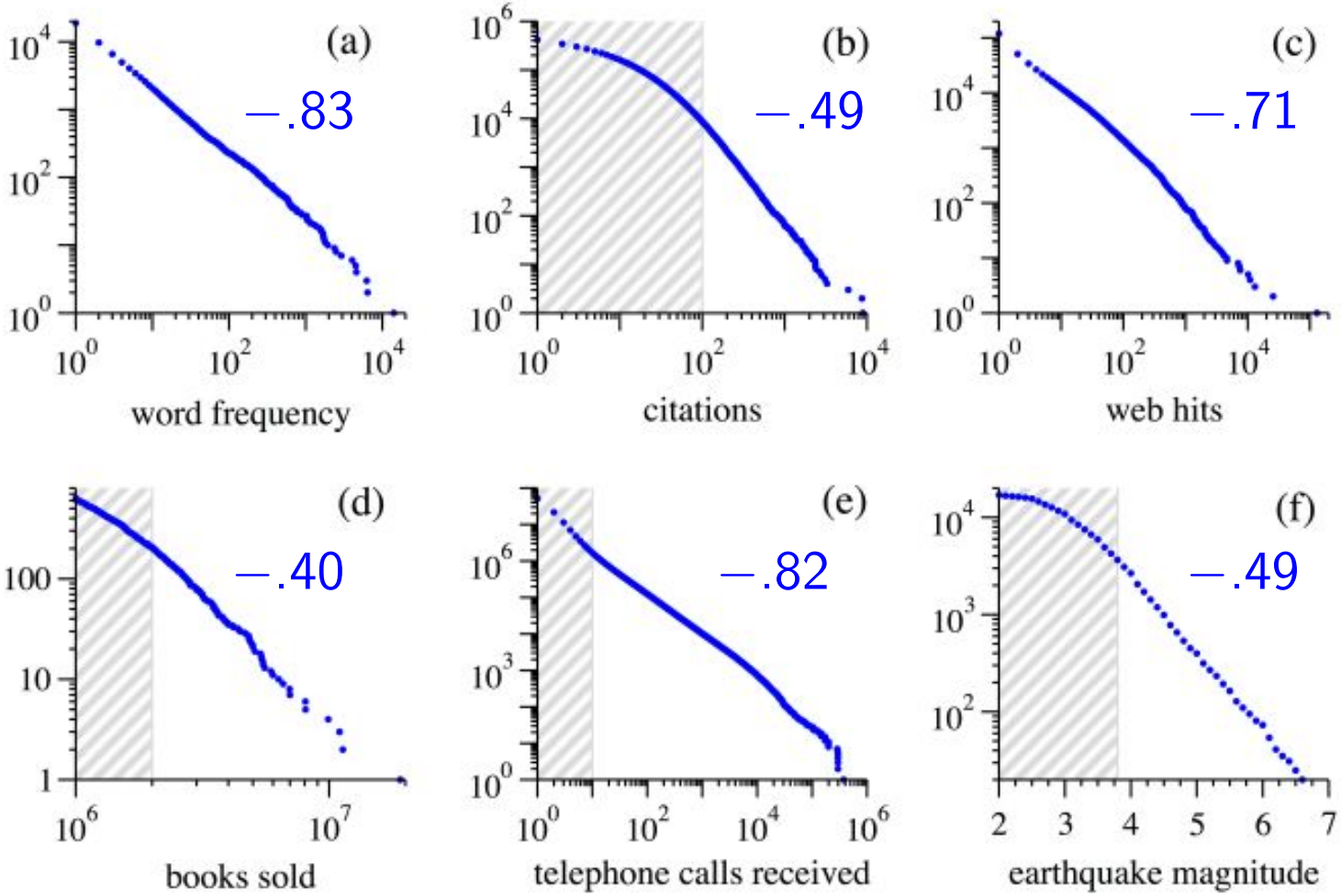
$$\mathbb{E}\left[\frac{ngX_{(n)}(t)}{X_{[n]}(t)}\right] = \begin{cases} O(1) & \text{for } \sigma^2/2 < g, \\ O(1/\log n) & \text{for } \sigma^2/2 = g, \\ O\left(n^{(1-\sigma^2/2g)}\right) & \text{for } \sigma^2/2 > g. \end{cases}$$

Hence,

$$\text{(A)} \quad \lim_{n\to\infty} \mathbb{E}\left[\frac{dX_{[n]}(t)}{X_{[n]}(t)}\right] = 0 \quad \text{plus} \quad \text{(B)} \quad \lim_{n\to\infty} \mathbb{E}\left[\frac{ngX_{(n)}(t)}{X_{[n]}(t)}\right] = 0,$$

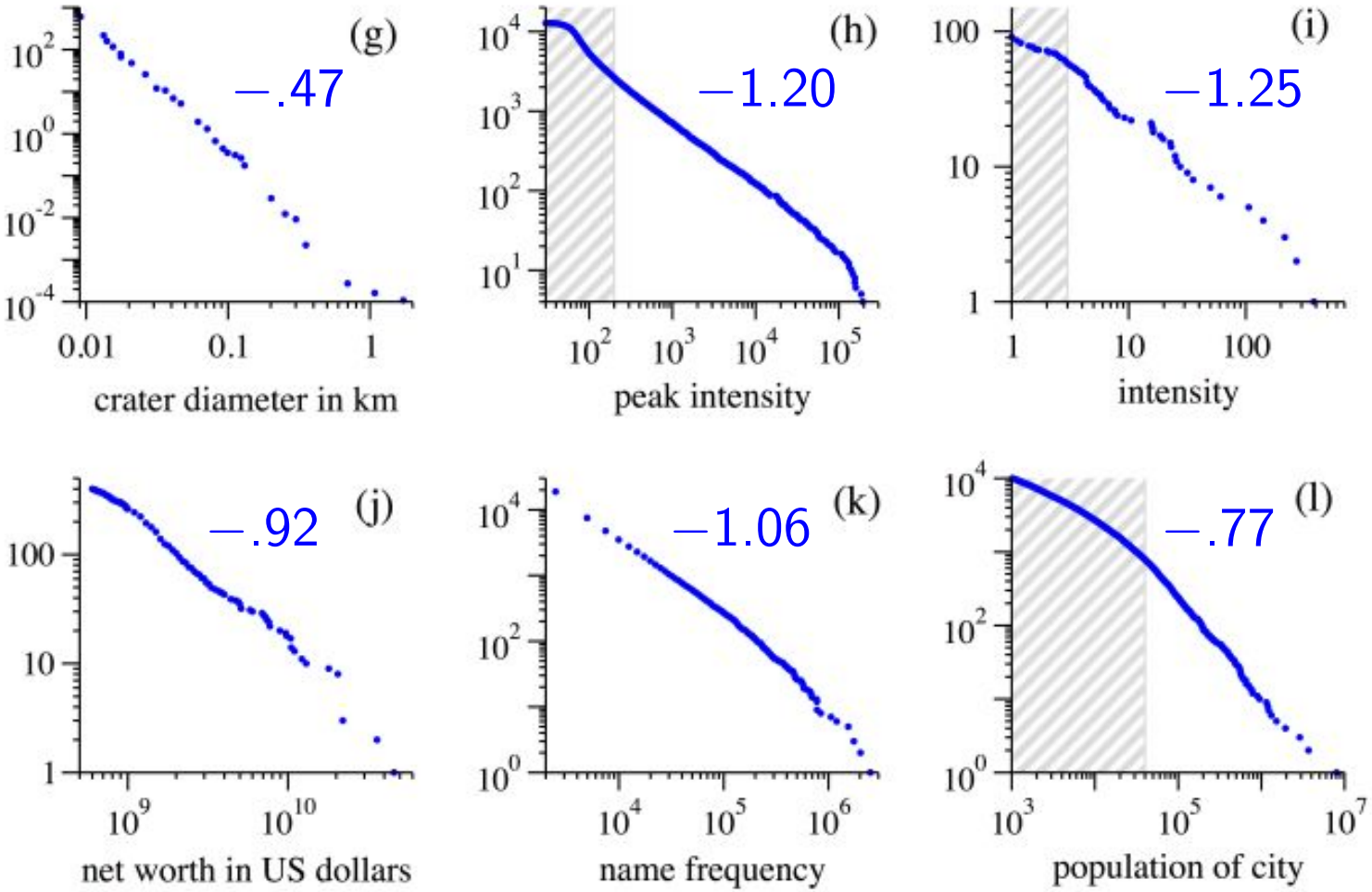
is equivalent to $\sigma^2/2 = g$, and this is equivalent to Zipf's Law.

Examples of Pareto distributions



Log-log slopes in blue (From Newman (2006))

Examples of Pareto distributions



Log-log slopes in blue (From Newman (2006))

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 (Banner et al., 2005). First-order models are asymptotically stable with

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

and

$$\sigma^2_{k,k+1} = \sigma_k^2 + \sigma_{k+1}^2, \quad \text{a.s.}$$

First-Order Approximation

Suppose that $\{Z_1(t), Z_2(t), \ldots\}$ is an asymptotically stable system of time-dependent data of indefinite size with parameters $\boldsymbol{\lambda}_{k,k+1}$ and $\boldsymbol{\sigma}^2_{k,k+1}$. Then the *first-order approximation* for the top n ranks of this system is the first-order model $X_1, \ldots, X_n$ with parameters

$$g_k = \frac{1}{2}\boldsymbol{\lambda}_{k-1,k} - \frac{1}{2}\boldsymbol{\lambda}_{k,k+1}, \quad \text{for } k = 1, \ldots, n-1$$

$$g_n = \frac{1}{2}\boldsymbol{\lambda}_{n-1,n}$$

$$\sigma_1^2 = \frac{1}{2}\boldsymbol{\sigma}^2_{1,2}$$

$$\sigma_k^2 = \frac{1}{4}\left(\boldsymbol{\sigma}^2_{k-1,k} + \boldsymbol{\sigma}^2_{k,k+1}\right), \quad \text{for } k = 2, \ldots, n.$$

In this manner, we can construct a first-order approximation for any asymptotically stable system.

First-Order Approximation

The first-order approximation $\{X_1, \ldots, X_n\}$ satisfies

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

a.s., with parameters

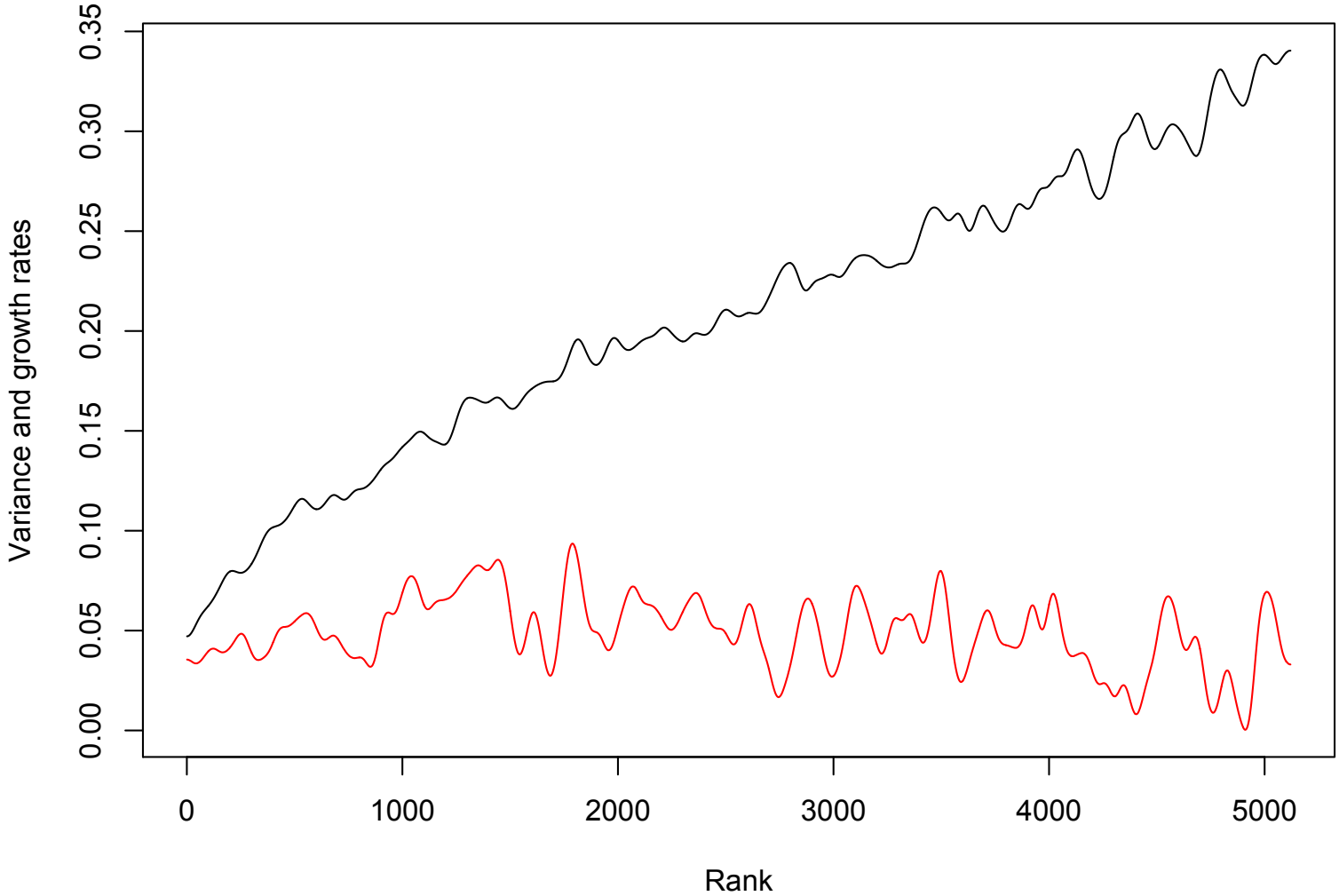
$$\lambda_{k,k+1} = \boldsymbol{\lambda}_{k,k+1}, \quad \sigma_1^2 = \frac{1}{2}\boldsymbol{\sigma}_{1,2}^2, \quad \sigma_k^2 = \frac{1}{4}\big(\boldsymbol{\sigma}_{k-1,k}^2 + \boldsymbol{\sigma}_{k,k+1}^2\big).$$

If the data $\{Z_1(t), Z_2(t), \ldots\}$ satisfy

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \big(\log Z_{(k)}(t) - \log Z_{(k+1)}(t) \big) dt = -\frac{\boldsymbol{\sigma}_{k,k+1}^2}{2\boldsymbol{\lambda}_{k,k+1}}, \qquad (*)$$

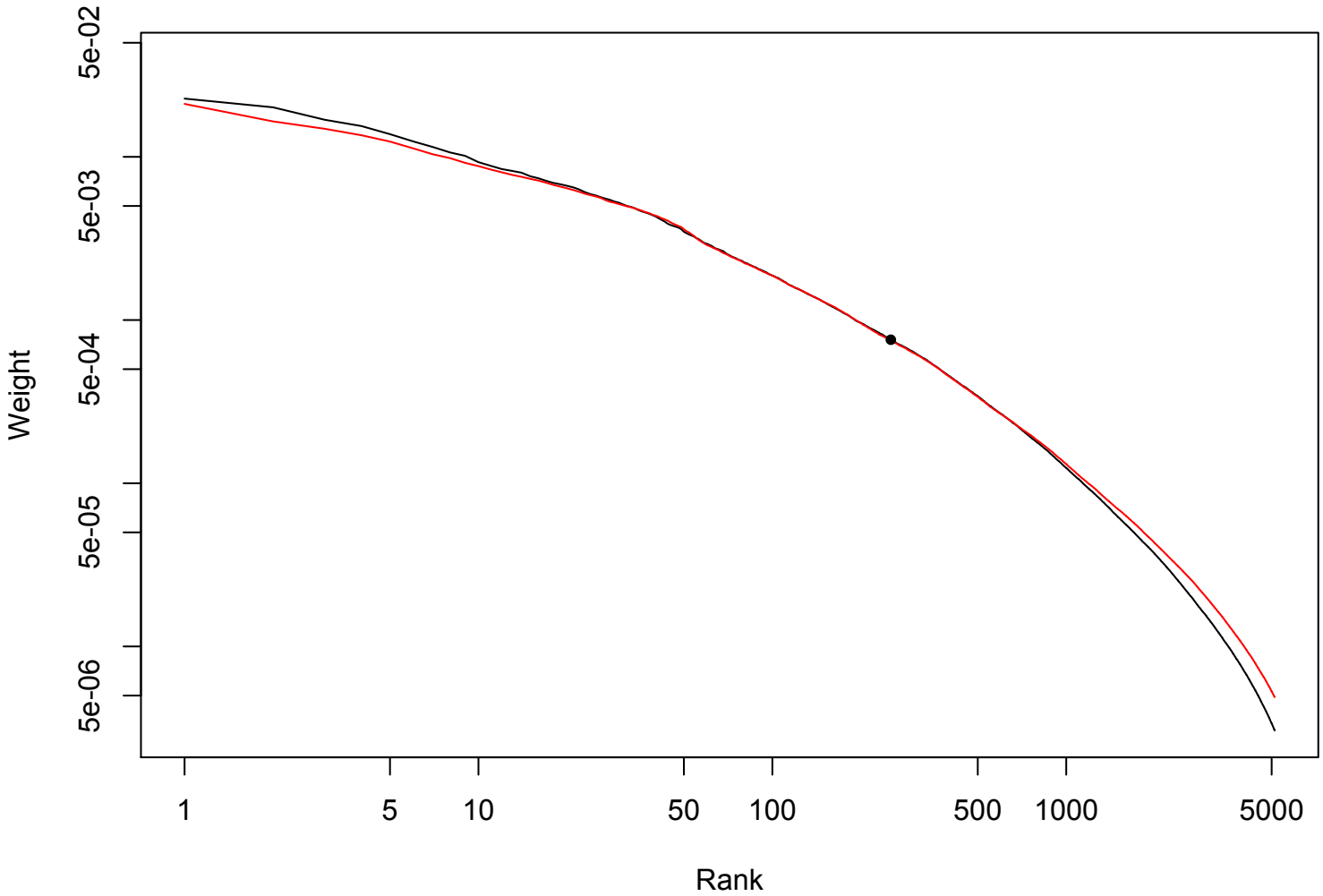
then the X distribution is a smoothed version of the Z distribution.

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 (red)

First-Order Families

A *first-order family* is a class of first-order models $\{X_1, \ldots, X_n\}$, for $n \in \mathbb{N}$, defined as in

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

with the common parameters g_k and σ_k^2 such that, for $k \in \mathbb{N}$,

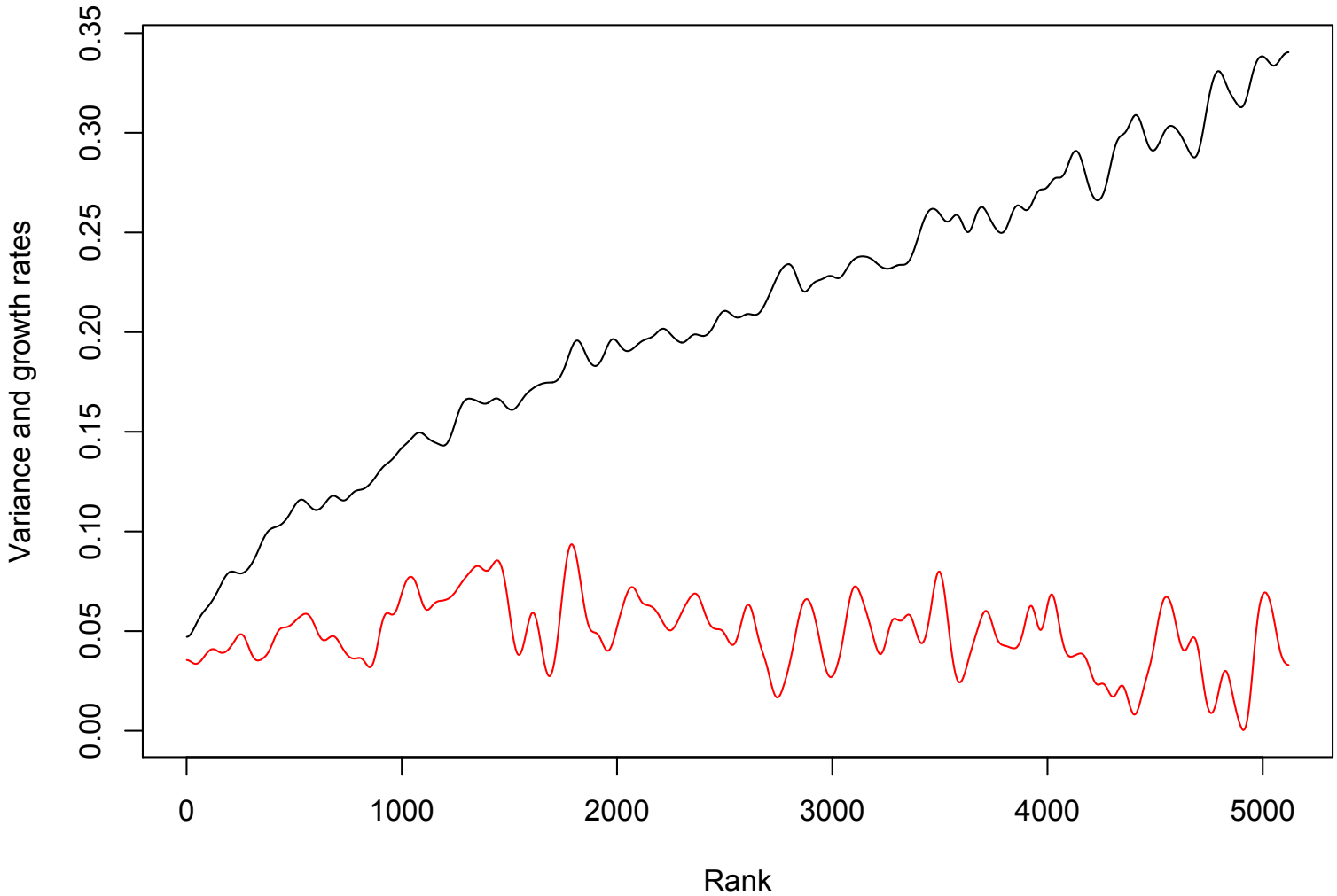
$$g_1 + \cdots + g_k < 0,$$
$$\sigma_k^2 > 0,$$

and $G_n = -(g_1 + \cdots + g_n)$. A first-order family is *simple* if it is of the form

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

where $\sigma_1^2 \leq \cdots \leq \sigma_n^2$.

First-Order Approximation of U.S. Capital Distribution



σ_k^2 (black), $-g_k$ (red)

Simple First-Order Families

For a simple first-order family, the slope of the log-log plot of $X_{(k)}$ versus rank k,

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

is increasingly negative, so the distribution curve is concave.

We wish to show that a simple first-order family is *quasi-Zipfian*—the distribution curve is concave with a tangent of slope -1 at some point—if the family is conservative, complete, and satisfies

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

Zipf's Law

If our family is of the form

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

with $\sigma_1^2 \leq \cdots \leq \sigma_n^2$, then

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

Hence, if

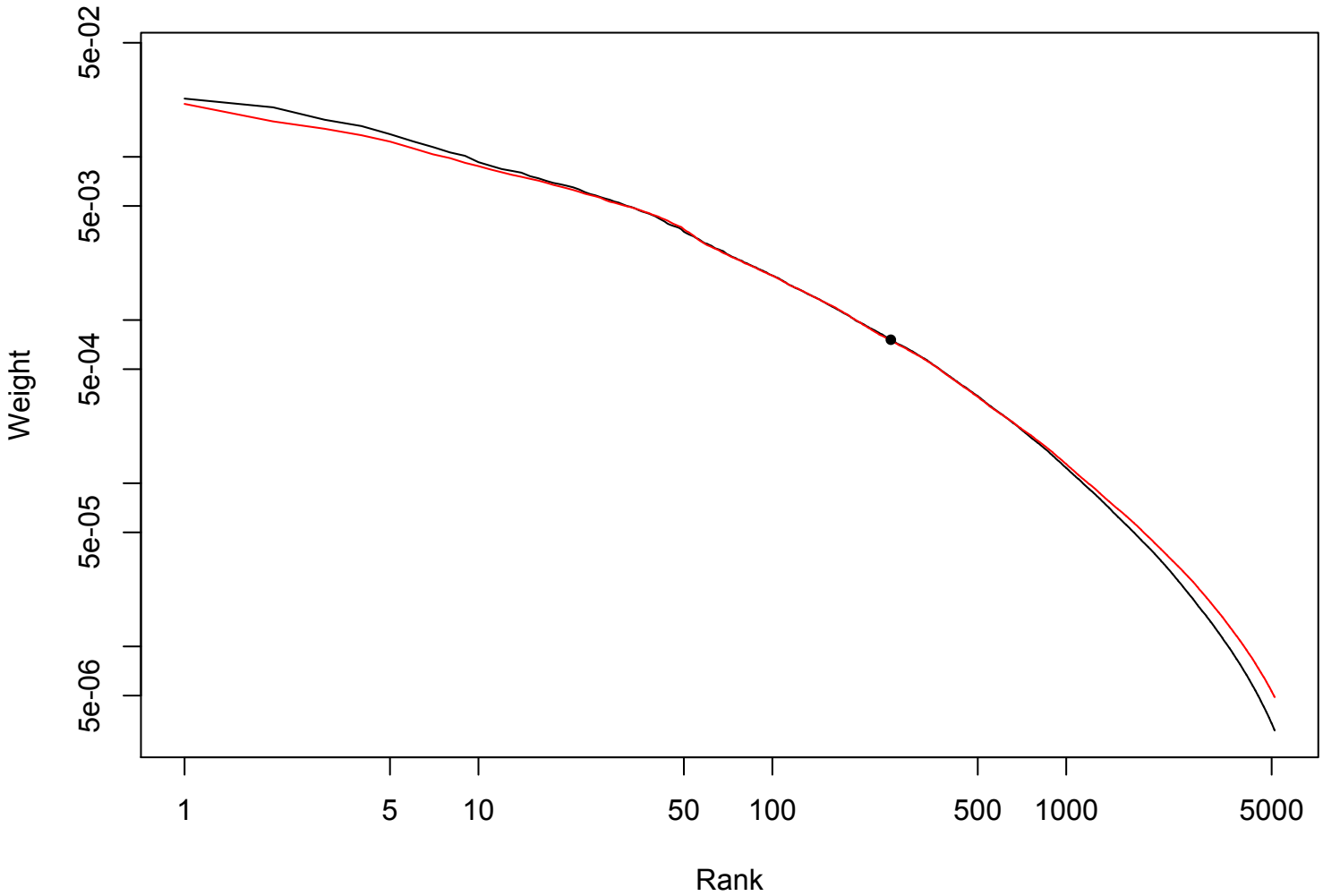
$$\text{(A)}\quad \lim_{n\to\infty}\mathbb{E}\left[\frac{dX_{[n]}(t)}{X_{[n]}(t)}\right] = 0 \quad\text{and}\quad \text{(B)}\quad \lim_{n\to\infty}\mathbb{E}\left[\frac{ngX_{(n)}(t)}{X_{[n]}(t)}\right] = 0,$$

then,

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

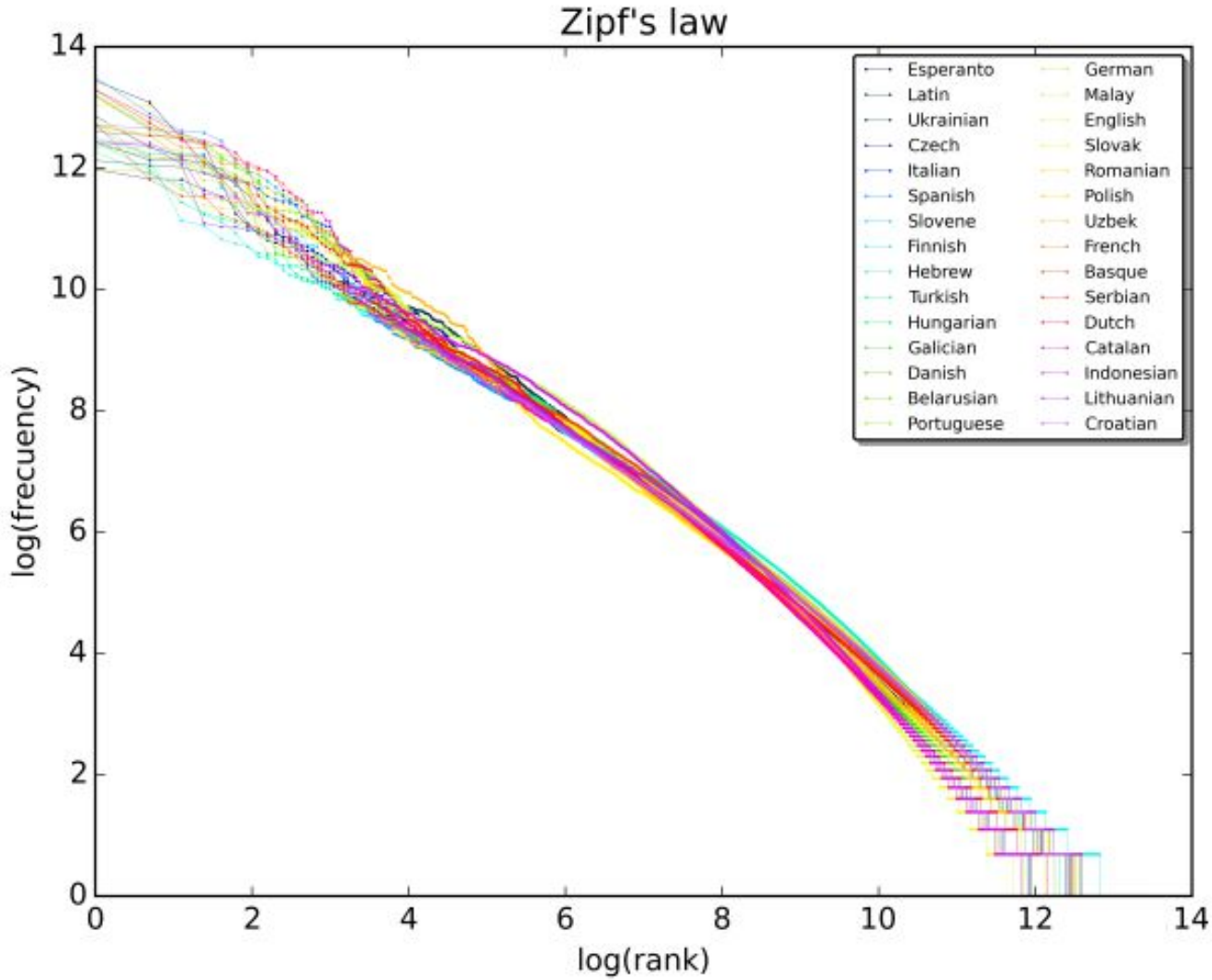
The family is quasi-Zipfian, having a tangent of slope -1 at some point.

U.S. Capital Distribution, 1990-1999



Actual (black), first-order (red)

Word Count from Wikipedia



The End

Thank You