Zipf's Law

Robert Fernholz
INTECH

Joint research with Ricardo Fernholz

Thera Stochastics
Santorini, Greece
May 31 – June 2, 2017

This talk is dedicated to Ioannis Karatzas
on the occasion of his 65th birthday.

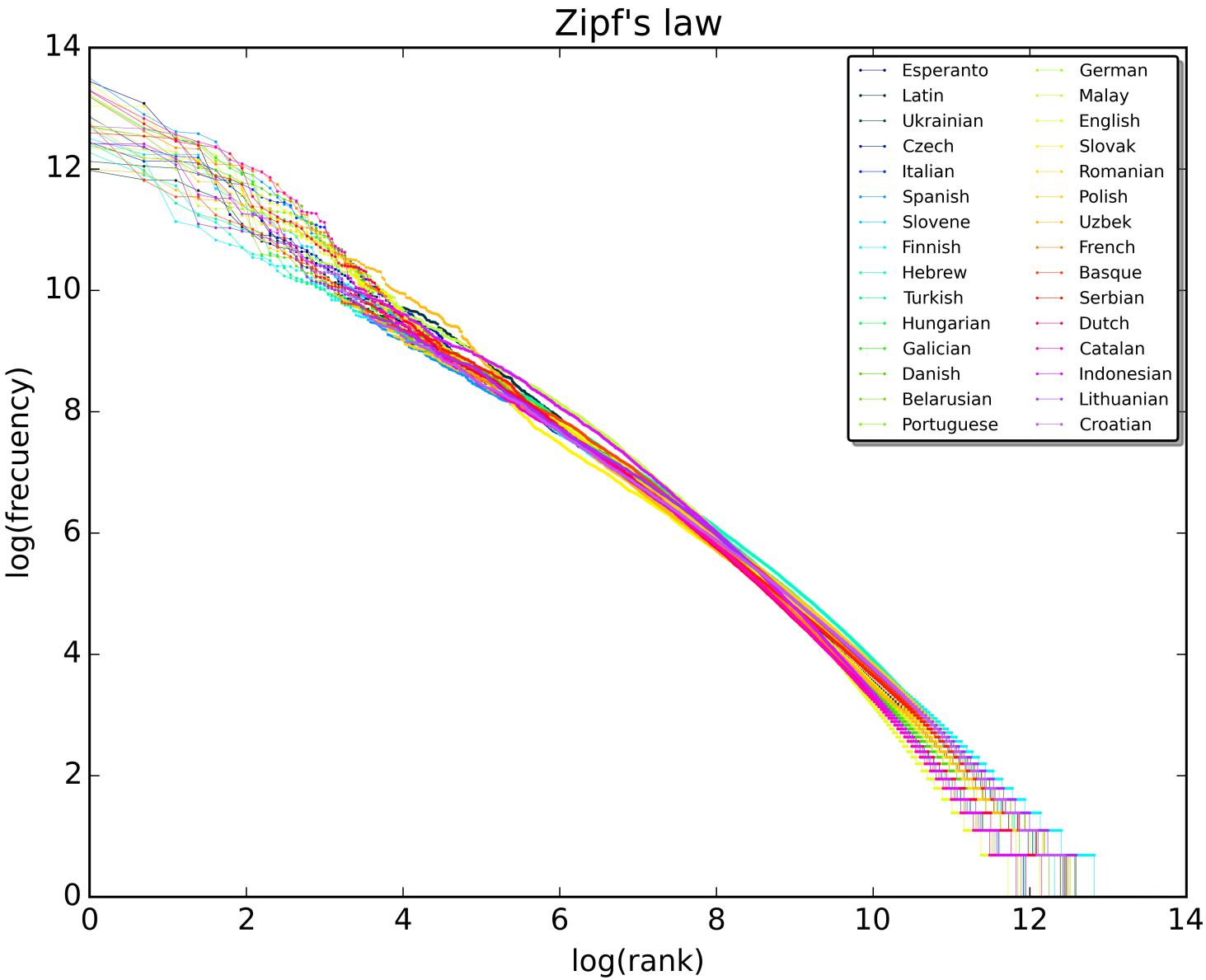
Introduction

"*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." (From Wikipedia (2017).)

Introduction

"*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." (From Wikipedia (2017).)

Word count from Wikipedia

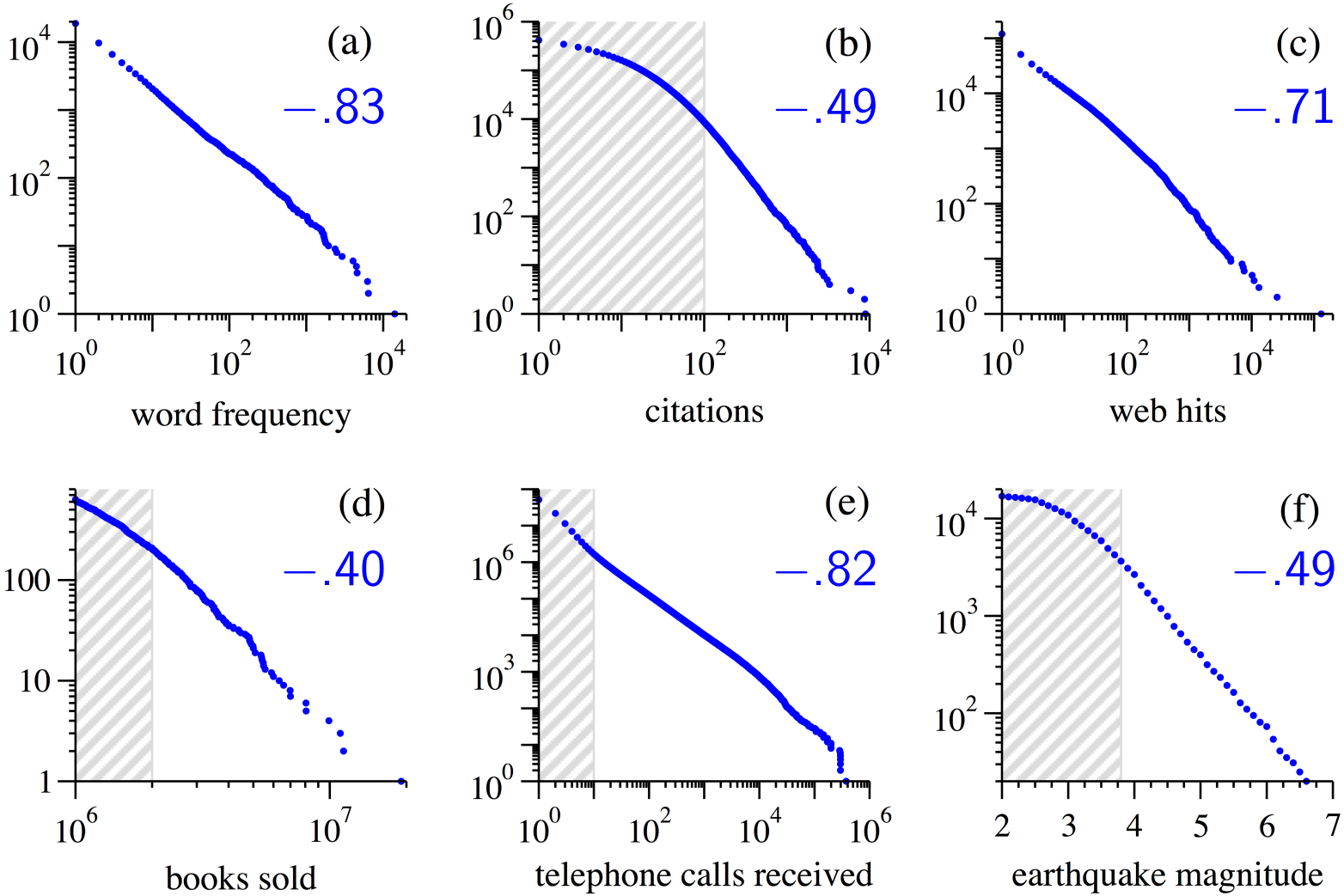


Power laws and the Pareto distribution

Data follow a *power law* or *Pareto distribution* if a log-log plot of the data versus rank is approximately a straight line. Pareto distributions can result from *self-organized criticality* or from time-dependent systems.

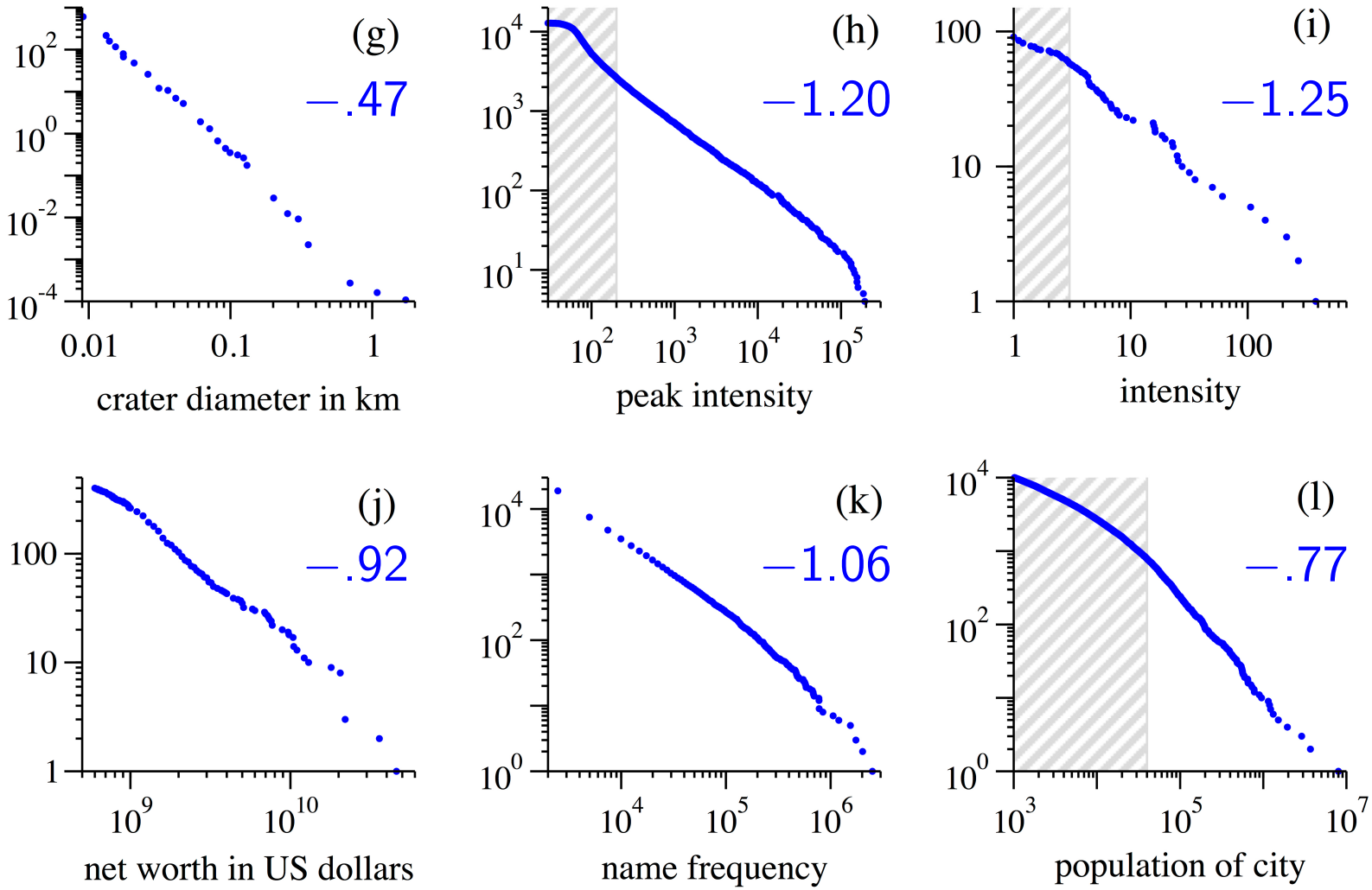
A Pareto distribution follows *Zipf's law* if the slope of the log-log plot is -1. Zipf's law is a form of *universality,* since many classes of data seem to follow this distribution. Specifically, certain time-dependent, rank-based systems seem to follow Zipf's law, and we shall try to characterize these systems.

Examples of Pareto distributions



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

Examples of Pareto distributions



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

Members and families

We wish to model systems of positive-valued, time-dependent data $\{\Xi_1(t), \Xi_2(t), \ldots\}$ of indefinite size. These data represent two classes of objects, *members* and *families*. The members are contained within the families, and $\Xi_i(t)$ indicates the number of members contained within the ith family at time t. Examples of members within families are:

- people within cities;
- occurrences within words;
- dollars within family fortunes;
- individuals within surnames;
- dollars within company capitalizations;
- birds within species.

Trends and sampling

The data we consider $\{\Xi_1(t), \Xi_2(t), \ldots\}$ might have a common global trend of the form $G(t)dt$, e.g., population growth, Wikipedia growth, GDP growth, etc. We shall study log-differences, so a global trend does not affect us, and it is convenient to assume it to be zero.

Alternatively, we can *sample* the total population with a constant number of people, words, dollars, etc., in our sample over time. This could introduce *sampling error* but should not materially affect the shape of the distribution curve.

In any case, to simplify the exposition, we shall assume henceforth that the total population we observe is free of trends.

Continuous semimartingales

To model the data $\{\Xi_1(t), \Xi_2(t), \ldots\}$ we shall use continuous semimartingales $X_1, X_2, \ldots$ of the form

$$d \log X_i(t) = \gamma_i(t) dt + \sigma_i(t) dW_i(t),$$

where W is a Brownian motion and the processes γ_i and σ_i are measurable and adapted to the Brownian filtration.

A model of this form might be reasonable if, e.g.,

1. the changes $d\Xi_i(t)$ are proportional to the values $\Xi_i(t)$;
2. the log-changes $d \log \Xi_i(t)$ are composed of many small, independent perturbations;
3. the changes in the different Ξ_i are independent.

Rank processes

For a system of positive continuous semimartingales $X_1, \ldots, X_n$ we define the *rank function* to be the 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 $i < j$. The *rank processes* $X_{(1)} \geq \cdots \geq X_{(n)}$ are defined by $X_{(r_t(i))}(t) = X_i(t)$.

If the X_i satisfy certain regularity conditions, e.g., they spend no local time at triple points, then the rank processes satisfy,

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

where $\Lambda^X_{k,k+1}$ is the local time at the origin for $\log(X_{(k)}/X_{(k+1)})$, with $\Lambda^X_{0,1} = \Lambda^X_{n,n+1} \equiv 0$ (Fernholz (2002)).

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.;

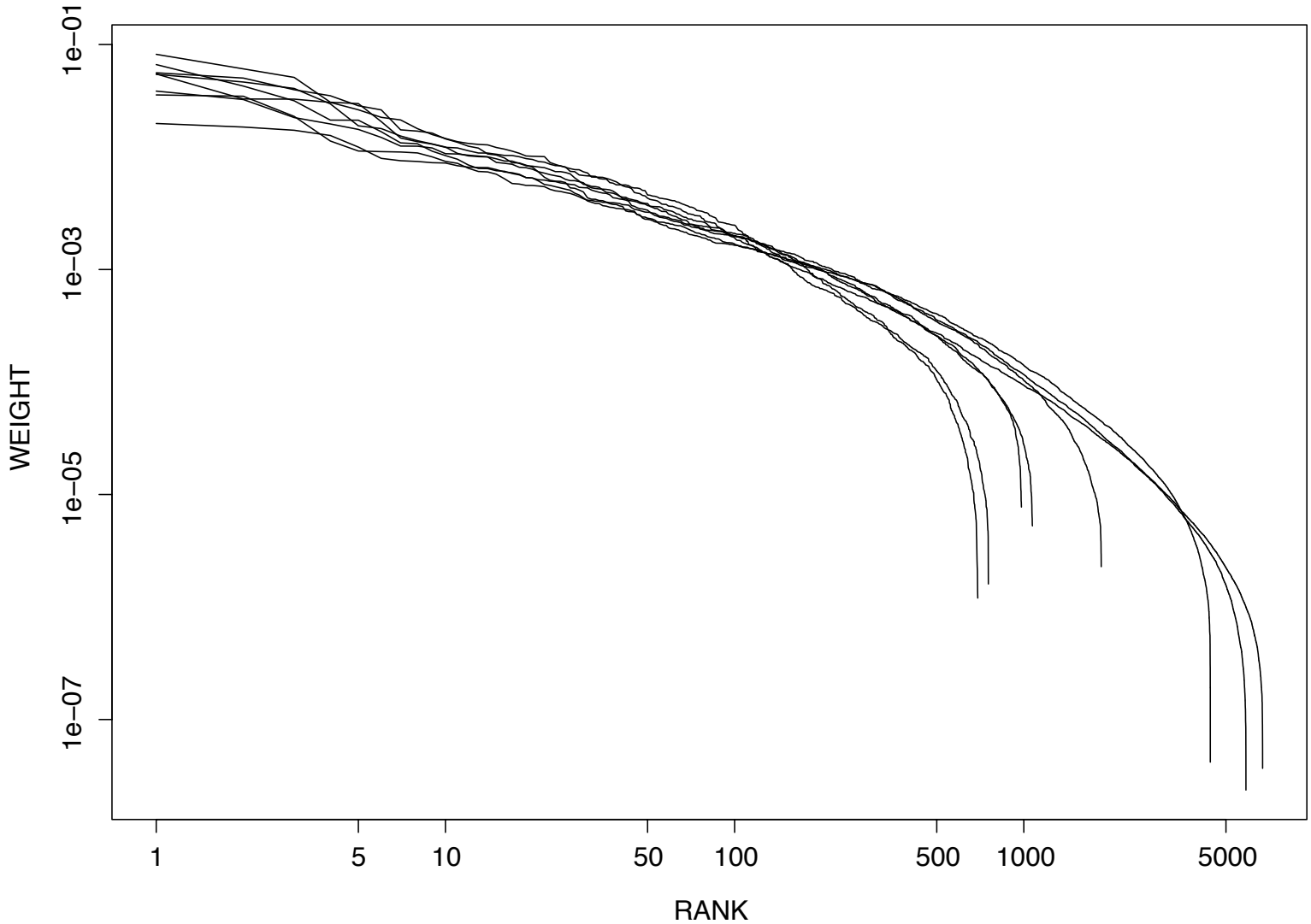
for $k = 1, \ldots, n-1$, where $\lambda_{k,k+1}$ and $\sigma^2_{k,k+1}$ are constants.

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

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

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

U.S. Capital Distribution, 1929 to 1999



Market weight curves (From Fernholz (2002)).

Conservation of 'mass'

Suppose that for the data $\{\Xi_1(t), \Xi_2(t), \ldots\}$ the "total mass"

$$\Xi_{(1)}(t) + \Xi_{(2)}(t) + \cdots$$

remains constant.

The mass of the top n ranks $\Xi_{(1)}, \ldots, \Xi_{(n)}$ is defined by

$$\Xi_{[n]}(t) \triangleq \Xi_{(1)}(t) + \cdots + \Xi_{(n)}(t),$$

and since the sample has constant total mass, for large enough n the mass of the top n ranks should also be approximately constant. Hence, we impose the condition on the model $X_1, \ldots, X_n$ that

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

Behavior of ranked systems

Let us suppose for the moment that the data processes Ξ_i are continuous semimartingales that spend no local time at triple points. In this case, the rank processes $\Xi_{(k)}$ will satisfy

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

for all k. By Itô's rule, for all k, a.s.,

$$\frac{d\Xi_{(k)}(t)}{\Xi_{(k)}(t)} = \sum_{i=1}^{\infty} \mathbb{1}_{\{r_t(i)=k\}} \frac{d\Xi_i(t)}{\Xi_i(t)} + \frac{1}{2} d\Lambda^{\Xi}_{k,k+1}(t) - \frac{1}{2} d\Lambda^{\Xi}_{k-1,k}(t)$$
$$= \sum_{i=1}^{\infty} \mathbb{1}_{\{r_t(i)=k\}} \frac{d\Xi_i(t)}{\Xi_{(k)}(t)} + \frac{1}{2} d\Lambda^{\Xi}_{k,k+1}(t) - \frac{1}{2} d\Lambda^{\Xi}_{k-1,k}(t).$$

Behavior of ranked systems

Hence,

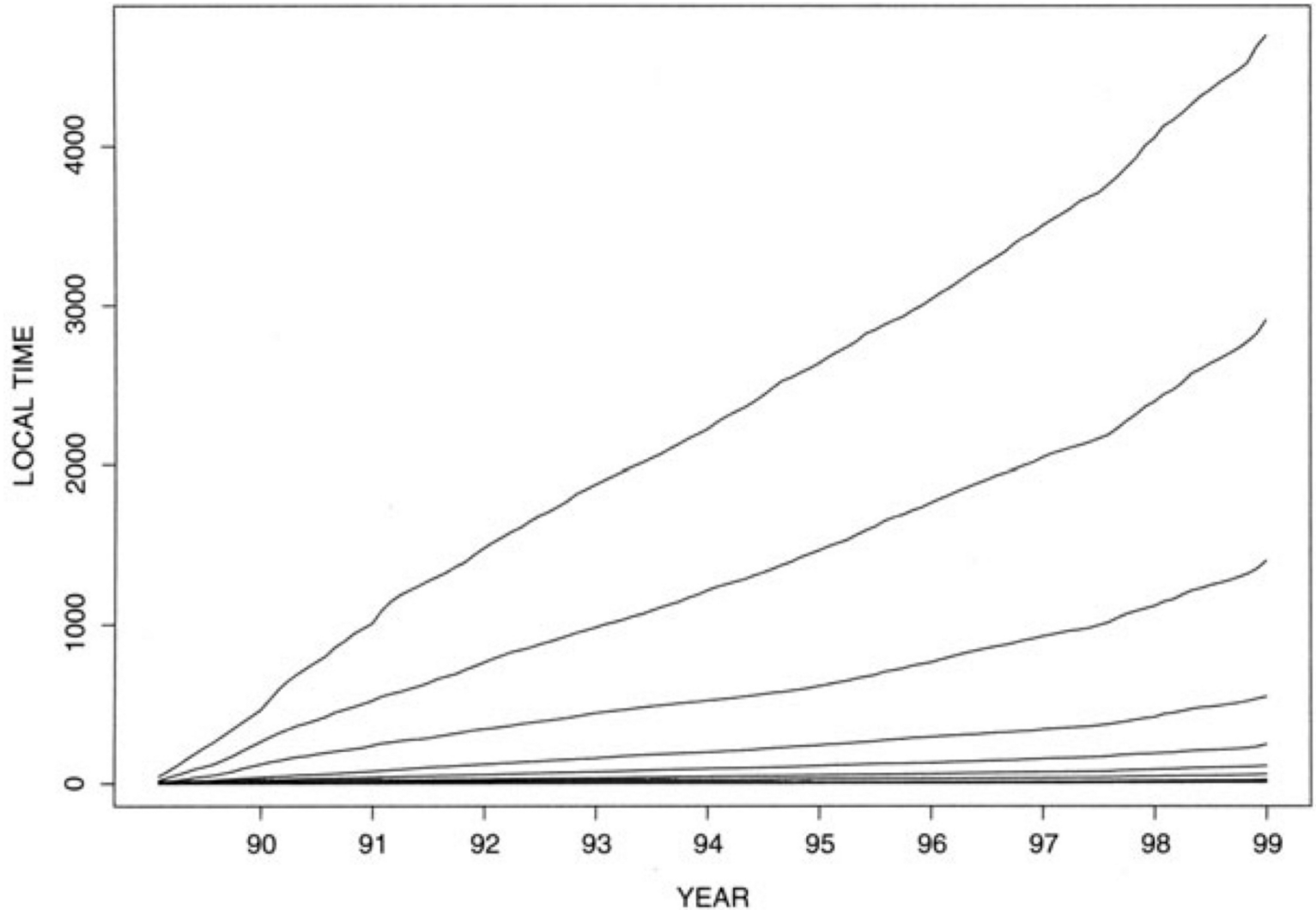
$$
\begin{aligned}
d\Xi_{(k)}(t) &= \sum_{i=1}^{\infty} \mathbb{1}_{\{r_t(i)=k\}} d\Xi_i(t) + \frac{1}{2}\Xi_{(k)}(t) d\Lambda^{\Xi}_{k,k+1}(t) \\
&\qquad - \frac{1}{2}\Xi_{(k)}(t) d\Lambda^{\Xi}_{k-1,k}(t) \\
&= \sum_{i=1}^{\infty} \mathbb{1}_{\{r_t(i)=k\}} d\Xi_i(t) + \frac{1}{2}\Xi_{(k)}(t) d\Lambda^{\Xi}_{k,k+1}(t) \\
&\qquad - \frac{1}{2}\Xi_{(k-1)}(t) d\Lambda^{\Xi}_{k-1,k}(t), \qquad \text{a.s.},
\end{aligned}
$$

so we can add up the $d\Xi_{(k)}(t)$ to obtain

$$
d\Xi_{[n]}(t) = \sum_{i=1}^{\infty} \mathbb{1}_{\{r_t(i)\le n\}} d\Xi_i(t) + \frac{1}{2}\Xi_{(n)}(t) d\Lambda^{\Xi}_{n,n+1}(t), \qquad \text{a.s.}
$$

This serves to define the *local time* $\Lambda^{\Xi}_{n,n+1}(t)$ for the data.

$\Lambda^{[\Xi]}_{k,k+1}(t)$ for U.S. capital distribution



$k = 10, 20, 40, \ldots, 5120$ (From Fernholz (2002)).

Leakage

For the data $\{\Xi_1(t), \Xi_2(t), \ldots\}$ we have the representation

$$d\Xi_{[n]}(t) = \sum_{i=1}^{\infty} \mathbb{1}_{\{r_t(i)\leq n\}} d\Xi_i(t) + \frac{1}{2}\Xi_{(n)}(t) d\Lambda^{\Xi}_{n,n+1}(t).$$

The final term compensates for the "leakage" from $\Xi_{[n]}$.

In order that the system not depend on mass replenished from outside, we impose the condition that the (relative) leakage tends to zero:

(B) $$\lim_{n\to\infty} \mathbb{E}\left[\frac{X_{(n)}(t)}{X_{[n]}(t)} d\Lambda^{X}_{n,n+1}(t)\right] = 0.$$

A conservation law

Conditions (A) and (B) together are a form of *conservation law* that ensures that the total mass of the system is autonomously maintained:

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

and

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

We shall now study the effects of conditions (A) and (B) on our continuous semimartingale model $X_1, \ldots, X_n$.

Atlas models

Perhaps the simplest model for the systems we consider is an *Atlas model*, a system of positive continuous semimartingales $X_1, \ldots, X_n$ defined by

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

where g and σ are positive constants, and $(W_1, \ldots, W_n)$ is a Brownian motion. Atlas models are asymptotically stable, and since the processes X_i are exchangeable, they asymptotically spend equal time in each rank. Hence, each of the X_i has zero asymptotic log-drift, so the entire system has zero asymptotic log-drift (Fernholz (2002), Banner et al. (2005)).

We shall assume that Atlas models are in their steady-state distributions.

The asymptotic distribution of Atlas models

The asymptotic parameters for Atlas models are

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

and these models satisfy

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

a.s., for $k = 1, \ldots, n-1$. 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 $\sigma^2/2 = g$. We wish to characterize this in terms of conditions (A) and (B).

The behavior of Atlas models

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 same role as leakage in the system Ξ. 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 Atlas models

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

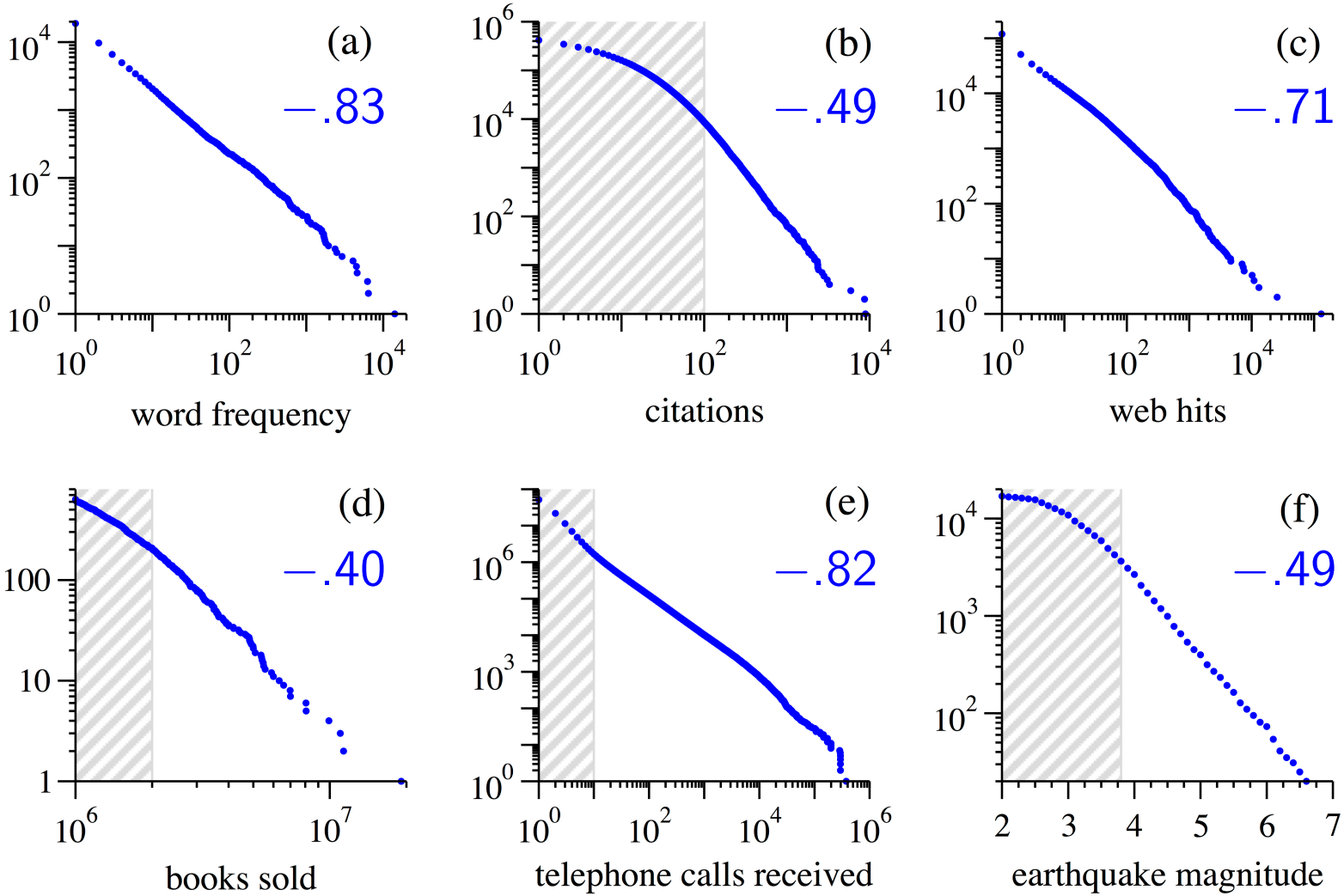
$$\mathbb{E}\left[\frac{ngX_{(n)}(t)}{X_{[n]}(t)}\right] = \begin{cases} O(1) & \text{for } \sigma^2/2 < g, \\ O(1/\log n) & \text{for } \sigma^2/2 = g, \\ O\left(n^{(1-\sigma^2/2g)}\right) & \text{for } \sigma^2/2 > g. \end{cases}$$

Hence,

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

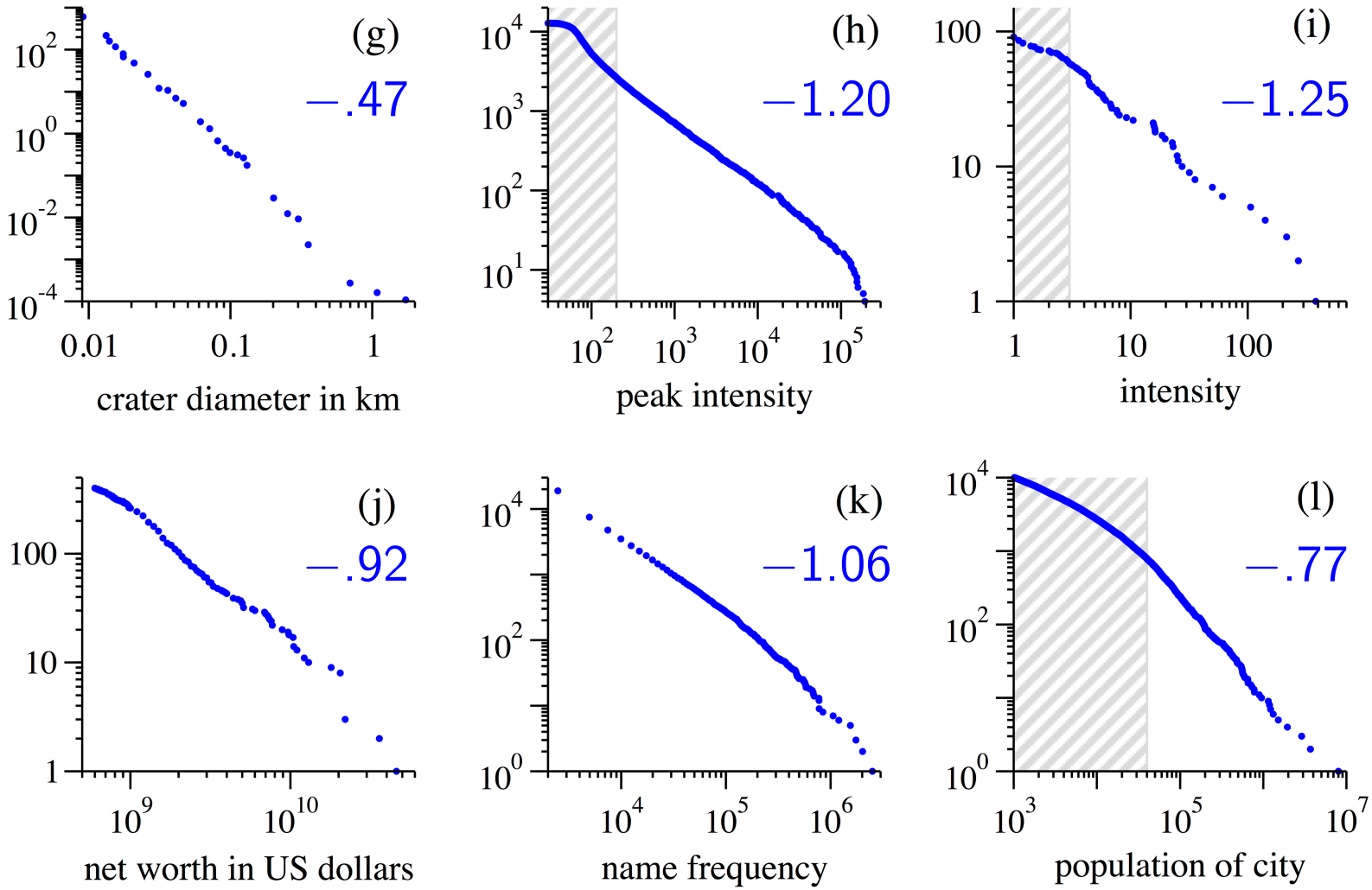
is equivalent to $\sigma^2/2 = g$, and this is equivalent to Zipf.

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 continuous semimartingales $X_1, \ldots, X_n$ with

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

where the g_k and σ_k are constants such that $\sigma_k^2 > 0$, with

$$g_1 + \cdots + g_n = 0 \text{ and } g_1 + \cdots + g_k < 0 \text{ for } k < n$$

(Fernholz (2002), Banner et al. (2005)). As usual, $(W_1, \ldots, W_n)$ is a Brownian motion. 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 $\{\Xi_1(t), \Xi_2(t), \ldots\}$ 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

$$
\begin{aligned}
g_k &= \frac{1}{2}\boldsymbol{\lambda}_{k-1,k} - \frac{1}{2}\boldsymbol{\lambda}_{k,k+1}, \qquad \text{for } k = 1, \ldots, n-1 \\
g_n &= \frac{1}{2}\boldsymbol{\lambda}_{n,n+1} \\
\sigma_1^2 &= \frac{1}{2}\boldsymbol{\sigma}^2_{1,2} \\
\sigma_k^2 &= \frac{1}{4}\left(\boldsymbol{\sigma}^2_{k-1,k} + \boldsymbol{\sigma}^2_{k,k+1}\right), \qquad \text{for } k = 2, \ldots, n.
\end{aligned}
$$

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

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

a.s., with parameters

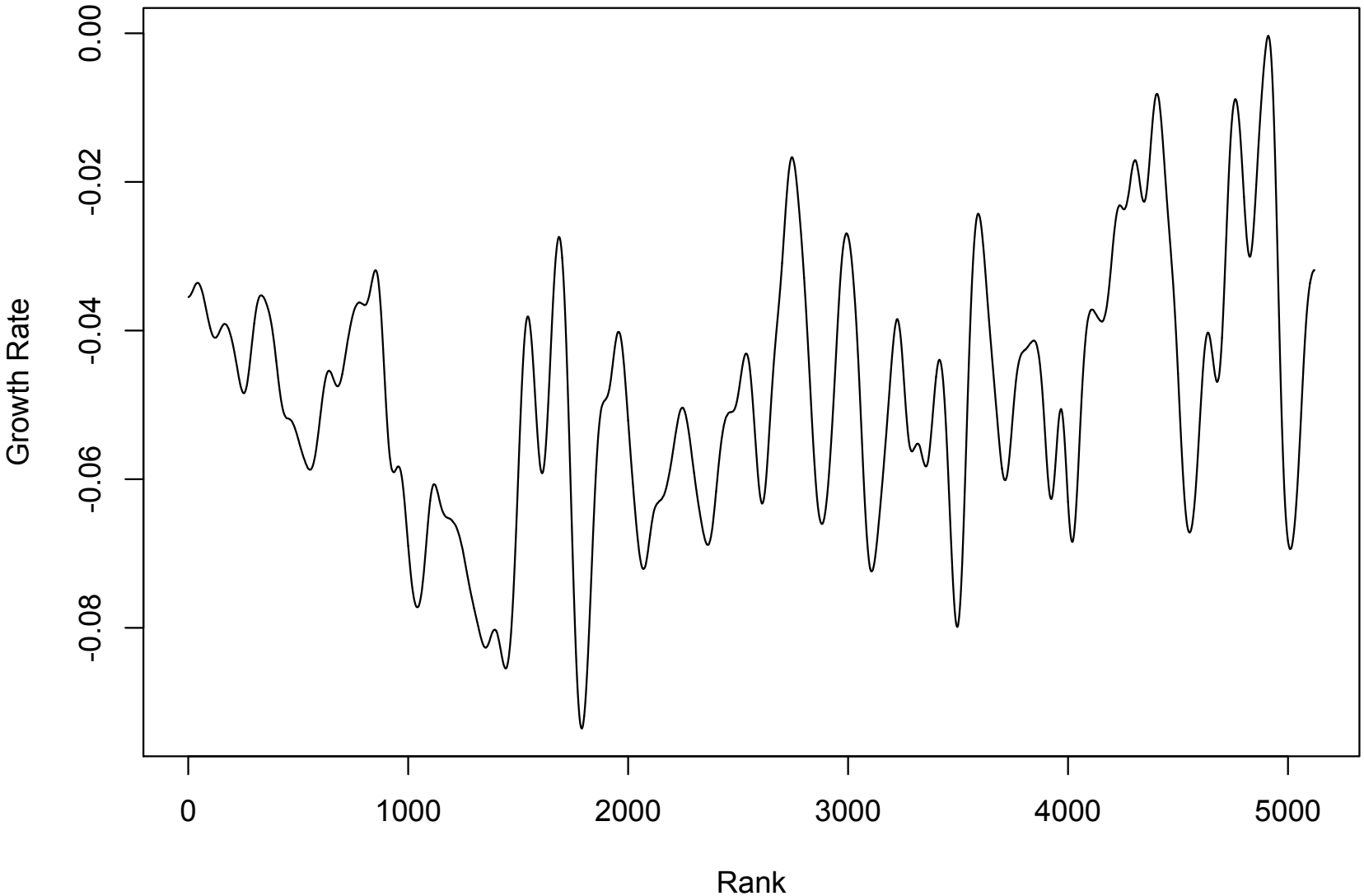
$$\lambda_{k,k+1} = \boldsymbol{\lambda}_{k,k+1}, \quad \sigma_1^2 = \frac{1}{2}\boldsymbol{\sigma}_{1,2}^2, \quad \sigma_k^2 = \frac{1}{4}\big(\boldsymbol{\sigma}_{k-1,k}^2 + \boldsymbol{\sigma}_{k,k+1}^2\big).$$

Let us suppose that the data $\{\Xi_1(t), \Xi_2(t), \ldots\}$ satisfy

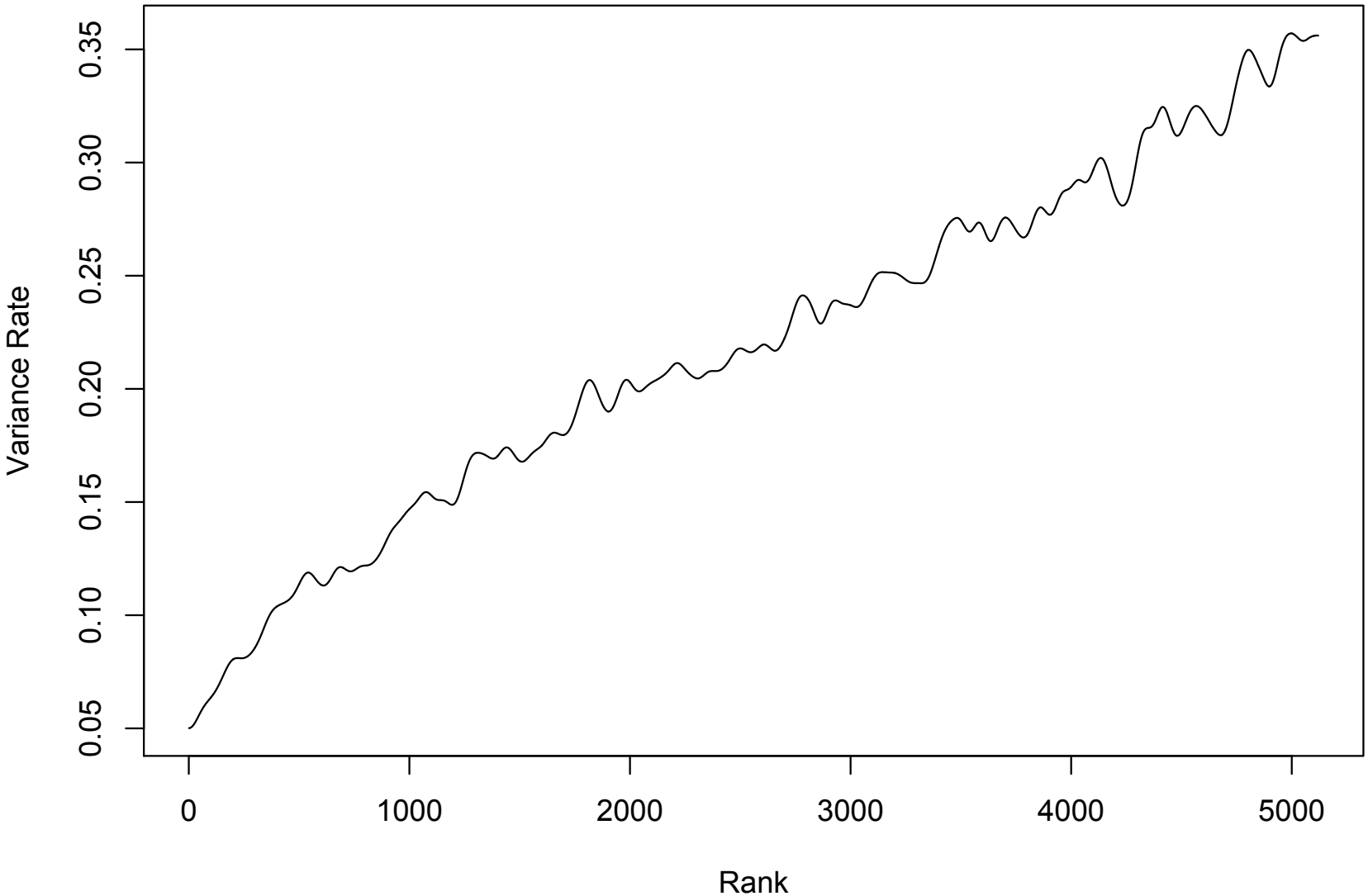
$$(*) \quad \lim_{T\to\infty} \frac{1}{T} \int_0^T \big(\log \Xi_{(k)}(t) - \log \Xi_{(k+1)}(t) \big) dt = -\frac{\boldsymbol{\sigma}_{k,k+1}^2}{2\boldsymbol{\lambda}_{k,k+1}},$$

so the X distribution is a smoothed version of the Ξ distribution.

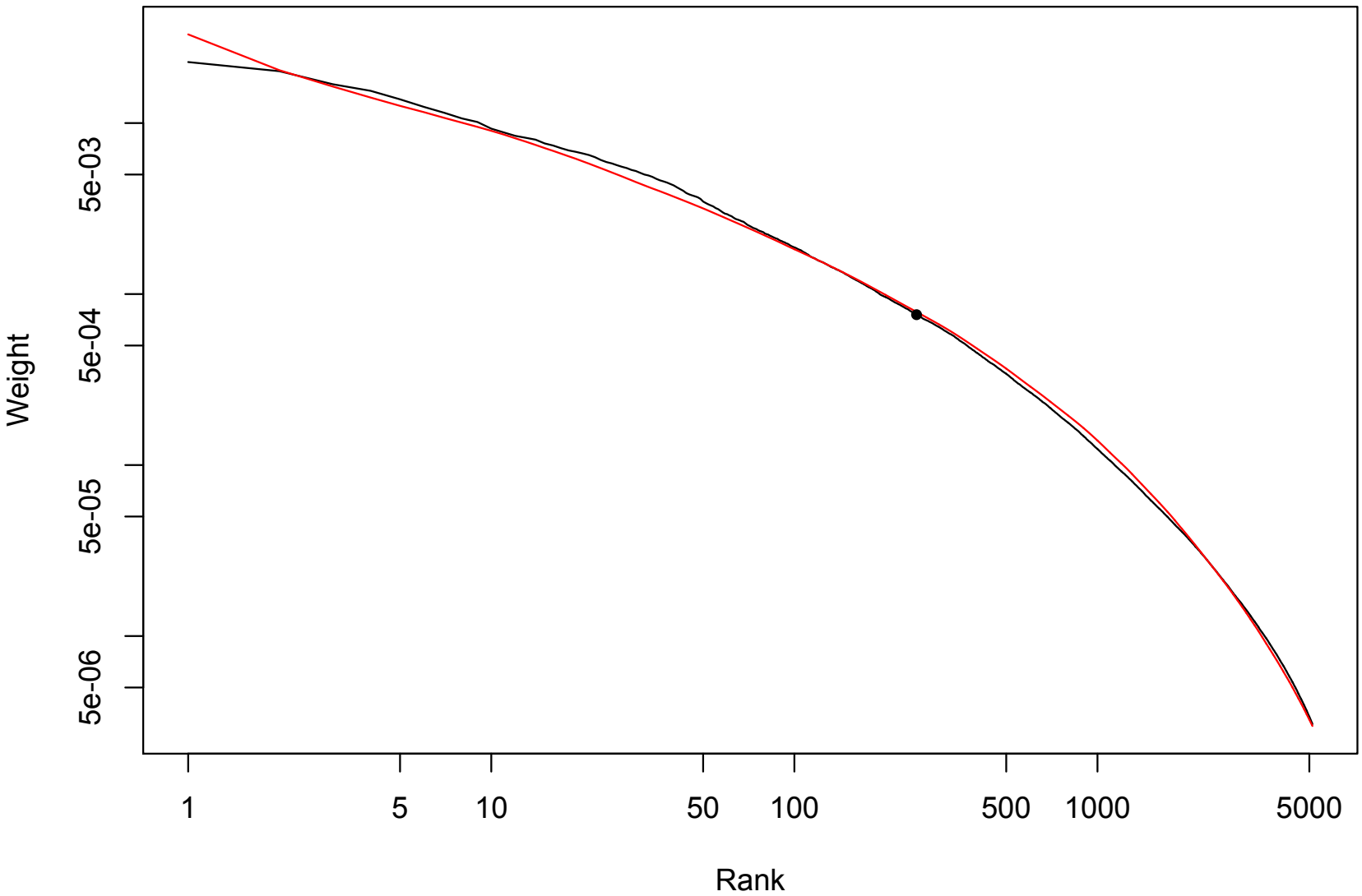
Parameters g_k for U.S. capital distribution



Parameters σ_k^2 for U.S. capital distribution



U.S. capital distribution, 1990 to 1999



Actual (black). First-order (red).

First-order approximation

Perhaps the simplest first-order model is of the form

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

where the σ_k^2 increase with rank, $\sigma_1^2 \leq \cdots \leq \sigma_n^2$. Indeed, this increasing variance would probably have occurred with the original Brownian motion, where the small pollen particles would vibrate more vigorously than the big ones. In this case, the slope of the tangent

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

will be increasingly negative, so the distribution curve will be concave.

Weakly Zipfian systems

Suppose our model 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 $\sigma_1^2 \le \cdots \le \sigma_n^2$. 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

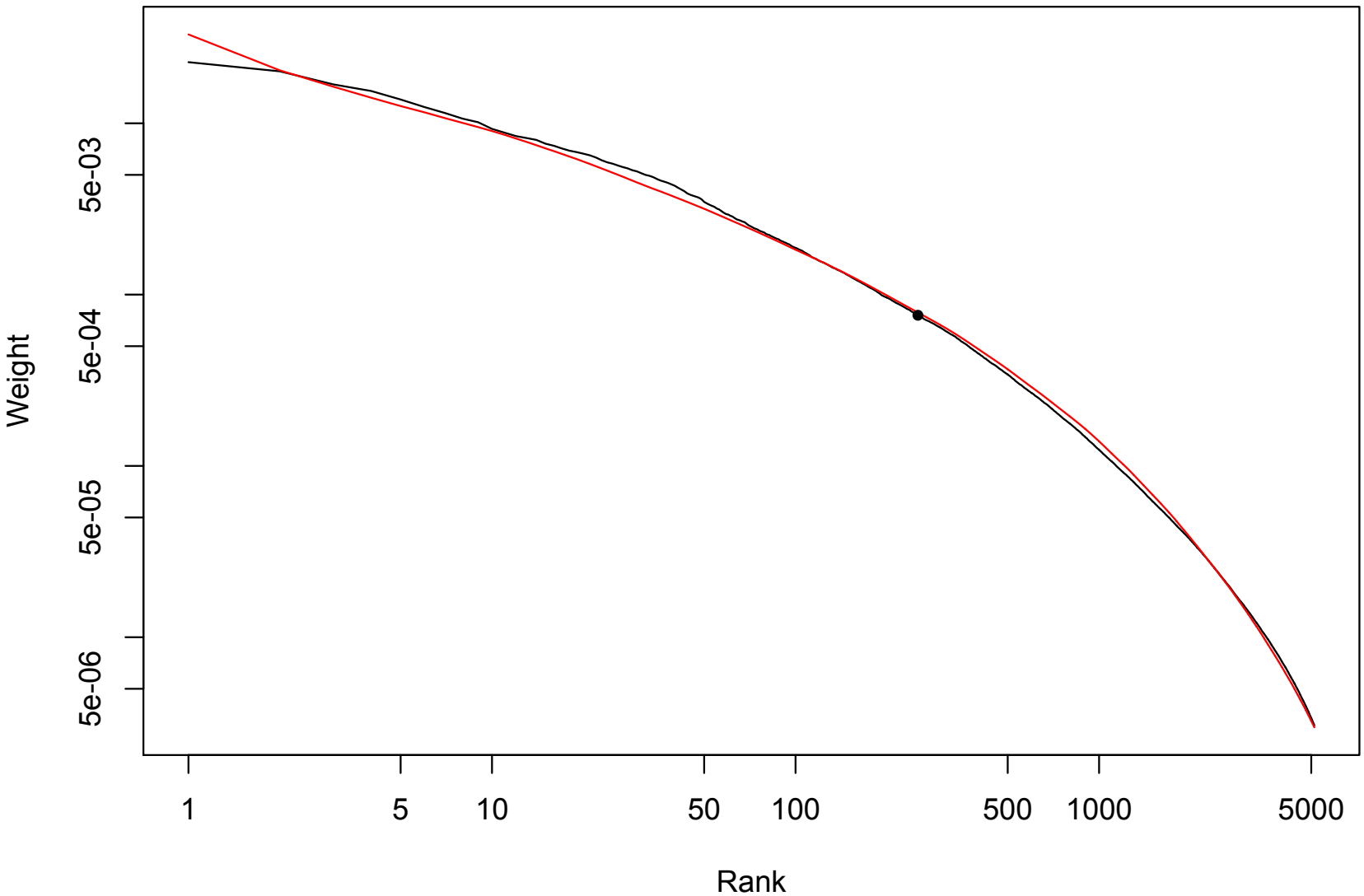
$$\text{(A)}\quad \lim_{n\to\infty}\mathbb{E}\left[\frac{dX_{[n]}(t)}{X_{[n]}(t)}\right] = 0 \quad\text{and}\quad \text{(B)}\quad \lim_{n\to\infty}\mathbb{E}\left[\frac{ngX_{(n)}(t)}{X_{[n]}(t)}\right] = 0,$$

then,

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

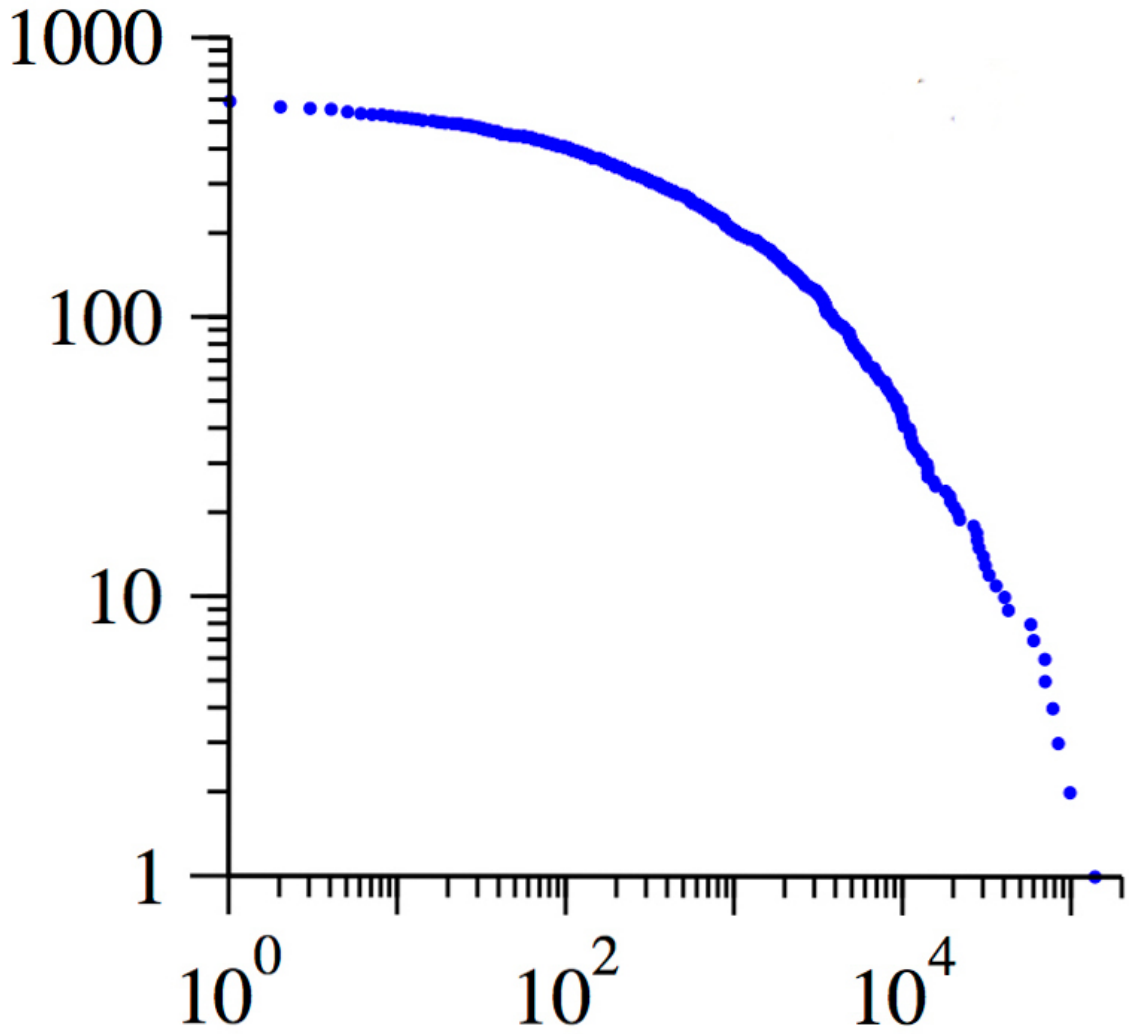
This system will be *weakly Zipfian*, having a distribution that is concave with tangent slope -1 somewhere in the middle ranks.

U.S. capital distribution, 1990 to 1999



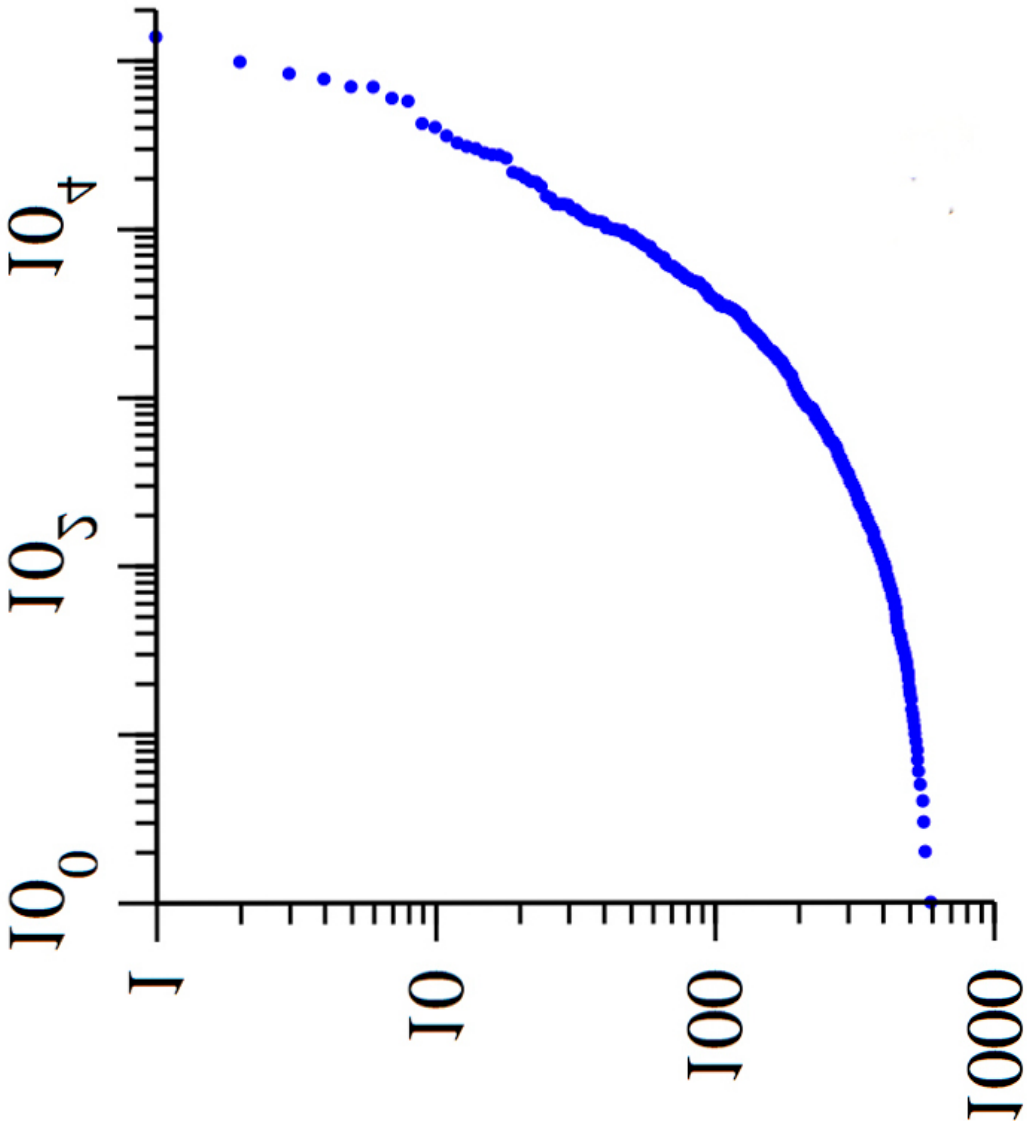
Actual (black). First-order (red).

Birds



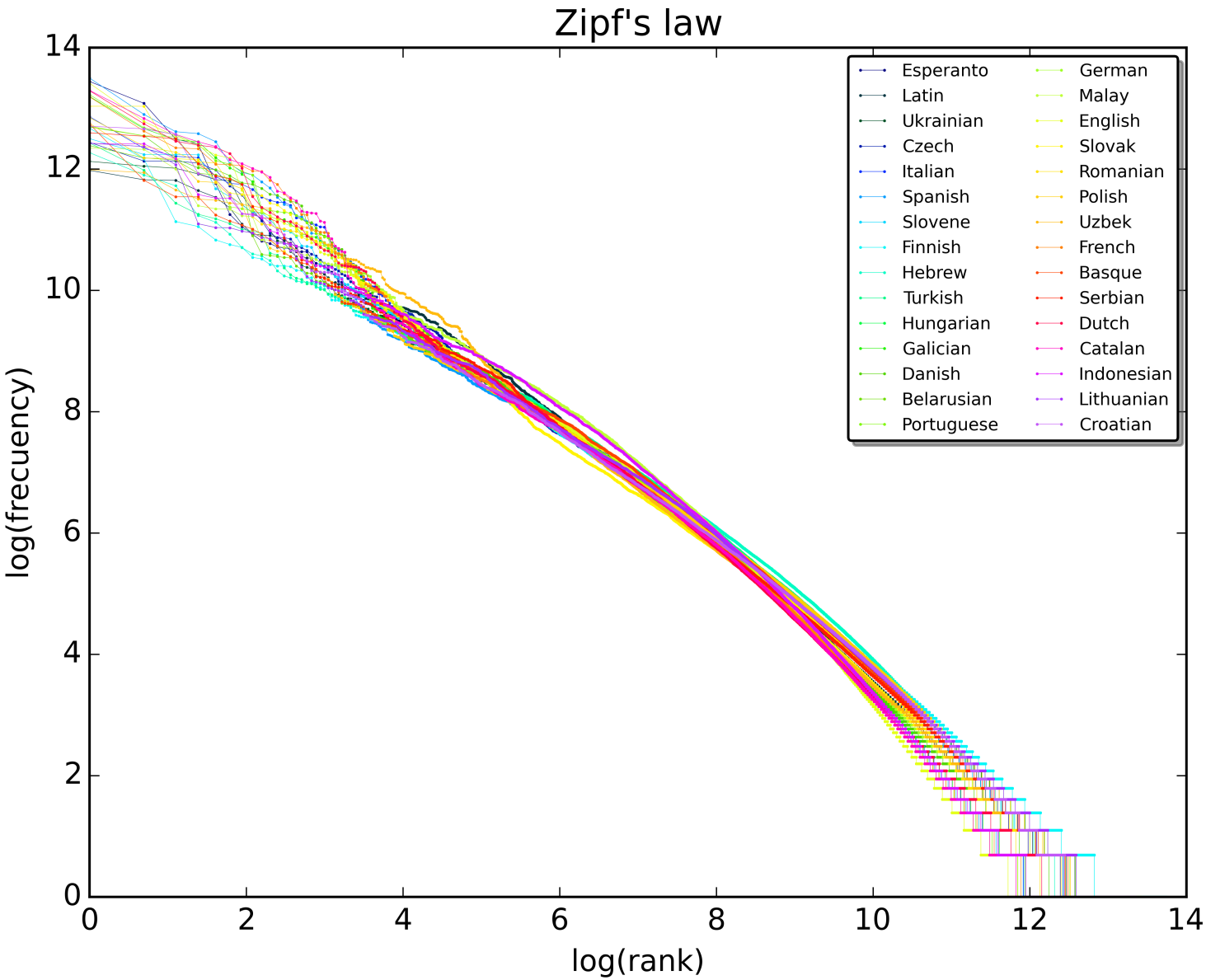
North American Bird Survey 2003 (From Newman (2006)).

Birds



North American Bird Survey 2003 (From Newman (2006)).

Word count from Wikipedia



References

- Banner, A., R. Fernholz, and I. Karatzas (2005). Atlas models of equity markets. *Annals of Applied Probability 15(4),* 2296–2330.
- Fernholz, R. (2002). *Stochastic Portfolio Theory.* Springer.
- Newman, M. E. J. (2006). Power laws, Pareto distributions and Zipf's law. *ArXiv 04120004v3,* 1–28.
- Zipf, G. K. (1935). *The Psychobiology of Language.* Houghton-Mifflin.
- Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley.

Thank you!