Mathematics in Medicine Journal Club:

"Power law distributions in empirical data"

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman

Cory Brunson
Center for Quantitative Medicine

March 7, 2017

Background: Scale-free biological networks

Background: Scale-free biological networks

- Barabási, A-L and Oltvai, Z N (2004). "Network biology: understanding the cell's functional organization". *Nature Reviews Genetics* **5**(2), 101–113.
- Albert, R (2005). "Scale-free networks in cell biology". *Journal of Cell Science* **118**(21), 4947–4957.
- Zhang, B and Horvath, S (2005). "A General Framework for Weighted Gene Co-Expression Network Analysis". *Statistical Applications in Genetics and Molecular Biology* **4**(1), Article 17.

Terminology

- **continuous power-law** probability distributions

$$P(X > x) \propto x^{-\alpha}, \quad x \geq x_{\min} \quad \exists\, x_{\min}, \alpha > 0$$

Terminology

- **continuous power-law** probability distributions

$$P(X > x) \propto x^{-\alpha}, \quad x \geq x_{\min} \quad \exists\, x_{\min}, \alpha > 0$$

- . . . are **scaling** in the sense that*

$$P(X > sx \mid X > x) = s^{-\alpha} \quad \forall\, x \geq x_{\min}$$

Terminology

- **discrete power-law** probability distributions

$$p_x \propto x^{-\alpha-1}, \quad x > x_{\min} \quad \exists\, x_{\min}, \alpha > 0$$

Terminology

- **discrete power-law** probability distributions

$$p_x \propto x^{-\alpha-1}, \quad x > x_{\min} \quad \exists\, x_{\min}, \alpha > 0$$

- …produce **scaling sequences***

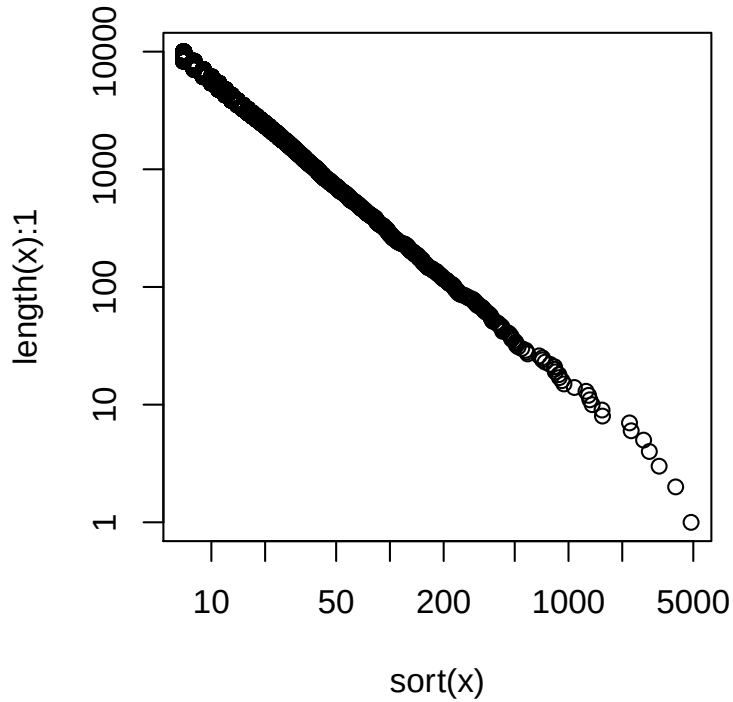
$$x_1 \geq x_2 \geq \cdots \geq x_n, \quad k \propto {x_k}^{-\alpha}$$

Diagnostics

Size–rank plots:

$$k = C x_k^{-\alpha} \quad \Rightarrow \quad \log k = \log(C) - \alpha \log x_k$$

```
library(poweRlaw)
m <- displ$new()
m$setXmin(7)
m$setPars(2.3)
x <- dist_rand(m, 1e4)
plot(x = sort(x),
     y = length(x):1,
     log = "xy")
```

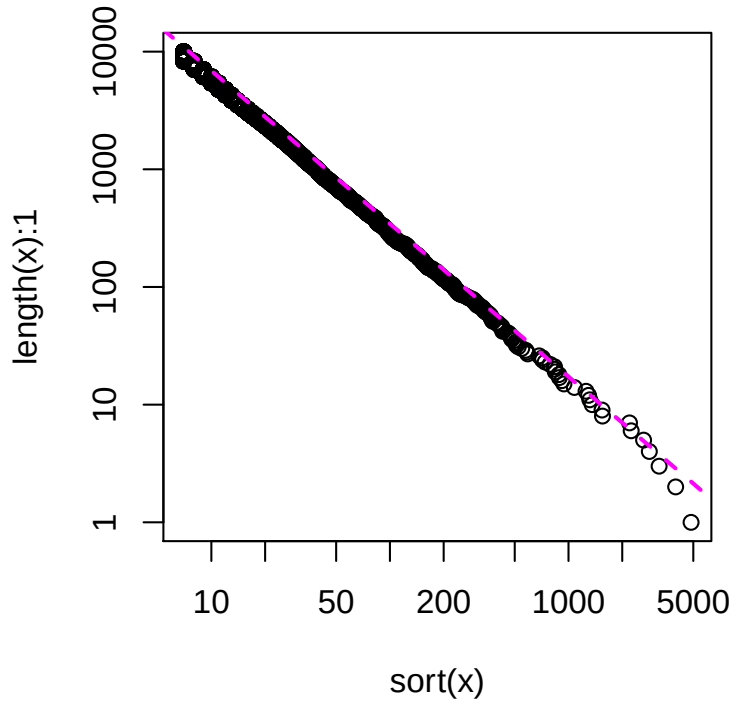


Diagnostics

Size–rank plots:

$$k = C x_k^{-\alpha} \quad \Rightarrow \quad \log k = \log(C) - \alpha \log x_k$$

```
library(poweRlaw)
m <- displ$new()
m$setXmin(7)
m$setPars(2.3)
x <- dist_rand(m, 1e4)
plot(x = sort(x),
     y = length(x):1,
     log = "xy")
```

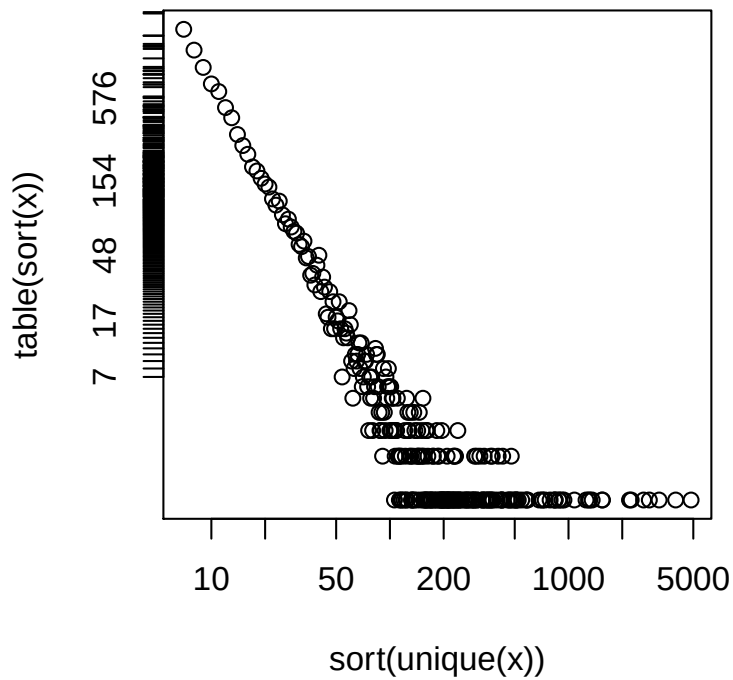


Diagnostics

Size–frequency plots:

$$p_x = Cx^{-\alpha-1} \quad \Rightarrow \quad \log p_x = \log(C) - (\alpha+1)\log(x)$$

```
library(poweRlaw)
m <- displ$new()
m$setXmin(7)
m$setPars(2.3)
x <- dist_rand(m, 1e4)
plot(x = sort(unique(x)),
     y = table(sort(x)),
     log = "xy")
```

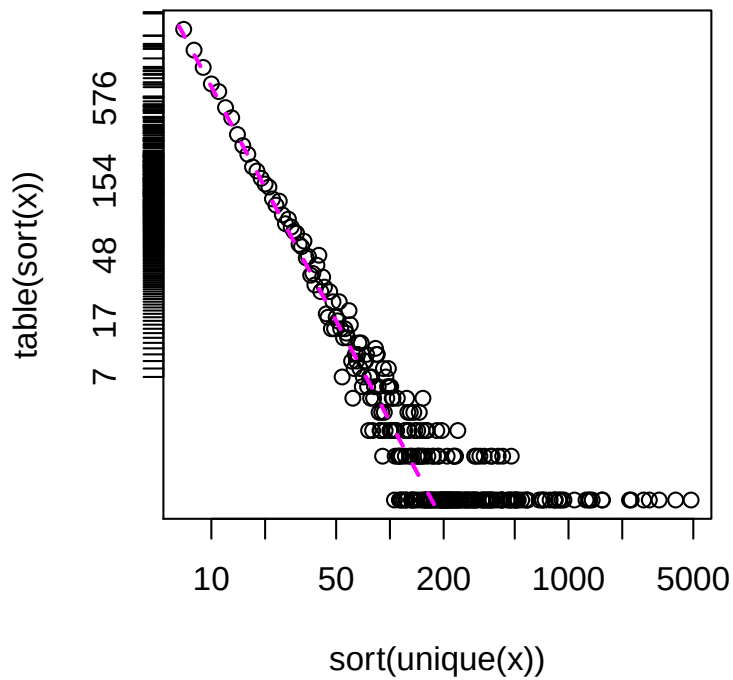


Diagnostics

Size–frequency plots:

$$p_x = Cx^{-\alpha-1} \quad \Rightarrow \quad \log p_x = \log(C) - (\alpha+1)\log(x)$$

```
library(poweRlaw)
m <- displ$new()
m$setXmin(7)
m$setPars(2.3)
x <- dist_rand(m, 1e4)
plot(x = sort(unique(x)),
     y = table(sort(x)),
     log = "xy")
```



Implications

Networks with scaling degree sequences are often called "scale-free", assumed to have specific origins:

- evolution via cumulative advantage (preferential attachment)
- emergence at phase transitions / criticality

Implications

Networks with scaling degree sequences are often called "scale-free", assumed to have specific origins:

- evolution via cumulative advantage (preferential attachment)
- emergence at phase transitions / criticality

and asserted to have many topological properties:

- highly central high-degree nodes (hubs)
- self-similarity / hierarchicality
- invariance under degree-preserving rewiring
- robustness to failure but vulnerability to attack

Implications

Networks with scaling degree sequences are often called "scale-free", assumed to have specific origins:

- evolution via cumulative advantage (preferential attachment)
- emergence at phase transitions / criticality

and asserted to have many topological properties:

- highly central high-degree nodes (hubs)
- self-similarity / hierarchicality
- invariance under degree-preserving rewiring
- robustness to failure but vulnerability to attack

The apparent ubiquity of scaling degree distributions in empirical networks then has profound implications:

- significance to natural order (*surprising*)
- universality to complex systems (*determinitive*)

Example

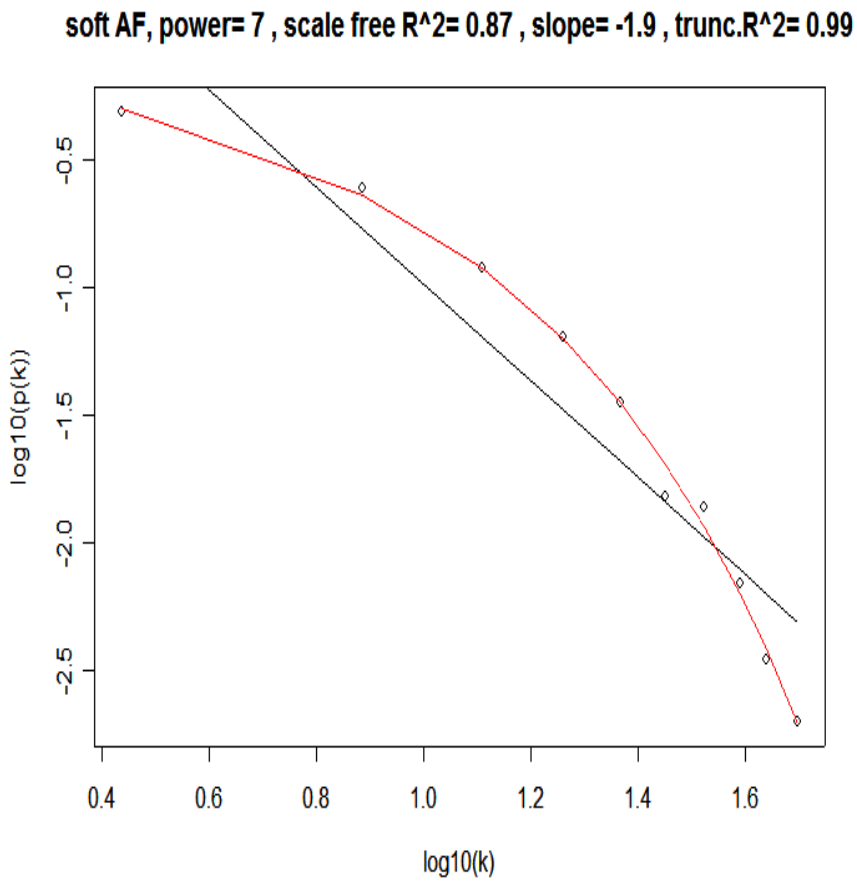
To visually inspect whether approximate scale-free topology is satisfied, one plots $\log_{10}(p(k))$ *versus* $\log_{10}(k)$*. A straight line is indicative of scale-free topology[.]*

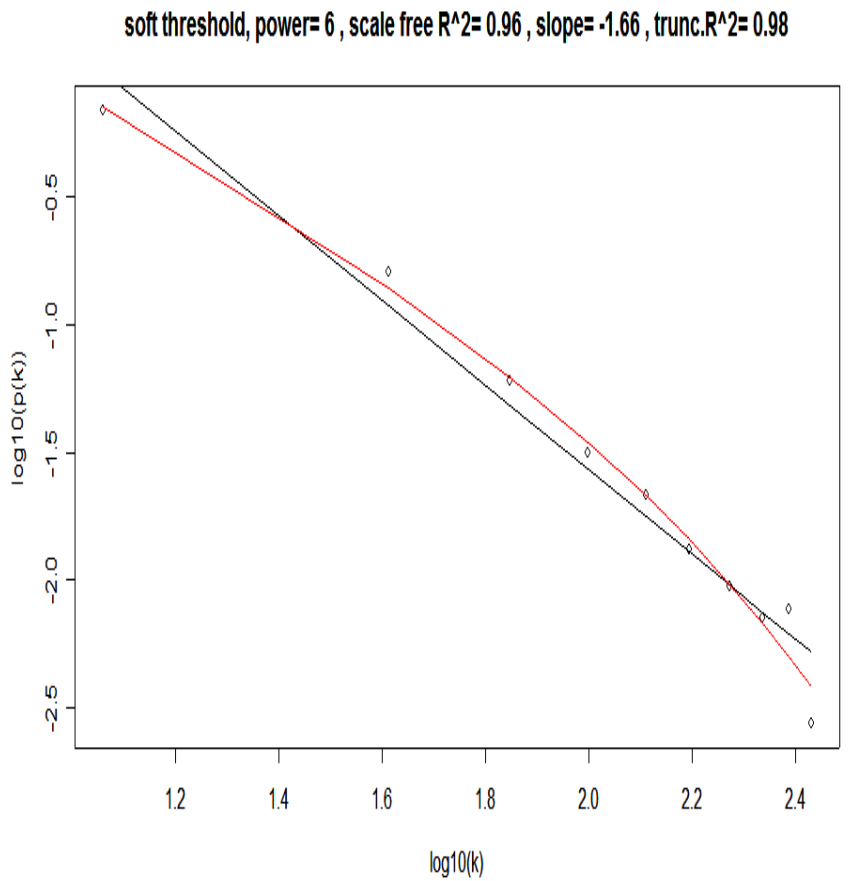
To measure how well a network satisfies a scale-free topology, we propose to use the square of the correlation between $\log(p(k))$ *and* $\log(k)$*, i.e. the model fitting index* R^2 *of the linear model that regresses* $\log(p(k))$ *on* $\log(k)$*.*

*

Zhang and Horvath, 2005

Example





Zhang and Horvath, 2005

Context

NETWORK BIOLOGY: UNDERSTANDING THE CELL'S FUNCTIONAL ORGANIZATION

Albert-László Barabási & Zoltán N. Oltvai‡*

Commentary

Scale-free networks in cell biology

Réka Albert

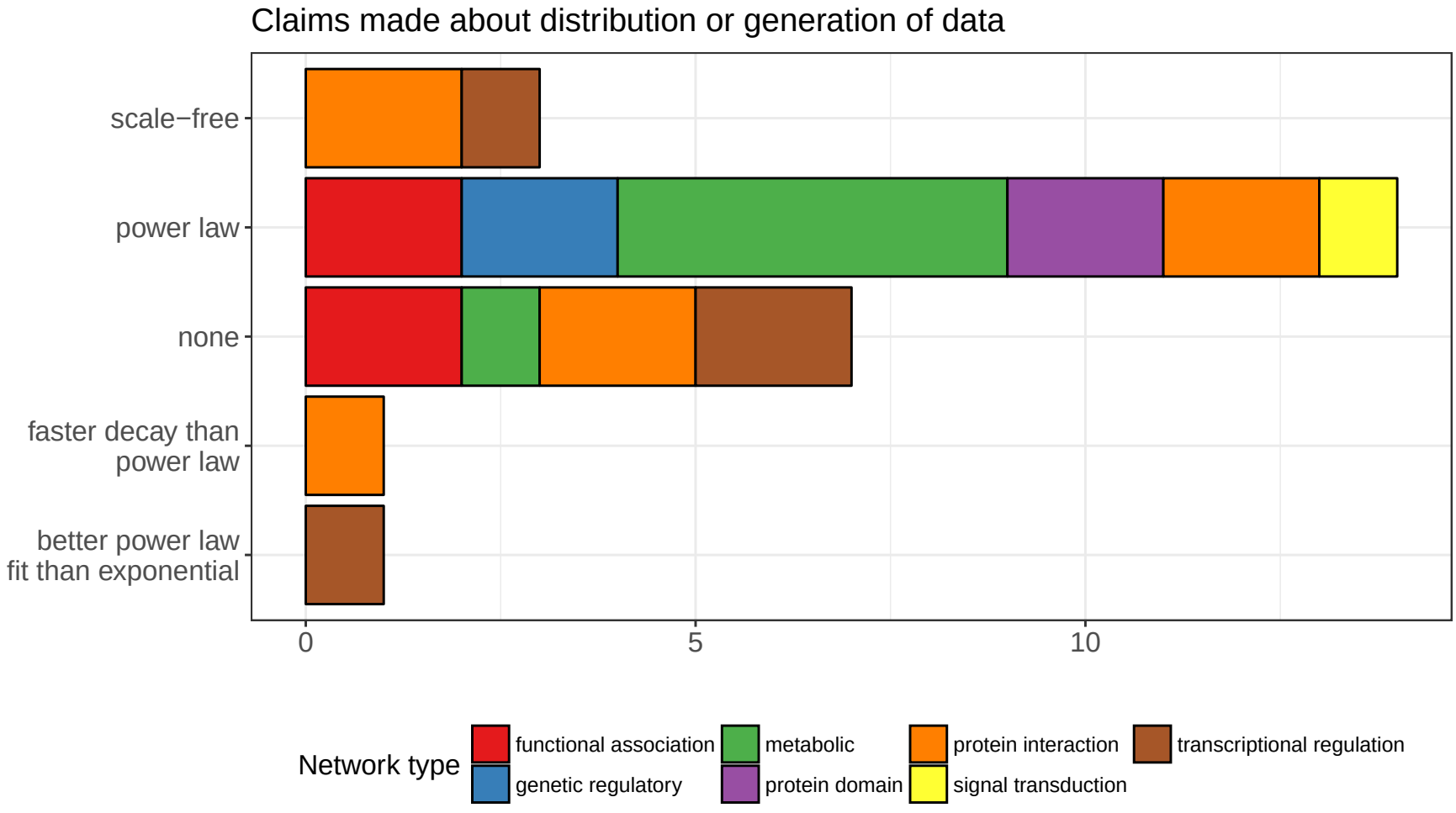
Department of Physics and Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA
(e-mail: ralbert@phys.psu.edu)

Journal of Cell Science 118, 4947-4957 Published by The Company of Biologists 2005
doi:10.1242/jcs.02714

Context

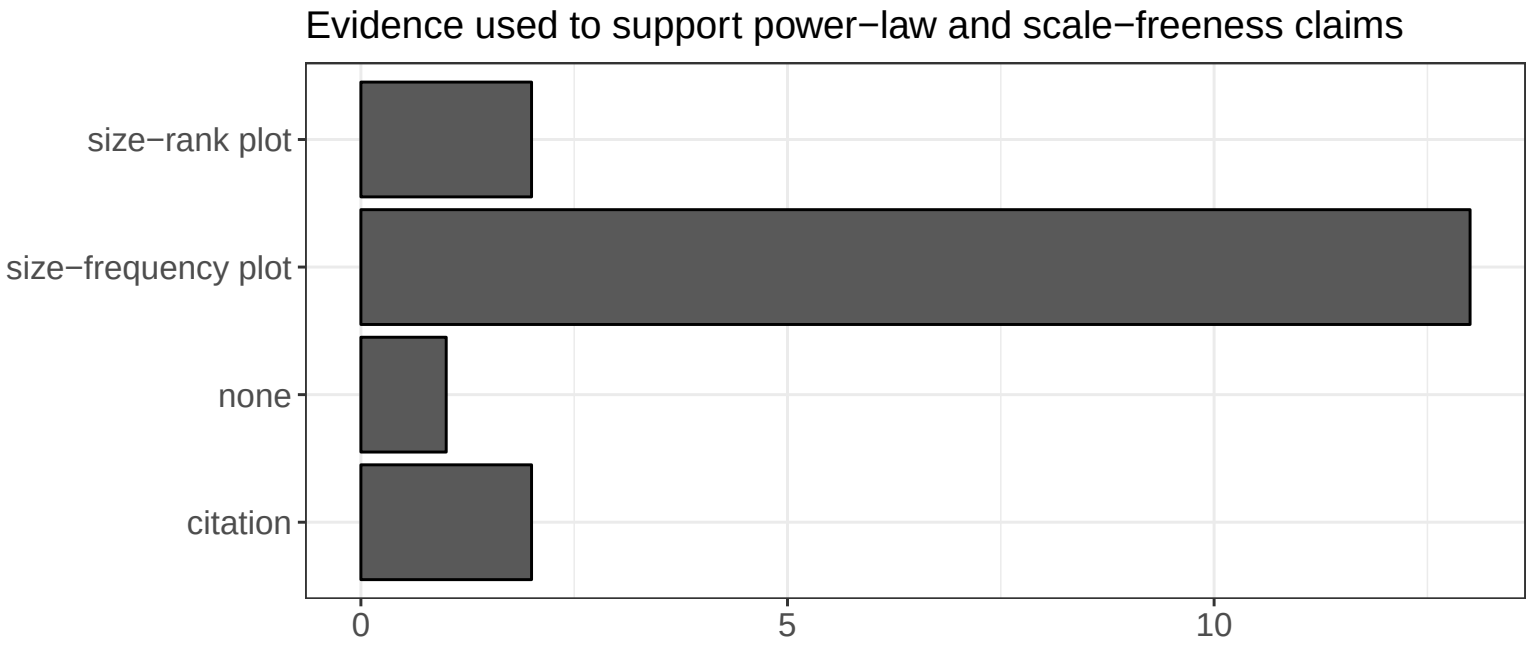
1. How widely do biologists observe scale-free networks?
2. What evidence supports these observations?
3. What implications follow from these observations?

Claims and evidence



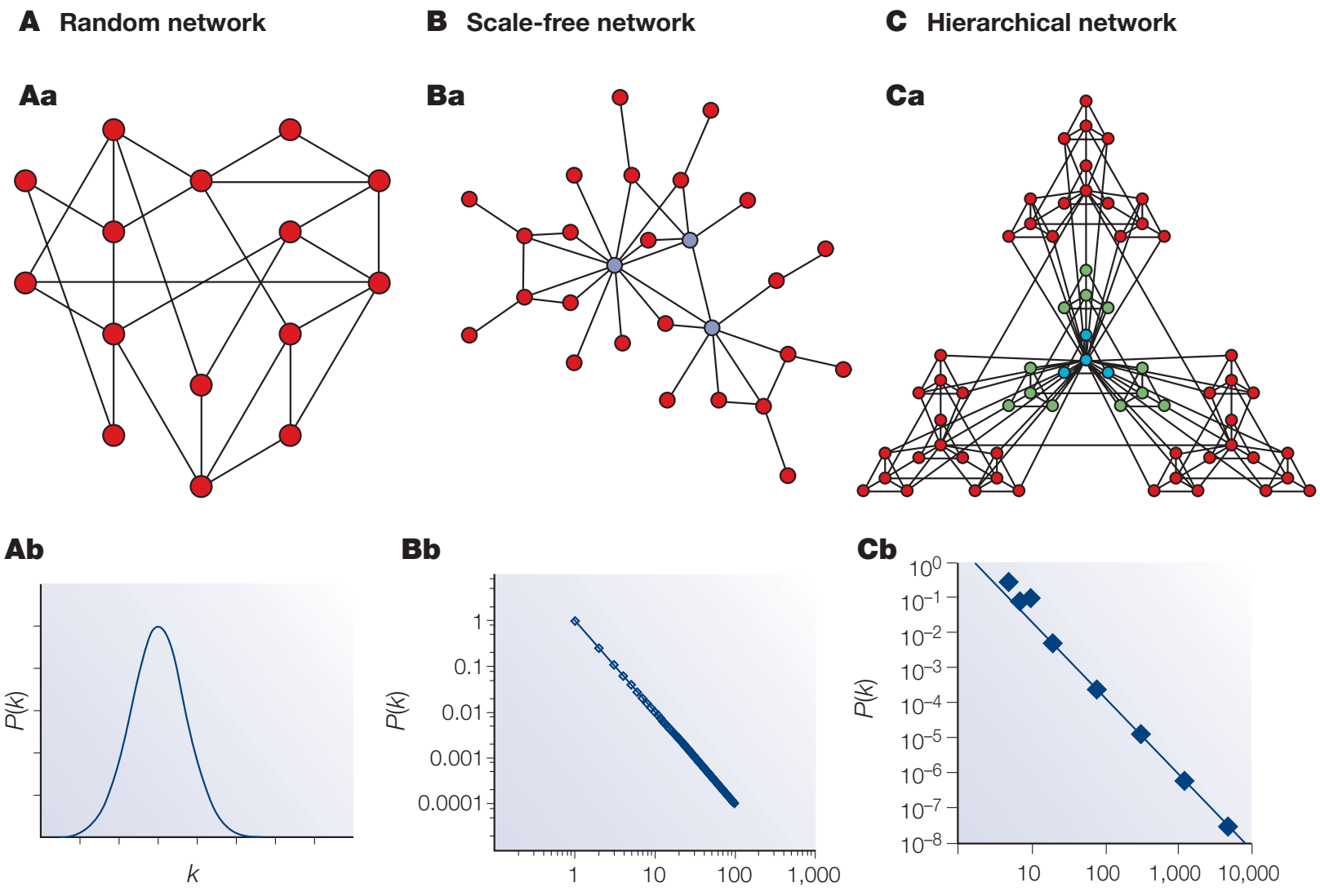
literature cited in Barabási & Oltvai, 2004 and in Albert, 2005

Claims and evidence



literature cited in Barabási & Oltvai, 2004 and in Albert, 2005

Implications



Barabási & Oltvai, 2004

Implications

- "Presence of scale-free behavior indicates a high degree of self-organization in the system and is known to be a characteristic of natural systems. . . . The scale-free character of coexpressed gene networks means that these networks are extremely inhomogeneous and contain few genes that are very highly connected and a large number of genes with low connectivity." (Agrawal, 2002)
- "Since hubs are rare relative to other nodes, random elimination of nodes has minimal effect on network topology because statistically, a randomly eliminated node is likely to have low connectivity. . . . We propose that organisms' general ability to compensate for individual mutations is largely a result of the scale-free properties of the gene expression network." (Featherstone & Broadie, 2002)
- "This result suggests the existence of a selective force in the overall design of genetic pathways to maintain a highly connected class of genes." (Stuart et al, 2003)

literature cited in Barabási & Oltvai, 2004 and in Albert, 2005

Confusion

- "Surprisingly, genes such as calcium/calmodulin-dependant kinase (CAM- KII) and the signaling GTPase Ras1, which are central to intracellular signaling and therefore would be most expected to 'tie' together various aspects of cell biology, are not hubs. . . . It is also somewhat surprising that transcription factors do not head the list of hubs, since deletion of a transcription factor would be expected to alter the expression of all its downstream genes. (Featherstone & Broadie, 2002)
- "The distribution of interactions per protein decays faster than the power law predicted by a "rich-get-richer" model of scale-free networks[.] This rapid decay suggests that highly connected proteins may be suppressed in biological networks and supports a previous observation that connections between highly connected proteins are also suppressed." (Giot et al, 2003)
- "[W]ith our current metabolic information, the [average path length] of the *E. coli* network remains ≈ 8, much larger than that of a random graph. The metabolic world of *E. coli* is therefore not small with respect to biosynthesis/degradation pathways on the traditional metabolic map." (Masanori, 2004)

literature cited in Barabási & Oltvai, 2004 and in Albert, 2005

Theory: Power laws and scale-free graphs

Theory: Power laws and scale-free graphs

- Mitzenmacher, M (2004). "A Brief History of Generative Models for Power Law and Lognormal Distributions". *Internet Mathematics* **1**(2), 226–251.
- Willinger, W, Alderson, D, Doyle, J C, and Li, L (2004). "More "normal" than normal: scaling distributions and complex systems". *Proceedings of the 2004 Winter Simulation Conference*, 130–141.
- Fox Keller, E (2005). "Revisiting "scale-free" networks". *BioEssays* **27**(10), 1060–1068.
- Li, L, Alderson, D, Doyle, J C, Willinger, W (2005). "Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications". *Internet Mathematics* **2**(4), 431–523.
- Clauset, A, Shalizi, C R, Newman, M E J (2009). "Power-Law Distributions in Empirical Data" *SIAM Review* **51**(4), 661–703.

Implications

Networks with scaling degree sequences are often called "scale-free", **assumed** to have specific origins:

- evolution via cumulative advantage (preferential attachment)
- emergence at phase transitions / criticality

and **asserted** to have many topological properties:

- highly central high-degree nodes (hubs)
- self-similarity / hierarchicality
- invariance under degree-preserving rewiring
- robustness to failure but vulnerability to attack

The **apparent** ubiquity of scaling degree distributions in empirical networks then has profound implications:

- significance to natural order (*surprising*)
- universality to complex systems (*determinitive*)

Li et al, 2005; Keller, 2005

GENERATIVE MODELS

Power-law behavior arises from many processes.

- Cumulative advantage
- Optimization $\longrightarrow$
- Multiplicative processes
- Durations between events
- Double Pareto distributions

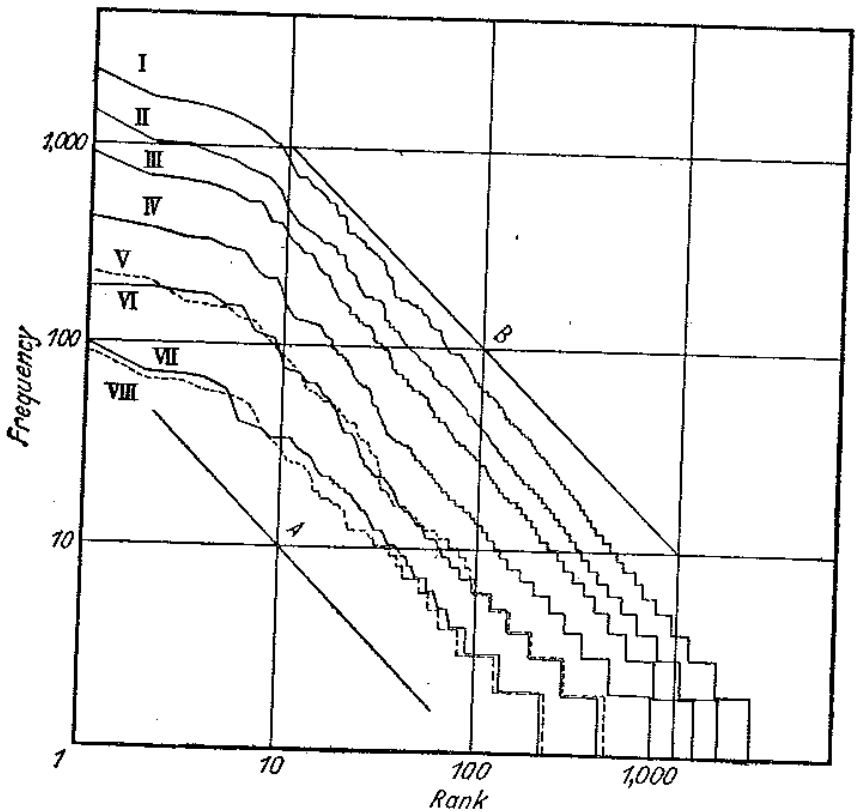


Figure 1. Rank-frequency distributions of Helen B. (with paranoid schizophrenia) in samples of (I) 50,000 words; (II) 30,000 words; (III) 20,000 words; (IV) 10,000 words; (V and VI) 5,000 words, and (VII and VIII) 2,000 words. (From *Arch. Neurol. Psychiat.* 49 (1943) 831)

Mitzenmacher, 2004

Invariance properties

Power-law behavior survives many transformations.

If $X_1, \ldots, X_n$

- follow scaling distributions $X_k \sim P_k$
- with common scaling parameter $1 < \alpha < 2$

then so do

$$S_n = \sum_k X_k \qquad \text{(aggregation; Central Limit Theorem)}$$

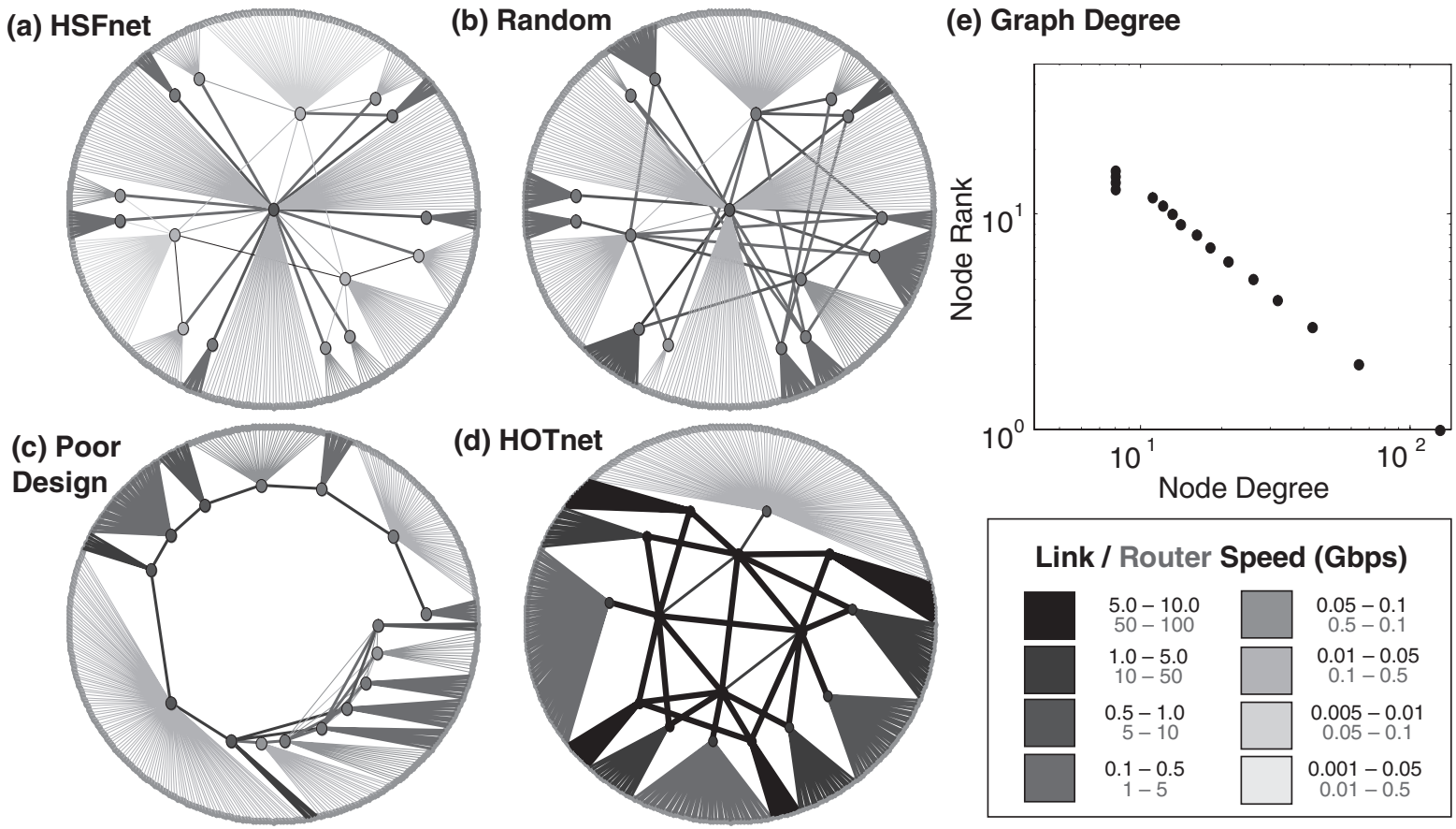
$$M_n = \max(X_n) \qquad \text{(maximizing choices)}$$

$$W_n \sim \sum_k w_k P_k \qquad \text{(weighted mixtures)}$$

Willinger et al, 2004

TOPOLOGICAL DIVERSITY

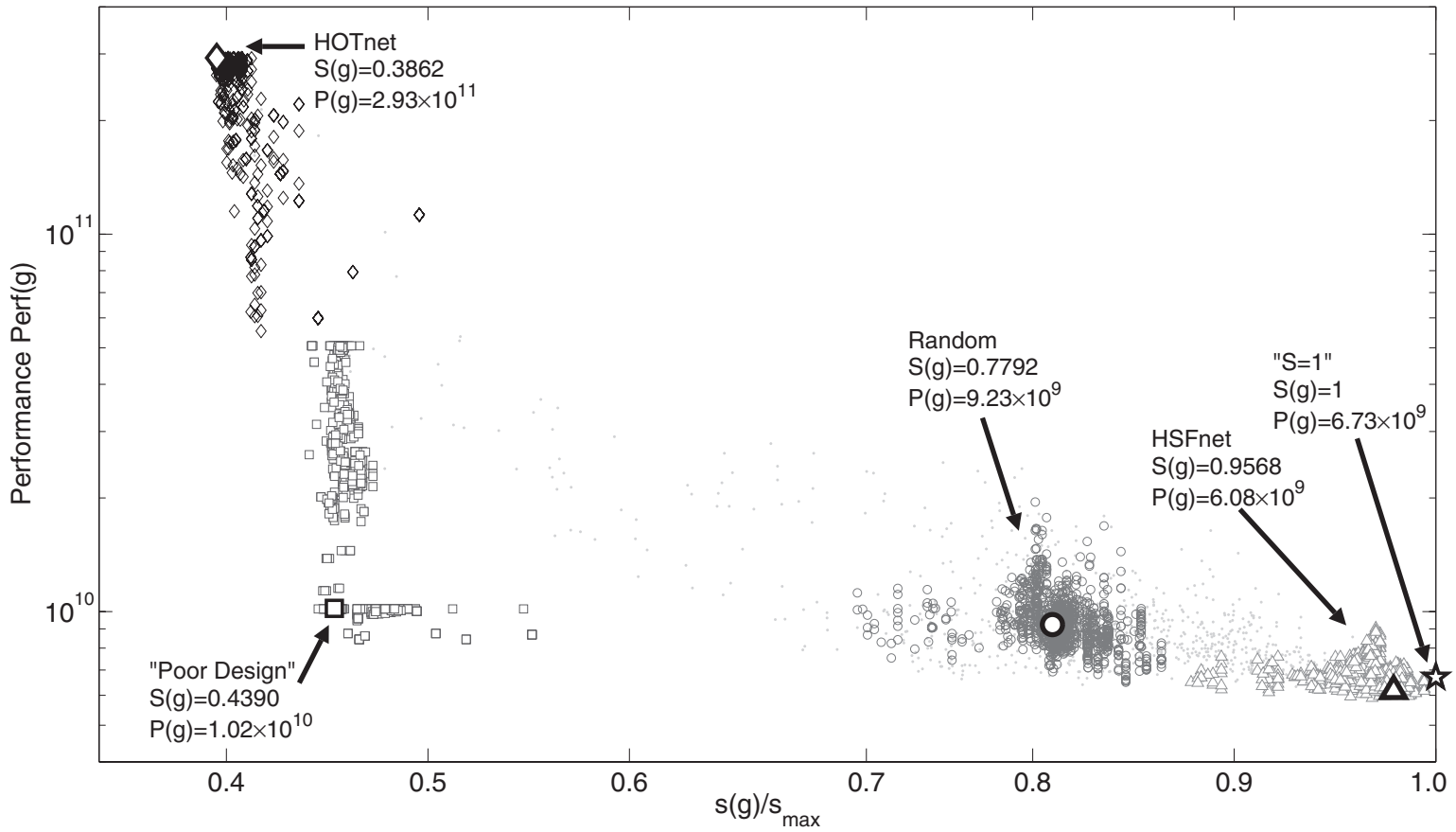
Power-law behavior admits many network topologies.



Li et al, 2005

TOPOLOGICAL DIVERSITY

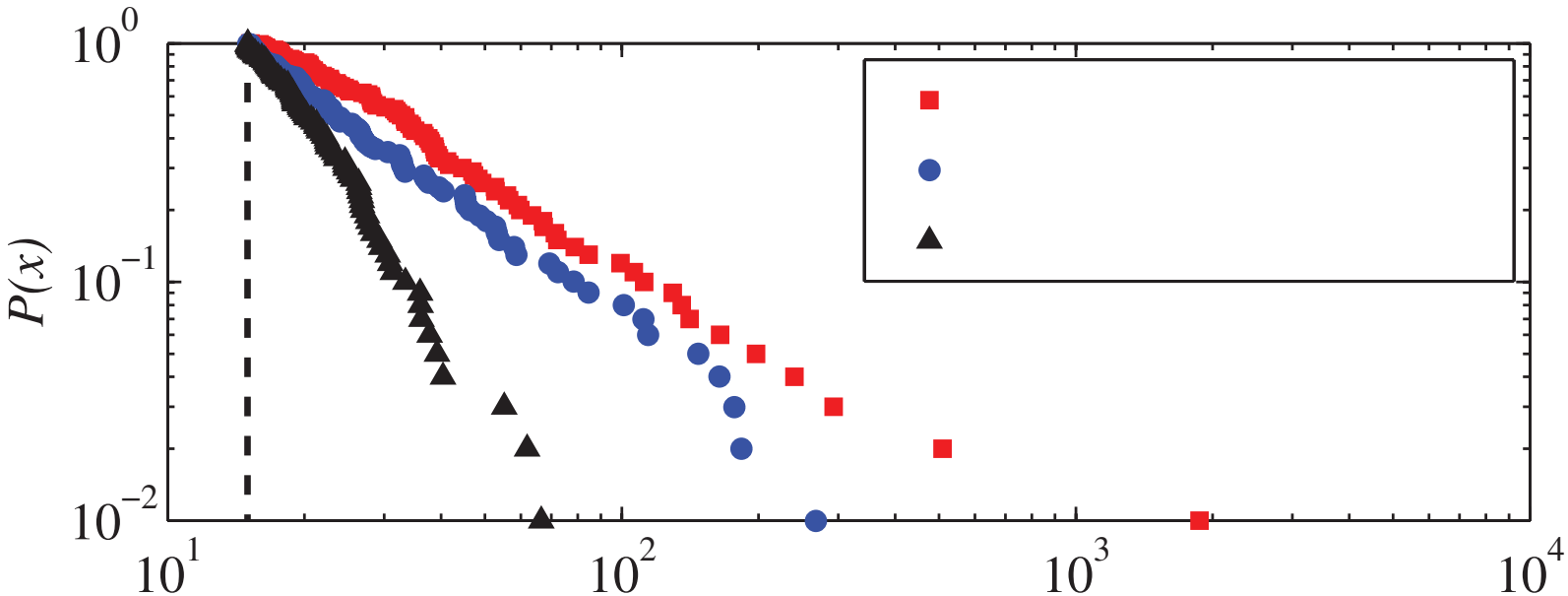
Power-law behavior admits many network topologies.



Li et al, 2005

Statistical inference

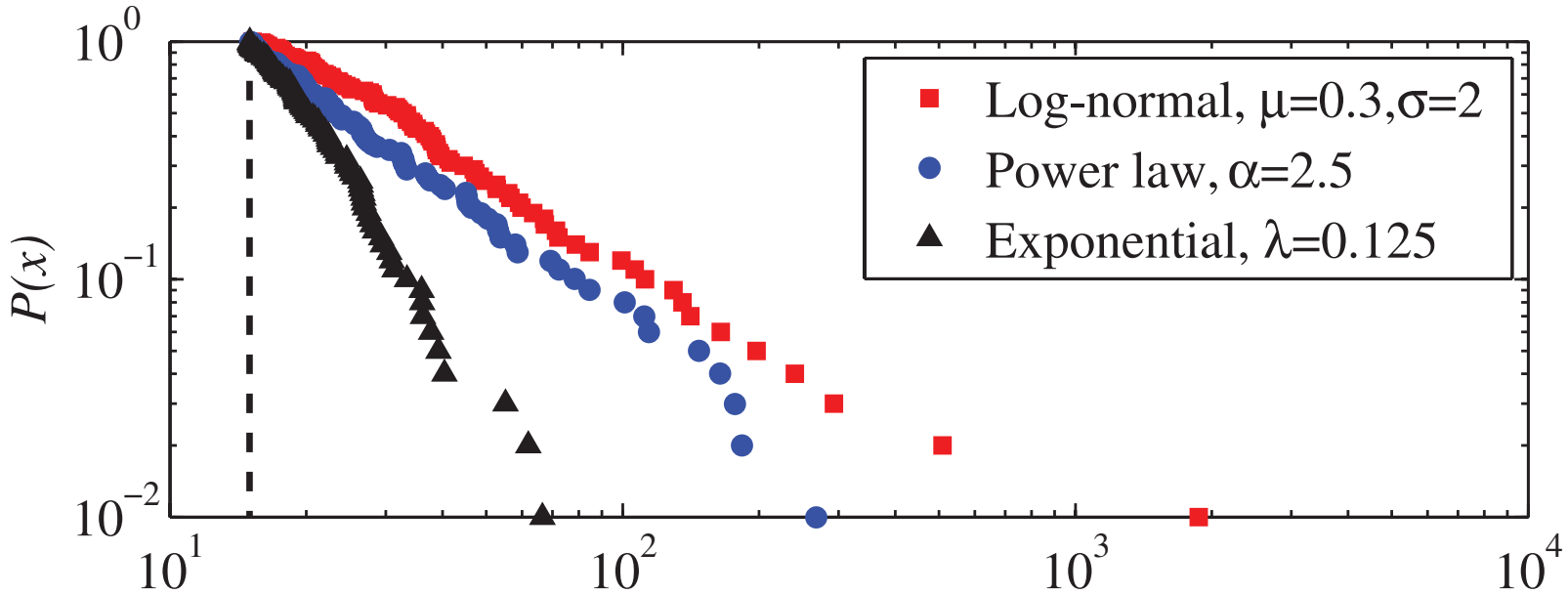
Power-law behavior cannot be reliably eyeballed.



Clauset et al, 2009

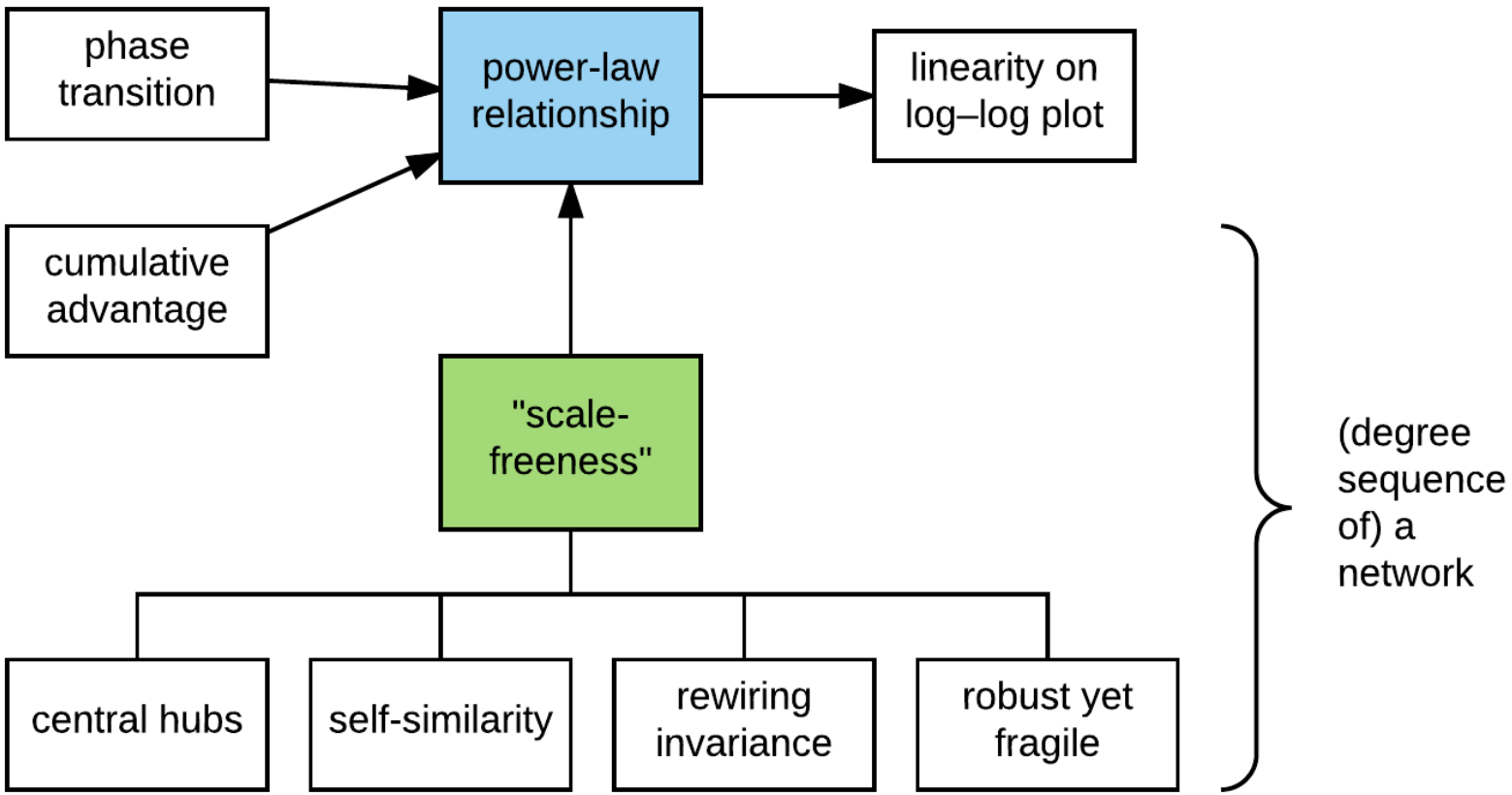
Statistical inference

Power-law behavior cannot be reliably eyeballed.

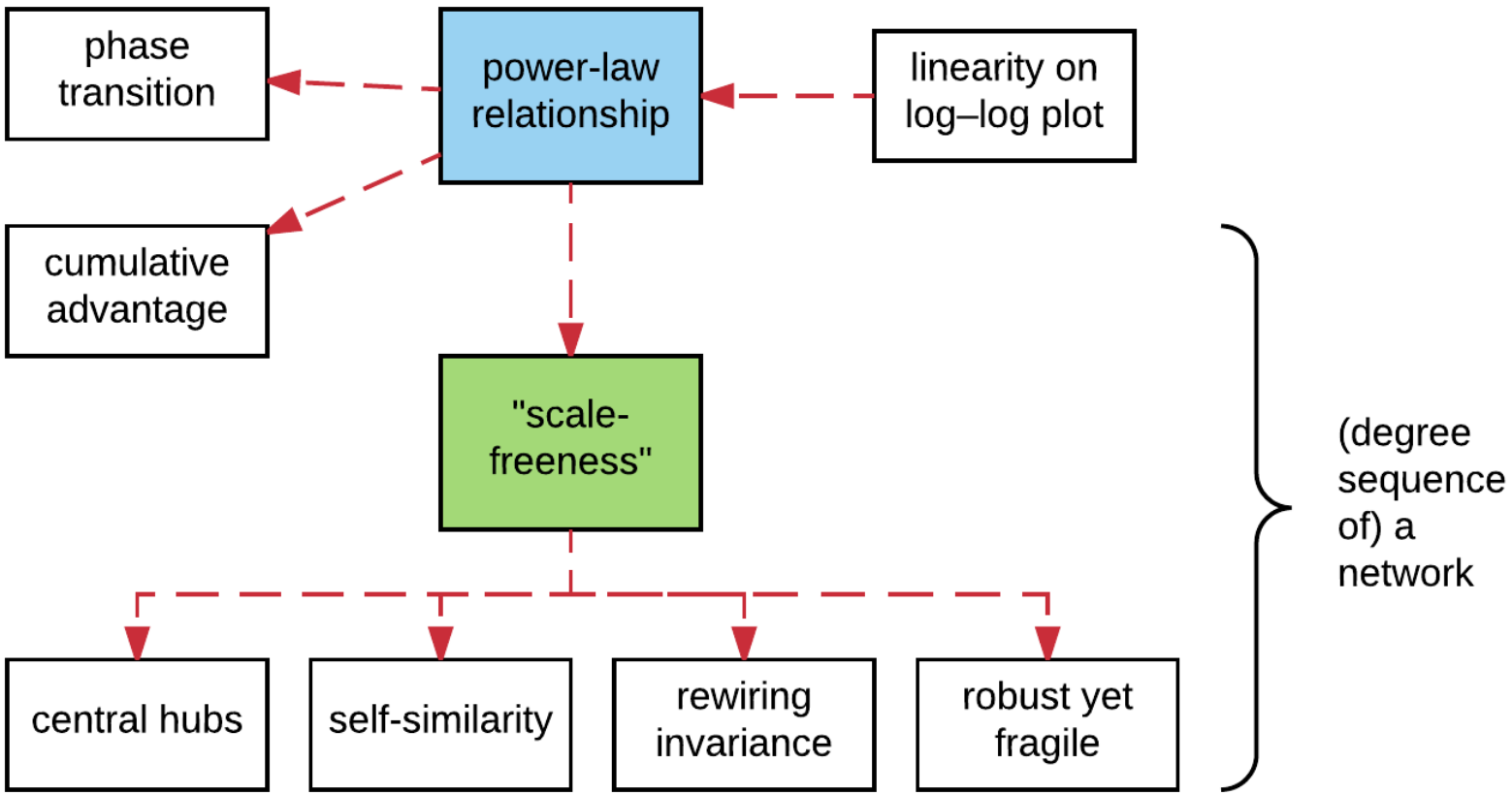


Clauset et al, 2009

Implications



Implications



Practice: Hypothesizing power laws

Practice: Hypothesizing power laws

- Vuong, Q H (1989). "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses". *Econometrica* **57**(2), 307–333.
- Clauset, A, Young, M, (2007). "On the frequency of severe terrorist events" *Journal of Conflict Resolution* **51**(1), 58–87.
- Clauset, A, Shalizi, C R, Newman, M E J (2009). "Power-Law Distributions in Empirical Data" *SIAM Review* **51**(4), 661–703. `http://tuvalu.santafe.edu/~aaronc/powerlaws/`
- Gillespie, C S (2015). "Fitting Heavy Tailed Distributions: The poweRlaw Package". *Journal of Statistical Software* **64**(2), 1–16. `http://www.jstatsoft.org/v64/i02/`

Recipe

How to analyze (discrete) empirical data hypothesized as

$$y = Cx^{-\alpha}, \quad x \geq x_{\min}$$

1. Estimate lower bound $x_{\min}$ and scaling parameter α
2. Assess goodness-of-fit (sampling + discernibility test)
3. Compare power-law to alternative hypotheses (LRT)

Clauset et al, 2009

Estimate the scaling parameter

For a fixed $x_{\min}$, either

- numerically maximize the likelihood function

$$-n \log \sum_{n=0}^{\infty} (n + x_{\min})^{-\alpha} - \alpha \sum_{i=1}^{n} \log x_i$$

or

- use the approximation

$$\hat{\alpha} \simeq 1 + n \left[\sum_{i=1}^{n} \log \frac{x_i}{x_{\min} - \frac{1}{2}} \right]^{-1}$$

(best when $x_{\min} \gtrsim 6$)

Clauset et al, 2009

ESTIMATE THE LOWER BOUND

1. For each $x_{\min}$, define

   $S(x) =$ CDF of data satisfying $x \geq x_{\min}$

   $P(x) =$ CDF of best-fit power-law model over $x \geq x_{\min}$*

2. Pick a distribution distance measure D
3. Select $x_{\min}$ that minimizes $D(S(x), P(x))$

Clauset et al, 2007

ESTIMATE THE LOWER BOUND

1. For each $x_{\min}$, define

   $S(x) =$ CDF of data satisfying $x \geq x_{\min}$

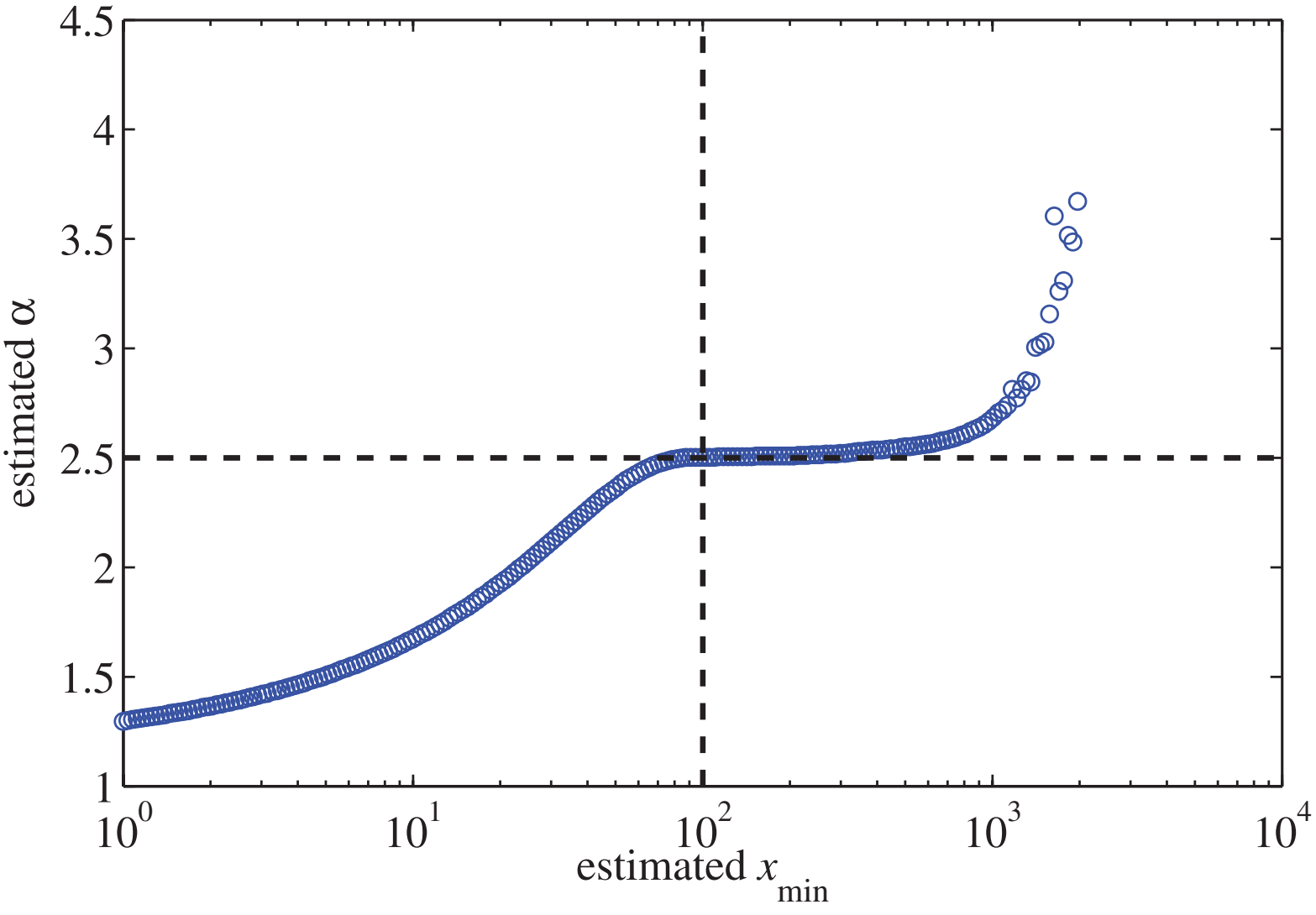
   $P(x) =$ CDF of best-fit power-law model over $x \geq x_{\min}$*

2. Pick a distribution distance measure D
3. Select $x_{\min}$ that minimizes $D(S(x), P(x))$

* Decide how to handle $P(X = x)$ for $x < x_{\min}$

Clauset et al, 2007

Evaluating an $x_{\min}$ estimate



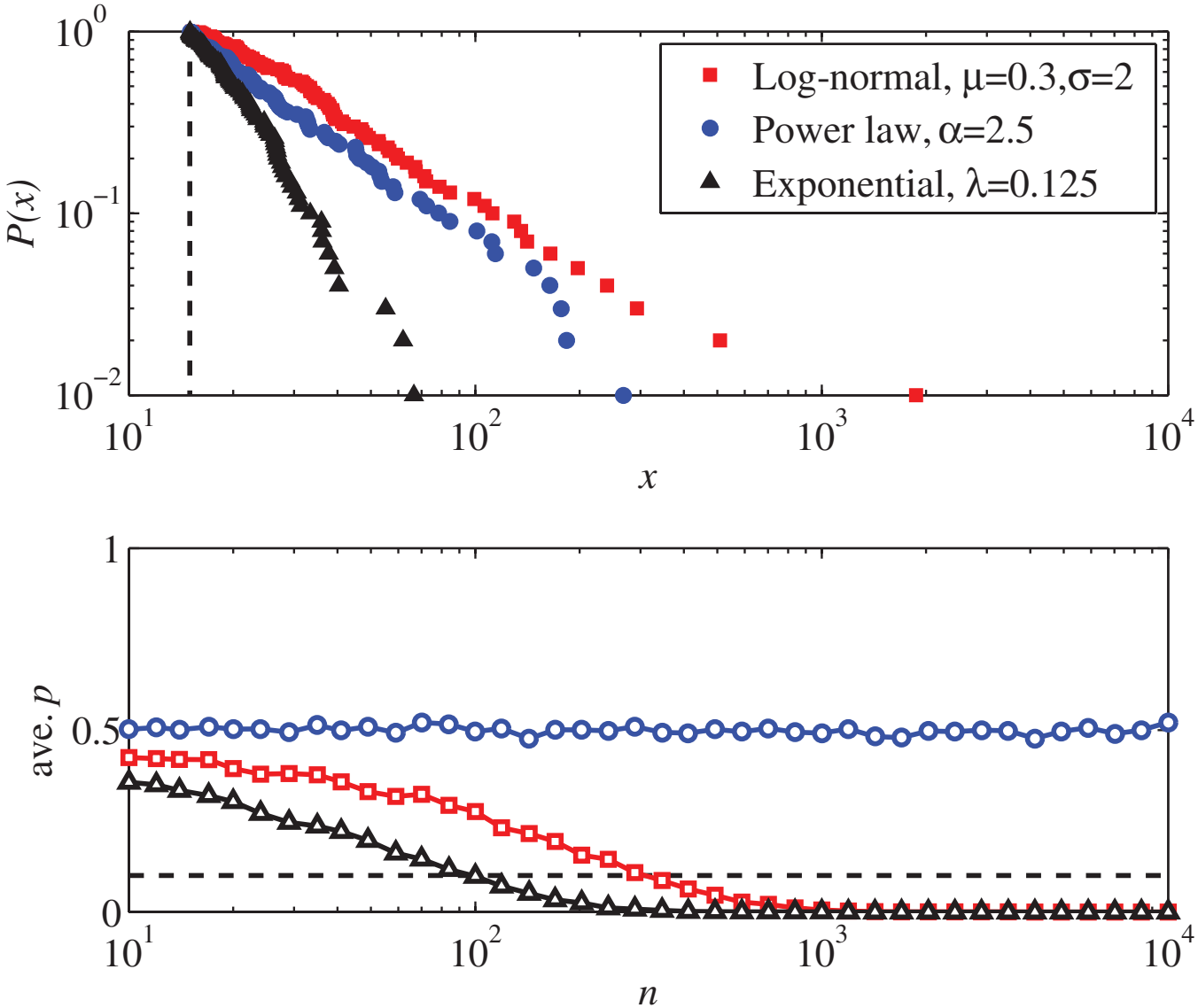
Clauset et al, 2009

Calculate the goodness-of-fit

-2. Pick D
-1. Estimate a power-law fit $P(x)$ from the data $S(x)$*
 0. Calculate $D = D(P(x), S(x))$
 1. Generate lots of artificial datasets $S'_i(x)$ from $P(x)$
 2. Calculate $D_i = D(P(x), S'_i(x)$
 3. Estimate $p \approx P(D_i > D)$

Clauset et al, 2009

EVALUATING THE GOODNESS-OF-FIT TEST



Clauset et al, 2009

Compare against alternative hypotheses

Given the best-fit power-law PDF and best-fit alternative PDF

$$p(x) \quad \text{and} \quad q(x)$$

calculate the likelihoods

$$L_p = \prod_{i=1}^{n} p(x_i) \text{ and } L_q = \prod_{i=1}^{n} q(x_i)$$

and consider the hypotheses

$$H_0 : E[L_p/L_q] = 0, \quad H_p : E[L_p/L_q] > 0, \quad H_q : E[L_p/L_q] < 0$$

Vuong, 1989; Clauset et al, 2009

Compare against alternative hypotheses

Given the best-fit power-law PDF and best-fit alternative PDF

$$p(x) \quad \text{and} \quad q(x)$$

calculate the likelihoods

$$L_p = \prod_{i=1}^{n} p(x_i) \text{ and } L_q = \prod_{i=1}^{n} q(x_i)$$

and consider the hypotheses

$$H_0 : E[L_p/L_q] = 0, \quad H_p : E[L_p/L_q] > 0, \quad H_q : E[L_p/L_q] < 0$$

Then the log-likelihood ratio

$$\mathcal{R} = \sum_{i=1}^{n} [\log p(x_i) - \log q(x_i)]$$

is asymptotically normal with $\sigma_{\mathcal{R}}^2 \overset{n\to\infty}{\longrightarrow} 0$

Vuong, 1989; Clauset et al, 2009

Evaluation of recipe

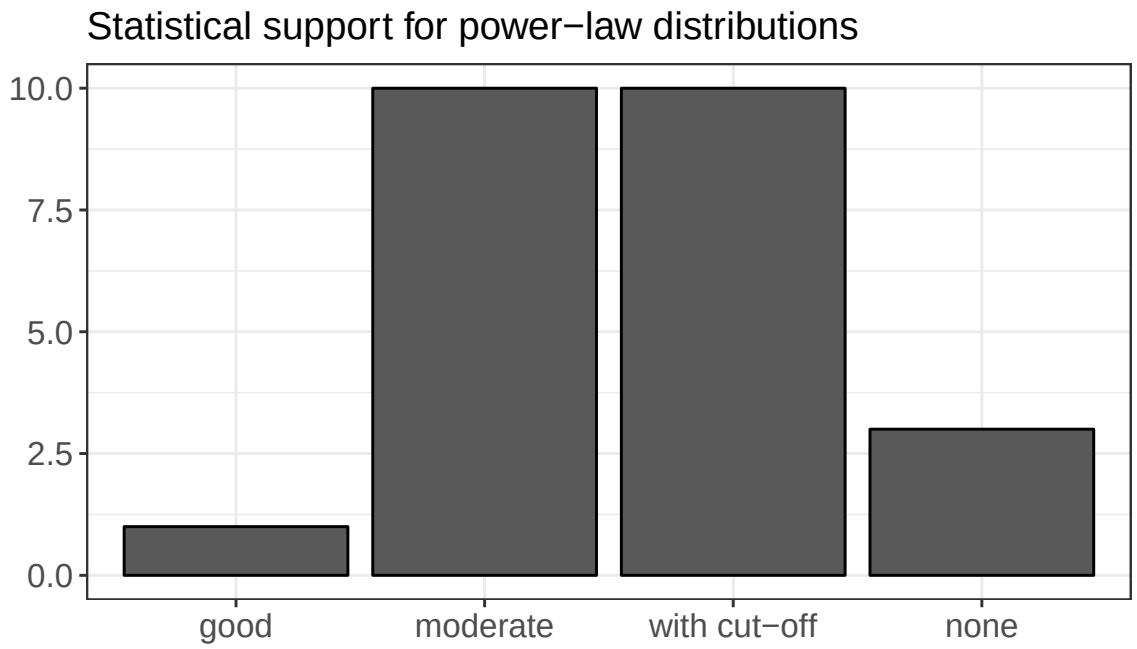
Compare 24 "power-law" datasets for power-law, log-normal, and exponential behavior…

(a) The frequency of occurrence of unique words in the novel *Moby Dick* by Herman Melville [43].
(b) The degrees (i.e., numbers of distinct interaction partners) of proteins in the partially known protein-interaction network of the yeast *Saccharomyces cerevisiae* [28].
(c) The degrees of metabolites in the metabolic network of the bacterium *Escherichia coli* [26].
(d) The degrees of nodes in the partially known network representation of the Internet at the level of autonomous systems for May 2006 [25]. (An autonomous system is a group of IP addresses on the Internet among which routing is handled internally or "autonomously," rather than using the Internet's large-scale border gateway protocol routing mechanism.)
(e) The number of calls received by customers of AT&T's long distance telephone service in the United States during a single day [1, 5].
(f) The intensity of wars from 1816–1980 measured as the number of battle deaths per 10 000 of the combined populations of the warring nations [53, 49].

Clauset et al, 2009

Statistical inference

Power-law behavior is not ubiquitous.

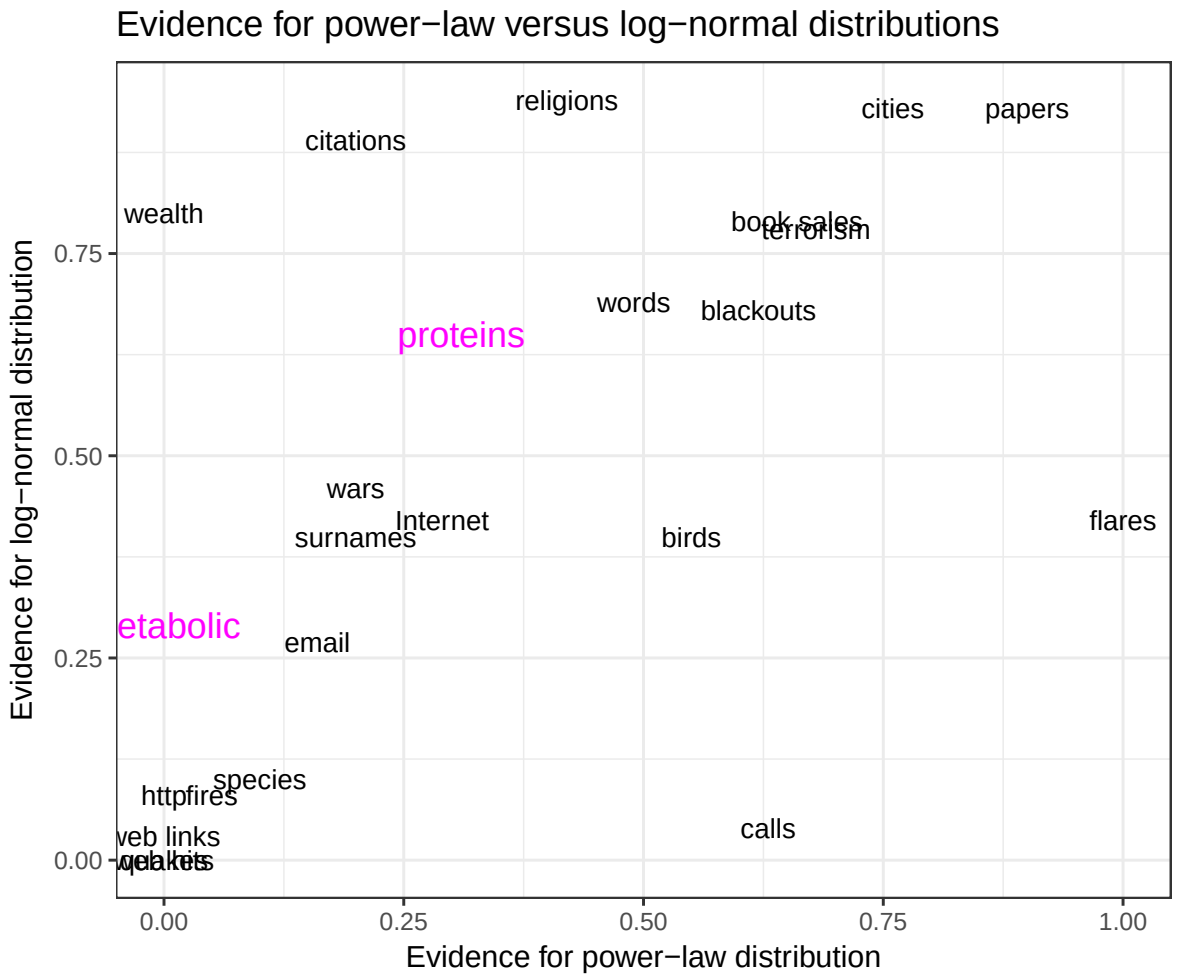


- "good": power-law is plausible; others are not
- "moderate": power-law is plausible; others are also
- "with cut-off": power-law with exponential cut-off is plausible
- "none": power-law is not plausible

Clauset et al, 2009

STATISTICAL INFERENCE

Power-law behavior is not ubiquitous.



Clauset et al, 2009

TUTORIAL IN `poweRlaw` PACKAGE FOR R

CICATS Study Group
20 March (Monday) @ noon