Redes Complexas
Aula 6

Roteiro

- Distribuição de Pareto
- Medindo lei de potência
- Estimando expoente
- Exemplos reais

Distribuição em Lei de Potência

- X é uma v.a. discreta ou contínua
- Distribuição de lei de potência
  - função de probabilidade

$$f_X(x) \sim c\, x^{-a}$$

c > 0, a > 1, constantes

- **Cauda pesada**: valores ordens de grandeza maior que a média podem ocorrer
- **Livre de escala**: razão entre probabilidades não depende da escala

Distribuição de Pareto

- Lei de potência para va. contínuas
  - Zeta e Zipf usadas para va. discretas
- Originalmente utilizada para caracterizar a distribuição da riqueza de pessoas em um país (por Vilfredo Pareto, na Itália do século 19)
  - atualmente usada para modelar diversos fenômenos
- Função densidade de probabilidade

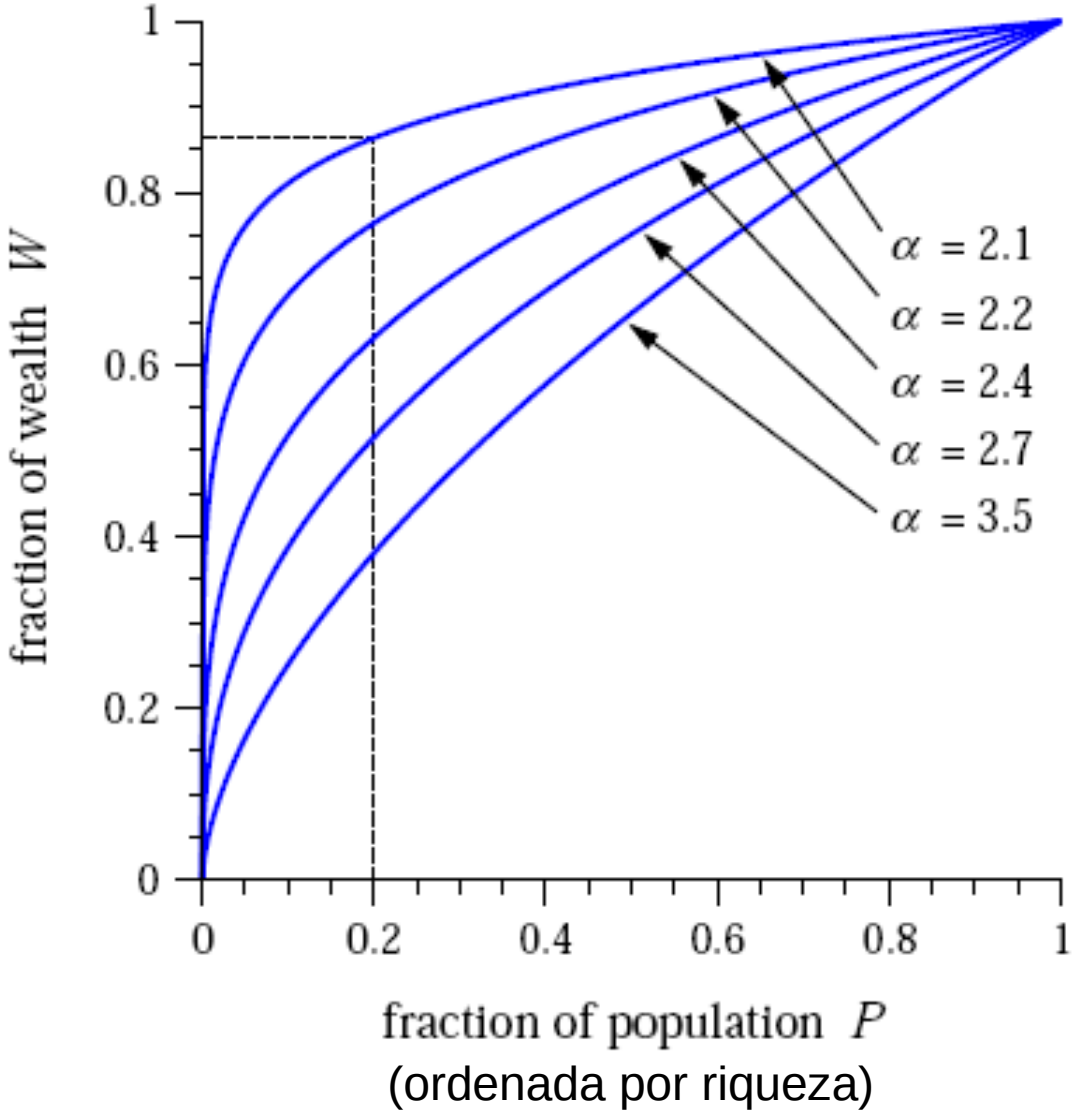
$$f_X(x) = \frac{a\, x_0^{\,a}}{x^{a+1}}$$

Parâmetros $a > 0$ e $x_0 > 0$

Definida para valores $x > x_0$

80-20 Rule

- Princípio de Pareto: 80 % dos recursos estão concentrado em 20% da população



- Fração mais rica da população versus fração de riqueza acumulada
- Curvas de Lorenz
- Usada para calcular coeficiente de Gini
- Princípio pode ser aplicado em outros contextos
  - ex. distribuição de seguidores no Twitter

Medindo Lei de Potência

- Muitos fenômenos parecem seguir lei de potência
- Dados empíricos, obtidos na prática
  - ex. renda, graus, praias, terremotos, estrelas, ...

Como identificar lei de potência?

- Plotar distribuição empírica

Muito cuidado!

Dados Reais

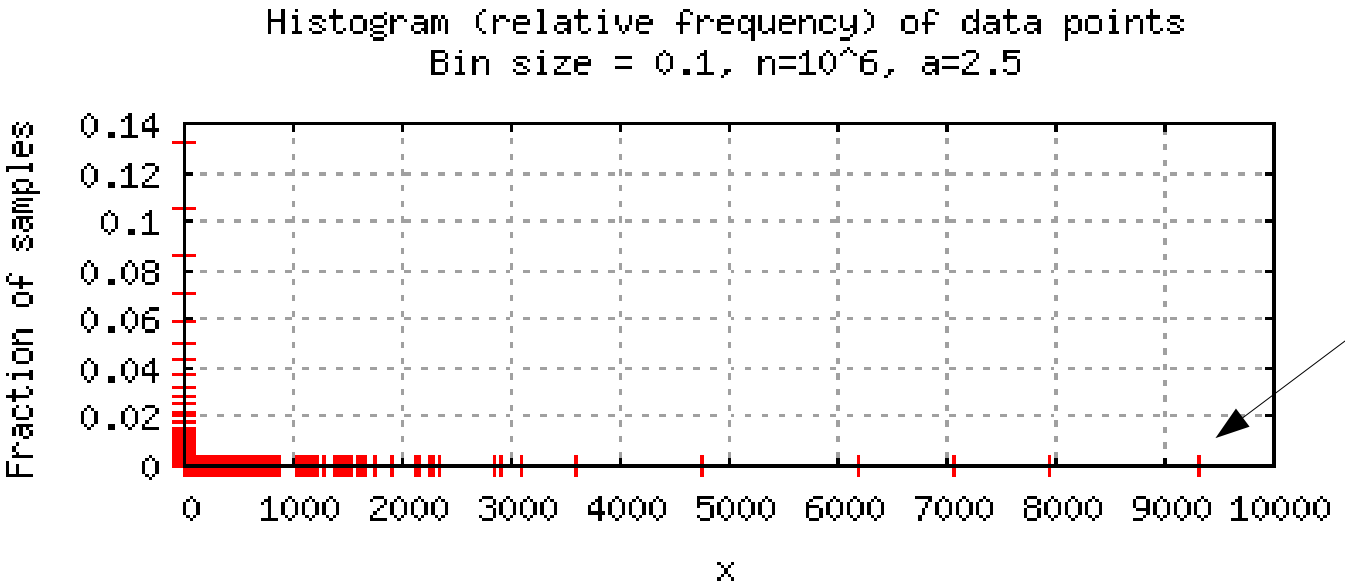
- Amostras geradas por simulação
  - 10^6 amostras
- Gerador pseudo-aleatório, método da transformada inversa
- Distribuição de Pareto com parâmetros
  - $a = 2.5$, $x_0 = 1$

Como apresentar resultados?

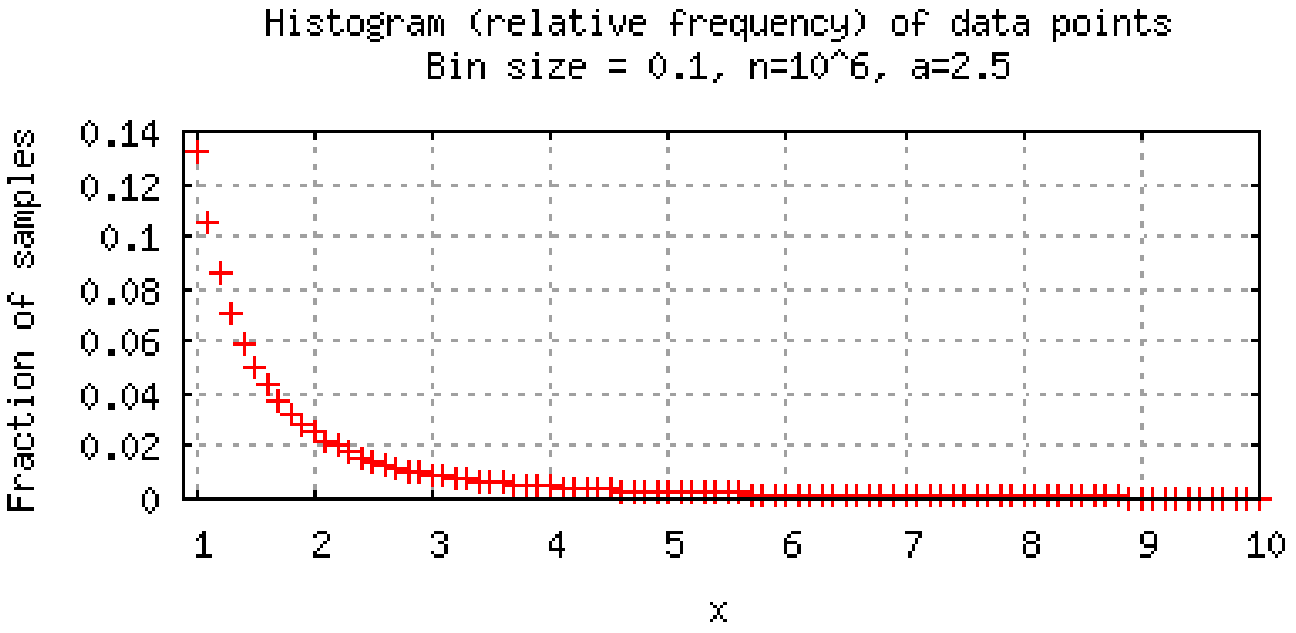
Histograma

- Definir intervalos de tamanho fixo
  - ex. b = 0.1
  - i-ésimo intervalo $[x_0+(i-1)*b , x_0+i*b)$
- Contar número de amostras em cada intervalo
- Dividir pelo total de amostras
  - frequência relativa

Resultados

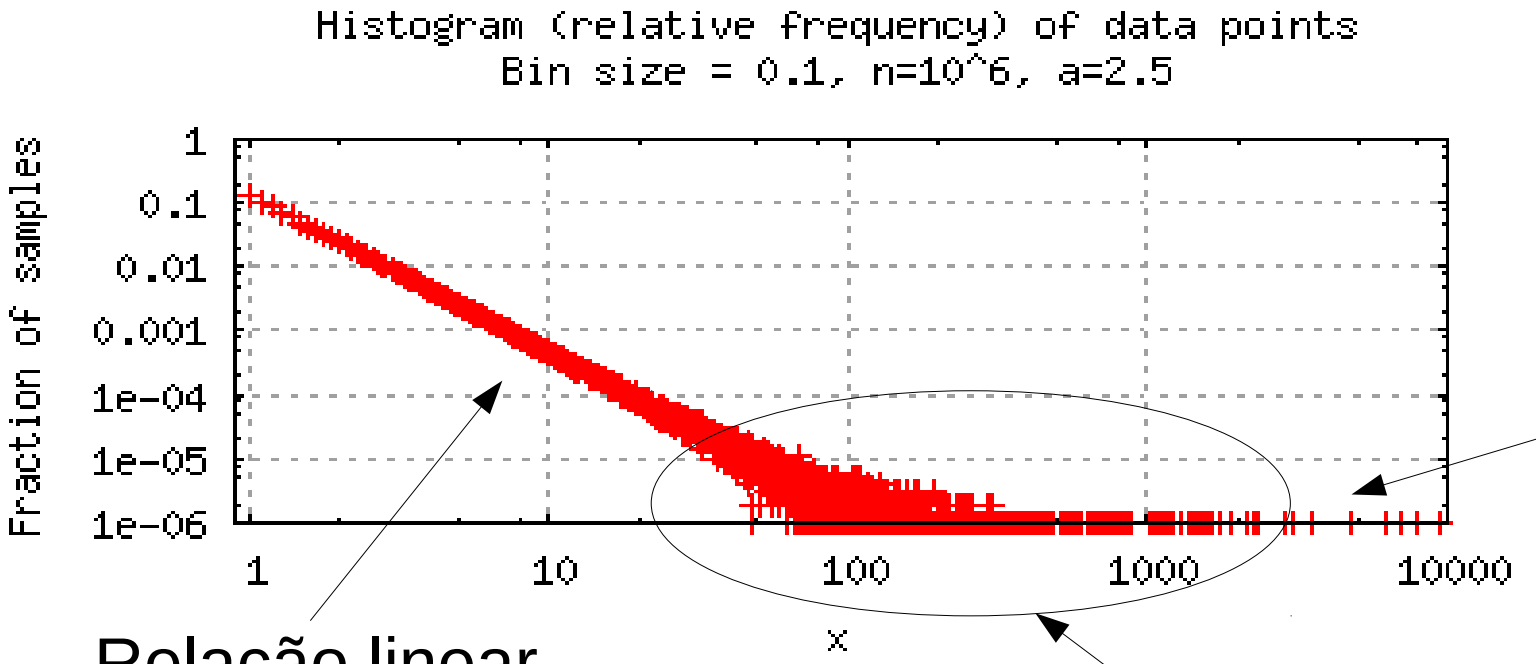


Valores muito grande ocorrem!



- Restringindo o eixo x

Resultados em Log-Log



Intervalos com apenas uma amostra (10^{-6})

Relação linear começa aparecer

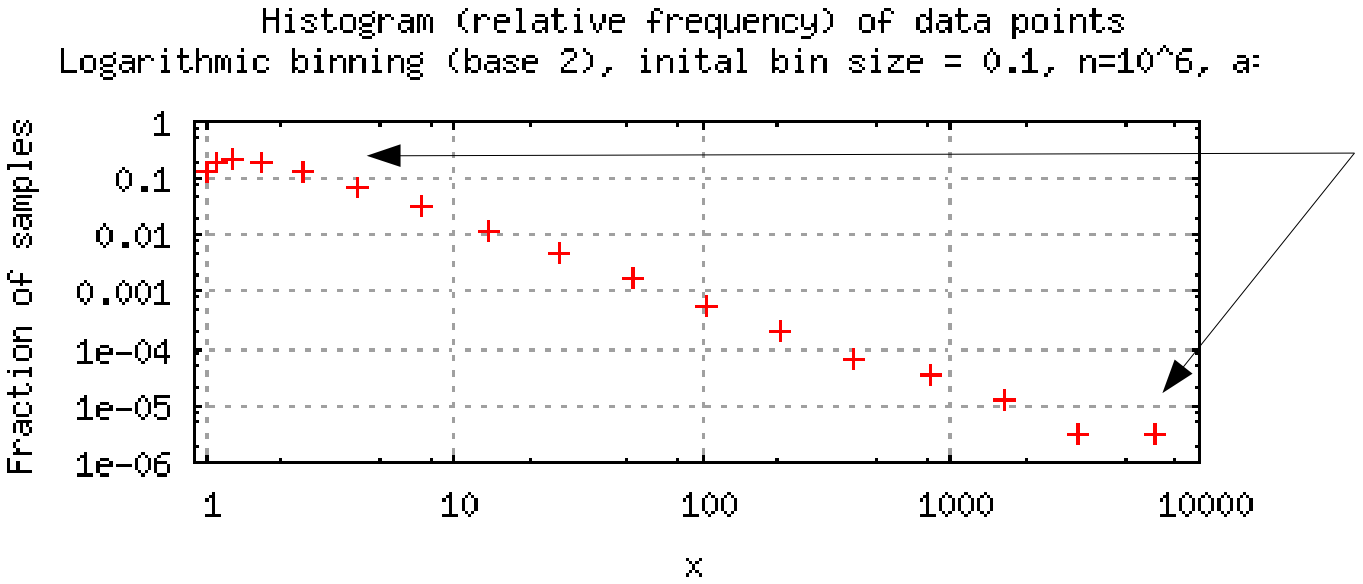
Ruído? Por que? Ignorar?

Outra idéia para visualizar?

Histograma Logarítmico

- **Problema**: intervalos contém poucos pontos quando x é grande
  - intervalo se torna relativamente ínfimo
- **Idéia:** Definir intervalos de tamanho *variável*
- Intervalos com crescimento exponencial
  - b tamanho do primeiro intervalo, 2b tamanho do segundo, 4b do terceiro, ...
  - i-ésimo intervalo $[x_0+2^{i-1}*b , x_0+2^{i}*b)$
- Intervalos espaçados uniformemente em escala log
- Calcular frequência relativa em cada intervalo

Resultados

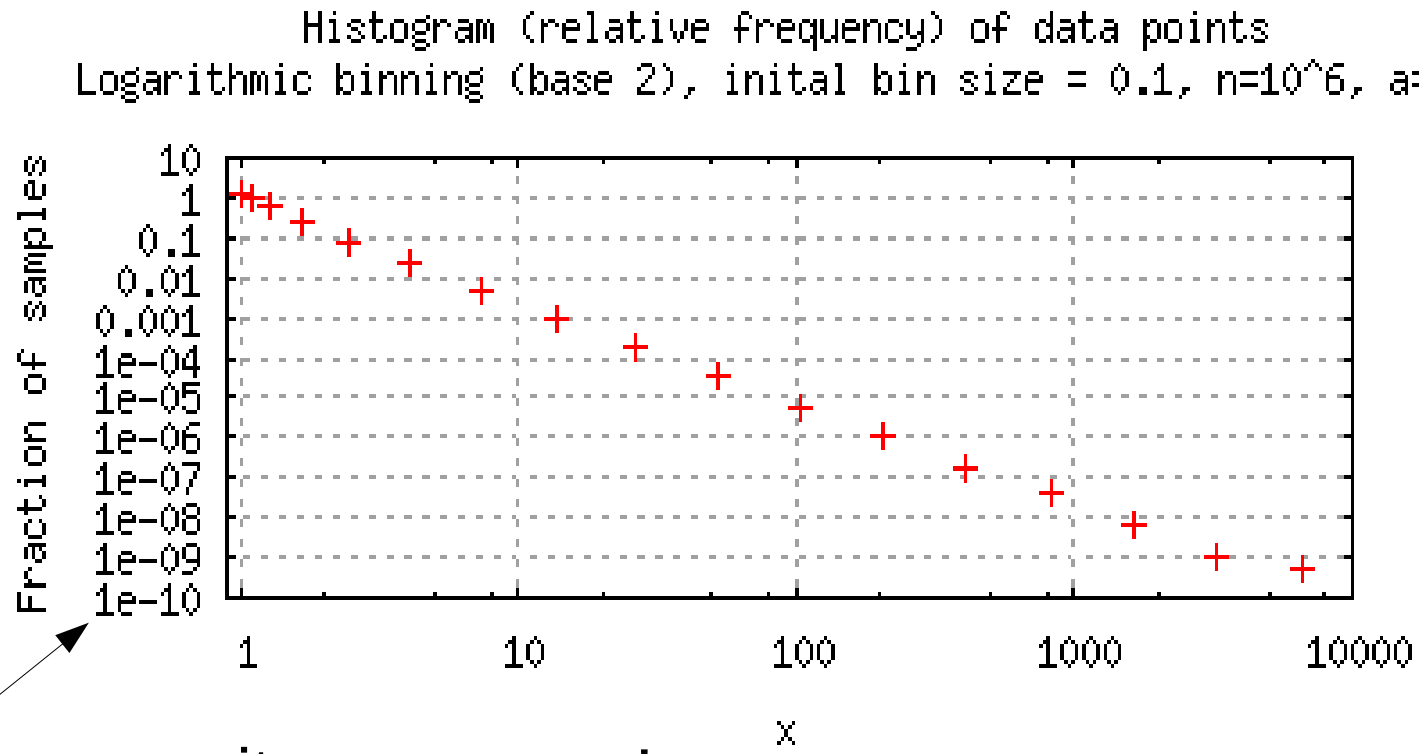


Problemas?

Intervalo maiores tem mais chance de ter pontos

- **Idéia**: normalizar pelo tamanho do intervalo
- Dividir número de amostras no *i*-ésimo intervalo pelo seu tamanho, 2^i
- Frequência relativa por unidade de valor
  - e não mais no intervalo

Intervalo Normalizado



Valores muito pequenos!

- Mas como estimar expoente?

Problemas com Histograma

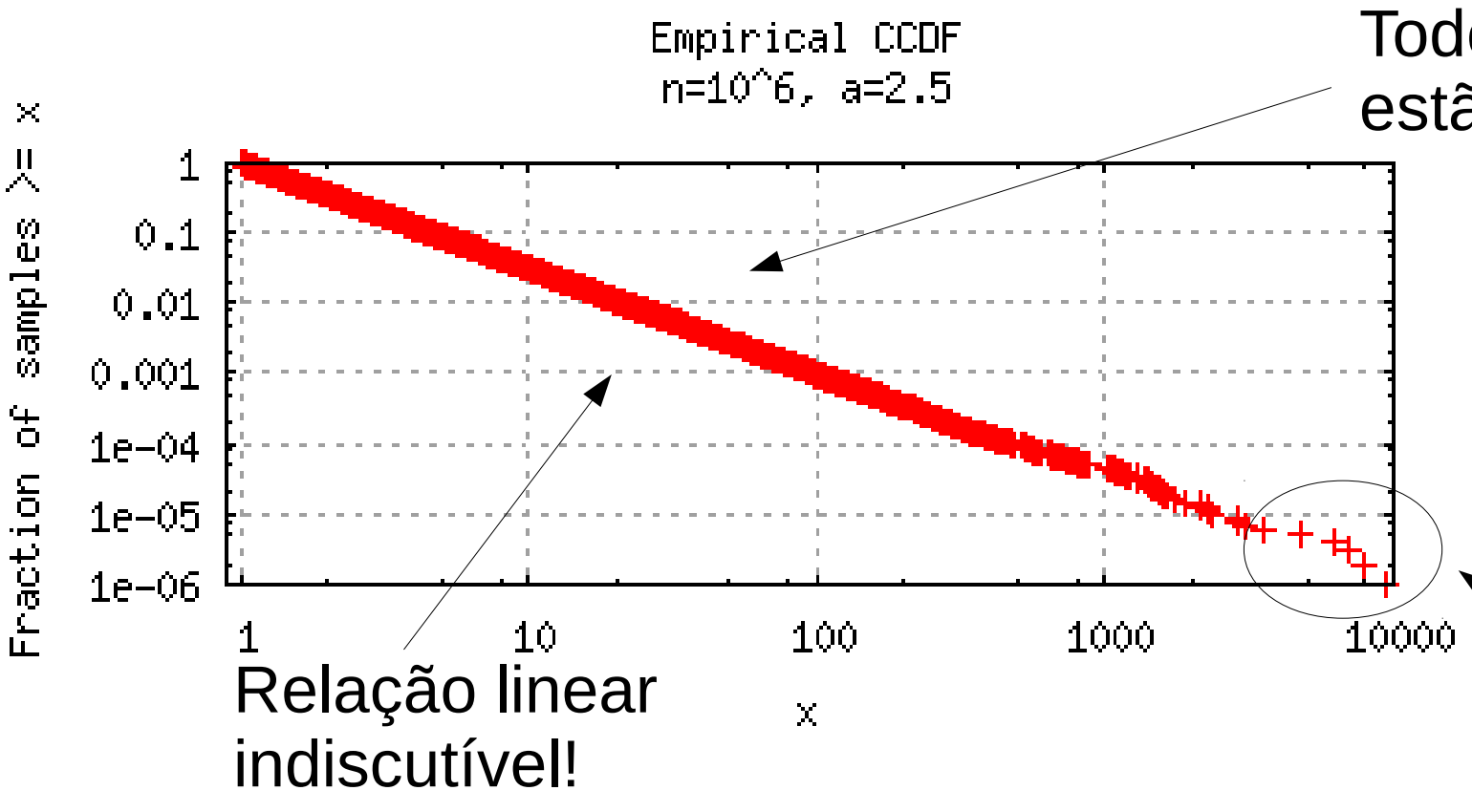
- Determinar tamanho do intervalo inicial
  - e potência, no caso logarítmico
- Valor do intervalo pode influenciar
- Número de amostras por intervalo diminui
  - mesmo no caso logarítmico
- **Agrega informação em intervalos!**
  - trabalha com “média”
- Perde informação das amostras

Outra idéia?

CCDF Empírica

- CCDF (Complementary Cumulative Distribution Function)
  - $P[X >= x] = 1 - P[X < x] = 1 - F_x(x)$
- Empírica
  - fração das amostras que são maiores que um valor
- Considerar todas as amostras
  - **não há intervalos**
- Ordenar amostras em ordem crescente
- Fração das amostras que são maiores ou iguais ao primeiro valor, ao segundo valor, etc.
- Visualizar em log-log

Resultado



Todos os pontos estão representados

Relação linear indiscutível!

Ruído? Porque?

- Método de visualização mais adequado
- Relação direta com exponente da PDF

Relação entre CCDF e PDF

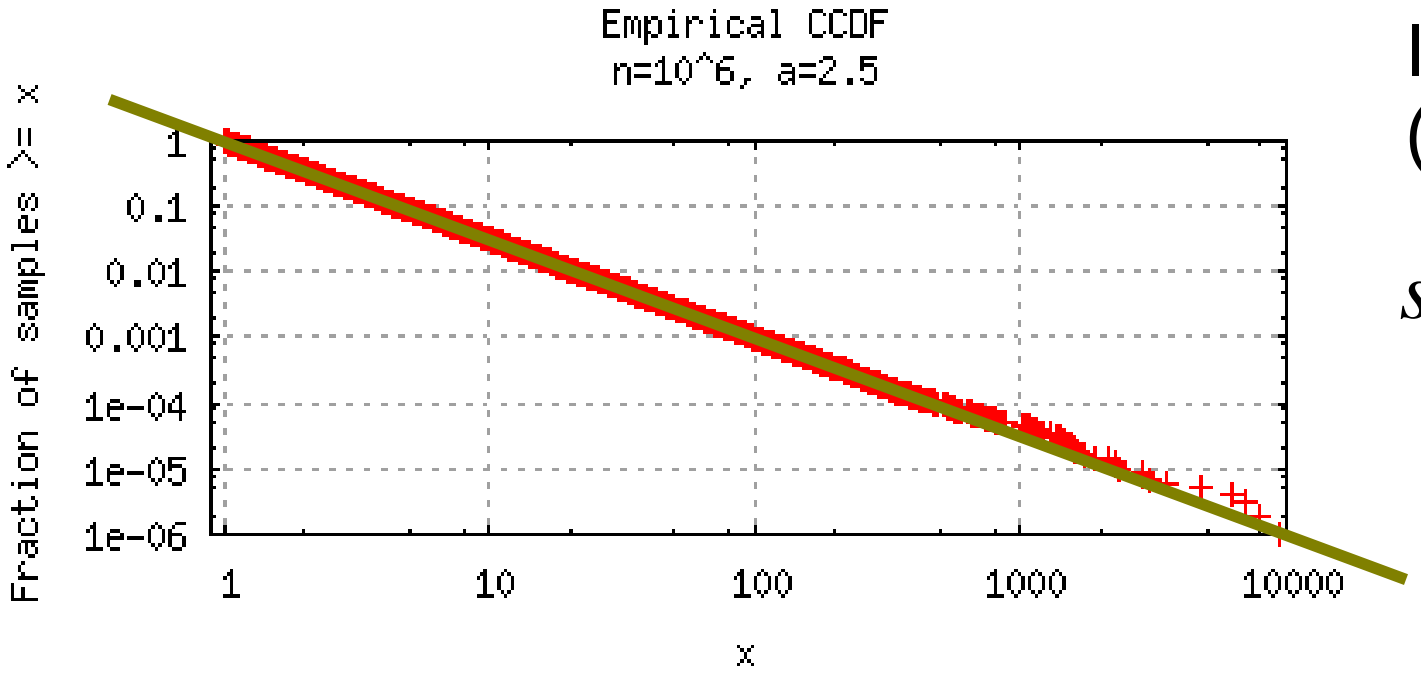
- Lembrando $f_X(x) = \frac{a\, x_0^a}{x^{a+1}}$
- F(y) para representar a CCDF

$$F(y) = \int_y^{\infty} f_X(x)\, dx \longrightarrow F_X(y) = \left(\frac{x_0}{y} \right)^a$$

- CCDF também segue lei de potência
- Exponente é uma unidade menor (em valor absoluto)
  - Ex: expoente CCDF = 2.1, expoente PDF = 3.1

Estimando o Expoente

- Regressão linear no gráfico log-log
  - usando todos os pontos?
- Usar inclinação da reta como expoente



Inclinação
(na escala log)?

$$s = \frac{\Delta y}{\Delta x} = \frac{6}{4} = 1.5$$

Correto!
Pois a = 2.5

- Forma mais comum, mas menos adequada
- Estimador pode ser muito ruim

Estimando o Expoente

- Forma mais adequada, via MLE
  - Maximum Likelihood Estimation
- **Idéia:** obter *a* para o qual as amostras geradas seja mais provável

$L(x_1,\ldots,x_n|a)$ ←

- Prob. de de gerar as n amostras dado um expoente a
- Likelihood function

$$L(x_1,\ldots,x_n|a)=\Pi_{i=1}^{n} f_X(x_i)=\Pi_{i=1}^{n} \frac{a\,x_0^a}{x_i^{a+1}}$$

- Trabalhar com a log likelihood funcion
  - $l(x_1,\ldots,x_n|a)=\log L(x_1,\ldots,x_n|a)$

Estimador MLE

- Obter o valor máximo da função l(x|a)
  - derivar com relação a a, igualar a zero e resolver
- Precisamos determinar também x_0
  - menor valor dentre as amostras irá maximizar L
- Estimadores

$$\hat{x}_0 = min_i\, x_i$$

$$\hat{a} = n\left[\sum_{i=1}^{n} \ln \frac{x_i}{\hat{x}_0}\right]^{-1}$$

← Estimador é uma v.a.

Erro do Estimador

- Erro do estimador dado por seu desvio padrão
- Podemos calcular E[$\hat{a}$] e E[$\hat{a}^2$] para v.a. $\hat{a}$ que é o estimador

$$E[\hat{a}] = a$$

Valor esperado do estimador é o parâmetro que queremos estimar

$$Var[\hat{a}] = \frac{(a-1)^2}{n}$$

Variância do estimador decresce com n (número de amostras)

$$\sigma_{\hat{a}} = \sqrt{Var[\hat{a}]} = \frac{a-1}{\sqrt{n}}$$

Desvio padrão do estimador usado como medida de erro

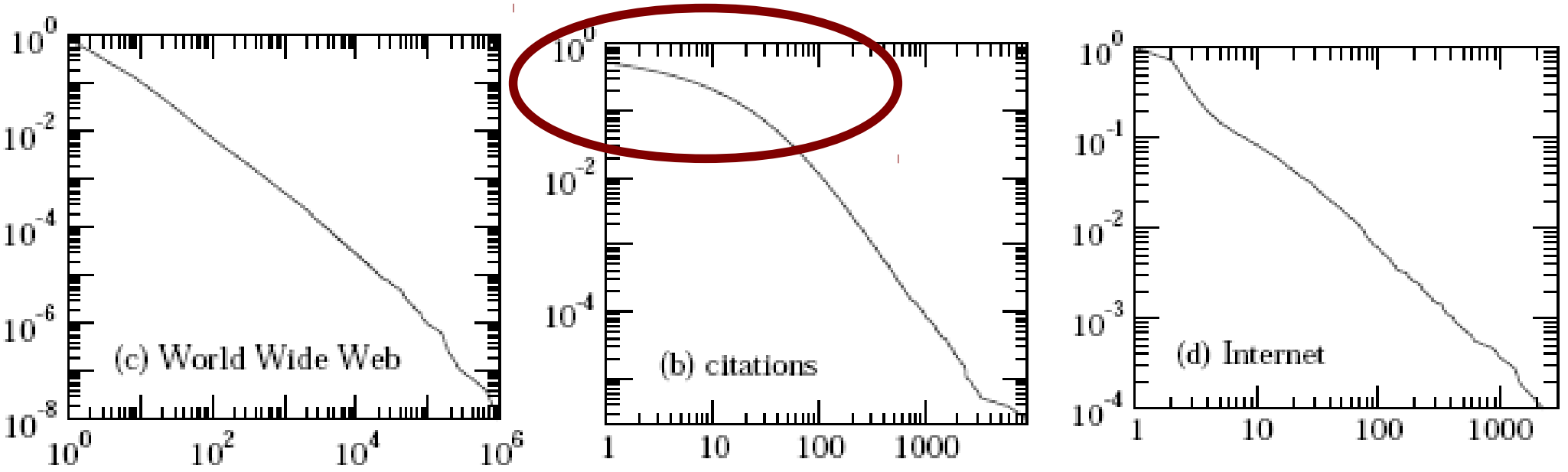
$$\log L(x_1, \ldots, x_n | \hat{a})$$

Medida de qualidade do estimador (valor da *likelihood function*)

Determinando o Início

- Na prática, distribuição empírica não segue lei de potência sobre todas as escalas
  - ruídos e outros fenômenos em escala menores
- Lei de potência para valores grandes, a partir de certo x_0
- Ignorar valores pequenos, onde distribuição desvia de lei de potência
- **Problema**: determinar x_0
  - onde começa a lei de potência?

Exemplos Reais

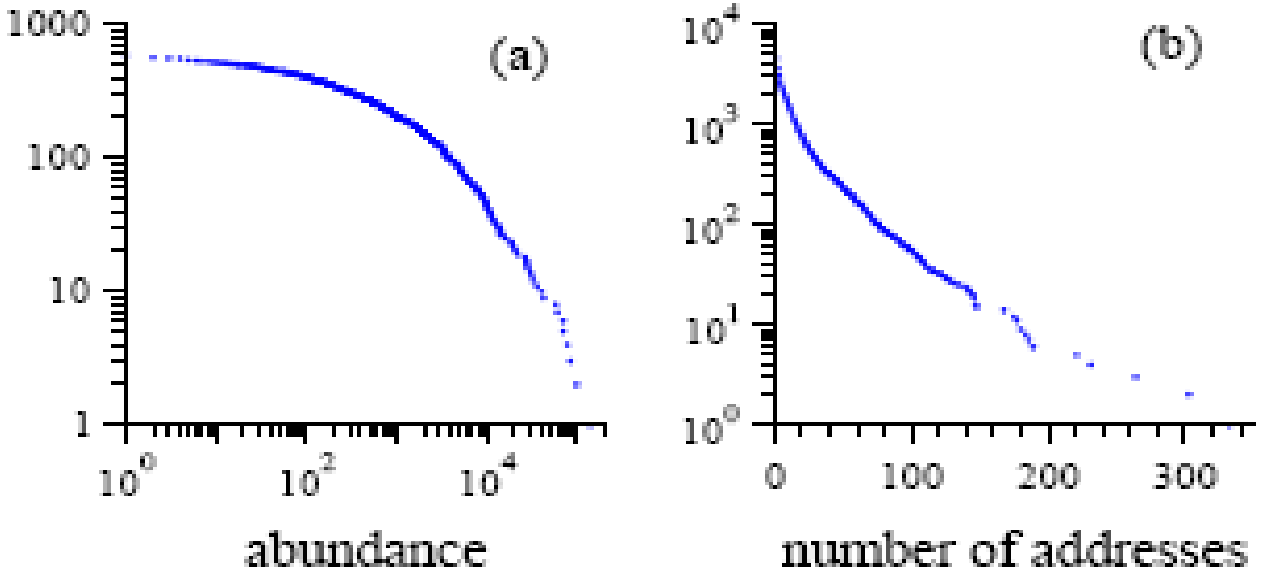


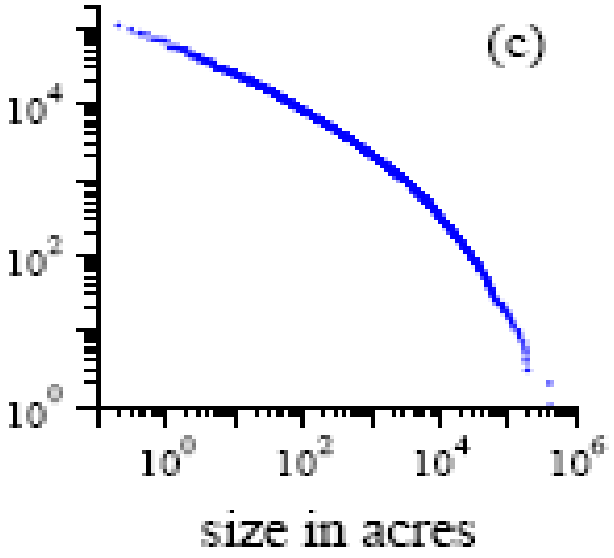
- *Cauda* segue lei de potência
  - “Power law tail”
  - x_0 pode ser relativamente grande

A Arte de Determinar x_0

- x_0 muito pequeno
  - ruídos perto de zero influenciam estimativa do expoente
- x_0 muito grande
  - perda de informação, ruído no final da cauda
- Expoente estimado depende de x_0
- Variar x_0 e avaliar função de likelihood ou algum teste de divergência (KS)
  - automatizar a inferência de x_0
- Usar outra técnica para estimar expoente
  - extreme value theory

Nem tudo é Lei de Potência





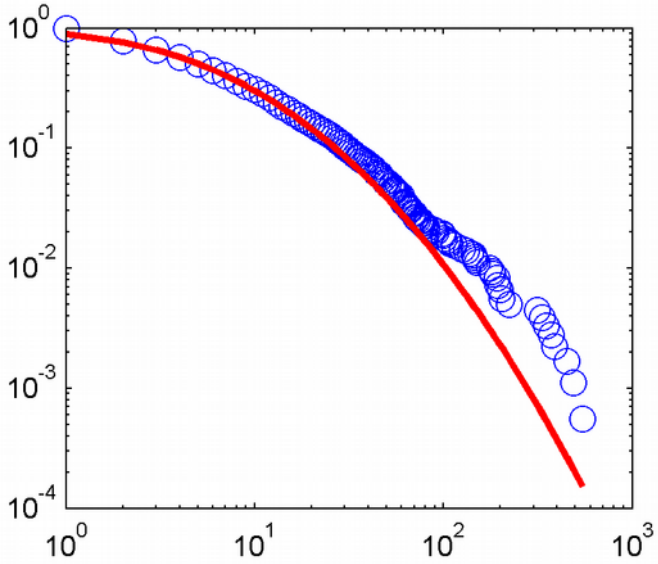
- Algumas va. assumem valores grandes
  - longe da média
- Distribuição não segue lei de potência
  - nem na cauda!
- Muitos casos são inconclusivos

Distribuição Log-Normal

- X va contínua, x > 0
- X tem distribuição log-normal se logaritmo de X tem distribuição Normal
  - se Y = log (X) tem distribuição Normal
- Dois parâmetros
  - média (μ) e variância (σ^2) da Normal Y
- Função densidade de probabilidade

$$f_X(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \quad x > 0$$

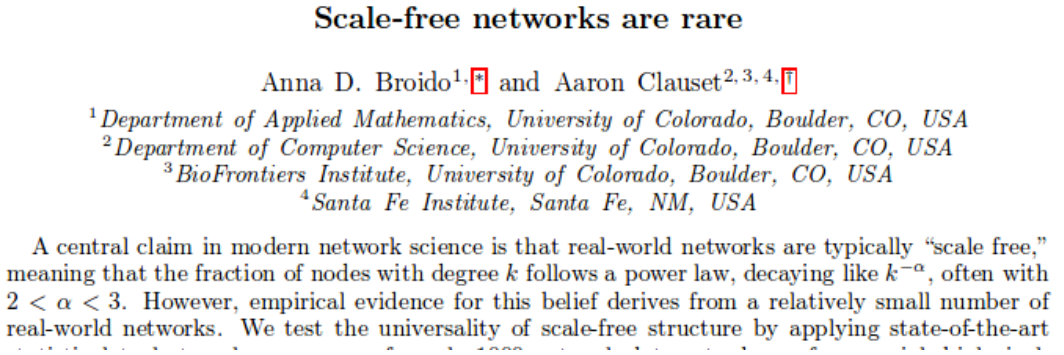
Distribuição Log-Normal



- Cauda pesada
  - valores muito longe da média com probabilidade não desprezível
- *Parece* com lei de potência
  - decaimento *sustentado* em log-log, mas não é lei de potência
  - Decaimento não é linear para valores arbitrariamente grandes de x

Motivo para muita discussão!

Debate Recente



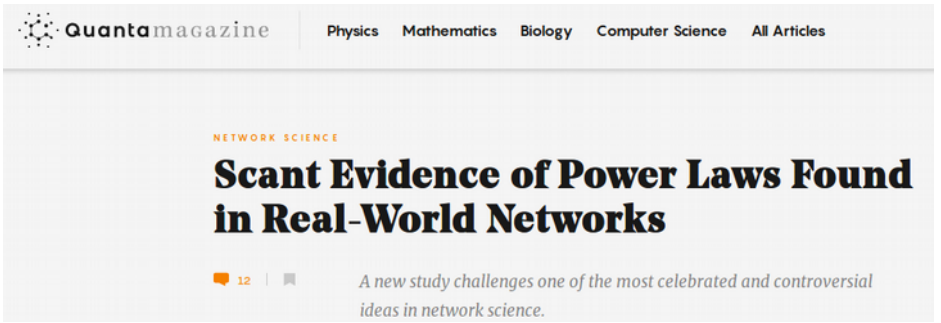
Scale-free networks are rare

Anna D. Broido[1,*] and Aaron Clauset[2,3,4,†]

[1] *Department of Applied Mathematics, University of Colorado, Boulder, CO, USA*
[2] *Department of Computer Science, University of Colorado, Boulder, CO, USA*
[3] *BioFrontiers Institute, University of Colorado, Boulder, CO, USA*
[4] *Santa Fe Institute, Santa Fe, NM, USA*

A central claim in modern network science is that real-world networks are typically "scale free," meaning that the fraction of nodes with degree k follows a power law, decaying like $k^{-\alpha}$, often with $2 < \alpha < 3$. However, empirical evidence for this belief derives from a relatively small number of real-world networks. We test the universality of scale-free structure by applying state-of-the-art

- ArXiv, 9 Jan 2018
- A. Clauset: *Rising Star* em Network Science (Erdós-Rény Prize 2016)



Quantamagazine — Physics | Mathematics | Biology | Computer Science | All Articles

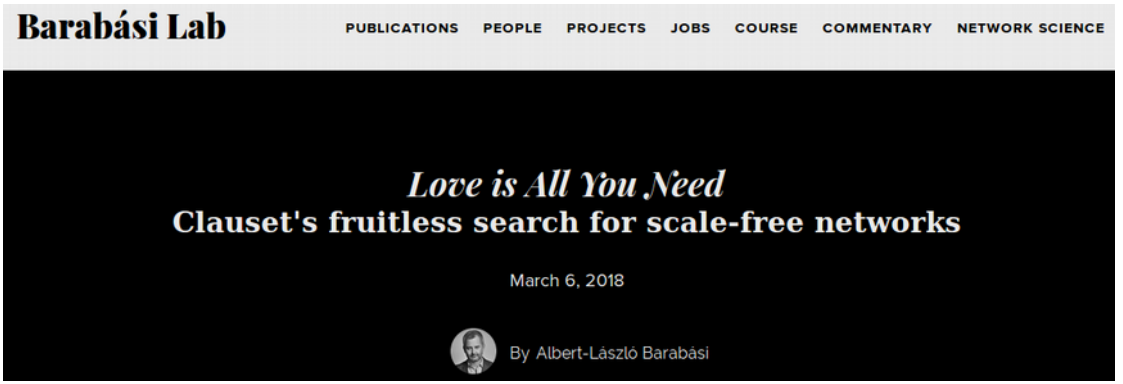
NETWORK SCIENCE

Scant Evidence of Power Laws Found in Real-World Networks

12

A new study challenges one of the most celebrated and controversial ideas in network science.

- *Quanta Mag*, Fev 2018
- Repercusão na mídia comum – *The Atlantic*



Barabási Lab — PUBLICATIONS | PEOPLE | PROJECTS | JOBS | COURSE | COMMENTARY | NETWORK SCIENCE

Love is All You Need

Clauset's fruitless search for scale-free networks

March 6, 2018

By Albert-László Barabási

- Blogpost do Barabási, Mar 2018, com críticas
- Artigo em 2019 (do mesmo grupo)

Scale-free networks well done

Ivan Voitalov, Pim van der Hoorn, Remco van der Hofstad, and Dmitri Krioukov
Phys. Rev. Research **1**, 033034 – Published 18 October 2019

Segue o debate!

Expoente de Redes Reais

Network Name	n	$\bar{k}$	$\hat{\gamma}^{Hill}$	$\hat{\gamma}^{Mom}$	$\hat{\gamma}^{Kern}$
CAIDA (IN)	26,475	4.03	2.1	2.11	2.11
Skitter (SK)	1,696,415	13.08	2.38	2.36	2.43
Actor collaborations (CL)	382,219	173.28	3.71	$6.7 \cdot 10^3$	2.36
Amazon (CA)	334,863	5.53	3.99	3.48	3.44
arXiv (AP)	18,771	21.1	4.41	5.78	7.29
Bible names (MN)	1,773	10.3	3.09	3.36	2.88
Brightkite (BK)	58,228	7.35	3.51	3.8	2.96
Catster (Sc)	149,684	72.8	2.09	2.06	1.98
Catster/Dogster (Scd)	623,748	50.33	2.1	2.11	2.04
Chicago roads (CR)	1,467	1.77	77.92	∞	∞
DBLP (CD)	317,080	6.62	6.59	13.99	3.06
Dogster (Sd)	426,816	40.03	2.15	2.15	2.12
Douban (DB)	154,908	4.22	4.42	6.88	1.86
U. Rovira I Virgili (A@)	1,133	9.62	6.49	∞	∞
Euro roads (ET)	1,174	2.41	4.73	44.48	29.57
Flickr (LF)	1,715,254	18.13	3.94	4.29	5.02
Flickr (FI)	105,938	43.74	6.18	1.79	1.65
Flixster (FX)	2,523,386	6.28	53.63	1.93	1.95
Gowalla (GW)	196,591	9.67	2.8	2.8	2.86
Hamsterster (Shf)	1,858	13.49	4.45	8.09	3.51
Hamsterster (Sh)	2,426	13.71	4.57	25.39	6.32
Hyves (HY)	1,402,673	3.96	2.98	2.23	1.99
LiveJournal (Lj)	5,203,763	18.72	3.86	4.04	3.15
Livemocha (LM)	104,103	42.13	9.13	∞	2.39
Orkut (OR)	3,072,441	76.28	3.58	2.65	3.35
Power grid (UG)	4,941	2.67	6.62	7.76	9.2
Proteins (Mp)	1,846	2.39	3.09	3.31	3.87
Reactome (RC)	6,229	46.93	4.86	34.33	∞
Roads CA (RO)	1,965,206	2.82	18.86	∞	∞
Roads PA (RD)	1,088,092	2.83	18.24	∞	∞
Roads TX (R1)	1,379,917	2.79	21.83	∞	∞
Route views (AS)	6,474	3.88	2.13	2.16	2.14
WordNet (WO)	146,005	9.00	2.86	2.68	2.61
Youtube (CY)	1,134,890	5.27	2.48	2.58	2.17
Human PPI (MV)	3,023	4.07	3.04	3.4	3.03

- Três estimadores (baseados em *extreme value theory*)
- Vermelho = não é LP
- Amarelo = quase não é LP
- Verde = LP com segundo momento infinito
- Azul = LP com segundo momento finito
- Muitas redes reais com LP tem exponente entre 2 e 3 (variância infinita)
- Muitas redes reais não exibem LP

Scale-free networks well done

Ivan Voitalov, Pim van der Hoorn, Remco van der Hofstad, and Dmitri Krioukov
Phys. Rev. Research **1**, 033034 – Published 18 October 2019