Redes Complexas
Aula 7

Aula retrasada

- Lei de potência
- Distribuição Zeta
- Propriedades
- Distribuição Zipf
- Exemplo Wikipedia

Aula de hoje

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

Distribuição de 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 muito longe da 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 entre indivíduos num país (por Vilfredo Pareto, na Itália do século 19)
  - atualmente usada para modelar diversos fenômenos
- Função densidade de probabilidade

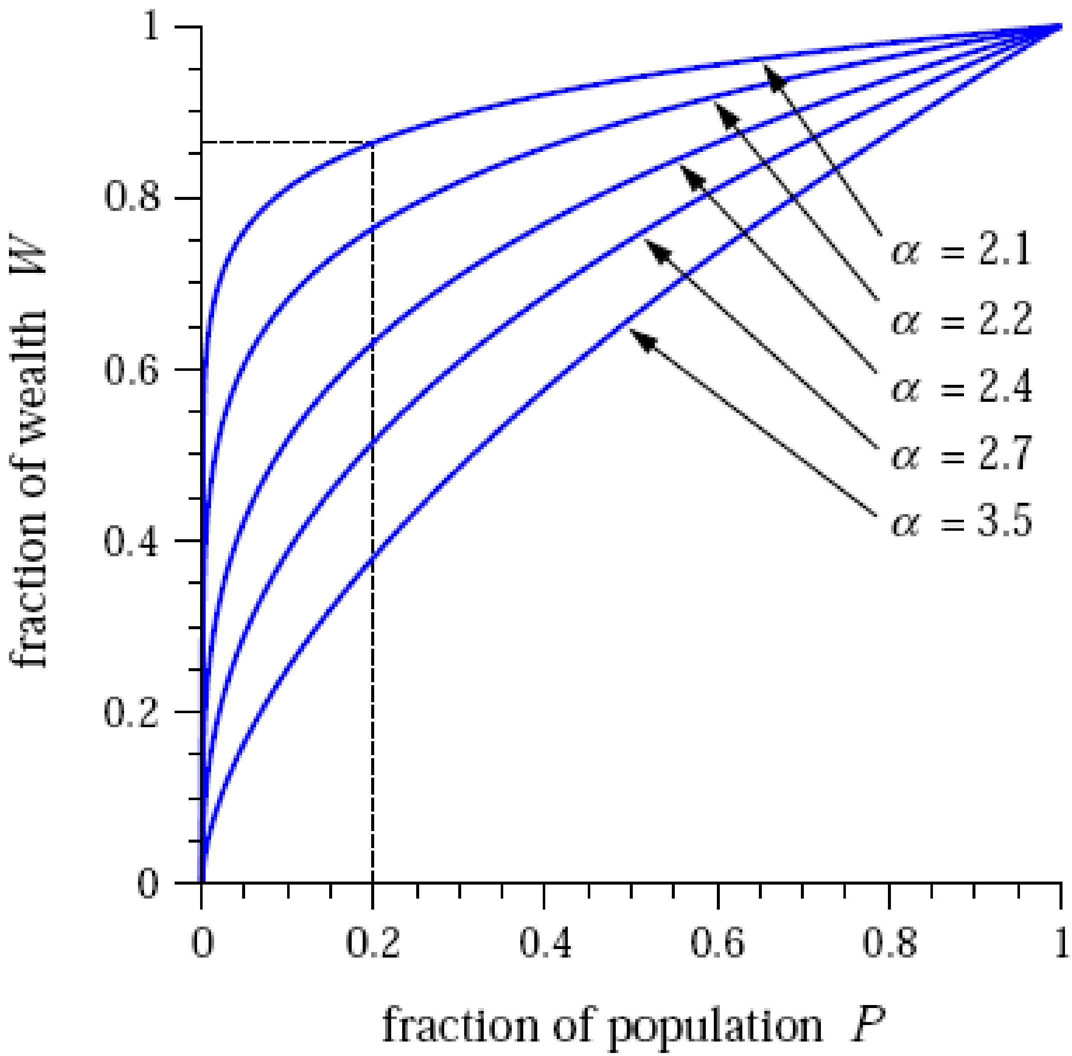
$$f_X(x) = \frac{a\, x_0^a}{x^{a+1}}$$

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 da população versus fração de riqueza acumulada
- Curvas de Lorenz
- Usada também para calcular coeficiente de Gini (índice de desigualdade na distribuição de renda)
- Brasil: um dos piores do mundo!

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

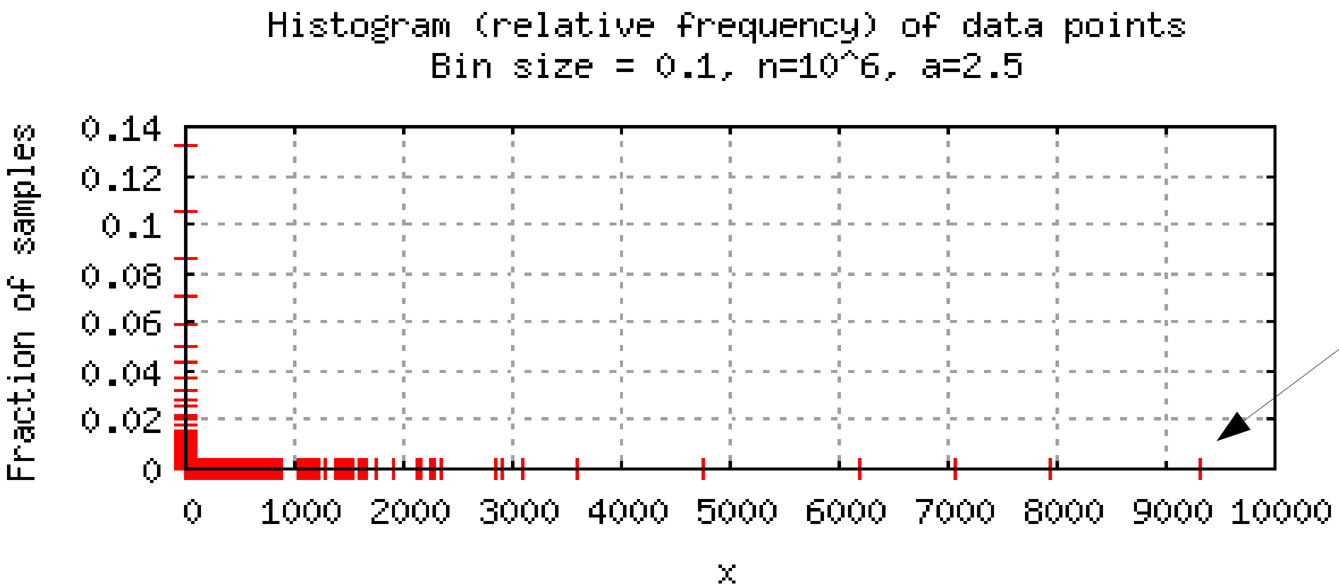
- Amostras geradas numericamente
  - 10^6 amostras
- Gerador pseudo-aleatório, método da transformada inversa
- Distribuição de Pareto com parâmetros
  - $a = 2.5$, $x_0 = 1$

Como apresentar resultados?

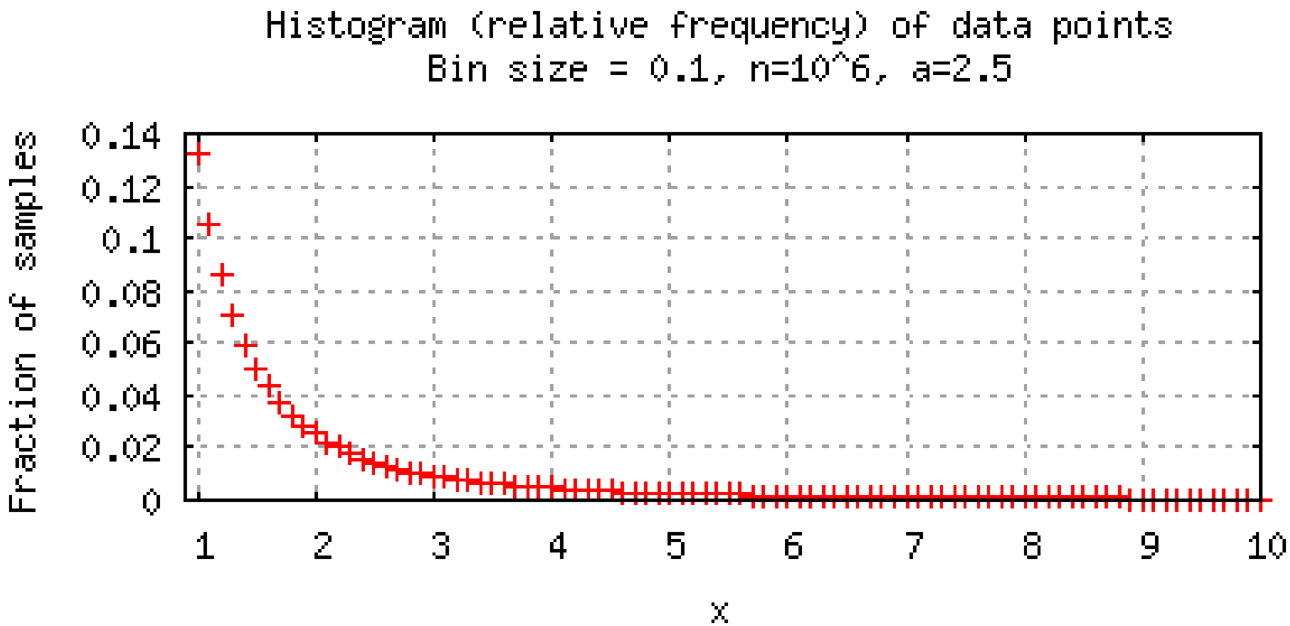
Histograma

- 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

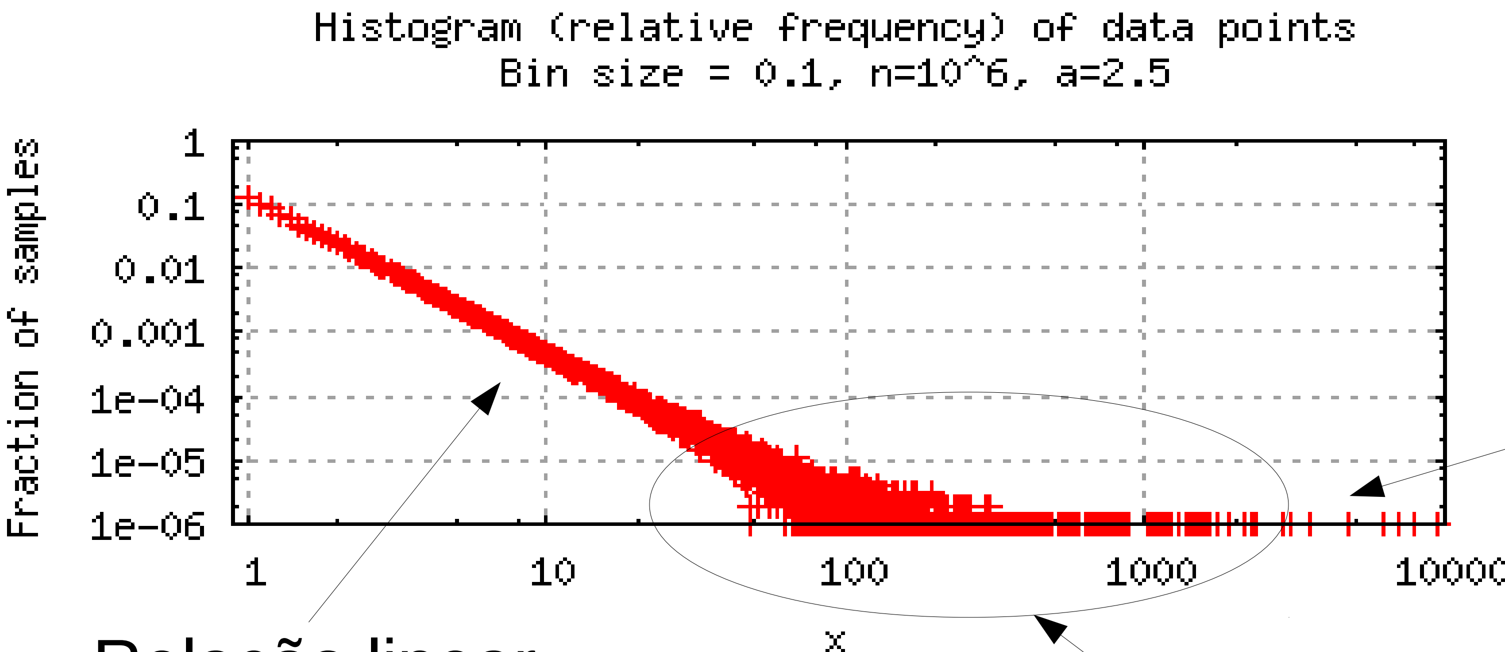


Valores muito grande ocorrem!



- Restringindo o eixo x

Resultados em Log-Log



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

Relação linear começa aparecer

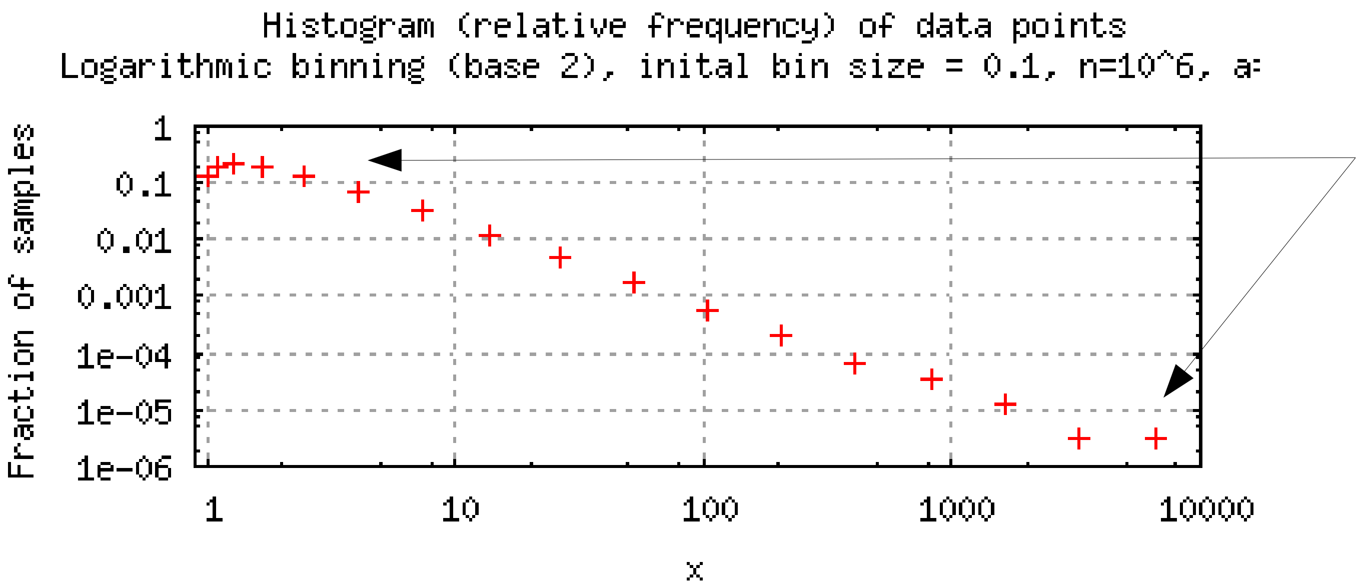
Ruído? Por que? Ignorar?

Outra idéia para visualizar?

Histograma Logarítmico

- **Problema**: intervalos contém poucos pontos quando x é grande
  - intervalo relativamente pequeno (fixo)
- **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

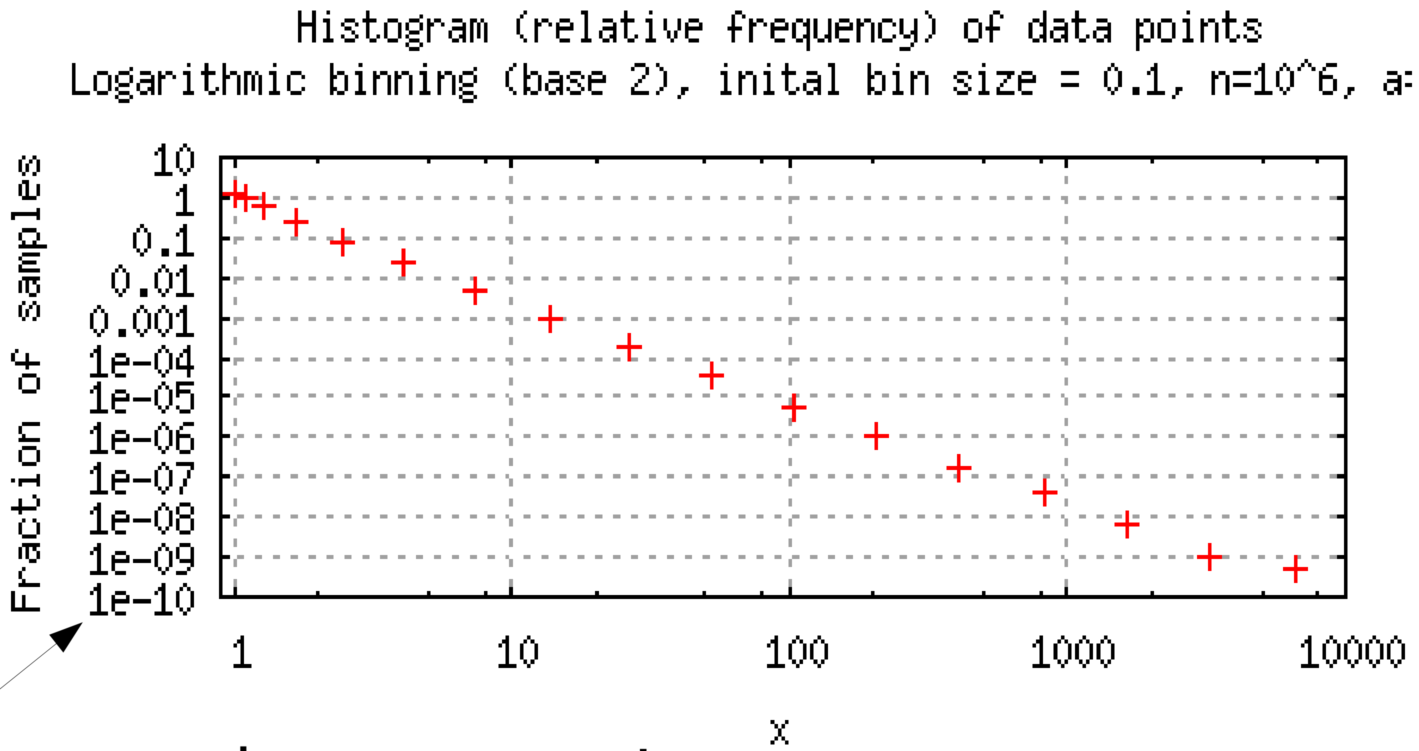


Problema?

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!

- Metodo muito usado pelos físicos
- Como estimar expoente?

Problemas com Histograma

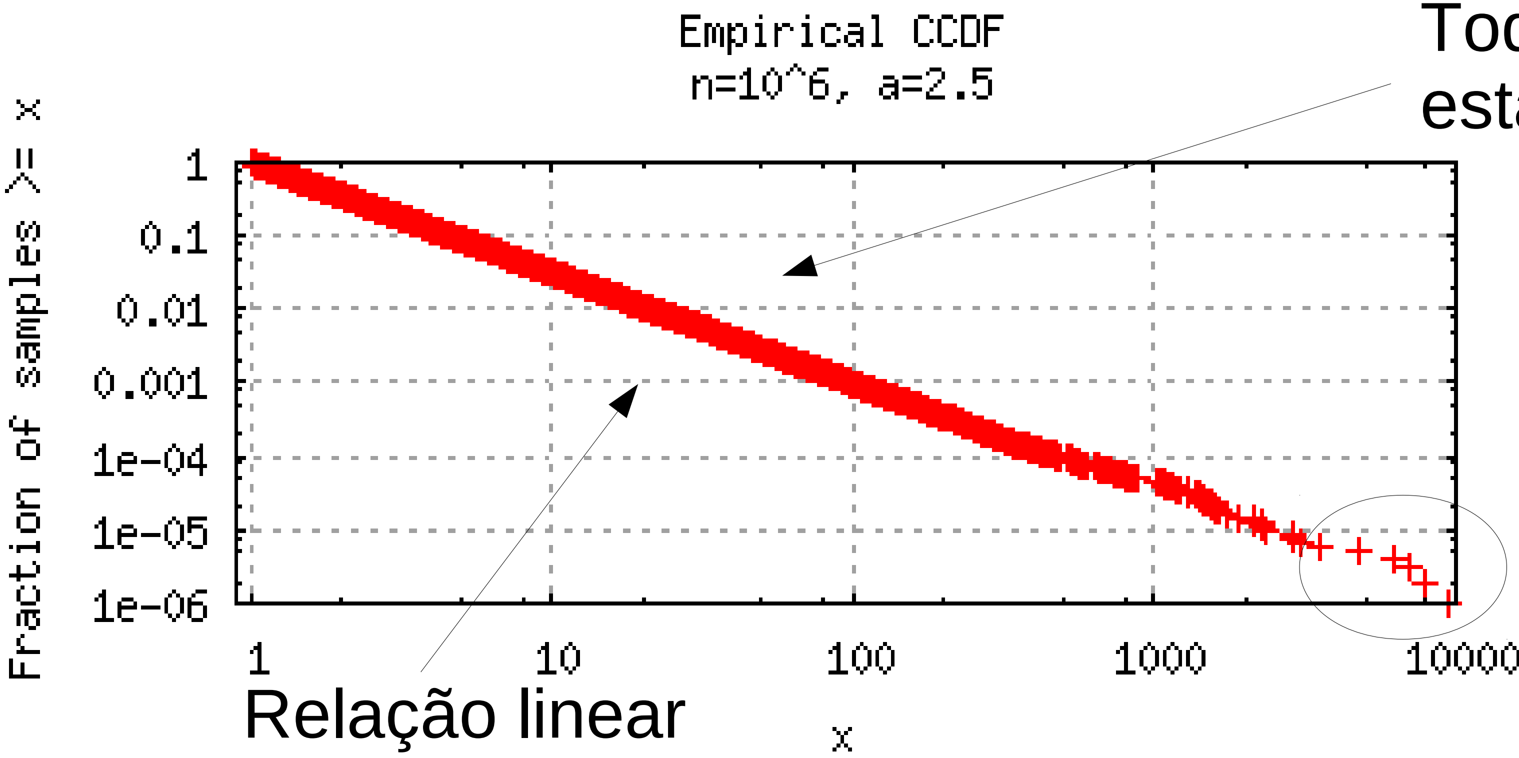
- Determinar tamanho do intervalo inicial
  - e potência, no caso logarítmico
- Valor do intervalo pode influenciar
- Número de amostras por intervalo diminui
  - mesmo no caso logarítmico
- **Agrega informação em intervalos!**
  - trabalha com “média”
- Perde informação das amostras

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 que o primeiro valor, que o 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 p(x)

Relação entre CCDF e PDF

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

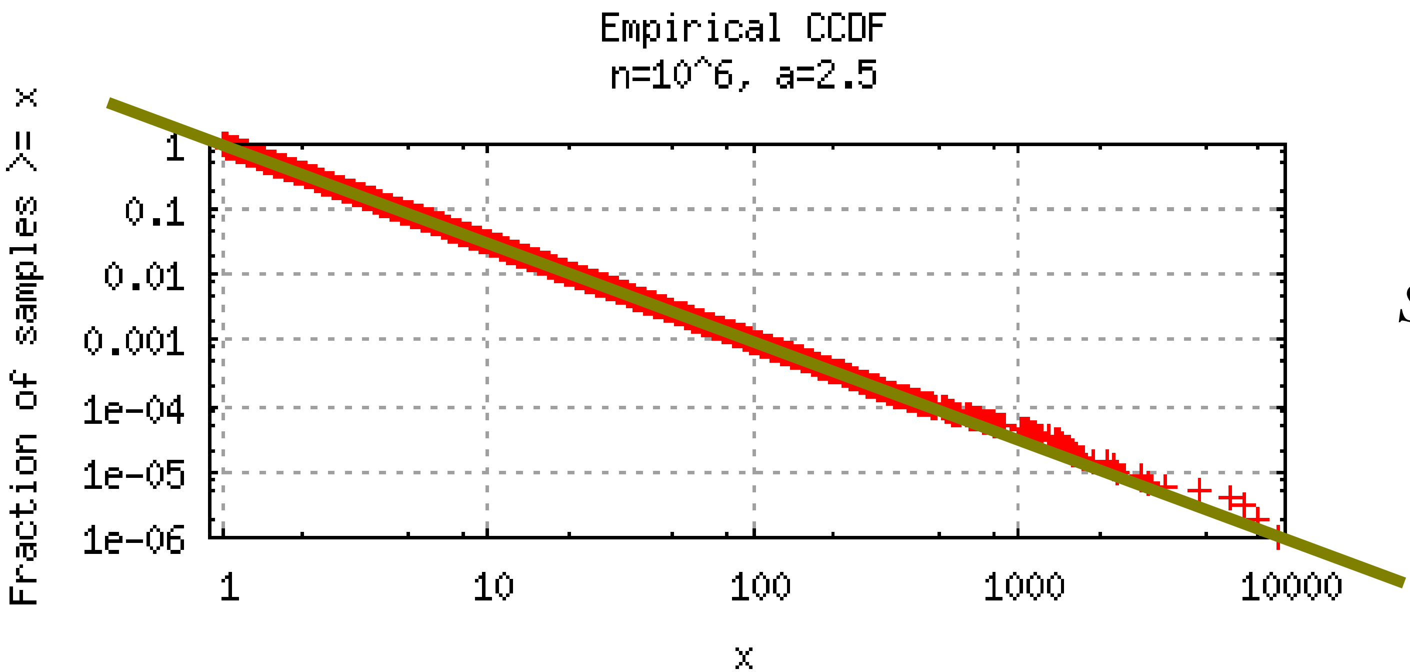
- F(y) : CCDF

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

- 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?

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

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

$L(x_1, \ldots, x_n | a) \longleftarrow$

- 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 maximiza 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[X] e E[X^2] para a v.a. X 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

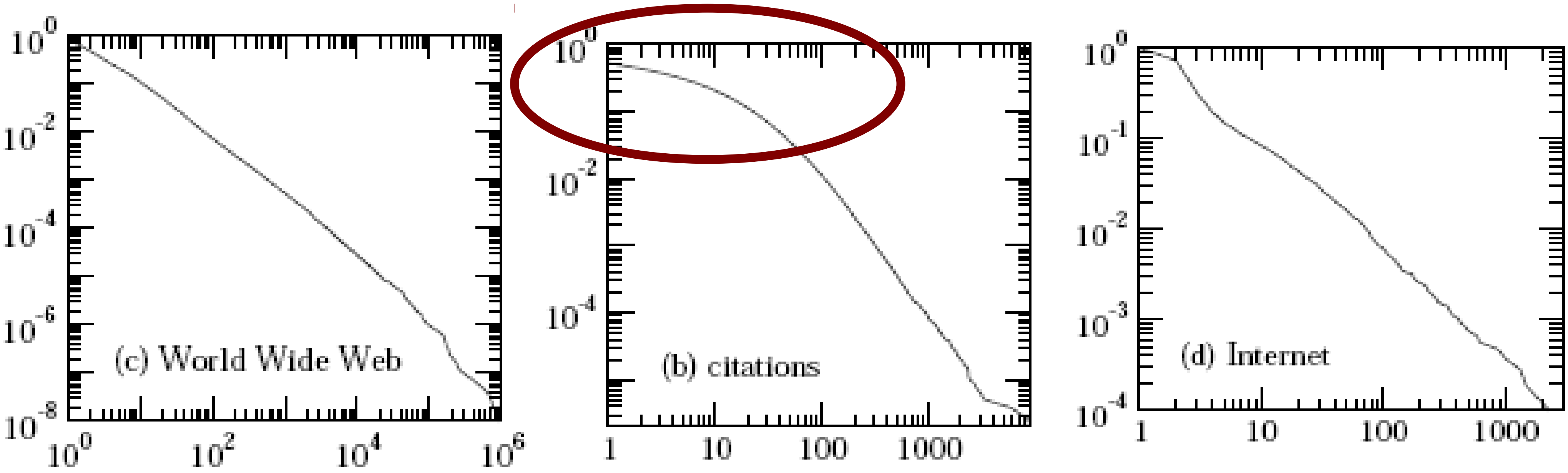
$$\log L(x_1, \ldots, x_n | \hat{a})$$

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

Determinando o Início

- Na prática, distribuição não segue lei de potência sobre todas as escalas
  - mistura de distribuições, no início
- Lei de potência para valores grandes, a partir de certo x_0
  - não vale para x pequeno
- Ignorar valores pequenos, onde distribuição desvia de lei de potência
- **Problema**: determinar x_0
  - onde começa a lei de potência?

Exemplos Reais

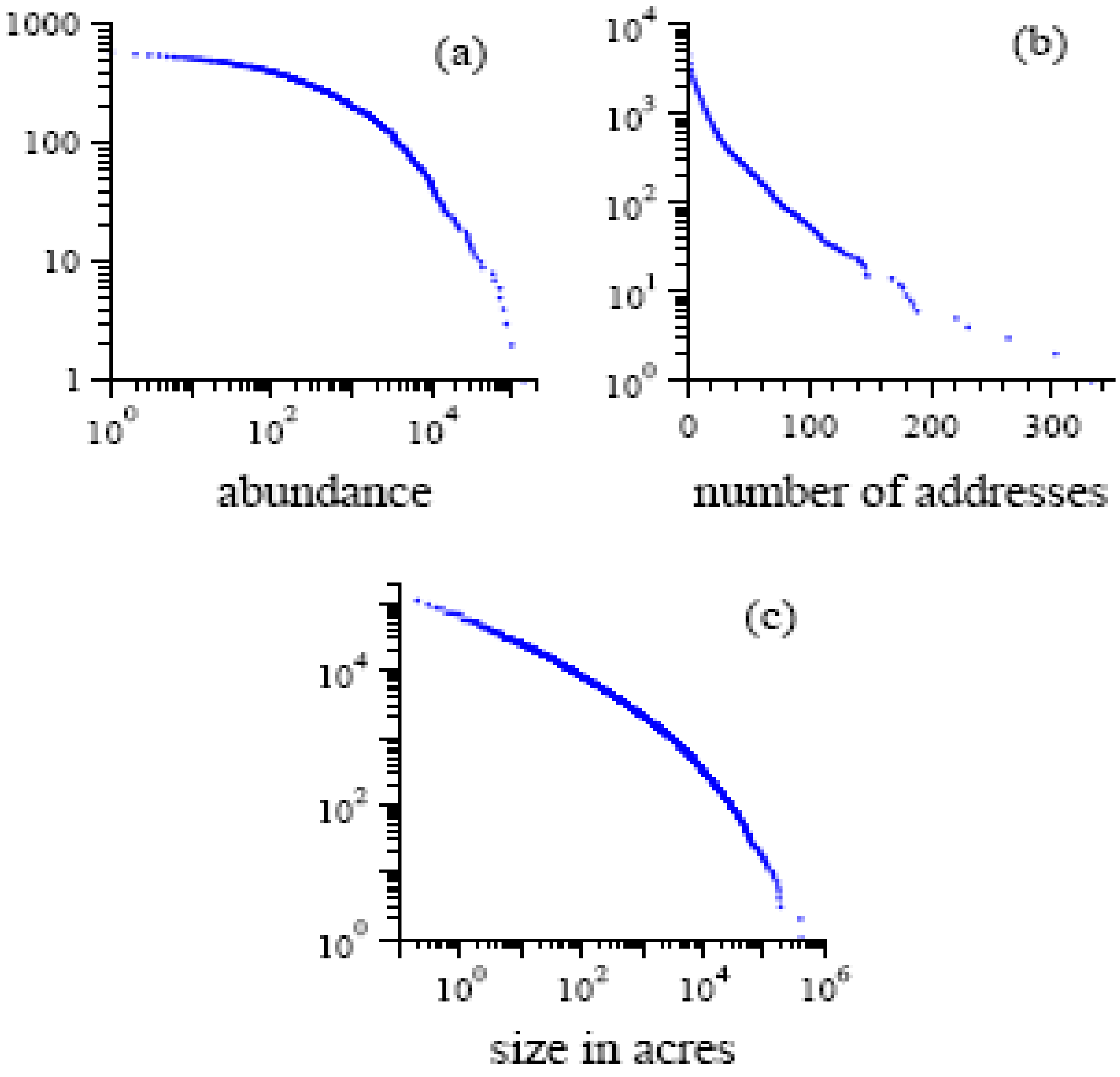


- Alguns casos, distribuição é mistura de funções
- *Cauda* segue lei de potência
  - "Power law tail"
  - x_0 pode ser relativamente grande

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 depende de x_0
  - influência direta nos resultados
- Usar o bom senso!
- Verificar variação do erro é uma boa idéia

Nem tudo é Lei de Potência



- Algumas va. assumem valores grandes
  - longe da média
- Distribuição não segue lei de potência
  - nem na cauda!
- Cuidado ao tirar conclusões!
  - muitos casos são inconclusivos

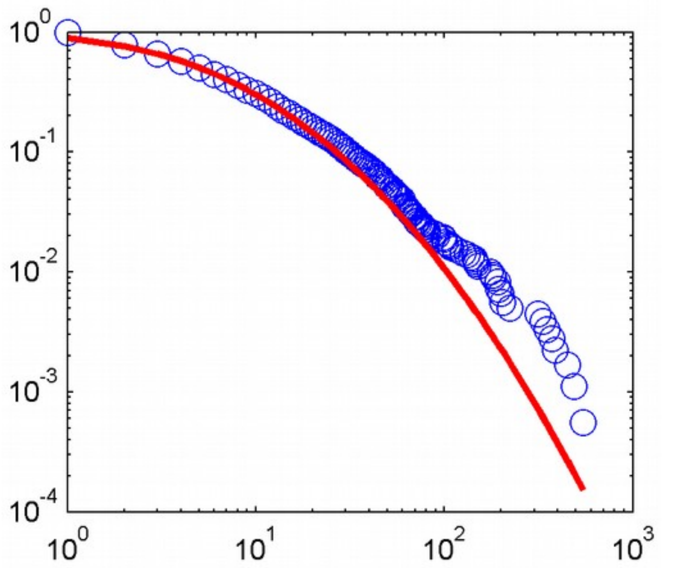
Figueiredo – 2018

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
- Ou seja, X = exp(Y) onde Y tem distribuição Normal
- Dois parâmetros
  - média (μ) e variância (σ^2) da Normal Y
- Densidade

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

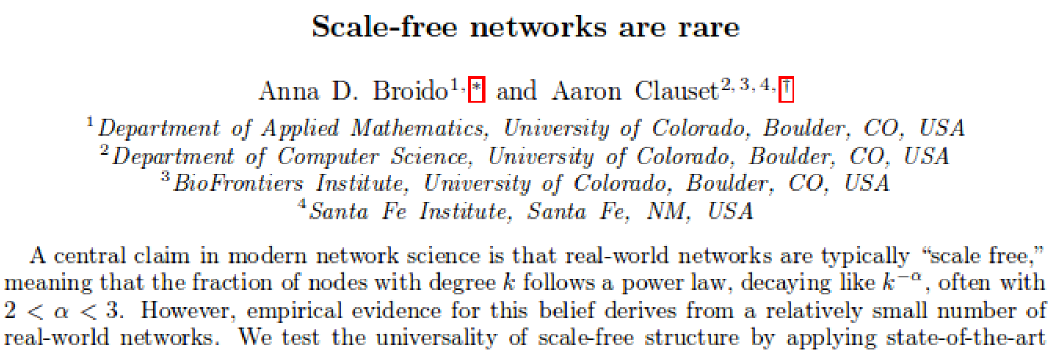
Distribuição Log-Normal



- Cauda pesada
  - Assume 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!

Recente Debate



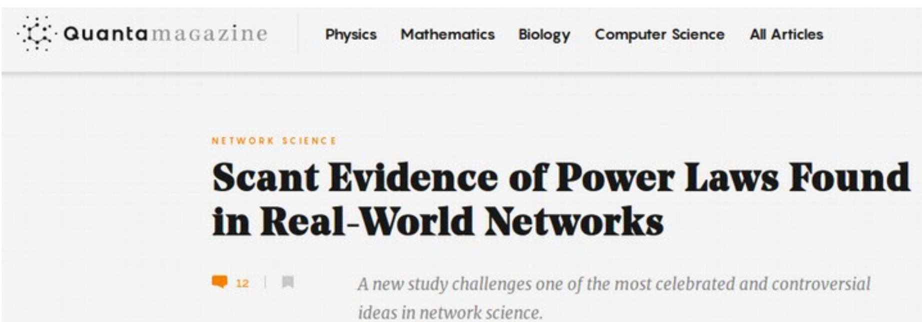
Scale-free networks are rare

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)



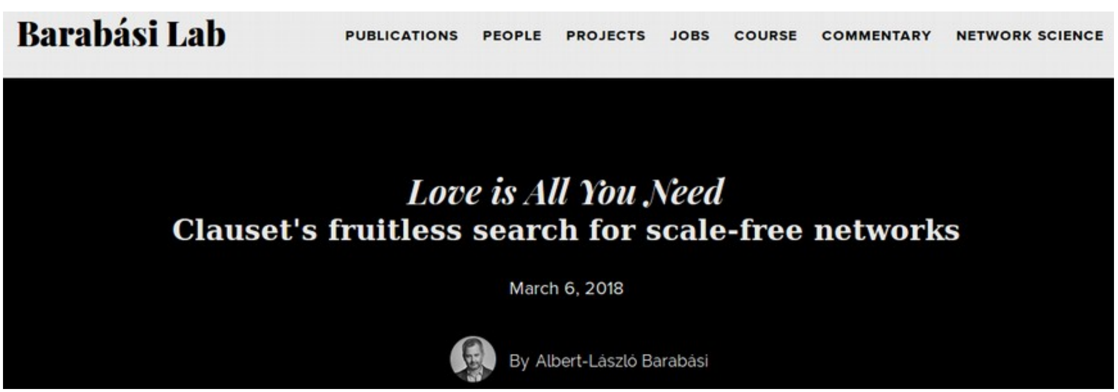
NETWORK SCIENCE

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*



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
- Duras críticas, e bem contundentes

Quem tem razão? E segue o debate!