Power laws

Leonid E. Zhukov

School of Data Analysis and Artificial Intelligence
Department of Computer Science
National Research University Higher School of Economics

Network Science

NATIONAL RESEARCH
UNIVERSITY

Table of contents

Continuous distribution

- Continuous random variable X
- Probability density function $p(x)$ (PDF):

$$
\begin{aligned}
Pr(a \leq X \leq b) &= \int_a^b p(x)dx \\
p(x) &\geq 0 \\
\int_{-\infty}^{\infty} p(x)dx &= 1
\end{aligned}
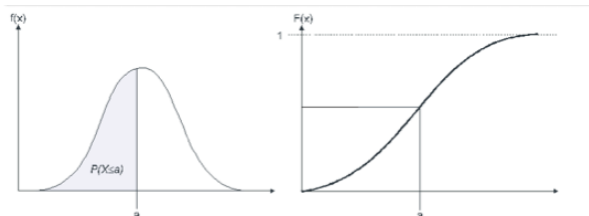$$

- Cumulative distribution function (CDF)

$$F(x) = Pr(X \leq x) = \int_{-\infty}^{x} p(x)dx; \quad \frac{d}{dx}F(x) = p(x)$$

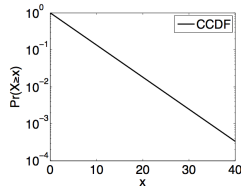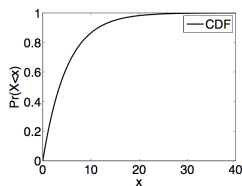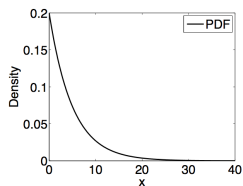- Complementary cumulative distribution function (cCDF)

$$\bar{F}(x) = Pr(X > x) = 1 - F(x) = \int_{x}^{\infty} p(x)dx$$

Continuous distribution

- Gaussian: $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $\quad F(x) = \frac{1}{2}[1 + erf(\frac{x-\mu}{\sigma\sqrt{2}})]$

- Exponential $(x \geq 0)$: $p(x) = \lambda e^{-\lambda x}$, $F(x) = 1 - e^{-\lambda x}$, $\bar{F}(x) = e^{-\lambda x}$

Discrete distribution

- Discrete random variable X_i
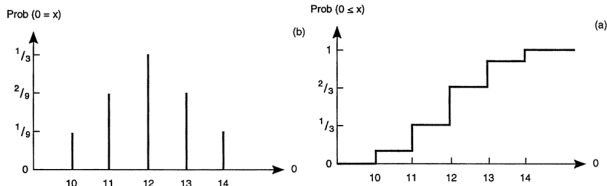- Probability mass function (PMF) $p(x)$:

$$
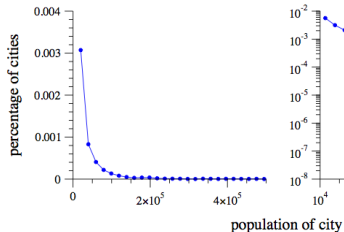\begin{aligned}
p(x) &= Pr(X_i = x) \\
p(x) &\geq 0 \\
\sum_x p(x) &= 1
\end{aligned}
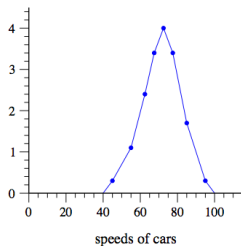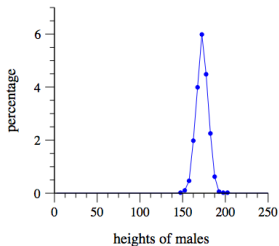$$

- Cumulative distribution function (CDF)

$$
F(x) = Pr(X_i \leq x) = \sum_{x' \leq x} p(x')
$$

Empirical distributions

Newman et.al, 2005

Power Laws

Continuous approximation

- Power law
$$p(x) = Cx^{-\alpha} = \frac{C}{x^{\alpha}}, \quad \text{for } x \geq x_{min}$$
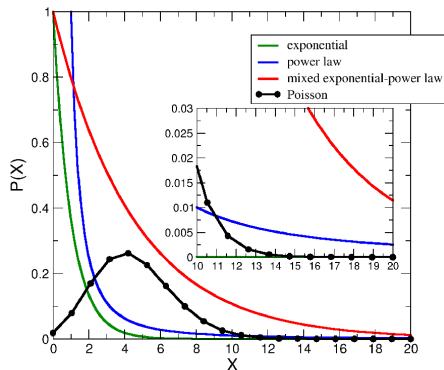
- Normalization ($\alpha > 1$)
$$1 = \int_{x_{min}}^{\infty} p(x)dx = C \int_{x_{min}}^{\infty} \frac{dx}{x^{\alpha}} = \frac{C}{\alpha - 1} x_{\min}^{-\alpha+1}$$

$$C = (\alpha - 1)x_{\min}^{\alpha-1}$$

- Power law PDF
$$p(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\alpha}$$

Power Laws

poisson: $p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$, exponent: $p(x) = Ce^{-\lambda x}$, power law:
$p(x) = Cx^{-\alpha}$
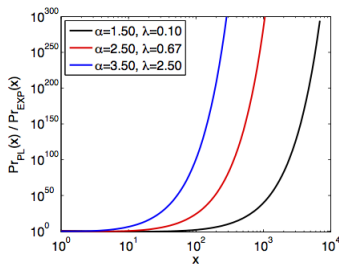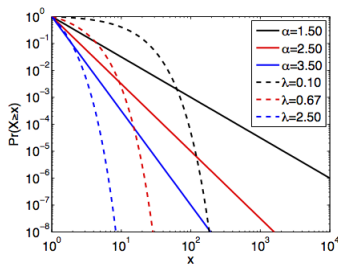
Power Laws

- Power law PDF

$$p(x) = Cx^{-\alpha} = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\alpha}$$

- Complimentary cumulative distribution function cCDF

$$\bar{F}(x) = Pr(X > x) = \int_x^\infty p(x)dx$$

$$\bar{F}(x) = \bar{C}x^{-(\alpha-1)} = \frac{C}{\alpha - 1} x^{-(\alpha-1)} = \left(\frac{x}{x_{min}} \right)^{-(\alpha-1)}$$
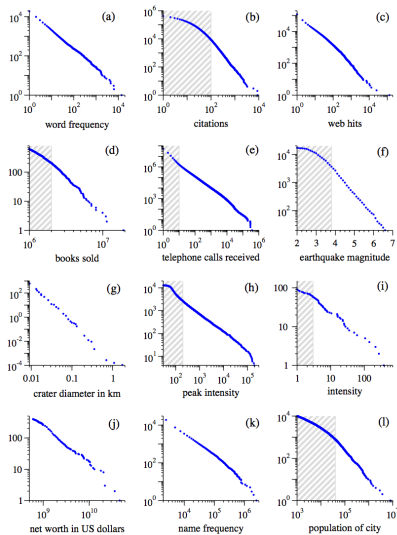
Power Laws

Power law:

$$p(x) = Cx^{-\alpha}, \quad \bar{F}(x) = \bar{C}x^{-(\alpha-1)}$$

$$\log p(x) = \log C - \alpha \log x, \quad \log \bar{F}(x) = \log C - (\alpha - 1) \log x$$

Empirical distributions

log-log scale

Moments

- PDF

$$p(x) = \frac{C}{x^\alpha}, \ x \geq x_{min}$$

- First moment (mean value), $\alpha > 2$:

$$\langle x \rangle = \int_{x_{min}}^{\infty} x p(x) dx = C \int_{x_{min}}^{\infty} \frac{dx}{x^{\alpha-1}} = \frac{\alpha-1}{\alpha-2} x_{min}$$
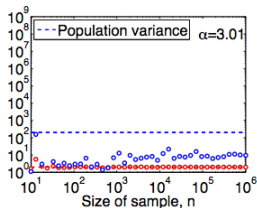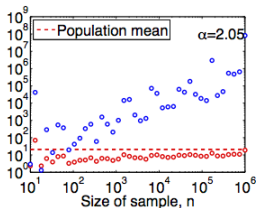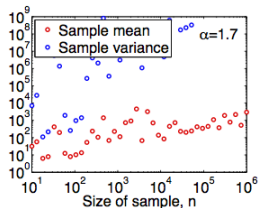
- Second moment, $\alpha > 3$:

$$\langle x^2 \rangle = \int_{x_{min}}^{\infty} x^2 p(x) dx = C \int_{x_{min}}^{\infty} \frac{dx}{x^{\alpha-2}} = \frac{\alpha-1}{\alpha-3} x_{min}^2$$

- k-th moment, $\alpha > k + 1$:

$$\langle x^k \rangle = \frac{\alpha-1}{\alpha-1-k} x_{min}^k$$

Moments

Fisrt moment (mean):

$$\langle x \rangle = C \int_{x_{\min}}^{x_{\max}} \frac{dx}{x^{\alpha-1}} = \frac{\alpha-1}{\alpha-2}\left(x_{\min} - \frac{x_{\min}^{\alpha-1}}{x_{\max}^{\alpha-2}}\right)$$
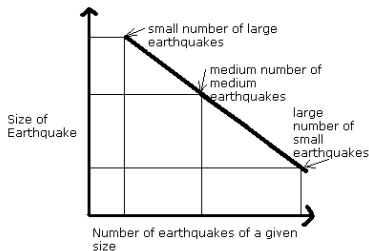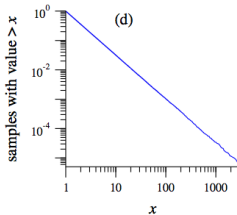
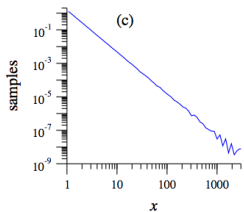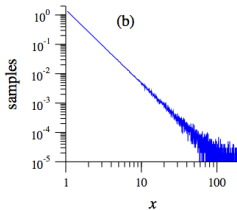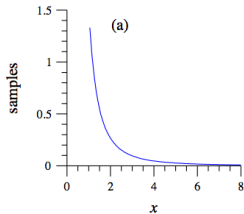Clauset et.al, 2009

Scale invariance

- Scaling of the density

$$x \rightarrow bx, \quad p(bx) = C(bx)^{-\alpha} = b^{-\alpha}Cx^{-\alpha} \propto p(x)$$

- Scale invariance

$$\frac{p(100x)}{p(10x)} = \frac{p(10x)}{p(x)}$$

Power law histograms

Newman et.al, 2005

Scale-free networks
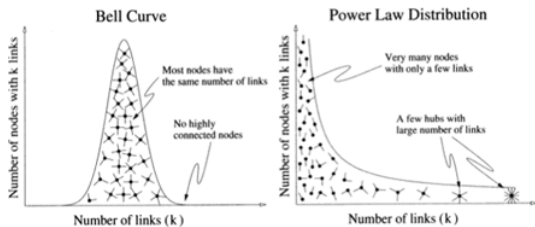
Node degree distribution

- k_i - node degree, i.e. number of nearest neighbors, $k_i = 1, 2, \ldots k_{\max}$
- n_k - number of nodes with degree k, $n_k = \sum_i \mathcal{I}(k_i == k)$
- total number of nodes $n = \sum_k n_k$
- Degree distribution $P(k_i = k) \equiv P(k)$

$$P(k) = \frac{n_k}{\sum_k n_k} = \frac{n_k}{n}$$

- CDF

$$F(k) = \sum_{k' \leq k} P(k') = \frac{1}{n} \sum_{k' \leq k} n_{k'}$$

- cCDF

$$F(k) = 1 - \sum_{k' \leq k} P(k') = \frac{1}{n} \sum_{k' > k} n_{k'}$$

Discrete power law distribution

- Power law distribution

$$P(k) = Ck^{-\gamma} = \frac{C}{k^{\gamma}}$$

- Normalization

$$\sum_{k=1}^{\infty} P(k) = C \sum_{k=1}^{\infty} k^{-\gamma} = C\zeta(\gamma) = 1; \quad C = \frac{1}{\zeta(\gamma)}$$

- Riemann zeta function, $\gamma > 1$
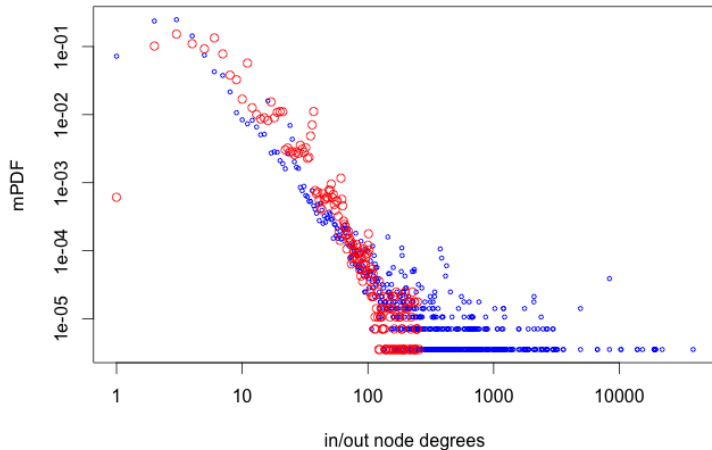
$$P(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}$$

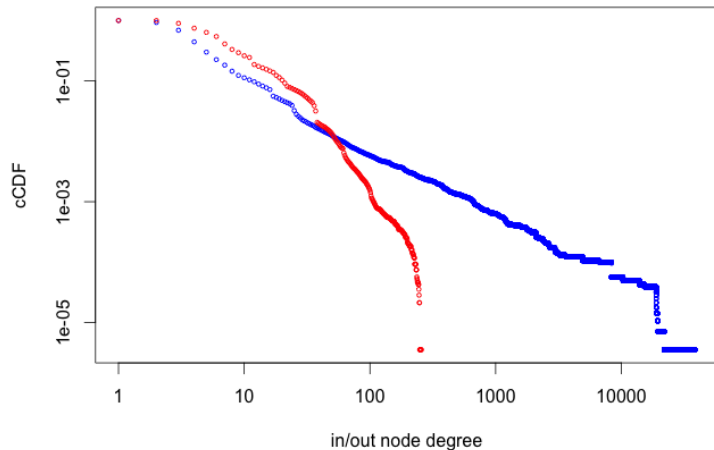- Log-log coordinates

$$\log(P(k)) = -\gamma \log k + \log C$$

Power law networks

Probability mass function PMF/mPDF

Power law networks

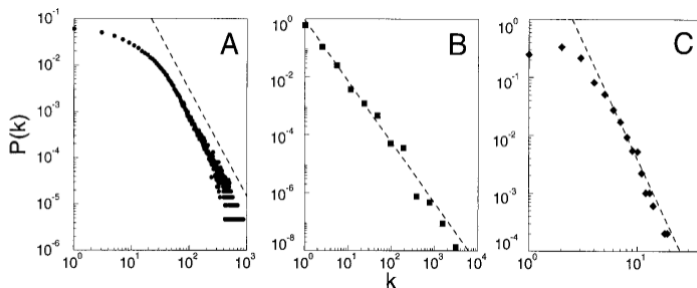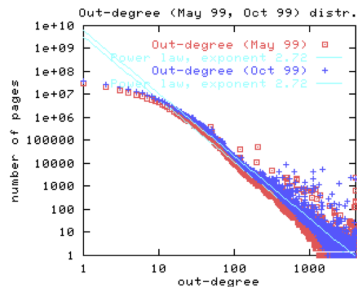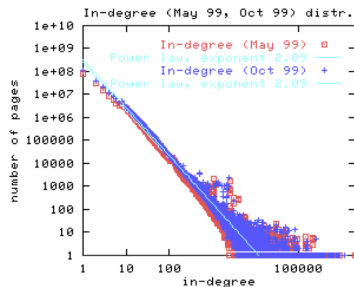Complementary cumulative distribution function cCDF

Power law networks

Actor collaboration graph, N=212,250 nodes, $\langle k \rangle = 28.8$, $\gamma = 2.3$
WWW, N = 325,729 nodes, $\langle k \rangle = 5.6$, $\gamma = 2.1$
Power grid data, N = 4941 nodes, $\langle k \rangle = 5.5$, $\gamma = 4$

Barabasi et.al, 1999

Power law networks

In- and out- degrees of WWW crawl 1999

Broder et.al, 1999

Parameter estimation: α

Maximum likelihood estimation of parameter α

- Let $\{x_i\}$ be a set of n observations (points) independently sampled from the distribution

$$P(x_i) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha}$$

- Probability of the sample

$$P(\{x_i\}|\alpha) = \prod_i^n \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha}$$

- Bayes' theorem

$$P(\alpha|\{x_i\}) = P(\{x_i\}|\alpha) \frac{P(\alpha)}{P(\{x_i\})}$$

Maximum likelihood

- log-likelihood

$$\mathcal{L} = \ln P(\alpha|\{x_i\}) = n \ln(\alpha - 1) - n \ln x_{\min} - \alpha \sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}$$

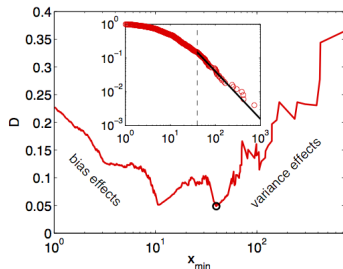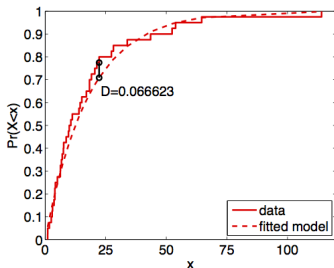- maximization $\frac{\partial \mathcal{L}}{\partial \alpha} = 0$

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

- error estimate

$$\sigma = \sqrt{n} \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right]^{-1} = \frac{\alpha - 1}{\sqrt{n}}$$
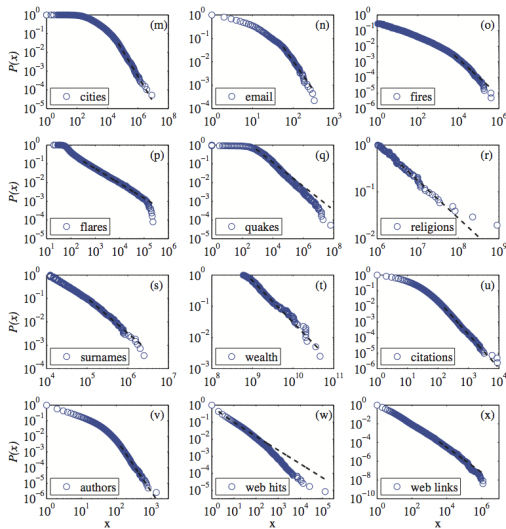
- Kolmogorov-Smirnov test (compare model and experimental CDF)

$$D = \max_x |F(x|\alpha, x_{min}) - F_{exp}(x)|$$

Clauset et.al, 2009

Empirical models

Clauset et.al, 2009

Word counting

Word frequency table (6318 unique words, min freq 800, corpus size $> 85mln$):

```
6187267 the
4239632 be
3093444 of
2687863 and
2186369 a
1924315 in
1620850 to
........
801 incredibly
801 historically
801 decision-making
800 wildly
800 reformer
800 quantum
```
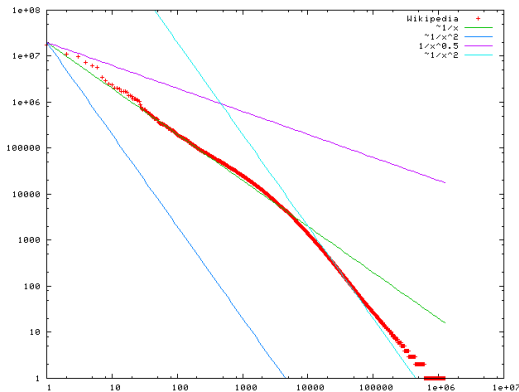
Zips'f law

Zipf's law - the frequency of a word in an natural language corpus is inversely proportional to its rank in the frequency table $f(k) \sim 1/k$.

$$f(k) = \frac{1/k^s}{\sum_{k=1}^{N}(1/k^s)}$$

Rank-frequency plot

- Sort items by their frequency in decreasing order (frequency table)
- Fraction of the words with frequencies higher or equal to the k-th word is cCDF $\bar{F}(k) = Pr(X \geq k)$. The number of the words with frequency above k-th word is its rank k!
- Plot word rank as a function of the word frequency: rank k - y axis, frequency - x axis.
- Use rank-frequency plot instead of computing and plotting cumulative distribution of a quantity.

```
6187267 the
4239632 be
3093444 of
2687863 and
2186369 a
1924315 in
```

Rank-frequency plot

References
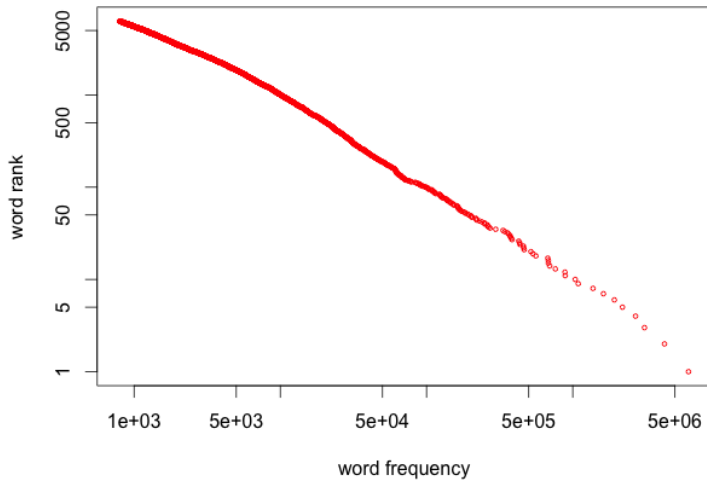
- Power laws, Pareto distributions and Zipfs law, M. E. J. Newman, Contemporary Physics, pages 323351, 2005.
- Power-Law Distribution in Empirical Data, A. Clauset, C.R. Shalizi, M.E.J. Newman, SIAM Review, Vol 51, No 4, pp. 661-703, 2009.
- A Brief History of Generative Models for Power Law and Lognormal Distributions, M. Mitzenmacher, Internet Mathematics Vol 1, No 2, pp 226-251.