Plotting Power-laws and the Degree Exponent

Presenter: Xiaoyan Lu

RPI

2015

Overview

- Plotting Power-Laws
- Estimating the Degree Exponent

Plotting Degree Distribution

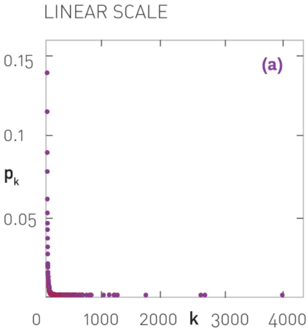
From N_k, the number of nodes with degree k, we calculate

$$p_k = \frac{N_k}{N}$$

Using linear k-axis to plot the distribution?

Linear-Linear Scale

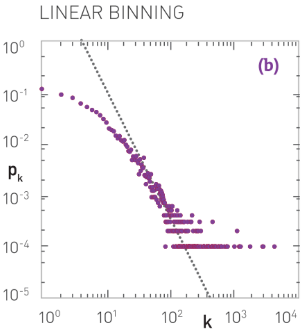
A degree distribution of the form $p_k \sim (k + k_0)^{-\gamma}$, with $k_0 = 10$ and $\gamma = 2.5$



It is impossible to see the distribution on a lin-lin scale.
$\Rightarrow$ log-log plot for scale-free networks.

Log-Log Scale, Linear Binning

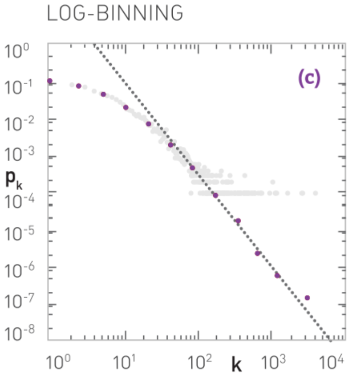
In the high-k region, we either have $N_k = 0$ (no node with degree k) or $N_k = 1$ (a single node with degree k).



linear binning, each bin has the same size $\Delta k = 1$.

Log-Log Scale, Log Binning

n-th bin contains all nodes with degrees from 2^{n-1} to $2^n - 1$.



$$p_{<k_n>} = \frac{N_n}{b_n}$$

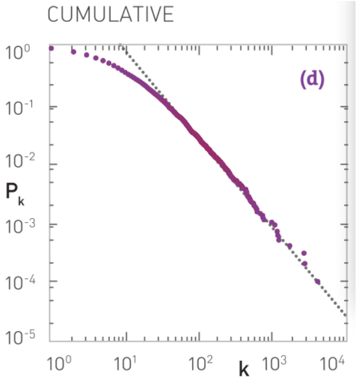
- N_n is number of nodes in bin n
- $< k_n >$ is average degree of nodes in bin n

With log-binning the plateau disappears. Show linear binning as light grey the data.

Log-Log Scale, Cumulative Binning

Plot the complementary cumulative distribution $P_k = \sum_{q=k+1}^{\infty} p_q$
If $p_k \sim k^{-\gamma}$, then

$$P_k \sim k^{-\gamma+1}$$



Estimating the Degree Exponent

Power-law distribution obeys

$$\ln p_k = -\gamma \ln k + constant$$

We need to determine the value of γ.

Estimating the Degree Exponent

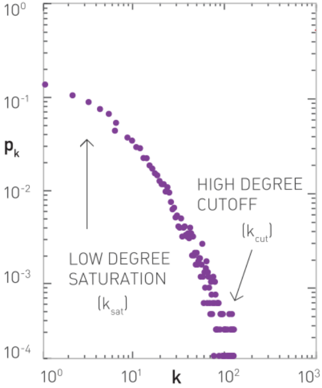
Power-law distribution obeys

$$\ln p_k = -\gamma \ln k + constant$$

We need to determine the value of γ.

However,

- Low-degree saturation: fewer small degree nodes than expected
- High-degree cutoff: fewer high-degree nodes than expected



Maximum Likelihood Estimate

Assume data follows a power law exactly for $k \geq K_{min}$
MLE maximizes the probability generating observed data $\{k_i\}$ with the provided model:

$$\max \prod_{i=1}^{n} p(k_i|\gamma) = \max \prod_{i=1}^{n} C p_{k_i}^{\gamma}$$

Maximum Likelihood Estimate

Assume data follows a power law exactly for $k \geq K_{min}$
MLE maximizes the probability generating observed data $\{k_i\}$ with the provided model:

$$\max \prod_{i=1}^{n} p(k_i|\gamma) = \max \prod_{i=1}^{n} Cp_{k_i}^{\gamma}$$

The maximum likelihood estimate for the scaling parameter:

$$\gamma = 1 + N \Big[\sum_{i=1}^{N} ln \frac{k_i}{K_{min} - \frac{1}{2}} \Big]^{-1}$$

K_{min} is the minimum value at which power-law behavior holds.

Maximum Likelihood Estimate

Assume data follows a power law exactly for $k \geq K_{min}$
MLE maximizes the probability generating observed data $\{k_i\}$ with the provided model:

$$\max \prod_{i=1}^{n} p(k_i|\gamma) = \max \prod_{i=1}^{n} Cp_{k_i}^{\gamma}$$

The maximum likelihood estimate for the scaling parameter:

$$\gamma = 1 + N \left[\sum_{i=1}^{N} ln \frac{k_i}{K_{min} - \frac{1}{2}} \right]^{-1}$$

K_{min} is the minimum value at which power-law behavior holds.

How to find best K_{min}?

Kolmogorov-Smirnov Test (KS statistic)

Quantifying the distance between two probability distributions:

$$D = max_{k \geq K_{min}} |S(k) - P_k|$$

- $S(k)$: Cumulative Distribution Function (CDF) of the data
- P_k: CDF provided by the fitted model

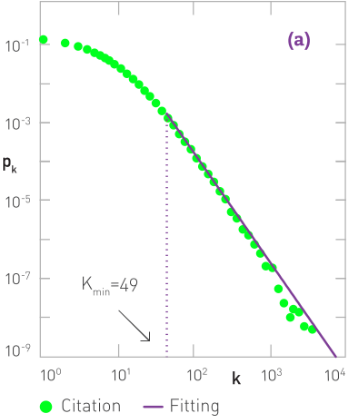
$$P_k = 1 - \sum_{x=k}^{\infty} Cx^{-\gamma}$$

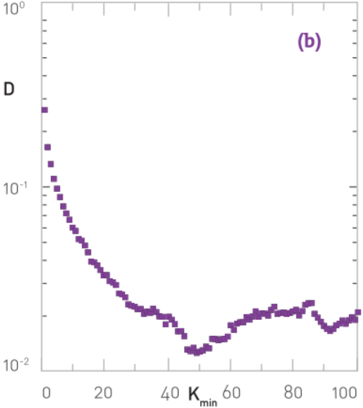
Scanning the whole K_{min} range from k_{min} to k_{max} to identify the K_{min} value for minimal D.

Citation Network

2,353,984 citations among 384,362 research papers published in journals published by the American Physical Society

- γ = 2.79
- K_{min} = 49





Goodness-of-Fit Test

Is power law itself is a good model for the studied distribution?

Goodness-of-Fit Test

Is power law itself is a good model for the studied distribution?

Hypothesis : power law is the right model

p-value

- 1. Obtain D^{real}, KS distance between the real data and the best fit

p-value

- 1. Obtain D^{real}, KS distance between the real data and the best fit
- 2. Generate a sequence of degrees according to obtained power law, and obtain $D^{synthetic}$ for this hypothetical degree sequence

p-value

- 1. Obtain D^{real}, KS distance between the real data and the best fit
- 2. Generate a sequence of degrees according to obtained power law, and obtain $D^{synthetic}$ for this hypothetical degree sequence
- 3. Repeat step2 M times ($M >> 1$) and obtain the distribution of $D^{synthetic}$

p-value

- 1. Obtain D^{real}, KS distance between the real data and the best fit
- 2. Generate a sequence of degrees according to obtained power law, and obtain $D^{synthetic}$ for this hypothetical degree sequence
- 3. Repeat step2 M times ($M >> 1$) and obtain the distribution of $D^{synthetic}$
- 4. p-value is

$$p = \int_{D^{real}}^{\infty} p(D^{synthetic}) dD^{synthetic}$$

p-value

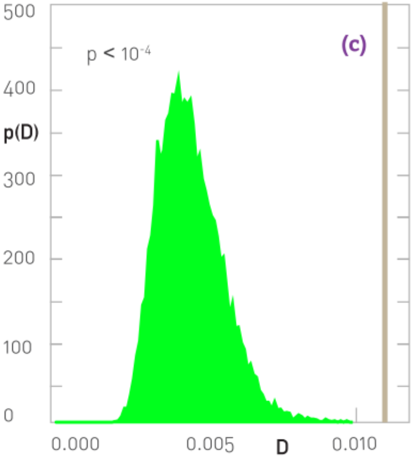
- 1. Obtain D^{real}, KS distance between the real data and the best fit
- 2. Generate a sequence of degrees according to obtained power law, and obtain $D^{synthetic}$ for this hypothetical degree sequence
- 3. Repeat step2 M times ($M >> 1$) and obtain the distribution of $D^{synthetic}$
- 4. p-value is

$$p = \int_{D^{real}}^{\infty} p(D^{synthetic}) dD^{synthetic}$$

If p is very small, the model is not a plausible fit to the data.

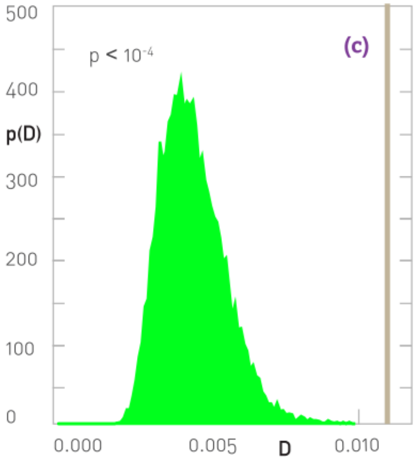
Goodness-of-Fit in Citation Network

For the citation network, $p < 10^{-4}$



Goodness-of-Fit in Citation Network

For the citation network, $p < 10^{-4}$



Choosing K_{min} = 49 forces us to discard over 96% of the data points

Not a Pure Power Law

Use a function that offers a better fit

Not a Pure Power Law

Use a function that offers a better fit

Degree distribution of many real networks, like the citation network, does not follow a pure power law

$$p_k = \frac{1}{\sum_{k'=1}^{\infty}(k' + k_{sat})^{-\gamma} e^{-k'/k_{cut}}}(k + k_{sat})^{-\gamma} e^{-k/k_{cut}}$$

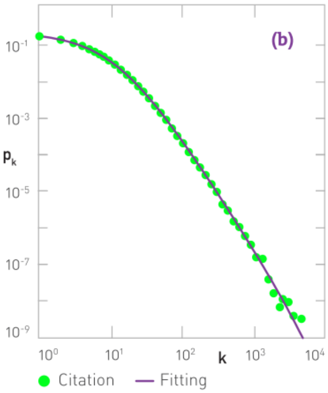
- k_{sat} low-k saturation
- k_{cut} large-k cutoff

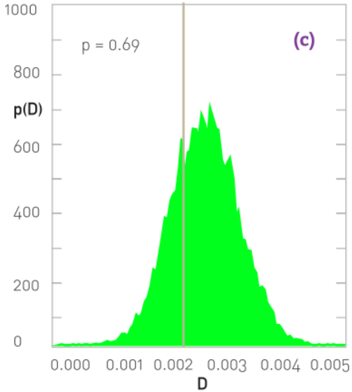
Procedure:

- 1. For fixed (k_{sat}, k_{cut}), estimate γ using MLE
- 2. Identify (k_{sat}, k_{cut}) values for which KS distance is minimal.
- 3. Obtain p-value using the Goodness-of-Fit test

Optimal Fit of Citation Network

$k_{sat} = 12, k_{cut} = 5,691, \gamma = 3.028$, p-value = 69%





Systematic Fitting Issues

- 1. A pure power law is an idealized distribution. In reality, if p_k does not follow a pure power law, methods described above will inevitably fail to detect statistical significance.

Systematic Fitting Issues

- 1. A pure power law is an idealized distribution. In reality, if p_k does not follow a pure power law, methods described above will inevitably fail to detect statistical significance.
- 2. Kolmogorov-Smirnov criteria: a single point deviating from the curve will affect the fit's statistical significance.

Systematic Fitting Issues

- 1. A pure power law is an idealized distribution. In reality, if p_k does not follow a pure power law, methods described above will inevitably fail to detect statistical significance.
- 2. Kolmogorov-Smirnov criteria: a single point deviating from the curve will affect the fit's statistical significance.
- 3. Remove a huge fraction of the nodes to obtain a statistically significant fit ?

Thanks

Thanks