The degree distribution

Ramon Ferrer-i-Cancho & Argimiro Arratia

Universitat Politècnica de Catalunya

Version 0.4
Complex and Social Networks (2016-2017)
Master in Innovation and Research in Informatics (MIRI)

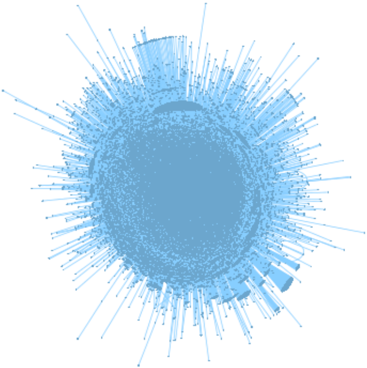
Official website: `www.cs.upc.edu/~csn/`

Contact:

- Ramon Ferrer-i-Cancho, rferrericancho@cs.upc.edu, `http://www.cs.upc.edu/~rferrericancho/`
- Argimiro Arratia, argimiro@cs.upc.edu, `http://www.cs.upc.edu/~argimiro/`

The limits of visual analysis

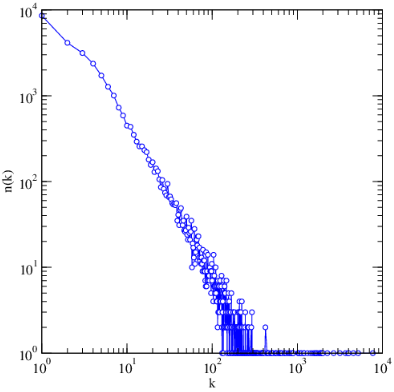
A syntactic dependency network [Ferrer-i-Cancho et al., 2004]



The empirical degree distribution

- N: finite number of vertices / k vertex degree
- $n(k)$: number of vertices of degree k.
- $n(1),n(2),...,n(N)$ defines the *degree spectrum* (loops are allowed).
- $n(k)/N$: the proportion of vertices of degree k, which defines the *(empirical) degree distribution*.
- $p(k)$: function giving the probability that a vertex has degree k, $p(k) \approx n(k)/N$.
- $p(k)$: probability mass function (pmf).

Example: degree spectrum

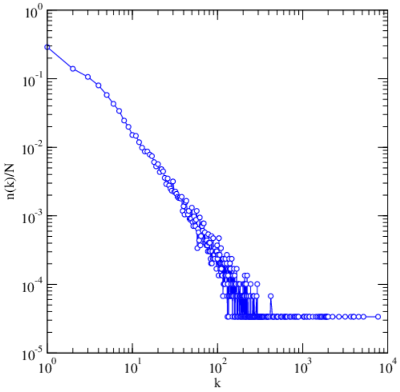


- Global syntactic dependency network (English)
- Nodes: words
- Links: syntactic dependencies

Not as simple:

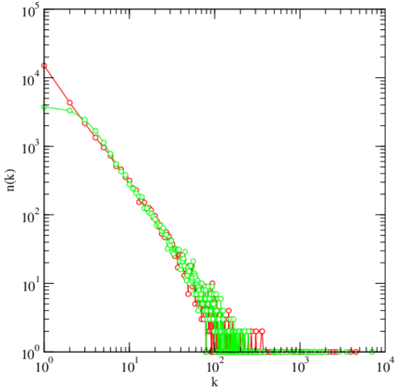
- Many degrees occurring just once!
- Initial bending or hump: power-law?

Example: empirical degree distribution



- Notice the scale of the y-axis.
- Normalized version of the degree spectrum (dividing over N).

Example: in-degree (red) degree versus out-degree (green)



- The distribution of in-degree and that of out-degree do not need to be identical!
- Similar for global syntactic dependency networks? Differences in the distribution or the parameters?
- Known cases of radical differences between in and out-degree distributions (e.g., web pages, wikipedia articles). **In-degree more power-law like than out degree.**

What is the mathematical form of $p(k)$?

Possible degree distributions

- The typical hypothesis: a *power-law* $p(k) = ck^{-\gamma}$ but what exactly? How many free parameters?
  - Zeta distribution: 1 free parameter.
  - Right-truncated zeta distribution: 2 free parameters.
  - ...

Motivation:

- Accurate data description (looks are deceiving).
- Help to design or select dynamical models.

Zeta distributions I

Zeta distribution:

$$p(k) = \frac{1}{\zeta(\gamma)} k^{-\gamma},$$

being

$$\zeta(\gamma) = \sum_{x=1}^{\infty} x^{-\gamma}$$

the Riemann zeta function.

- (here it is assumed that γ is real) $\zeta(\gamma)$ converges only for $\gamma > 1$ ($\gamma > 1$ is needed).
- γ is the only free parameter!
- Do we wish $p(k) > 0$ for $k > N$?

Zeta distributions I

Right-truncated zeta distribution

$$p(k) = \frac{1}{H(k_{max}, \gamma)} k^{-\gamma},$$

being

$$H(k_{max}, \gamma) = \sum_{x=1}^{k_{max}} x^{-\gamma}$$

the generalized harmonic number of order k_{max} of γ.
Or why not

$$p(k) = ck^{-\gamma} e^{-k\beta}$$

(modified power-law, Altmann distribution,...) with 2 or 3 free parameters?
Which one is best? (standard model selection)

What is the mathematical form of $p(k)$?

Possible degree distributions

- The null hypothesis (for a Erdös-Rényi graph without loops)

$$p(k) = \binom{N-1}{k} \pi^k (1-\pi)^{N-1-k}$$

  with π as the only free parameter (assuming that N is given by the real network).
  Binomial distribution with parameters $N-1$ and π, thus $\langle k \rangle = (N-1)\pi \approx N\pi$.
- Another null hypothesis: random pairing of vertices with constant number of edges E.

The problems II

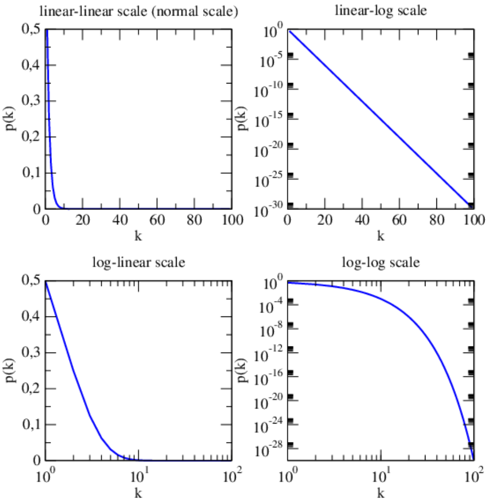
- Is $f(k)$, a good candidate? Does $f(k)$ fit the empirical degree distribution well enough?
- $f(k)$ is a (candidate) model.
- How do we evaluate goodness of a model? Three major approaches:
  - Qualitatively (visually).
  - The error of the model: the deviation between the model and the data.
  - The likelihood of the model: the probability that the model produces the data.

Visual fitting

Assume a two variables: a predictor x (e.g., k, vertex degree) and a response y (e.g., $n(k)$, the number vertices of degree k; or $p(k)$...).

- Look for a transformation of the at least one of the variables showing approximately a straight line (upon visual inspection) and obtain the dependency between the two original variables.
- Typical transformations: $x' = log(x)$, $y' = log(y)$.
  1. If $y' = log(y) = ax + b$ (linear-log scale) then $y = e^{ax+b} = ce^{ax}$, with $c = e^b$ (exponential).
  2. If $y' = log(y) = ax' + b = alog(x) + b$ (log-log scale) then $y = e^{alog(x)+b} = cx^a$, with $c = e^b$ (*power-law*).
  3. If $y = ax' + b = alog(x) + b$ (log-linear scale) then the transformation is exactly the functional dependency between the original variables (logarithmic).

What is this distribution?



Solution: geometric distribution

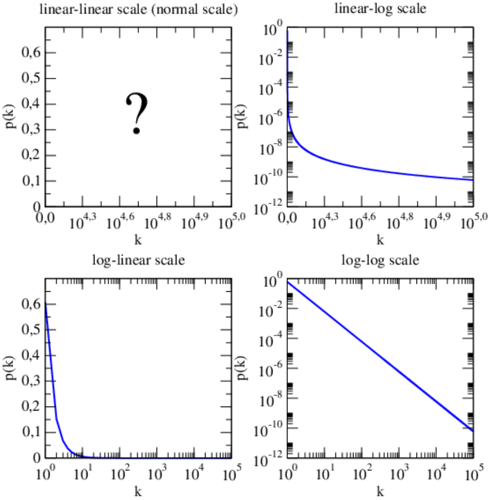
$y = (1-p)^{x-1}p$ (with $p = 1/2$ in this case).
In standard exponential form,

$$\begin{aligned} y &= (1-p)^x \frac{p}{1-p} = e^{x\log(1-p)} \frac{p}{1-p} \\ &= ce^{ax} \end{aligned}$$

with $a = \log(1-p)$ and $c = p/(1-p)$.
Examples:

- Random network models (degree is geometrically distributed).
- Distribution of word lengths in random typing (empty words are not allowed) [Miller, 1957].
- Distribution of projection lengths in real neural networks [Ercsey-Ravasz et al., 2013].

A *power-law distribution*



What is the exponent of the power-law?

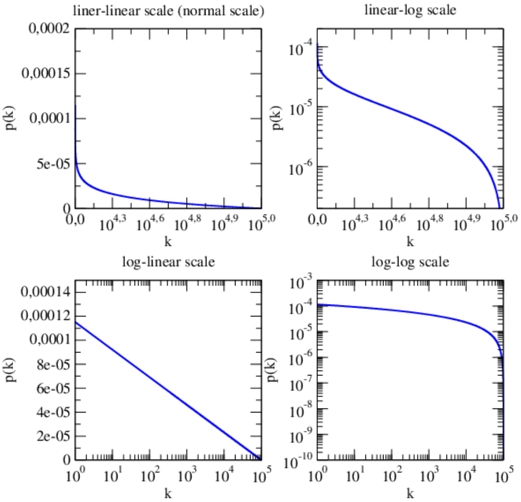
Solution: zeta distribution

$$y = \frac{1}{\zeta(a)} x^{-a}$$

with $a = 2$.
Formula for $\zeta(a)$ is known for certain integer values, e.g., $\zeta(2) = \pi^2/6 \approx 1.645$.
Examples:

- Empirical degree distribution of global syntactic dependency networks [Ferrer-i-Cancho et al., 2004] (but see also lab session on degree distributions).
- Frequency spectrum of words in texts [Corral et al., 2015].

What is this distribution?



Solution: a "logarithmic" distribution

$$y = c(log(x_{max}) - \log x))$$

with $x = 1, 2, ..., x_{max}$ and c being a normalization term, i.e.

$$c = \frac{1}{\sum_{x=1}^{x_{max}} (log(x_{max}) - \log x))}$$

.

The problems of visual fitting

- The right transformation to show linearity might not be obvious (taking logs is just one possibility).
- Looks can be deceiving with noisy data.
- A good guess or strong support for the hypothesis requires various decades.
- Solution: a quantitative approach.

Non-linear regression I [Ritz and Streibig, 2008]

- A univariate response y.
- A predictor variable x
- Goal: functional dependency between y and x.

Formally: $y = f(x, \beta)$, where

- $f(x, \beta)$ is the "model".
- K parameters.
- $\beta = (\beta_1, ..., \beta_K)$

Examples:

- Linear model: $f(x, (a, b)) = ax + b$ $(K = 2)$.
- A non-linear model (power-law): $f(x, (a, b)) = ax^b$ $(K = 2)$.

Non-linear regression II

Problem of regression:

- A data set of n pairs: $(x_1, y_1), ..., (x_n, y_n)$. Example: x_i is vertex degree (k) and y_i is the number of vertices of degree k ($n(k)$) of a real network.
- n is the sample size.
- $f(x, \beta)$ is unlikely to give a perfect fit. $y_1, y_2, ..., y_n$ may contain error.

Solution: the conditional mean response

$$E(y_i|x_i) = f(x_i, \beta)$$

($f(x, \beta)$ is not actually the model for the data points but a model for expectation given x_i).

Non-linear regression II

The full model is then

$$y_i = E(y_i|x_i) + \epsilon_i = f(x_i, \beta) + \epsilon$$

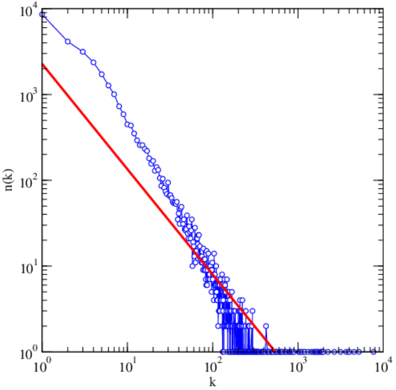
The quality of the fit of a model with certain parameters: the residual sums of squares

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2$$

The parameters of the model are estimated minimizing the RSS.
Non-linear regression: minimization of RSS.
Common metric of the quality of the fit: the residual standard error

$$s^2 = \frac{RSS(\beta)}{n-K}$$

Example of non-linear regression



- Non-linear regression yields $y = 2273.8x^{-1.23}$ (is the exponent that low?)
- Is the method robust? (=not distracted by undersampling, noise, and so on)
- Likely and unlikely events are weighted equally.
- Solution: weighted regression, taking likelihood into account,...

Likelihood I [Burnham and Anderson, 2002]

- A probabilistic metric of the quality of the fit.
- $L(parameters|data, model)$: likelihood of the parameters given the data (sample of size n) and a model.
  Example: $L(\gamma|data, \text{Zeta distribution with parameter}\gamma)$
- Best parameters: the parameters that maximize $L(parameters|data, model)$.

Likelihood II

- Consider a sample $x_1, x_2, ...x_n$ (e.g., the degree sequence of a network).
- Definition (assuming independence)

$$L(parameters|data, model) = \prod_{i=1} p(x_i; parameters)$$

- For a zeta distribution

$$\begin{aligned} L(\gamma|x_1, x_2, .., x_n; \text{Zeta distribution}) &= \prod_{i=1}^{n} p(x_i; \gamma) \\ &= \zeta(\gamma)^{-n} \prod_{i=1}^{n} x_i^{-\gamma} \end{aligned}$$

Log-likelihood

Likelihood is a vanishingly small number. Solution: taking logs.

$$\begin{aligned}\mathcal{L}(parameters|data, model) &= \log L(parameters|data, model) \\ &= \sum_{i=1} \log p(x_i; parameters)\end{aligned}$$

Example:

$$\begin{aligned}\mathcal{L}(\gamma|x_1, x_2, .., x_n; \text{Zeta distribution}) &= \sum_{i=1}^{n} \log p(x_i; \gamma) \\ &= \gamma \sum_{i=1}^{n} \log x_i - n \log(\zeta(\gamma))\end{aligned}$$

Question to the audience

What is the best model for data?

Cue: a universal method.

What is the best model for data?

- The best model of the data is the data itself. Overfitting!
- The quality of the fit cannot decrease if more parameters are added (wisely). Indeed, the quality of the fit normally increases when adding parameters.
- The metaphor of picture compression. Compressing a picture (with quality reduction). A good compression technique shows a nice trade-off between file size and image quality).
- Modelling is compressing a sample, the empirical distribution (e.g., compressing the degree sequence of a network).
  - Models with many parameters should be penalized!
  - Models compressing the data with a low quality should be also penalized.

How?

Akaike's information criterion (AIC)

$$AIC = -2\mathcal{L} + 2K,$$

with K being the number of parameters of the model. For small samples, a correction is necessary

$$AIC_c = -2\mathcal{L} + 2K\left(\frac{n}{n-K-1}\right),$$

or equivalently

$$\begin{aligned} AIC_c &= -2\mathcal{L} + 2K + \frac{2K(K+1)}{n-K-1} \\ &= AIC + \left(\frac{2K(K+1)}{n-K-1}\right) \end{aligned}$$

AIC_c is recommended if $n \gg K$ is not satisfied!

Model selection with AIC

- What is the best of a set of models? The model that minimizes AIC
- AIC_{best}: the AIC of the model with smallest AIC.
- Δ: "AIC difference", the difference between the AIC of the model and that of the best model ($\Delta = 0$ for the best model).

Example of model selection with *AIC*

Consider the case of model selection with three nested models:

Model 1 $p(k) = \frac{k^{-2}}{\zeta(2)}$ (zeta distribution with (-)2 exponent)

Model 2 $p(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}$ (zeta distribution)

Model 3 $p(k) = \frac{k^{-\gamma}}{H(k_{max}, \gamma)}$ (right-truncated zeta distribution)

Model i is nested model of $i-1$ if the model i is a generalization of model $i-1$ (adding at least one parameter).

Example of model selection with *AIC*

Model	K	$\mathcal{L}$	AIC	Δ
1	0	...	...	....
2	1	...	...	....
3	2	...	...	....

Imagine that the true model is a zeta distribution with $\gamma = 1.5$ and the sample is large enough, then

Model	K	$\mathcal{L}$	AIC	Δ
1	0	...	...	$\gg 0$
2	1	...	...	0
3	2	...	...	> 0

AIC for non-linear regression I

- RSS: "distance" between the data and fitted regression curve based on the the model fit.
- AIC: estimate of the "distance" from the model fit to the true but unknown model that generated the data.
- In a regression model one assumes that the error ϵ follows a normal distribution, the p.d.f. is

$$f(\epsilon) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left\{-\frac{(\epsilon-\mu)^2}{2\sigma^2}\right\}$$

The only parameter is σ as standard non-linear regression assumes $\mu = 0$.

AIC for non-linear regression II

- Applying $\mu = 0$ and $\epsilon_i = y_i - f(x_i, \beta)$

$$f(\epsilon_i) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left\{-\frac{(y_i - f(x_i, \beta))^2}{2\sigma^2}\right\}$$

- Likelihood in a regression model:

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} f(\epsilon_i)$$

- After some algebra one gets

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left\{-\frac{RSS(\beta)}{2\sigma^2}\right\}.$$

AIC for non-linear regression III

Equivalence between maximization of likelihood and minimization of error (under certain assumptions)

- If $\hat{\beta}$ is the **best estimate** of β then

$$L(\hat{\beta}, \hat{\sigma}^2) = \frac{1}{(2\pi RSS(\hat{\beta})/n)^{n/2}} exp(-n/2)$$

with $\hat{\sigma} = \frac{n-K}{n}s^2$ (recall $s^2 = \frac{RSS(\beta)}{n-K}$).

Models selection with regression models:

$$\begin{aligned} AIC &= -2\log L(\hat{\beta}, \hat{\sigma}^2)) + 2(K+1) \\ &= n\log(2\pi) + n\log(RSS(\hat{\beta}/n) + n + 2(K+1) \end{aligned}$$

Why the term for parsimony is $2(K+1)$ and not K?

Concluding remarks

- Under non-linear regression AIC is the way to go for model selection if the models are not nested (alternative methods do exist for nested models [Ritz and Streibig, 2008]).
- Equivalence between maximum likelihood and non-linear regression implies some assumption (e.g., homocedasticity).

Burnham, K. P. and Anderson, D. R. (2002).
Model selection and multimodel inference. A practical information-theoretic approach.
Springer, New York, 2nd edition.

Corral, A., Boleda, G., and Ferrer-i-Cancho, R. (2015).
Zipf's law for word frequencies: word forms versus lemmas in long texts.
PLoS ONE, 10:e0129031.

Ercsey-Ravasz, M., Markov, N., Lamy, C., VanEssen, D., Knoblauch, K., Toroczkai, Z., and Kennedy, H. (2013).
A predictive network model of cerebral cortical connectivity based on a distance rule.
Neuron, 80(1):184 – 197.

Ferrer-i-Cancho, R., Solé, R. V., and Köhler, R. (2004).

Patterns in syntactic dependency networks.
Physical Review E, 69:051915.

Miller, G. A. (1957).
Some effects of intermittent silence.
Am. J. Psychol., 70:311–314.

Ritz, C. and Streibig, J. C. (2008).
Nonlinear regression with R.
Springer, New York.