Analysis of the degree distribution

Ramon Ferrer-i-Cancho & Argimiro Arratia

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

1 Introduction

In this session, we are going to practice on information theoretic model selection for the degree distribution of global syntactic dependency networks. In these networks vertices are words (*word types*) and two vertices are linked if they have been linked at least once in a dependency treebank [Ferrer i Cancho et al., 2004].

Through some procedure, three groups of students will be formed:

- The undirected group. Its member will have to work on the on undirected degree distributions. Their dataset is `degree_sequences.tar.gz`.
- The in-degree group. Its members will have to work on in-degree distributions. Their dataset is `in-degree_sequences.tar.gz`.
- The out-degree group. Its members will have to work on out-degree distributions. Their dataset is `out-degree_sequences.tar.gz`.

Each member of the group works independently from other members but is allowed to compare results with other members of the group.

Each of the files contains the degree sequences from global syntactic dependency networks (undirected degree sequences, in-degree sequences or out-degree sequences from various languages). A degree sequence is sequence $k_1, ..., k_i, ..., k_N$, where k_i is the degree of the i-th node.

Download your .tar.gz and uncompress it, e.g. using the command

```
tar -xzvf degree_sequences.tar.gz
```

You can load the degree sequence of the English network with

```
degree_sequence = read.table("./data/English_degree_sequence.txt",
                  header = FALSE)
```

The number of nodes N, (i.e. the length of the degree sequence) can be computed requesting the number of rows of the $N \times 1$ matrix.

```
dim(degree_sequence)[1]
```

which is equivalent to

```
nrow(degree_sequence)
```

(the number of columns is obtained with `ncol(degree_sequence)`). The sum of the degrees can be computed with

```
sum(degree_sequence)
```

Thus the mean degree can be computed as

```
sum(degree_sequence)/dim(degree_sequence)[1]
```

For simplicity, we assume that $k_i \geq 1$. This means that unlinked nodes, if any, must be removed. One reason is that the family of zeta distributions we are going to consider cannot produce degree 0. Another reason is that unlinked nodes originate quite often from missing information and our goal here is not that of modelling missing information.

In the file `summary_table.R` you have an example of an `R` script that shows some elementary properties of the degree sequence for each language: N, the maximum degree, the mean degree and its inverse. The script can be executed with

```
source("summary_table.R")
```

With the output of that script, you have to produce a table with the format of Table 1. This table will be needed in the coming sections. The script can be adapted to produce the tables that you will have to produce in the coming sections.

2 Visualization

```
degree_sequence = read.table("./data/English_degree_sequence.txt", header = FALSE)
degree_spectrum = table(degree_sequence)
```

Table 1: Summary of the properties of the degree sequences. N is the number of nodes, M/N is mean degree where M is the sum of degrees.

Language	N	Maximum degree	M/N	N/M
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

The command produces a bar plot in normal scale

```
barplot(degree_spectrum, main = "English",
        xlab = "degree", ylab = "number of vertices")
```

while

```
barplot(degree_spectrum, main = "English",
        xlab = "degree", ylab = "number of vertices", log = "xy")
```

does it in log-log scale.

3 A toy ensemble of distributions

Let $p(k)$ be the probability that a vertex has degree k. $p(k)$ is the probability mass function of k and is one way of defining the degree distribution (the cumulative degree distribution might be another way of defining the degree distribution). In this lab session, the probability mass function of degree $p(k)$ and the degree distribution are treated as equivalent.

We consider a toy ensemble of degree distributions on which performing model selections (see [Stumpf et al., 2005, Li et al., 2010] for a richer and more powerful ensemble of degree distributions), where, in all cases $p(0) = 0$. The ensemble contains two distributions from null models of networks and three nested variants of the zeta distribution as possible models of *power-laws*. The distributions deriving from null models of networks are the Poisson distribution and the geometric distribution. The Poisson distribution is chosen for being a mathematically simple approximation to the binomial distribution characterizing Erdös-Rényi graphs [Newman, 2010]. There is a very importance difference between the Poisson and the geometric distribution: they exhibit an exponential tail while the zeta distribution exhibits a so-called heavy tail.

The displaced geometric distribution (with $p(0) = 0$) is defined as

$$p(k) = (1 - q)^{k-1} q, \tag{1}$$

where q is the only parameter of the distribution and $k \geq 0$. We will not use a perhaps more popular version of the distribution defined as

$$p(k) = (1-q)^k q. \tag{2}$$

for $k \geq 0$ because we do not allow for unlinked vertices.

For the same reason, we will not use the popular definition of the Poisson distribution with parameter λ

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3}$$

for $k > 0$. Imposing the condition $k \geq 1$, the displaced Poisson distribution is obtained (see Appendix)

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!(1-e^{-\lambda})} \tag{4}$$

The three nested variants of the zeta distribution that we consider are:

1. $p(k) = \frac{k^{-2}}{\zeta(2)}$ (zeta distribution with (-)2 exponent)
2. $p(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}$ (zeta distribution with exponent γ)
3. $p(k) = \frac{k^{-\gamma}}{H(k_{max},\gamma)}$ (right-truncated zeta distribution with exponent γ with right-truncation beyond k_{max})

4 Estimation of the parameters

Before applying standard model selection methods, the parameters giving the best fit must be obtained. This is done may maximizing the $\mathcal{L}$, the log-likelihood function. If the degree sequence of a network of N vertices is $k_1, k_2, ...k_N$, its log-likelihood is

$$\mathcal{L} = \log\left(\prod_{i=1}^{N} p(k_i)\right) = \sum_{i=1}^{N} \log p(k_i). \tag{5}$$

The parameters giving the best fit are those that maximize $\mathcal{L}$. Table 2 shows a summary of the log-likelihood function of each distribution. First, we will explain how these log-likelihoods have been derived and then we will explain how to use `R` to estimate the best parameters of a function.

Table 2: The log-likelihood $\mathcal{L}$ for each of the probability mass functions. K is the number of free parameters. M is the sum of degrees, i.e. $M = \sum_{i=1}^{N} k_i$. M' is the sum of degree logarithms, i.e. $M' = \sum_{i=1}^{N} \log k_i$ and C is the sum of logarithm of degree factorials, i.e. $C = \sum_{i=1}^{N} \sum_{j=2}^{k_i} \log j$.

Model	Function	K	$\mathcal{L}$
1	Displaced Poisson	1	$M \log \lambda - N(\lambda + \log(1 - e^{-\lambda})) - C$
2	Displaced geometric	1	$(M - N)\log(1 - q) + N \log q$
3	Zeta with $\gamma = 2$	0	$-2M' - N \log \frac{\pi^2}{6}$
4	Zeta	1	$-\gamma M' - N \log \zeta(\gamma)$
5	Right-trucanted zeta	2	$-\gamma M' - N \log H(k_{max}, \gamma)$

4.1 Derivation of the log-likelihood functions

- Speakers of Romance languages (Catalan, Spanish,...): do not confuse *derivation*, e.g., the deduction of a formula (*derivació*, *derivación*,...) with derivative (*derivada*,...), e.g., a formula deduced by differentiation (the tangent of a curve in a graphical sense).
- *Derive* in English (as well as *derivar* in Catalan or Spanish) has at least two meanings: obtain (applying some inferences) and a more restrictive meaning of obtaining by differentiation. The latter is not intended for this session!

For the displaced geometric distribution (Eq. 1),

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^{N} \log\left[(1-q)^{k_i - 1} q\right], \\ &= \sum_{i=1}^{N} \left[(k_i - 1)\log(1-q) + \log q\right] \\ &= (M - N)\log(1-q) + N \log q, \end{aligned} \tag{6}$$

where

$$M = \sum_{i=1}^{N} k_i. \tag{7}$$

Notice that the number of edges E, satisfies $M = 2E$ (in the absence of loops).

For the displaced Poisson distribution (Eq. 4), the log-likelihood is

$$\mathcal{L} = \sum_{i=1}^{N} \log \frac{\lambda^{k_i} e^{-\lambda}}{k_i!(1 - e^{-\lambda})}$$

$$= \sum_{i=1}^{N} \left[k_i \log \lambda - \lambda \log e - \log(1 - e^{-\lambda}) \right] - C \tag{8}$$

with

$$\begin{aligned} C &= \sum_{i=1}^{N} \log(k_i!) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{k_i} \log j \\ &= \sum_{i=1}^{N} \sum_{j=2}^{k_i} \log j \end{aligned} \tag{9}$$

Working further on Eq. 8, it is finally obtained

$$\begin{aligned} \mathcal{L} &= \log \lambda \sum_{i=1}^{N} k_i - \sum_{i=1}^{N} (\lambda + \log(1 - e^{-\lambda})) - C \\ &= \log \lambda \sum_{i=1}^{N} k_i - (\lambda + \log(1 - e^{-\lambda})) \sum_{i=1}^{N} 1 - C \\ &= M \log \lambda - N(\lambda + \log(1 - e^{-\lambda})) - C, \end{aligned} \tag{10}$$

where M is defined as before.

For the zeta distributions, will start from the log-likelihood for the right-truncated zeta distribution and then obtain the log-likelihood for the particular versions. The log-likelihood of the right-truncated zeta distribution is

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^{N} \log \frac{k_i^{-\gamma}}{H(k_{max}, \gamma)} \\ &= -\gamma M' - \sum_{i=1}^{N} \log H(k_{max}, \gamma), \end{aligned} \tag{11}$$

where

$$M' = \sum_{i=1}^{N} \log k_i. \tag{12}$$

Working further on Eq. 11, it is finally obtained

$$\mathcal{L} = -\gamma M' - N \log H(k_{max}, \gamma). \tag{13}$$

The result is analogous to the derivation in the context of the rank spectrum for word frequencies [Baixeries et al., 2013].

The log-likelihood for the zeta distribution is obtained applying $k_{max} \to \infty$ to Eq. 13, which gives

$$\mathcal{L} = -\gamma M' - N \log \zeta(\gamma). \tag{14}$$

The log-likelihood of the zeta distribution with $\gamma = 2$ is obtained applying $\zeta(2) = \pi^2/6$, which gives

$$\mathcal{L} \quad = \quad -2M' - N \log \frac{\pi^2}{6}. \tag{15}$$

4.2 How to use R to estimate the parameters by maximum likelihood

The procedure to obtain the best parameters by maximum likelihood needs an initial value for the parameters. For displaced geometric distribution, an initial value can be inferred from the fact that the expectation of k is $\mu_k = 1/q$. μ_k can be estimated by the mean degree, i.e. M/N. Thus, a suitable initial value of q is $q_0 = 1/\mu_k = N/M$. For the displaced Poisson distribution (see Appendix),

$$\mu_k = \frac{\lambda}{1 - e^{-\lambda}} \tag{16}$$

and then $\mu_k \approx \lambda$ for sufficiently large λ. Thus, a suitable initial value of λ is $\lambda_0 = M/N$.

For the two zeta distributions with γ as free parameter, a suitable initial value is $\gamma_0 = 2$ (alternatively, a better initial value can be inferred from the slope of a linear regression in log-log scale). For the right-truncated zeta distribution a suitable initial value for k_{max} is $k_{max,0} = N$ (alternatively, one could use the largest degree for $k_{max,0}$).

From this section onwards, the packages `stats4` and `VGAM` are needed. They are loaded with the commands

```
require("stats4") # for MLE
require("VGAM") # for the Riemann-zeta function
```

Before you may need to install them applying one of the following commands

```
install.packages("stats4")
install.packages("VGAM")
```

Let us see a concrete example for the zeta distribution. The minus log-likelihood function is defined through

```
x <- degree_sequence$V1
minus_log_likelihood_zeta <- function(gamma) {
     length(x) * log(zeta(gamma)) + gamma * sum(log(x))
}
```

where `minus_log_likelihood` is $-\mathcal{L}$ for the zeta distribution following Table 2.

To estimate the parameters by maximum likelihood, we call the command `mle(...)` indicating that the minus log-likelihood function is `log_likelihood_zeta`, that the initial value of the only parameter is 2, that the method to maximize the log-likelihood is "L-BFGS-B" (an minimization/maximization methods allowing one to define upper and lower bounds or the parameters) and that the minimum value of γ is a very small number above 1 (recall that the zeta distribution requires $\gamma > 1$).

```
mle_zeta <- mle(minus_log_likelihood_zeta,
            start = list(gamma = 2),
            method = "L-BFGS-B",
            lower = c(1.0000001))
```

> Notice that `mle` does not have $\mathcal{L}$ as parameter but $-\mathcal{L}$ instead. `mle` solves the problem of maximum likelihood by minimizing $-\mathcal{L}$. This has an important practical consequence: the sign of the formulae in Table 2 must be inverted before supplying them to `mle`.

The following command allows one to see a summary of the results of the maximum likelihood estimation

```
summary(mle_zeta)
```

The command `attributes(...)` allows one to see how to fish for the information

```
attributes(summary(mle_zeta))
```

Then the exponent giving the best fit can be retrieved with

```
attributes(summary(mle_zeta))$coef[1]
```

Check that $q \approx q_0$, $\lambda \approx \lambda_0$, with the help of Table 1 and 3. In contrast, (γ and γ_0 might differ substantially specially if they were obtained by linear regression in log-log scale).

5 Model selection

We are going to choose the best of the models according to AIC with a correction for sample size, which is defined as

$$AIC_c = -2\mathcal{L} + 2K\frac{N}{N-K-1}. \tag{17}$$

Table 3: Summary of the most likely parameters. γ_1 and γ_2 refer, respectively, to the exponent of the zeta distribution and the right-truncated distribution.

	Model				
	1	2	4		5
Language	λ	q	γ_1	γ_2	k_{max}
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

For the values of N and K that we are going to use ($N \gg K$ in our case), the correction is likely to not alter the conclusions of model selection with regard to the original AIC.

5.1 Model selection with `R`

$-2logL$ ($= -2\mathcal{L}$) can be fished from the object returned by the `mle` with

```
attributes(summary(mle_zeta))$m2logL
```

for the example of the zeta distribution above.

It is convenient to define a function that computes the AIC from the relevant information, e.g.,

```
get_AIC <- function(m2logL,K,N) {
   m2logL + 2*K*N/(N-K-1) # AIC with a correction for sample size
}
```

Following the example of the zeta distribution, the AIC can be obtained with

```
get_AIC(attributes(summary(mle_zeta))$m2logL, 1, N)
```

To complete the work, you have to calculate AIC_{best}, the smallest AIC of the ensemble of distributions and, for every model, calculate $\Delta = AIC - AIC_{best}$, the so-called AIC difference. Then, you have to produce a summary table with the format of Table 4. Calling `mle(...)` with the right-truncated zeta distribution is tricky. You will have to define the lower bound for the k_{max} parameter carefully.

Finally, investigate the consequences of adding a new probability distribution that is able to give a better fit than the best model so far. We suggest an Altmann function such as

$$p(k) = ck^{-\gamma}e^{-\delta k} \tag{18}$$

Table 4: The AIC difference (Δ) of a model on a given source. Models are numbered according to Table 2

	Model				
Language	1	2	3	4	5
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

if $1 \leq k \leq N$ and $p(k) = 0$ otherwise, with

$$c = \frac{1}{\sum_{k=1}^{N} k^{-\gamma} e^{-\delta k}} \tag{19}$$

6 Checking your methods

One important limitation of the real datasets we are providing you is that the true distribution is unknown. Furthermore, the true distribution may not belong to the ensemble of probability functions suggested above. Thus, it is difficult to be certain about the correctness of the results conditioned on that the ensemble of distributions. For this reasons, we are also providing artificial datasets where the true distribution is a priori known. The package

`http://www.cs.upc.edu/~rferrericancho/data/samples_from_discrete_distributions.tar.gz`

provides a collection of files produced with the geometric distribution and the zeta distribution. You have two check that your methods:

- Select the right distribution.
- Obtain the right parameters of the distribution.

Bear in mind that the file for zeta with $\gamma = 1.5$ is *a priori* problematic (do not worry if that is the only functions in Table 2 for which `R` crashes). Hypothesizing a distribution that is totally different from the real one can be problematic, e.g., a geometric distribution for a zeta distribution with $\gamma < 2$.

7 Deliverables

You have to prepare a report including the following sections (in this order): introduction, results, discussion and methods. Results includes all the tables

and figures (the preliminary plots and the plots of the best model to the real data) and some guiding text. Methods should include any relevant methods not explained in this guide (for instance, decisions that you had to made and might have an influence on the results). The discussion should include a summary of the results and your interpretation. For instance, you should discuss

- If there is a significance difference between the fit of the distributions from null models and those of the zeta family.
- Discuss if the distribution giving the best fit gives a reasonably good fit (e.g., checking visually that the best function provides a sufficiently good fit). Remember that the best function of an ensemble is not necessarily the best in absolute terms.
- The extent to which languages resemble or differ.

The discussion section should also include some conclusions.

Important rule: The lab session, and especially the report you have to hand in, are strictly individual work. Plagiarism will be prosecuted. Nevertheless, you are encouraged to ask the teacher as soon as possible if you think you do not understand what you are supposed to do, and also if you feel you are spending much more time than the rest of the group – sometimes a tiny error can be tricky to find and does not add much to your knowledge. Questions can be asked either in person or by email, and you will never be penalized by asking questions, no matter how stupid they look in retrospect.

To deliver: You must deliver the report explained above. The formats accepted for the report are, in principle, pdf, Word, OpenOffice, and Postscript. You also have to hand in the source code in `R` (or other languages) that you have used, including some minimal comments that can help the reader.

Procedure: Submit your work through the raco platform as a single zipped file.

Deadline: Work must be delivered within 2 weeks from the lab session you attend. Late deliveries risk being penalized or not accepted at all. If you anticipate problems with the deadline, please tell us as soon as possible.

References

[Baixeries et al., 2013] Baixeries, J., Elvevåg, B., and Ferrer-i-Cancho, R. (2013). The evolution of the exponent of zipf's law in language ontogeny. *PLoS ONE*, 8(3):e53227.

[Ferrer i Cancho et al., 2004] Ferrer i Cancho, R., Solé, R. V., and Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E*, 69:051915.

[Li et al., 2010] Li, W., Miramontes, P., and Cocho, G. (2010). Fitting ranked linguistic data with two-parameter functions. *Entropy*, 12(7):1743–1764.

[Newman, 2010] Newman, M. E. J. (2010). *Networks. An introduction*. Oxford University Press, Oxford.

[Stumpf et al., 2005] Stumpf, M., Ingram, P., Nouvel, I., and Wiuf, C. (2005). Statistical model selection methods applied to biological network data. *Trans. Comp. Syst. Biol.*, 3:6577.

Appendix

The standard Poisson distribution can be defined as

$$p(k|k \geq 0) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{20}$$

We want to derive a displaced Poisson distribution where $k \geq 1$. Its probability mass function is

$$\begin{aligned} p(k|k>0) &= \frac{p(k|k \geq 0)p(k \geq 0|k \geq 0)}{p(k>0|k \geq 0)} \\ &= \frac{p(k|k \geq 0)}{p(k>0|k \geq 0)} \\ &= \frac{p(k|k \geq 0)}{1-p(k=0|k \geq 0)} \end{aligned} \tag{21}$$

as $p(k \geq 0|k \geq 0) = 1$. Thus, applying the definition of the standard Poisson (Eq. 20) to Eq. 21, it is obtained

$$p(k|k>0) = \frac{\lambda^k e^{-\lambda}}{k!(1-e^{-\lambda})} \tag{22}$$

The expected k for the standard Poisson distribution is $E[k|k \geq 0] = \lambda$. The expected k for the displaced Poisson distribution above is

$$\begin{aligned} E[k|k>0] &= \sum_{k=1}^{\infty} p(k|k>0)k \\ &= \frac{1}{1-p(k=0|k \geq 0)} \sum_{k=1}^{\infty} p(k|k \geq 0)k \\ &= \frac{1}{1-p(k=0|k \geq 0)} \sum_{k=0}^{\infty} p(k|k \geq 0)k \\ &= \frac{E[k|k \geq 0]}{1-p(k=0|k \geq 0)} \\ &= \frac{\lambda}{1-e^{-\lambda}} \end{aligned} \tag{23}$$