# UPC-FIB Master's in Data Science
# Complex and Social Networks - CSN
# Lab 2 Report

Petar Novak, petar.novak@estudiantat.upc.edu
Giovanni Spisso, giovanni.spisso@estudiantat.upc.edu

September 25, 2024

## 1 Introduction

In this report, we investigate the in-degree distribution of a type of real-world networks known as global syntactic dependency networks [Ferrer i Cancho et al., 2004]. Our goal is to understand to what extent their in-degree distributions follow well-known probability distributions.

We analyze 10 different languages and compare the sampled data with the following distributions: Geometric, Poisson, Zeta with $\gamma = 2$, Zeta, and right-truncated Zeta. Additionally, to validate our procedure, we take samples from known distributions and verify whether we are able to predict the correct model.

Finally, we introduce the Altmann distribution, which is expected to yield better results compared to the best model identified so far.

## 2 Results

We begin our analysis by creating Table 1, which summarizes the most relevant characteristics of our networks.

Next, we calculate the log-likelihood functions for the different distributions, as shown in Table 2. This table also includes the log-likelihood function for the Altmann distribution, which is analyzed and compared to the other distributions at the end of Section 3.

For parameter estimation, we primarily follow the procedure outlined in the guide. We use the **mle(...)** command and, for each distribution, we specify initial parameter values and set appropriate ranges for them. Refer to Section 4 for the choices we have made. The parameter values that provide the best fit are reported in Table 3.

Finally, to evaluate which of our models performs best, we calculate the AIC with a correction for sample size, as outlined in the guide, and present the results in Table 4. From the table, we observe that the best AIC is consistently achieved by model 5 (Right-truncated Zeta distribution), though model 3 (Zeta distribution) also provides comparable values. In Section 3, we discuss the quality of these results in more detail.

| Language | N | Maximum degree | M/N | N/M |
|---|---|---|---|---|
| Arabic | 21065 | 2249 | 3.35 | 0.30 |
| Basque | 11868 | 576 | 2.18 | 0.46 |
| Catalan | 35524 | 5522 | 5.75 | 0.17 |
| Chinese | 35563 | 7645 | 5.20 | 0.19 |
| Czech | 66014 | 4727 | 3.97 | 0.25 |
| English | 29172 | 4547 | 6.86 | 0.15 |
| Greek | 12704 | 1081 | 3.52 | 0.28 |
| Hungarian | 34600 | 6540 | 3.10 | 0.32 |
| Italian | 13433 | 2678 | 4.23 | 0.24 |
| Turkish | 20403 | 6704 | 2.31 | 0.43 |

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.

| 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 | 1 | $-\gamma M' - N \log \zeta(\gamma)$ |
| 4 | Zeta with $\gamma = 2$ | 0 | $-2M' - N \log \frac{\pi^2}{6}$ |
| 5 | Right-truncated zeta | 1 | $-\gamma M' - N \log H(k_{max}, \gamma)$ |
| 6 | Altmann | 2 | $-\gamma M' - \delta M - N \log(\sum_{k=1}^{N} k^{-\gamma} e^{-\delta k})$ |

Table 2: The log-likelihood $\mathcal{L}$ for each of the probability mass function. $K$ 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. $C = \sum_{i=1}^{N} \sum_{j=2}^{k_i} \log(j)$.

To verify the correctness of our implementations, we tested our models on samples drawn from known distributions. The key characteristics of these samples, along with the distributions they were generated from, are reported in Table 5.

Our results are presented in Table 6 and Table 7. From these tables, we can clearly observe that our predictions are accurate. We consistently identify the correct parameters for the appropriate distribution. When the samples are drawn from a geometric distribution, we correctly select the geometric model as the best fit. However, when the samples are taken from a Zeta distribution, the right-truncated Zeta model yields slightly better results and is chosen as the best fit. This occurs because the truncated Zeta distribution is able to better model the specific samples on which it is trained, as the value of $k_{max}$ is optimized to best fit the sampled data. This behaviour is further confirmed by the fact that, for the sample taken from a Zeta distribution with $\gamma = 2$, the right-truncated and the Zeta distribution with free $\gamma$ obtain better results, since they are able to fit the specific data better. This behaviour is explained more precisely in Section 3.

# 3 Discussion

From our analysis of syntactic dependency networks, it is clear that the model that consistently best fits these networks, among those tested, is the right-truncated Zeta distribution. The Zeta

| Language | lambda | p | gamma 1 | gamma 2 | k max |
|---|---|---|---|---|---|
| Arabic | 3.22 | 0.30 | 2.10 | 2.10 | 2249 |
| Basque | 1.83 | 0.46 | 2.36 | 2.36 | 576 |
| Catalan | 5.73 | 0.17 | 1.92 | 1.92 | 5522 |
| Chinese | 5.17 | 0.19 | 1.89 | 1.88 | 7645 |
| Czech | 3.89 | 0.25 | 2.05 | 2.05 | 4727 |
| English | 6.85 | 0.15 | 1.80 | 1.79 | 4547 |
| Greek | 3.41 | 0.28 | 2.13 | 2.13 | 1081 |
| Hungarian | 2.93 | 0.32 | 2.34 | 2.34 | 6540 |
| Italian | 4.16 | 0.24 | 2.11 | 2.11 | 2678 |
| Turkish | 2.00 | 0.43 | 2.54 | 2.54 | 6704 |

Table 3: Summary of the most likely parameters. $\gamma_1$ and $\gamma_2$ refer, respectively, to the exponent of the zeta distribution and the right-truncated distribution.

| Language | Poisson | Geometric | Zeta | Zeta ($\gamma = 2$) | Right-truncated Zeta |
|---|---|---|---|---|---|
| Arabic | 240323.25 | 24190.94 | 4.90 | 177.28 | 0.00 |
| Basque | 50166.59 | 8366.11 | 2.23 | 847.40 | 0.00 |
| Catalan | 913882.08 | 61883.73 | 15.95 | 227.90 | 0.00 |
| Chinese | 618350.40 | 48680.82 | 16.69 | 505.99 | 0.00 |
| Czech | 940679.71 | 91101.39 | 10.85 | 149.28 | 0.00 |
| English | 739583.46 | 45744.17 | 47.04 | 1471.66 | 0.00 |
| Greek | 157139.23 | 17135.83 | 5.42 | 166.10 | 0.00 |
| Hungarian | 468252.49 | 53469.92 | 0.30 | 2220.58 | 0.00 |
| Italian | 245805.51 | 22879.90 | 2.49 | 120.37 | 0.00 |
| Turkish | 193345.06 | 24615.38 | 0.02 | 2740.49 | 0.00 |

Table 4: The AIC difference ($\Delta$) of a model on a given source.

distribution also yields comparable results in terms of the AIC. However, for theoretical reasons discussed in Section 4, the right-truncated Zeta distribution will always provide a better fit, as the maximum degree $k_{max}$ is selected to optimally fit the specific sample, potentially leading to overfitting. It would be worthwhile to conduct further experiments by applying the models we have identified to different global syntactic dependency networks for the same languages, to assess the quality of the fit and explore their ability to generalize.

To visually evaluate whether these models are truly able to fit the data, we plotted the degree distributions of the real networks alongside the predictions from our models. The plots reveal a very similar behavior across different languages, as illustrated in Figure 1. There are no visible differences between the right-truncated Zeta distribution and the Zeta distribution, as the models are overlapped. We use a log-log scale, and since the data exhibit a linear relationship with this scale, at least for nodes with a degree that is not too large, the Zeta distribution appears to be the appropriate model for this type of network, as they seem to follow a power-law distribution.

The main difference between the Zeta distribution and syntactic dependency networks is that these networks exhibit a much longer tail than predicted by the theoretical distribution. This behaviour is typical of real networks (see [Clauset et al., 2009]) and suggests that while the Zeta distribution may accurately model the degree distribution up to a certain point, it may underes-

| Distribution | N | Maximum degree | M/N | N/M |
|---|---|---|---|---|
| geo 0.05 | 1000 | 133 | 19.24 | 0.05 |
| geo 0.1 | 1000 | 95 | 10.30 | 0.10 |
| geo 0.2 | 1000 | 36 | 5.02 | 0.20 |
| geo 0.4 | 1000 | 14 | 2.48 | 0.40 |
| geo 0.8 | 1000 | 5 | 1.27 | 0.79 |
| zeta 1.5 | 1000 | 20070737 | 21418.12 | 0.00 |
| zeta 2.5 | 1000 | 53 | 1.83 | 0.54 |
| zeta 2 | 1000 | 362 | 5.08 | 0.20 |
| zeta 3.5 | 1000 | 36 | 1.27 | 0.79 |
| zeta 3 | 1000 | 41 | 1.36 | 0.73 |

Table 5: Summary of the properties of the samples. The labels have the same meaning of 1.

| Distribution | lambda | p | gamma 1 | gamma 2 | k max |
|---|---|---|---|---|---|
| geo 0.05 | 19.24 | 0.05 | 1.33 | 1.00 | 133 |
| geo 0.1 | 10.30 | 0.10 | 1.42 | 1.06 | 95 |
| geo 0.2 | 4.98 | 0.20 | 1.57 | 1.20 | 36 |
| geo 0.4 | 2.21 | 0.40 | 1.89 | 1.54 | 14 |
| geo 0.8 | 1.00 | 0.79 | 3.00 | 2.77 | 5 |
| zeta 1.5 | 21418.12 | 0.01 | 1.50 | 1.49 | 20070737 |
| zeta 2 | 5.04 | 0.20 | 1.98 | 1.97 | 362 |
| zeta 2.5 | 1.37 | 0.54 | 2.45 | 2.43 | 53 |
| zeta 3 | 1.00 | 0.73 | 3.00 | 2.99 | 41 |
| zeta 3.5 | 1.00 | 0.79 | 3.35 | 3.35 | 36 |

Table 6: Summary of the most likely parameters.

timate the presence of nodes with very high degrees in dependency networks. This behaviour is, more or less, common to all languages.

The various languages do not appear to differ significantly, as they are all well-fitted by a right-truncated Zeta distribution with $\gamma_2 \in [1.79, 2.54]$.

To compare the models more precisely, we calculate their Akaike weights, obtaining quantitative measures that tell how much better is our best model compared to the others. For the details about this method refer to 4. The results are reported in Table 8.

The table confirms that the right-truncated Zeta distribution consistently emerges as the best model, and for most languages, it results in Akaike weights $w \approx 1$, showing that it outclasses the other models. Notably, for Hungarian and Turkish, the Zeta distribution performs equally well, showing comparable effectiveness.

To further confirm our method empirically, we have also plotted the results for the random samples from discrete distributions. In Figure 2, we present two plots where it is visually evident that our best models accurately fit the data, as expected. For this plots we have chosen a linear-log scale for the geometric and a log-log scale for the Zeta distribution, expecting the data to be on a line. Notice that, in our random samples, we can see a similar "long tail" behaviour compared to the one we obtain in the real networks.

Finally, we add a new probability distribution, the Altmann function, which should be able to

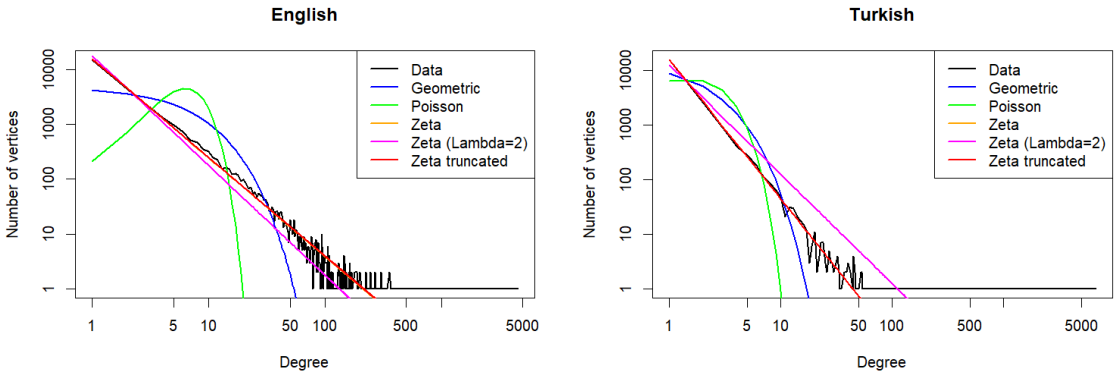


Figure 1: Plots of real and estimated distributions for english ($\gamma_2 = 1.79$) and turkish ($\gamma_2 = 2.54$).

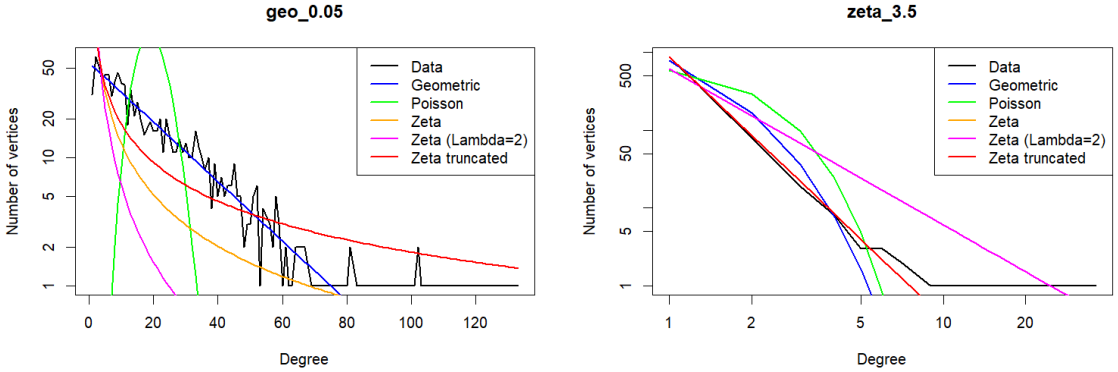


Figure 2: Plots of sample and estimated distributions for a geometric distribution ($p = 0.05$) and a zeta distribution ($\gamma_2 = 3.5$).

| Distribution | Poisson | Geometric | Zeta | Zeta ($\gamma = 2$) | Right-truncated Zeta |
|---|---|---|---|---|---|
| geo 0.05 | 11580.49 | 0.00 | 1334.45 | 3082.14 | 512.76 |
| geo 0.1 | 5322.28 | 0.00 | 957.30 | 1939.30 | 455.63 |
| geo 0.2 | 1643.99 | 0.00 | 686.56 | 1040.70 | 322.51 |
| geo 0.4 | 805.21 | 0.00 | 349.16 | 359.49 | 157.54 |
| geo 0.8 | 1310.06 | 0.00 | 56.20 | 351.58 | 21.98 |
| zeta 1.5 | 281022481.29 | 433220.32 | 0.37 | 596.11 | 0.00 |
| zeta 2 | 16940.07 | 1695.93 | 3.97 | 2.32 | 0.00 |
| zeta 2.5 | 2491.28 | 412.16 | 3.28 | 103.33 | 0.00 |
| zeta 3 | 1804.75 | 224.56 | 0.50 | 294.81 | 0.00 |
| zeta 3.5 | 2020.11 | 275.96 | 0.16 | 410.82 | 0.00 |

Table 7: The AIC difference ($\Delta$) of a model on a given distribution.

| Language | Zeta | Right-truncated Zeta |
|---|---|---|
| Arabic | 0.08 | 0.92 |
| Basque | 0.25 | 0.75 |
| Catalan | 0.00 | 1.00 |
| Chinese | 0.00 | 1.00 |
| Czech | 0.00 | 1.00 |
| English | 0.00 | 1.00 |
| Greek | 0.06 | 0.94 |
| Hungarian | 0.46 | 0.54 |
| Italian | 0.22 | 0.78 |
| Turkish | 0.50 | 0.50 |

Table 8: The Akaike weights of the best models (the Zeta and right-truncated Zeta distribution) for different languages.

give a better fit than the best model so far. After obtaining the log-likelihood (see Section 4), we follow the same procedure as before. We select the appropriate parameters using the **mle** function and then calculate the AIC. The final results, comparing the AIC of the Altmann distribution with the previously best AIC, are presented in Table 9.

From the table, we observe that the Altmann distribution yields results very similar to the right-truncated Zeta distribution. For a couple of languages (specifically Chinese and English), it provides a better AIC, while for others, it is slightly worse. This behavior can be attributed to the fact that, for all languages, the estimated parameter $\delta$ is small. When $\delta \approx 0$, we have $p(k) \approx ck^{-\gamma}$, which essentially corresponds to a right-truncated Zeta distribution with $k_{max} = N$. The worst results occur when $\delta$ is estimated to be very small, in which case the right-truncated Zeta provides better results, since we are using $k_{max} = \max(k)$ instead of N, which is better as discussed in Section 4. Conversely, when $\delta$ is larger (in our case, when $\delta = 0.003$), the exponential factor in the Altmann distribution has a greater impact on the fit, leading to better results. In Figure 3, we present the plots where the Altmann distribution achieves a better AIC than the previously best-fitting model. From the plots, there does not seem to be a significant difference between the Altmann distribution and the right-truncated Zeta distribution, as their curves overlap. To have more precise results, we repeat the same analysis previously done for the Akaike weights with the

| Language | gamma | delta | Alt.AIC - best.AIC |
|---|---|---|---|
| Arabic | 2.09 | 0.001 | -7.88 |
| Basque | 2.35 | 0.002 | 2.08 |
| Catalan | 1.90 | 0.001 | -29.15 |
| Chinese | 1.83 | 0.003 | -182.68 |
| Czech | 2.04 | 0.001 | -25.59 |
| English | 1.74 | 0.003 | -211.34 |
| Greek | 2.13 | 0.000 | 7.00 |
| Hungarian | 2.34 | 0.000 | 2.70 |
| Italian | 2.11 | 0.000 | 4.09 |
| Turkish | 2.54 | 0.000 | 2.21 |

Table 9: Best $\gamma$ and $\delta$ for every language and the difference between the Altmann AIC and best AIC obtained with previous models.

addition of the Altmann distribution, obtaining the results reported in Table 10.

| Language | Zeta | Right-truncated Zeta | Altmann |
|---|---|---|---|
| Arabic | 0.00 | 0.02 | 0.98 |
| Basque | 0.20 | 0.59 | 0.21 |
| Catalan | 0.00 | 0.00 | 1.00 |
| Chinese | 0.00 | 0.00 | 1.00 |
| Czech | 0.00 | 0.00 | 1.00 |
| English | 0.00 | 0.00 | 1.00 |
| Greek | 0.06 | 0.90 | 0.04 |
| Hungarian | 0.46 | 0.53 | 0.02 |
| Italian | 0.21 | 0.72 | 0.08 |
| Turkish | 0.46 | 0.47 | 0.07 |

Table 10: Updated Akaike weights for the best models, with the addition of the Altmann function.

From the table, we observe varying behaviour across different languages. For Arabic, Catalan, Chinese, Czech, and English, the Altmann distribution clearly provides the best results. In contrast, for Greek, the right-truncated Zeta emerges as the best model. For Basque and Italian, the situation is more uncertain, with a slight preference for the right-truncated Zeta. Finally, for Hungarian and Turkish, two models appear to fit equally well.

This highlights the greater precision of Akaike weights in comparing models. By relying solely on the plots, there are no clear differences between languages, as the three models seem equally effective, with their lines consistently overlapping.

From this analysis we conclude that for some languages the Altmann distribution models better the networks, while for others the right-truncated Zeta seems more appropriate, even if the plots do not show real practical differences between the three models.

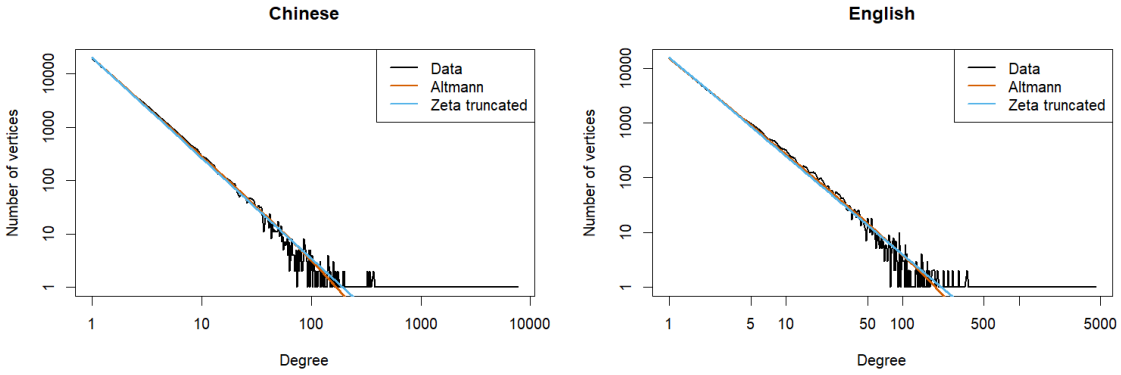


Figure 3: Comparison between Zeta truncated and Altmann distribution, for languages 4 and 6, where the Altmann distribution achieves a better AIC

# 4 Methods

In this report, we have primarily followed the suggestions provided in the guide. However, we have also made a few additional observations and calculations to enhance our analysis.

## 4.1 The choice of the parameters and of $k_{max}$

Using the notation present in the guide, a suitable initial value for the parameter in the geometric distribution is $q_0 = N/M$ and an obvious range is $q \in [0, 1]$, since it represents a probability (for numerical reasons we set the interval as $[0.01, 0.99]$, otherwise we get numerical errors for some languages). For the Poisson distribution we choose $\lambda_0 = M/N$, which only has to be greater than 0. For the two zeta distributions we choose $\gamma_0 = 2$ and for convergence reasons we have to impose $\gamma > 1$. The main difference from the guide is that, for the right-truncated Zeta distribution, we set $k_{max}$ = "the largest degree in the network." The reasoning behind this choice is explained by the following observation.

**Observation 1.** *The value of $k_{max}$ that maximizes the likelihood of the right-truncated zeta distribution*

$$p(k) = \begin{cases} \frac{k^{-\gamma}}{\sum_{i=1}^{k_{max}} i^{-\gamma}}, & 1 \leq k \leq k_{max} \\ 0 & otherwise \end{cases}$$

*for a given sample $(x_i, k_i)$, $1 \leq i \leq N$, is $k_{max} = \max_i(k_i)$.*

*Proof.* As stated in the guide, the log-likelihood of the right-truncated Zeta distribution is given by

$$\mathcal{L} = -\gamma \sum_{i=1}^{N} \log k_i - N \log \sum_{i=1}^{k_{max}} i^{-\gamma}.$$

Since maximizing the log-likelihood is equivalent to minimizing its negative, we reformulate our

problem as:

$$\underset{\gamma>1, k_{max} \geq max(k_i)}{\arg\min} \gamma \sum_{i=1}^{N} \log k_i + N \log \sum_{i=1}^{k_{max}} i^{-\gamma}$$

It is important to note that $k_{max}$ must be greater than or equal to $max(k_i)$ for the probability distribution to make sense. Otherwise, we would have $p(max(k_i)) = 0$, which is impossible since we have observed a sample with that value of $k$. Moreover, in the expression, only the second term depends on $k_{max}$. Specifically, the function is strictly increasing with respect to $k_{max}$, as $\sum_{i=1}^{k_{max}} i^{-\gamma} > 1$ implies that its logarithm is greater than 0.

This shows that to minimize the function with respect to $k_{max}$, we must always choose the smallest possible value, which, in our case, is $max_i(k_i)$. □

This theorem implies that the Zeta-truncated function will always provide a better fit compared to the Zeta function, since in the latter case we have $k_{max} = \infty$. For a comprehensive discussion on the theoretical differences on these distributions, see [Naldi, 2015].

As the optimization method, we chose "L-BFGS-B" as recommended in the guide ([Byrd et al., 2003]), noticing that, since $k_{max}$ is fixed, the parameters we need to estimate are continuous and satisfy the required constraints.

## 4.2 The use of Akaike weights

Another procedure we introduced, which is not present in the guide, is the use of the "Akaike weights" to rigorously validate our results.

Recall that, from the maximum likelihood theory, we are choosing the parameters that maximize the likelihood of a model given the data, that is $\mathcal{L}$(parameters|data,model). However, what we are really interested about is what is the probability of the model given the data, that is $\mathcal{L}$(model|data).

As seen in class this probability can be approximated by:

$$\mathbb{P}(model_i|data) \propto e^{-\frac{1}{2}\Delta_i},$$

where $\Delta_i = \text{AIC}_i - \text{AIC}_{\text{best}}$ (the difference of the AIC of model $i$ and the best model we have found). This approximation is crucial for having a quantitative way to measure how much better is the best model compared to the others we have tried.

For a tested model $i$, we define its Akaike weight as:

$$w_i(AIC) = \frac{exp\{-\frac{1}{2}\Delta_i(AIC)\}}{\sum_k \{exp\{-\frac{1}{2}\Delta_k(AIC)\}\}},$$

which is the probability of model $i$ given the data divided by the sum of the probabilities of all the models we have tested. This method of performing model evaluation is widely explained in [Wagenmakers and Farrell, 2004], where there is also an interesting comparison with the same procedure performed with the $\text{BIC} = -2\log(\mathcal{L}) + K\log(N)$, where the stronger penalization given to $k$ (the number of parameters) favors simpler models. It could be interesting to use BIC for further analysis on the networks and seeing if the estimates changes, keeping in mind that the BIC assumes that the true generation model is in the set of candidate models, which is not realistic in our case.

### 4.3 The Altmann distribution

The Altmann distribution is defined as

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

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

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

It corresponds to a power-law distribution combined with an exponential function, truncated at $N$. For the parameters it is sufficient that $\gamma > 0$ and $\delta > 0$ for having a valid probability function. To conclude, we show how we have derived the log-likelihood function for the Altmann distribution.

$$\begin{aligned}
\mathcal{L} &= \sum_{i=1}^{N} \log(ck_i^{-\gamma}e^{-\delta k_i}) \\
&= \sum_{i=1}^{N} \log(c) - \gamma \log(k_i) - \delta k_i \\
&= -\gamma M' - \delta M - N \log(\sum_{k=1}^{N} k^{-\gamma}e^{-\delta k}).
\end{aligned}$$

# References


[Byrd et al., 2003] Byrd, R., Lu, P., Nocedal, J., and Zhu, C. (2003). A limited memory algorithm for bound constrained optimization.

[Clauset et al., 2009] Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*.

[Ferrer i Cancho et al., 2004] Ferrer i Cancho, R., Solé, R. V., and Köhler, R. (2004). Patterns in syntactic dependency networks. *Phys. Rev. E*.

[Naldi, 2015] Naldi, M. (2015). Approximation of the truncated zeta distribution and zipf's law.

[Wagenmakers and Farrell, 2004] Wagenmakers, E.-J. and Farrell, S. (2004). Aic model selection using akaike weights.