Complex and Social Networks: Lab session 2

Model selection for degree distributions

Sergio H. Martínez Mateu & Egon Ferri

Contents

Introduction

In this report we analyse 10 different networks corresponding to real data of syntactic dependency in different languages. The goal is to apply a model selection process, based on maximum likelihood estimation and the AIC criterion, to choose among a set of 6 potential models for the out-degree distribution.

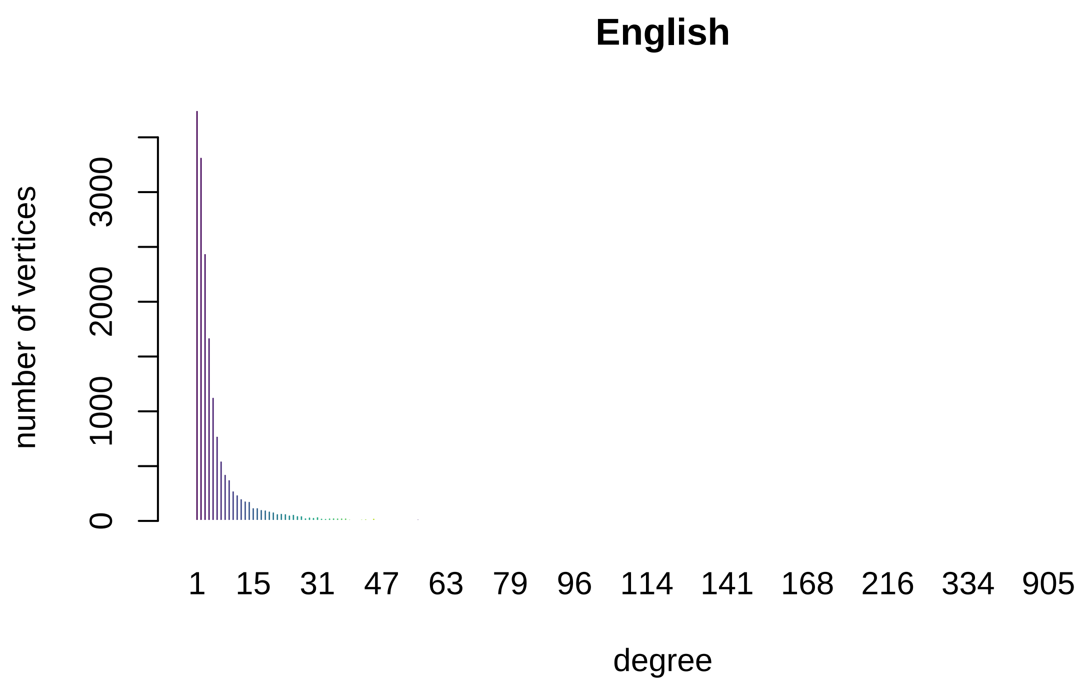
We hid a lot of the code for tidiness, it can be found mostly in the markdown, and in the script "checkthemethods" for the part that check the correctness of our code.

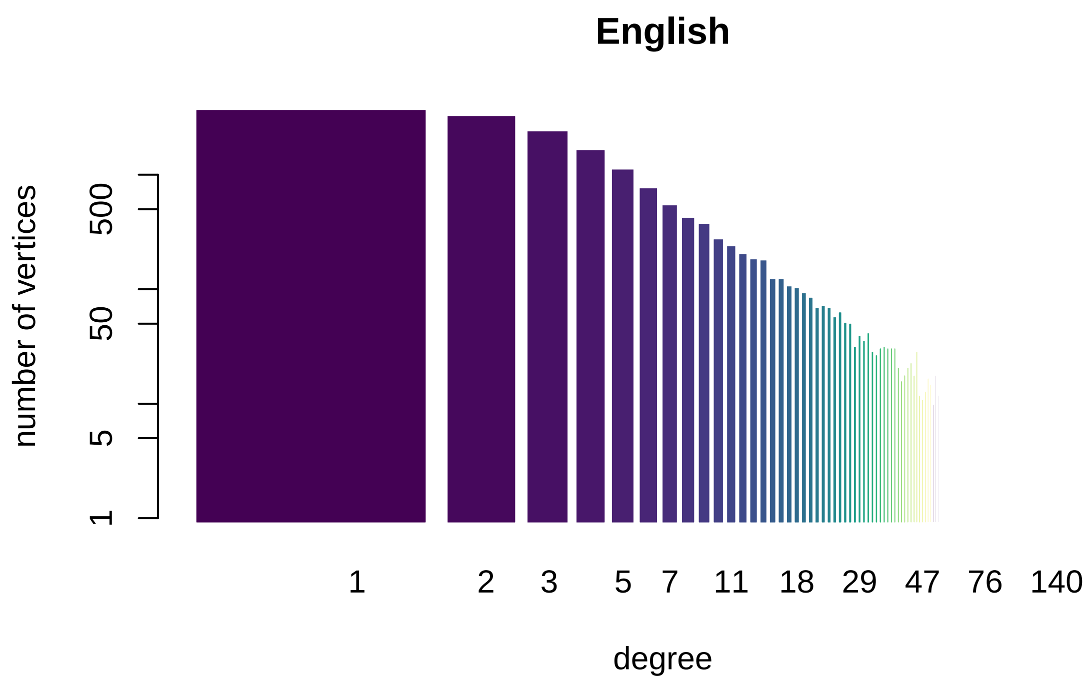
The data

The following table shows some network metrics of the examples under study; N, the maximum degree, the mean degree and its inverse.

```
## Arabic 21532 5743 6.387423 0.1565577
## Basque 12207 2447 4.187433 0.2388098
## Catalan 36865 9880 10.7049 0.0934152
## Chinese 40298 13182 8.987096 0.1112706
## Czech 69303 14551 7.42522 0.1346761
## English 29634 7701 13.03813 0.0766981
## Greek 13283 3317 6.621095 0.1510324
## Hungarian 36126 6586 5.907989 0.1692623
## Italian 14726 2955 7.6113 0.1313836
## Turkish 20409 10180 4.472733 0.223577
```

The following two plots show the out-degree distribution of the English network in the natural and in the log-log scale. The latter version, which shows a linear pattern, allows for a more in-depth inspection of the distribution.


English
number of vertices
0
1000
2000
3000
1 15 31 47 63 79 96 114 141 168 216 334 905
degree


English
number of vertices
1
5
50
500
1 2 3 5 7 11 18 29 47 76 140
degree

Methods

Minus log-likelihood of the models

Here we define some functions that will be used to compute the log-likelihood of the models we want to fit. These functions will be fed later to an optimization algorithm in order to obtain the maximum likelihood estimates of the parameters.

As an extra model to use in this exercise we have chosen the Menzerath-Altmann law model. Here is the derivation of the log-likelihood:

$$\mathcal{L} = \sum_{i=1}^{N} \log(p(k_i)) = \sum_{i=1}^{N} \log(ck_i^{-\gamma} e^{-\delta k_i}) = N \log(c) - \delta \sum_{i=1}^{N} k_i - \gamma \sum_{i=1}^{N} \log(k_i)$$

```
# Model 1: displaced Poisson distribution
minus_log_likelihood_displaced_pois <- function(lambda) {
  C <- sum(sapply(x, function(y) sum(log(2:y))))
  return(-sum(x)*log(lambda)+length(x)*(lambda+log(1-exp(-lambda)))+C)
}
# Model 2: displaced geometric distribution
minus_log_likelihood_geom_displaced <- function(q) {
-(sum(x)-length(x)) * log(1-q) - length(x) * log(q)
}
# Model 3: restricted Zeta distribution
minus_log_likelihood_zeta_restrict <- function() {
  M <- sum(log(x))
  return(2 * M + length(x) * log(pi^2/6))
}
# Model 4: Zeta distribution
minus_log_likelihood_zeta <- function(gamma) {
  M <- sum(log(x))
  return(gamma * M + length(x) * log(zeta(gamma, deriv = 0)))
}
# Model 5: right-truncated Zeta distribution
minus_log_likelihood_zeta_rtrunc <- function(gamma, kmax) {
  M <- sum(log(x))
  #kmax <- length(x)-1
  #kmax <- max(x) # perhaps we need a larger a value (n-1???)
  return(gamma * M + length(x) * log(sum((1:kmax)^(-gamma))))
}
# Model 6: Altmann function
minus_log_likelihood_altmann <- function(gamma, delta) {
  cinv <- sum(sapply(1:length(x),function(k) k^(-gamma)*exp(-delta*k)))
  return(delta * sum(x) + gamma * sum(log(x)) + length(x) * log(cinv))
}
```

Sample size corrected AIC

As suggested, we used the sample size corrected version of the AIC:

```
get_AIC <- function(m2logL,K,N) {
m2logL + 2*K*N/(N-K-1)
}
```

Optimization

In order to find the maximum likelihood estimates we basically followed the given instructions. We used the function `mle` and the method `L-BFGS-B`. The bounds of the parameters were used if they were known, and in most of cases are trivial, except for the kmax of the right-truncated zeta and for the parameters of the Menzerath-altmann. For the right-truncated Zeta we used the maximum degree as the lower bound (since it's the lowest value possible) and, after a lot of trials and repeat, we used again the maximum degree as a starting value. The optimization seems very hard, and this choice seems to return always that value as the optimum; but, comparing to other parameters set, this is the set-up that performed better. For the Menzerath-Altmann model we explored different initial seeds and found that delta had to be positive but close to 0, while gamma worked well starting from 1. The optimization worked without bounds for all languages except for Turkish. We solved this by specifying 0 as lower bound for both γ and δ.

Checking the methods

Before applying the selection procedure to the real data, we perform here a control analysis. Using networks that were generated according to given degree distributions, we want check two things:

- That the selected model(the one with minimum AIC) coincides with the correct one
- That the MLE parameter of the selected model is close to the correct one

The results are:

	AIC differences with respect to the best AIC in our ensemble of models
geometric_with_parameter_0.05	1.158049e+04
geometric_with_parameter_0.1	5.322284e+03
geometric_with_parameter_0.2	1.643985e+03
geometric_with_parameter_0.4	8.052070e+02
geometric_with_parameter_0.8	1.116876e+03
zeta_with_parameter_1.5	2.810225e+08
zeta_with_parameter_2.5	2.489267e+03
zeta_with_parameter_2	1.693806e+04
zeta_with_parameter_3.5	1.835160e+03
zeta_with_parameter_3	1.721096e+03

	Estimated parameter
geometric_with_parameter_0.05	0.0519696
geometric_with_parameter_0.1	0.0971062
geometric_with_parameter_0.2	0.1992826
geometric_with_parameter_0.4	0.4024145
geometric_with_parameter_0.8	0.7898894
zeta_with_parameter_1.5	1.4954914
zeta_with_parameter_2.5	2.4507700
zeta_with_parameter_2	1.9802979
zeta_with_parameter_3.5	3.3541069
zeta_with_parameter_3	2.9960233

In addiction, we saw that in some cases the right-truncated Zeta distribution is chosen instead of the Zeta distribution, although the AIC difference between them is generally pretty small. This might be explained by the fact that the right-truncated Zeta, which is a generalization of the Zeta, can be in practice very similar to the Zeta. Regarding the value of the optimized parameters, we see that indeed they are pretty close to the known values. This is a good indication, and gives us enough confidence to go to the next step.

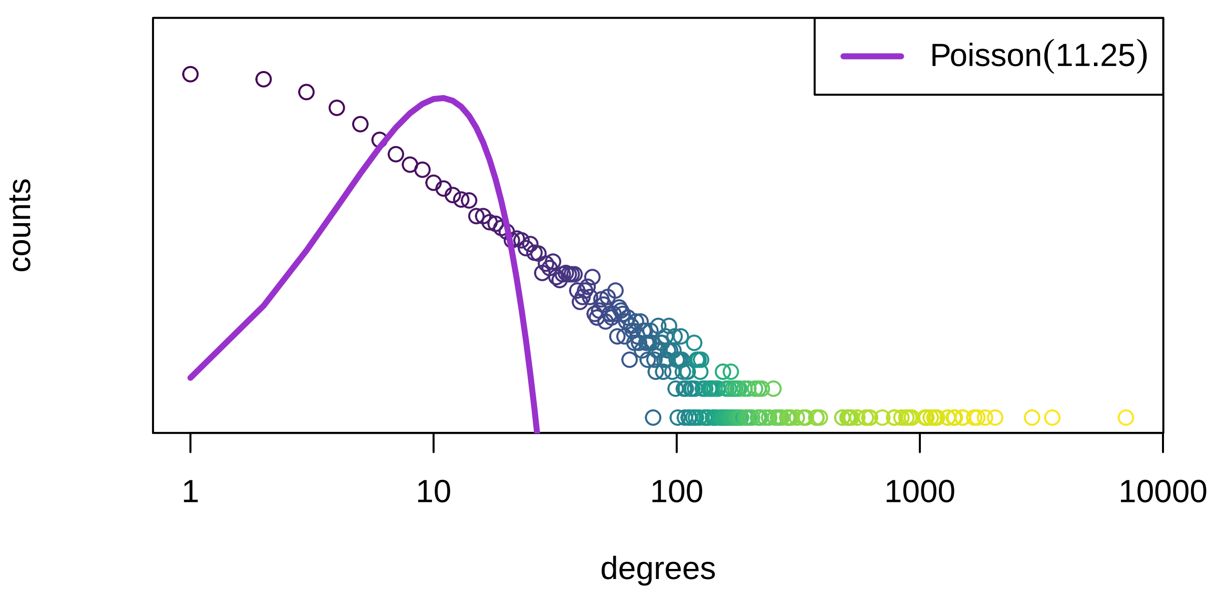
Results

First example: 'english' network

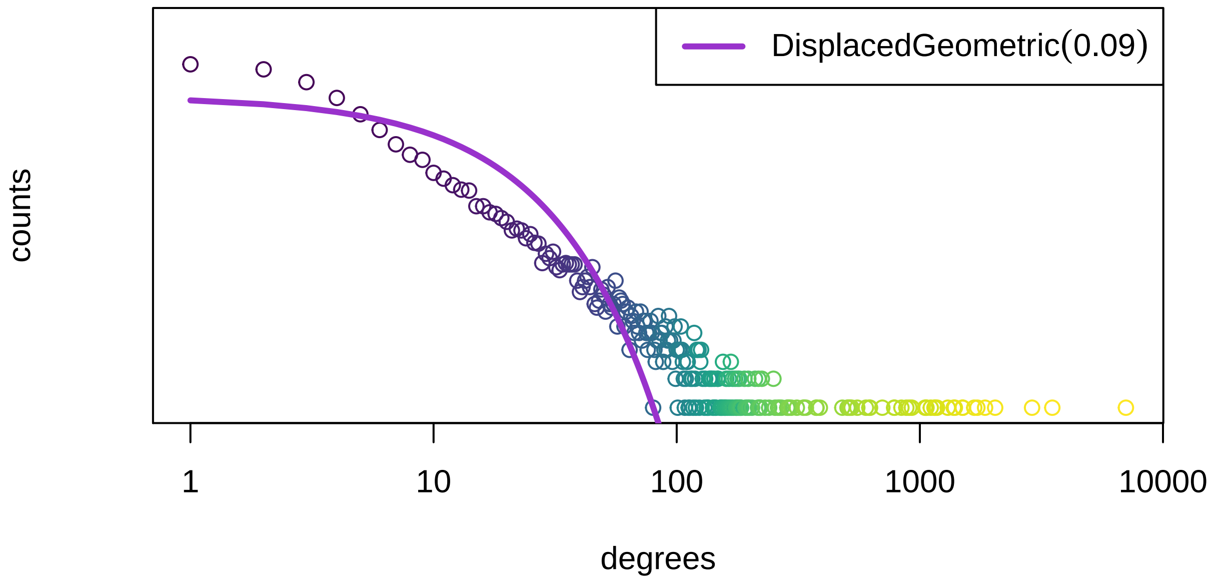
```
# 'english' network degree sequence
x <- read.table("data/data_out/English_out-degree_sequence.txt",
                                header = FALSE)$V1
# These will be used later in the plots
x_list <- 1:max(x)
degree_spectrum <- table(x)
counts <- unname(degree_spectrum)
degrees <- as.numeric(names(degree_spectrum))
```

Model 1: displaced Poisson

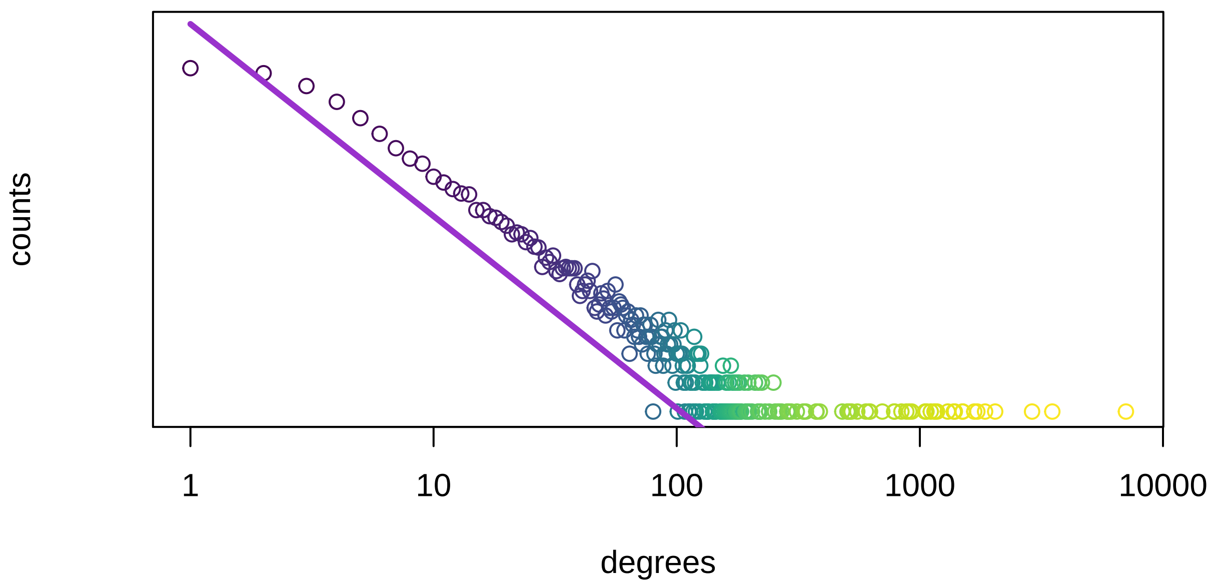
```
# Initial values
lambda0 <- list(sum(x)/length(x))

# Fit
displ_pois <- mle(minuslogl = minus_log_likelihood_displaced_pois,
              start = list(lambda = lambda0),
              method = "L-BFGS-B",
              lower = 1e-7)

# MLE estimate
lambda_opt <- coef(displ_pois)
```



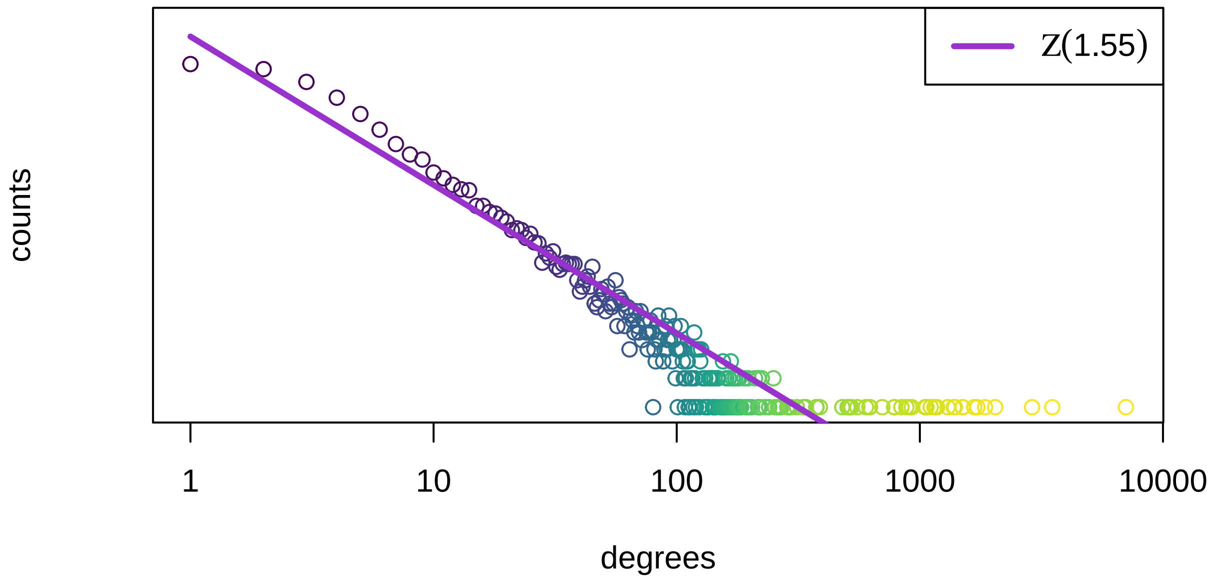
Model 2: Displaced geometric distribution



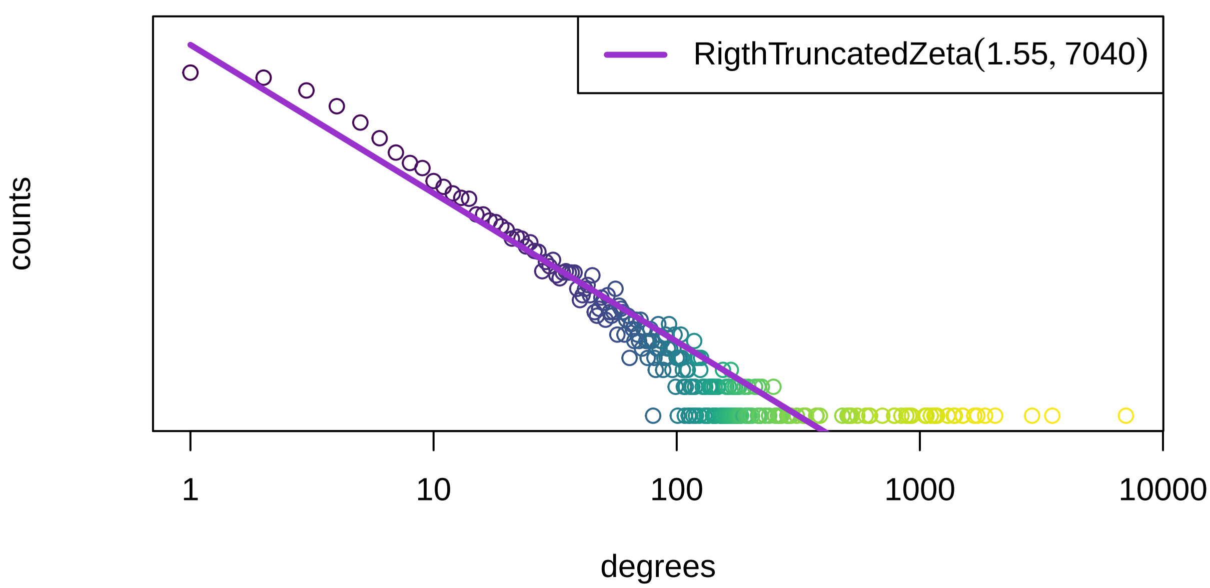
Model 3: Restricted Zeta



Model 4: Zeta



Model 5: right-truncated Zeta



Model 6: Menzerath-Altmann law

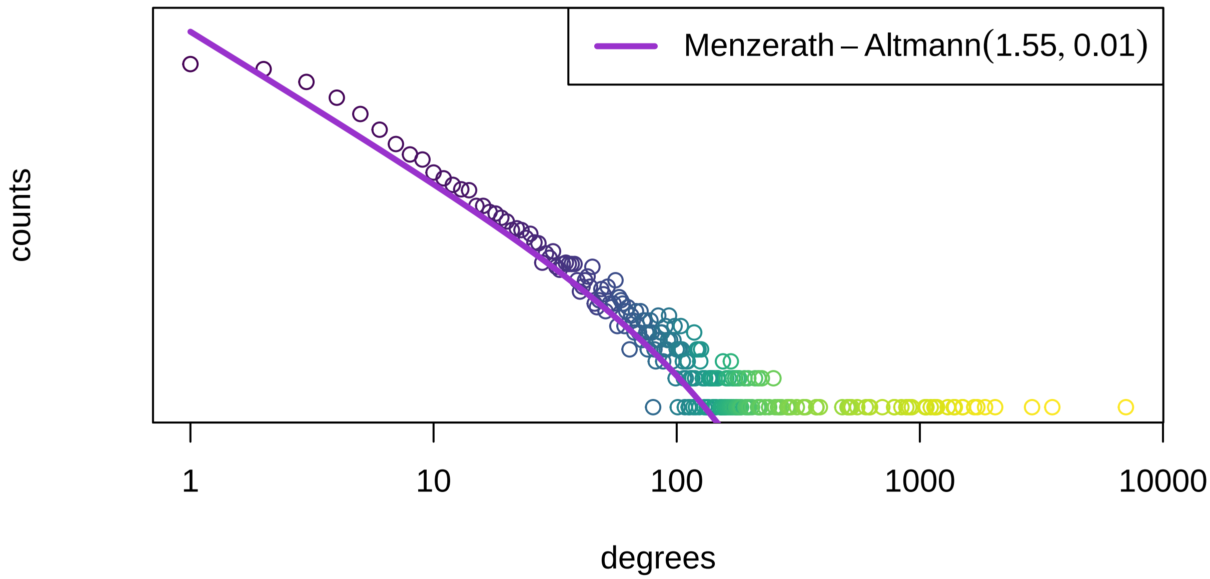


Table 3: best fit for model parametrs

	1: λ	2: q	4: γ	$5:\gamma_2$	$kmax$	6: γ_3	δ
English	11.25392	0.0888568	1.545278	1.524973	7040	1.256093	0.0114628

AIC comparison

Table 4: AIC differences with respect to the best AIC in our ensemble of models

	Displ. Poisson	Displ. Geom	Restricted Zeta	Zeta	R-T Zeta	Menzerath-Altmann
English	648870.5	16531.39	9920.595	2322.885	2091.383	0

This table represent the so-called AIC difference $\Delta = AIC - AIC_{best}$,

Is shown that for the English language the best fit is given by the Altman function, closely followed by the zeta function and the truncated zeta function. Let's now repeat the analysis for the other language to check if results hold.

Analysis for the 10 languages

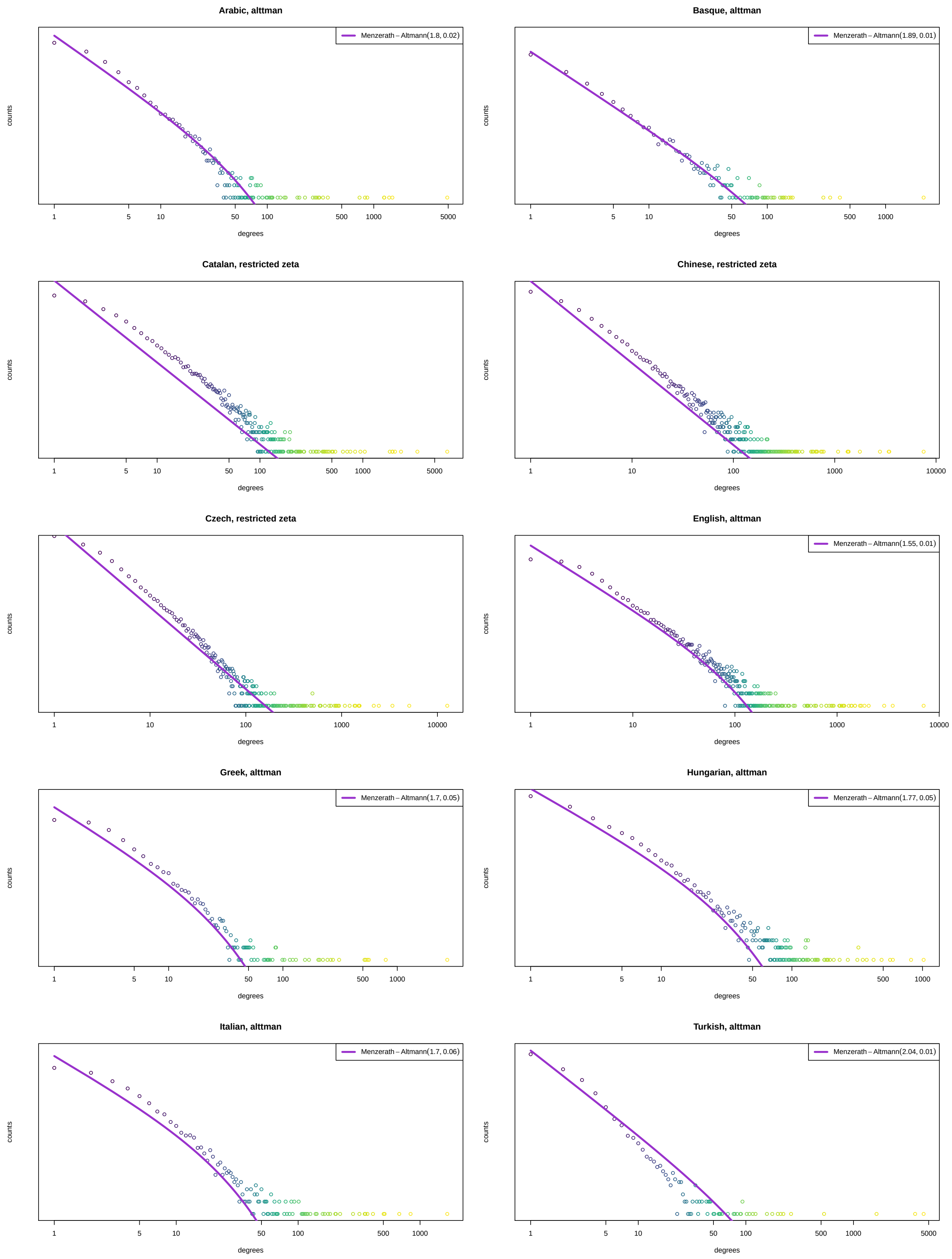


Table 5: AIC differences with respect to the best AIC in our ensemble of models

	Displ. Poisson	Displ. Geom	Restricted Zeta	Zeta	R-T Zeta	Menzerath-Altmann
Arabic	204373.27	10574.561	49180.588	753.5464	731.25678	0.000
Basque	67206.64	5568.399	90468.012	102.3919	95.26607	0.000
Catalan	564879.32	37343.652	0.000	23273.9218	23061.06465	19959.255
Chinese	610233.13	29203.043	0.000	5471.4876	5378.13894	4104.483
Czech	910635.60	117068.726	0.000	86504.2919	86418.24498	83599.295
English	648870.46	16531.394	9920.597	2322.8867	2091.38471	0.000
Greek	91832.75	3282.089	70955.551	1340.0342	1288.07204	0.000
Hungarian	166975.79	10498.974	6176.303	2447.7938	2272.23723	0.000
Italian	97322.06	3772.553	57808.101	1914.8306	1819.04030	0.000
Turkish	166595.28	11708.776	65638.358	114.2351	113.38155	0.000

Table 6: best fit for model parametrs

	1: λ	2: q	4: γ	$5: \gamma_2$	$kmax$	6: γ_3	δ
Arabic	4.449833	0.2221026	1.797628	1.792754	4896	1.554883	0.0223993
Basque	4.113253	0.2391405	1.887150	1.881876	2097	1.763402	0.0105570
Catalan	8.251780	0.1211544	1.590979	1.575434	6622	1.248778	0.0192140
Chinese	7.722839	0.1294287	1.662662	1.653665	7537	1.466832	0.0099200
Czech	6.244246	0.1598365	1.690866	1.685455	12671	1.439131	0.0168688
English	11.253920	0.0888568	1.545277	1.524973	7040	1.255839	0.0114606
Greek	4.783788	0.2072909	1.699111	1.685881	2737	1.195924	0.0528793
Hungarian	4.129952	0.2382392	1.769320	1.752150	1020	1.352568	0.0474952
Italian	4.578370	0.2161748	1.704723	1.687240	1671	1.156100	0.0614923
Turkish	2.920239	0.3239732	2.042634	2.041608	4488	1.949726	0.0105829

Discussion

The methods we have applied throughout the analysis worked well on simulated test data, which gives us enough confidence to draw the following conclusions about the out-degree distributions in the real data sets under study.

It can be observed that the languages form somehow two "clusters":

- A bigger one that, as seen with English language, is well described by an Altman function (and also by the zeta and the zeta truncated) and contains all the languages except three.
- A smaller one that contains only Catalan, Chinese and Czech that is fitted better by a restricted zeta.

Another interesting remarks is that There is always a small difference between the Zeta and the right-truncated Zeta models. This is in agreement with visual fitting observations in that they yield very similar fits. The later version improves with respect to the former one enough to compensate the additional parameter that it has. However, finding the MLE of the right-truncated Zeta can be problematic in terms of numerical optimization.

In the end, we can conclude by saying that for degree distributions that are close to linear in the log-log scale the Poisson and the geometric distributions seems not appropriate. Instead, either Zeta-related distributions or the Menzerath-Altmann law seem to perform way better.