UPC - Complex and Social Networks

# Lab 2: Analysis of the degree distribution

Michele Gentili and Simon Van den Eynde

October 13, 2016

## 1 Introduction

In this lab we will try to select a theoretic model, fitting the degree distribution of syntactic dependency networks in different languages. We will focus on in-degrees.

We considered 5 theoretical models and used the Akaike information criterions (AIC) to decide which model was most favourable. The models we considered were

1. a zeta distribution
2. a zeta distribution with fixed exponent 2
3. a right-truncated zeta distribution
4. a geometric distribution
5. a geometric displaced distribution
6. a displaced poisson distribution

After choosing the statistically best models, we plotted our models to our data to do a visual check of correctness.

And finally recompute the Akaike scores, introducing a better model: the Altmann distribution.

# 2 Results

In table 1 we find our main result, the difference in AIC from the best (of the 5) model. We notice that except for Hungarian and Turkish the right-truncated zeta always ends up best.

Table 1: Difference in AIC for the 5 different models in different languages

| | Models | | | | |
|---|---|---|---|---|---|
| Language | 1 | 2 | 3 | 4 | 5 |
| Arabic | 2.91 | 175.28 | 0.00 | 24188.95 | 240321.25 |
| Basque | 0.24 | 845.41 | 0.00 | 8364.11 | 50164.59 |
| Catalan | 13.96 | 225.90 | 0.00 | 61881.73 | 913880.09 |
| Chinese | 14.69 | 503.99 | 0.00 | 48678.82 | 618348.40 |
| Czech | 8.85 | 147.28 | 0.00 | 91099.39 | 940677.71 |
| English | 45.05 | 1469.66 | 0.00 | 45742.18 | 739581.47 |
| Greek | 3.43 | 164.11 | 0.00 | 17133.83 | 157137.24 |
| Hungarian | 0.00 | 2220.28 | 1.71 | 53469.62 | 468252.20 |
| Italian | 0.49 | 118.37 | 0.00 | 22877.90 | 245803.51 |
| Turkish | 0.00 | 2740.47 | 1.98 | 24615.35 | 193345.03 |

In table 2 we find the RSS-values for the zeta and right-truncated zeta distribution. We notice that in both table 1 and 2 most values for the zeta and right-truncated zeta are very alike.

Table 2: RSS values for zeta and right-truncated zeta distribution with optimised parameters

| | zeta | RT_zeta |
|---|---|---|
| Arabic | 9.54e-04 | 9.33e-04 |
| Basque | 1.68e-05 | 1.78e-05 |
| Catalan | 1.12e-04 | 1.03e-04 |
| Chinese | 5.21e-04 | 4.94e-04 |
| Czech | 4.12e-05 | 3.82e-05 |
| English | 3.46e-04 | 2.90e-04 |
| Greek | 4.38e-04 | 4.15e-04 |
| Hungarian | 3.93e-04 | 3.94e-04 |
| Italian | 1.82e-04 | 1.75e-04 |
| Turkish | 6.92e-05 | 6.92e-05 |

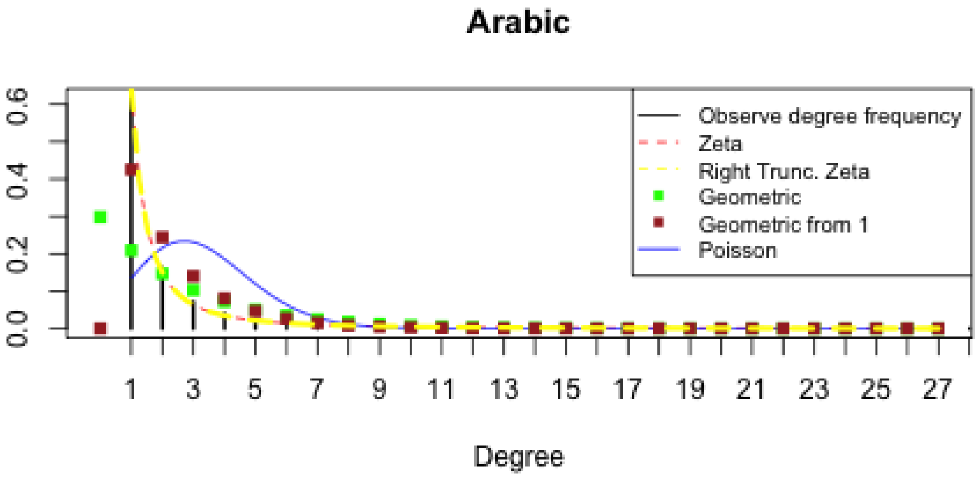
In table 3 we added two more models, the Altmann model and a corrected geometric model (see also Method section). As we expected has the Altmann function best score, so all the other values are shifted of the difference between the previous best model score and the Altmann score.

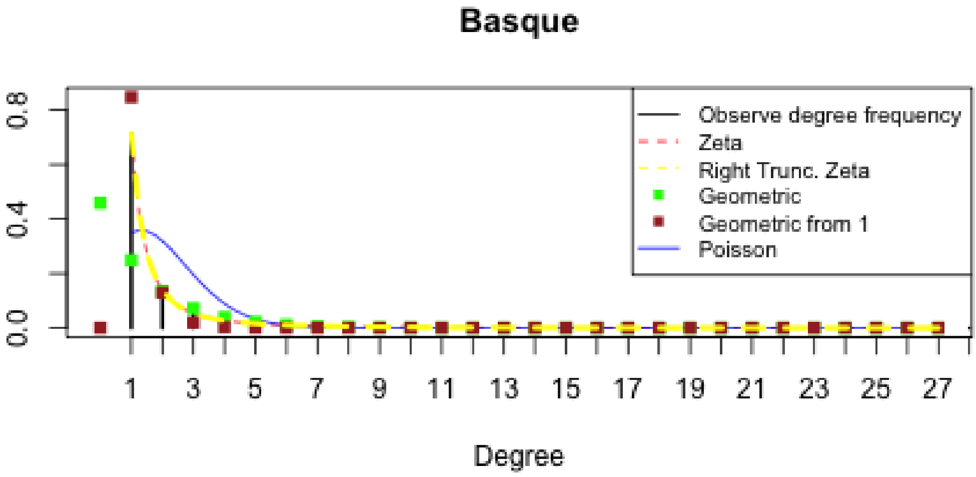
Table 3: Difference in AIC for the 5 different models + 2 extra models in different languages

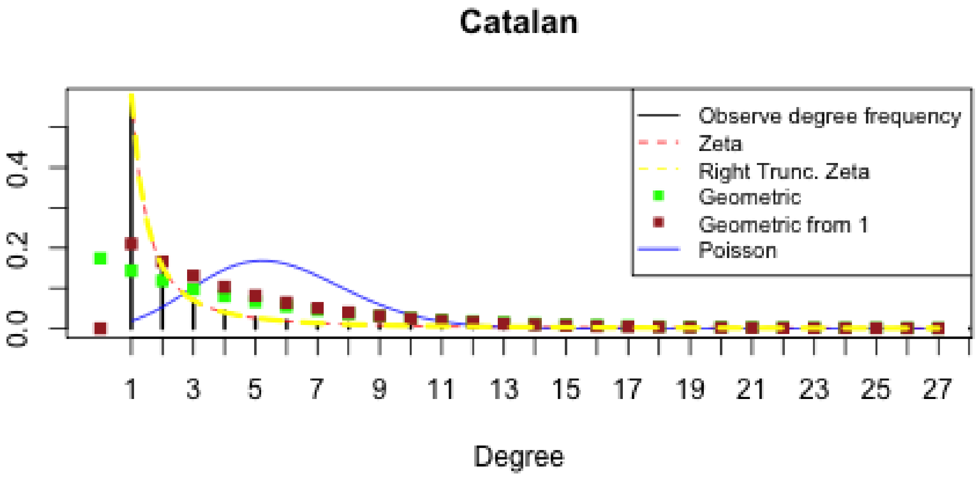
| | zeta | zeta_2 | RT_zeta | geom | poisson | geom_corrected | Altmann |
|---:|---:|---:|---:|---:|---:|---:|---:|
| Arabic | 132092 | 132264 | 132089 | 156278 | 372410 | 137774 | 0 |
| Basque | 58151 | 58996 | 58151 | 66515 | 108315 | 42796 | 0 |
| Catalan | 271389 | 271601 | 271375 | 333256 | 1185255 | 318171 | 0 |
| Chinese | 277583 | 278072 | 277568 | 326247 | 895917 | 309159 | 0 |
| Czech | 439807 | 439946 | 439799 | 530898 | 1380476 | 485598 | 0 |
| English | 251394 | 252819 | 251349 | 297091 | 990930 | 287081 | 0 |
| Greek | 78708 | 78869 | 78705 | 95839 | 235842 | 85490 | 0 |
| Hungarian | 185034 | 187254 | 185036 | 238504 | 653286 | 204135 | 0 |
| Italian | 86617 | 86735 | 86617 | 109495 | 332420 | 101040 | 0 |
| Turkish | 91486 | 94227 | 91488 | 116101 | 284831 | 80961 | 0 |

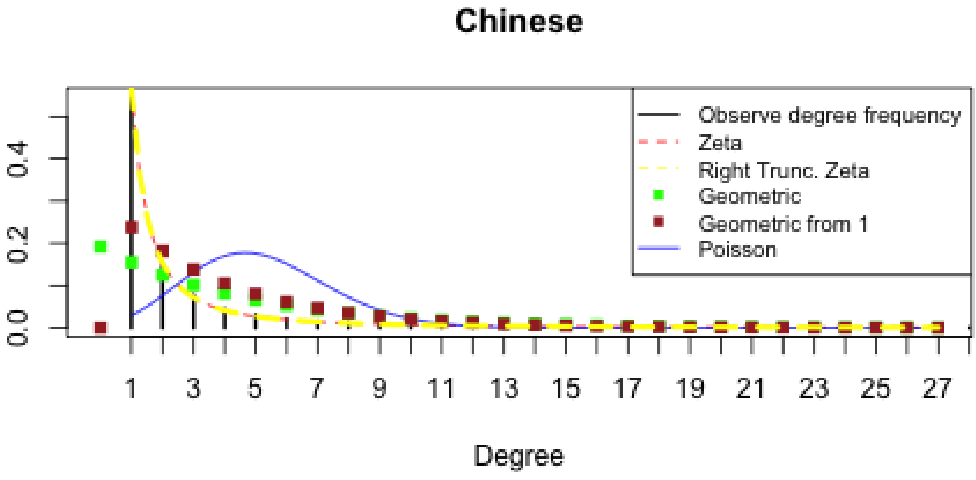
Hereafter we added a plot for every language, plotting the fit of some distributions on the data.

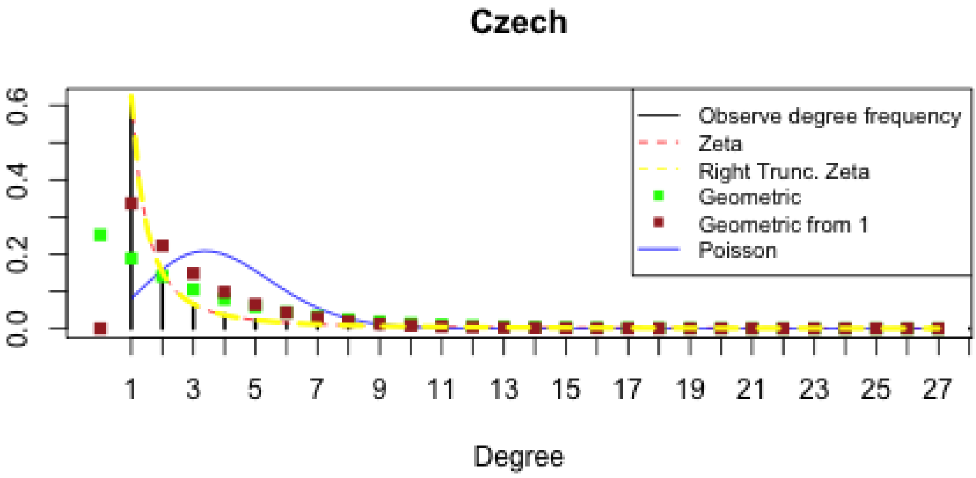
And we also added some extra plots regarding the Altmann distribution.

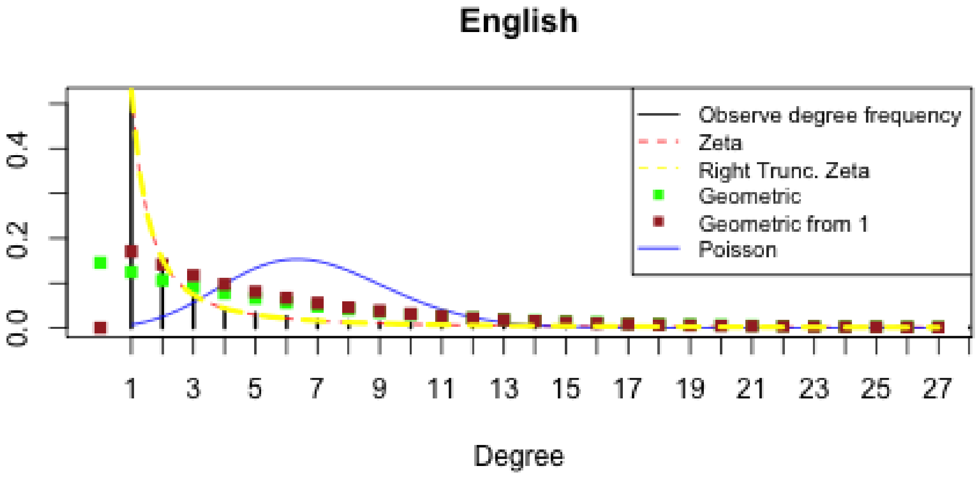

Arabic
0.0
0.2
0.4
0.6
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree
Observe degree frequency
Zeta
Right Trunc. Zeta
Geometric
Geometric from 1
Poisson

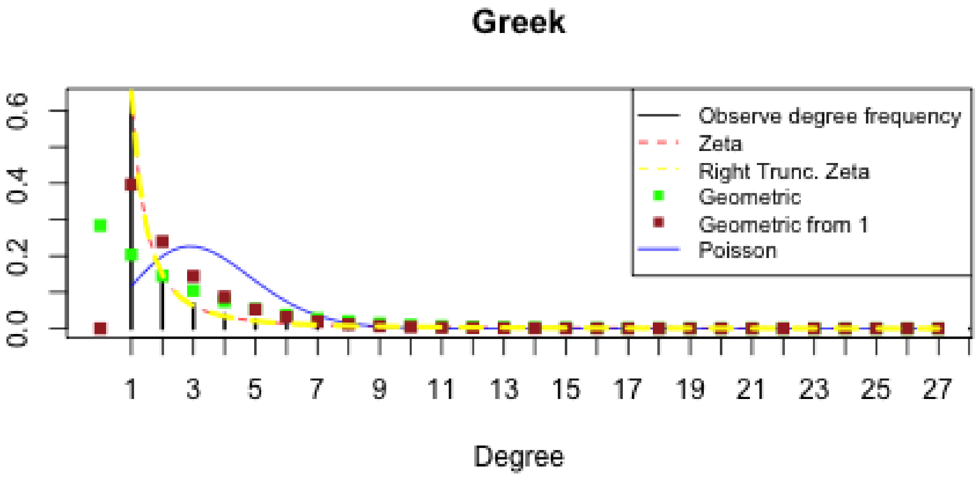

Basque
0.0
0.4
0.8
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree
Observe degree frequency
Zeta
Right Trunc. Zeta
Geometric
Geometric from 1
Poisson

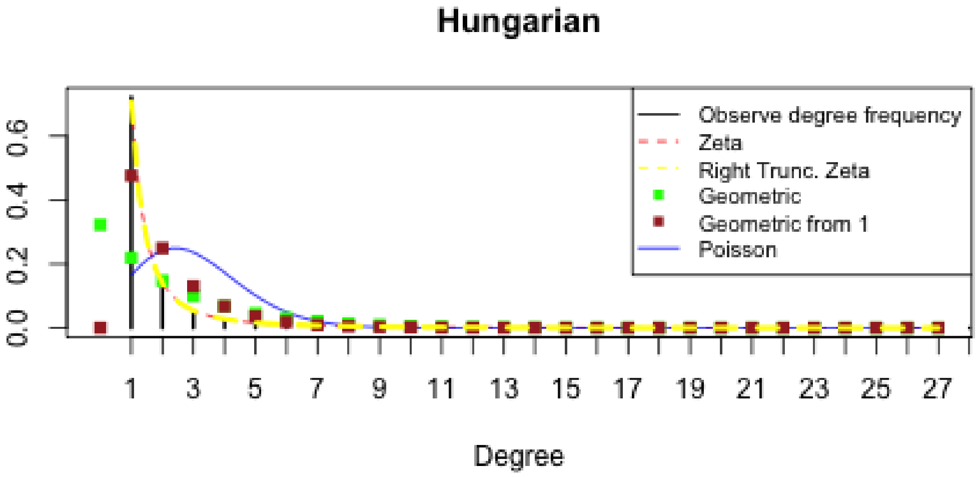

Catalan
Observe degree frequency
Zeta
Right Trunc. Zeta
Geometric
Geometric from 1
Poisson
0.0
0.2
0.4
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree

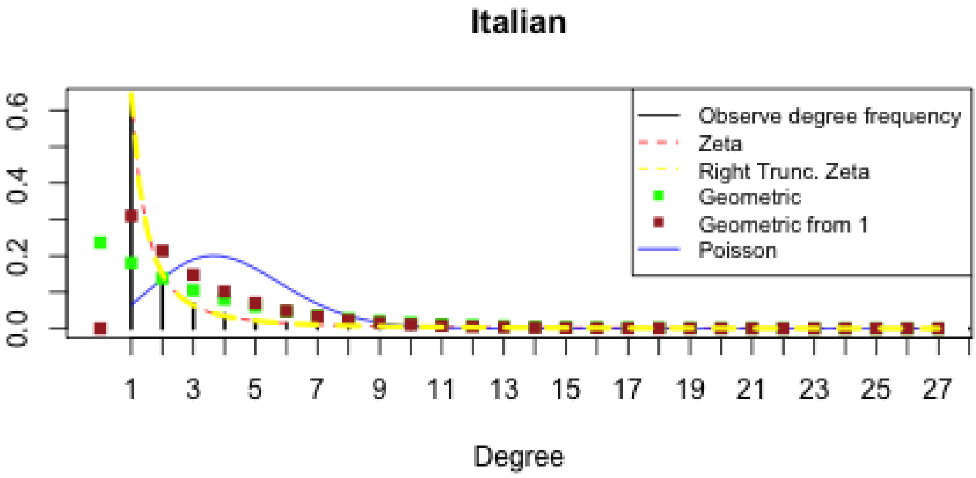

Chinese
Observe degree frequency
Zeta
Right Trunc. Zeta
Geometric
Geometric from 1
Poisson
0.0
0.2
0.4
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree

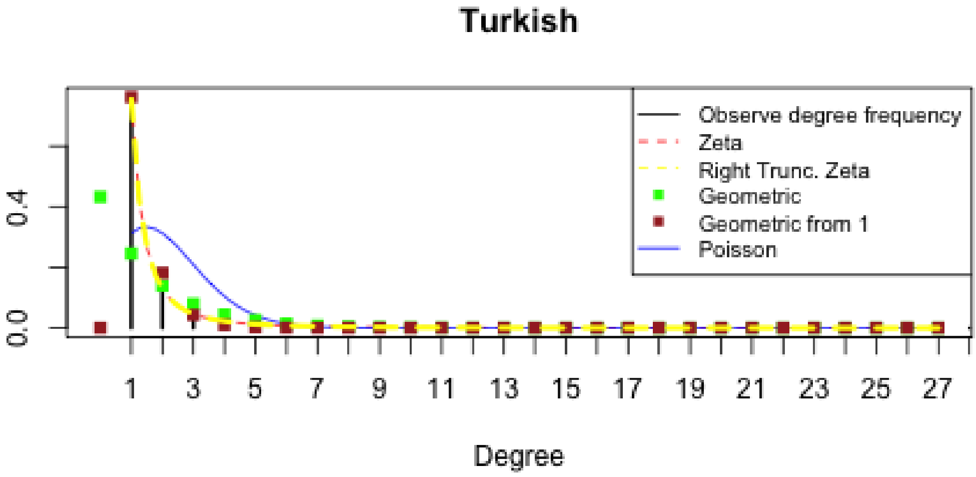

Czech
0.0
0.2
0.4
0.6
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree
Observe degree frequency
Zeta
Right Trunc. Zeta
Geometric
Geometric from 1
Poisson

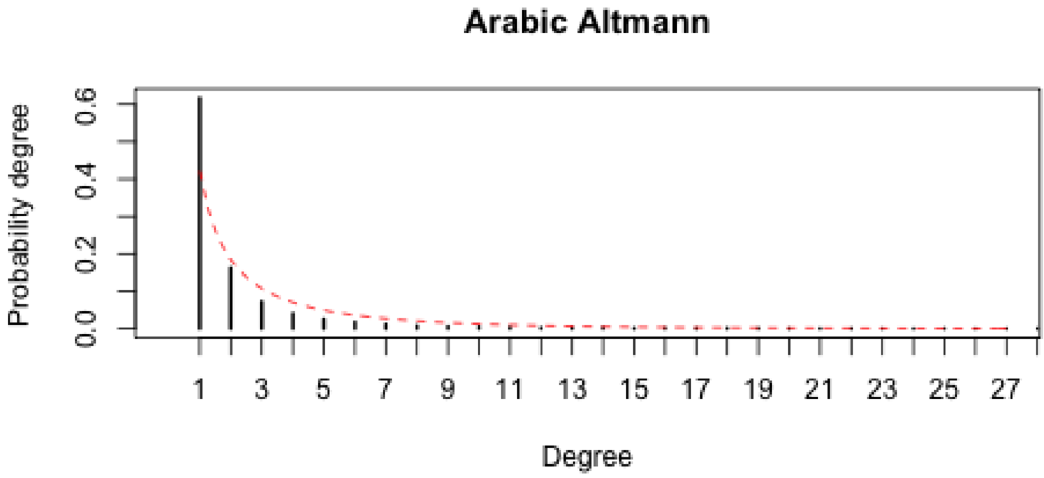

English
0.0
0.2
0.4
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree
Observe degree frequency
Zeta
Right Trunc. Zeta
Geometric
Geometric from 1
Poisson

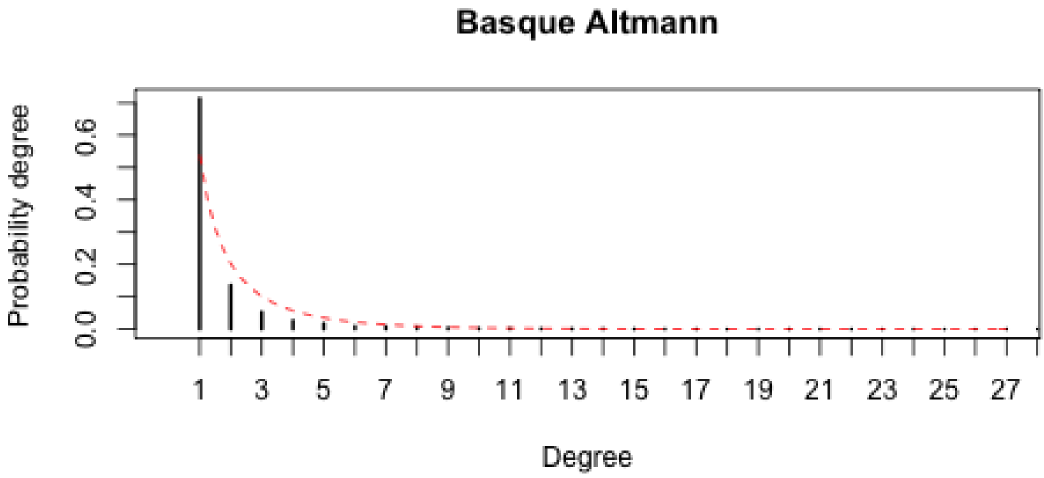

Greek
0.0
0.2
0.4
0.6
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree
Observe degree frequency
Zeta
Right Trunc. Zeta
Geometric
Geometric from 1
Poisson


Hungarian
0.0
0.2
0.4
0.6
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree
Observe degree frequency
Zeta
Right Trunc. Zeta
Geometric
Geometric from 1
Poisson

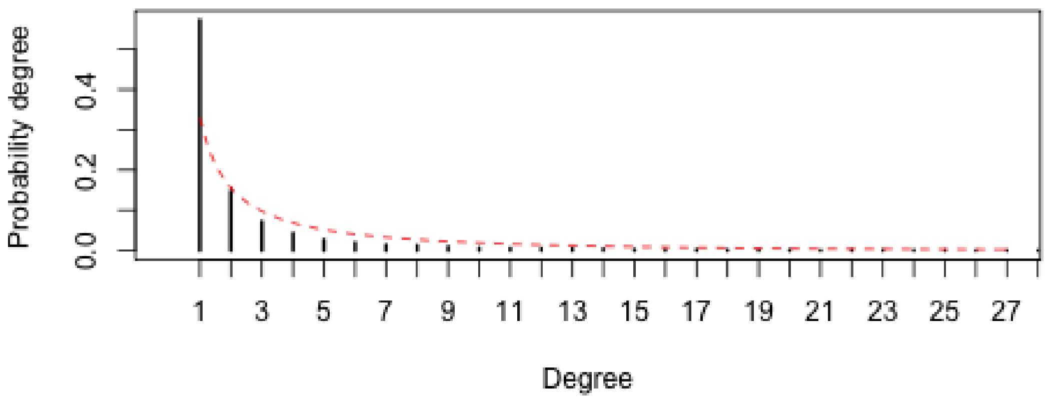

Italian
0.0
0.2
0.4
0.6
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree
Observe degree frequency
Zeta
Right Trunc. Zeta
Geometric
Geometric from 1
Poisson


Turkish
0.0
0.4
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree
Observe degree frequency
Zeta
Right Trunc. Zeta
Geometric
Geometric from 1
Poisson

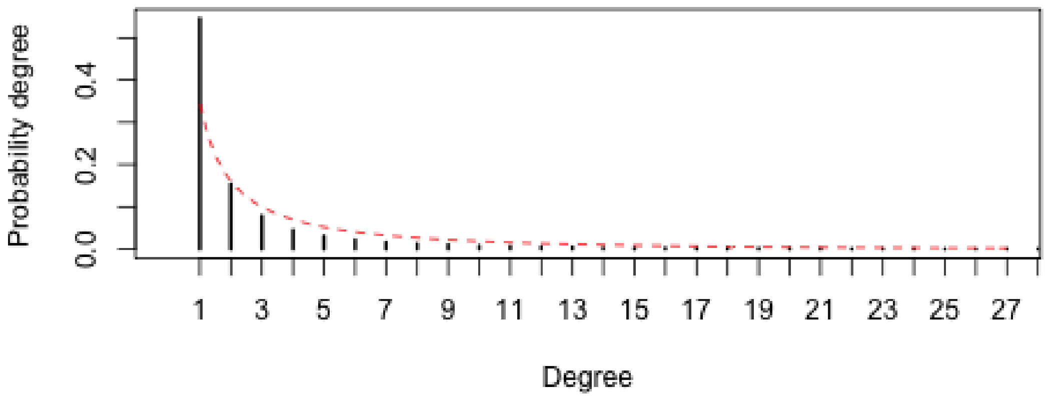

Arabic Altmann
Probability degree
0.0
0.2
0.4
0.6
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree

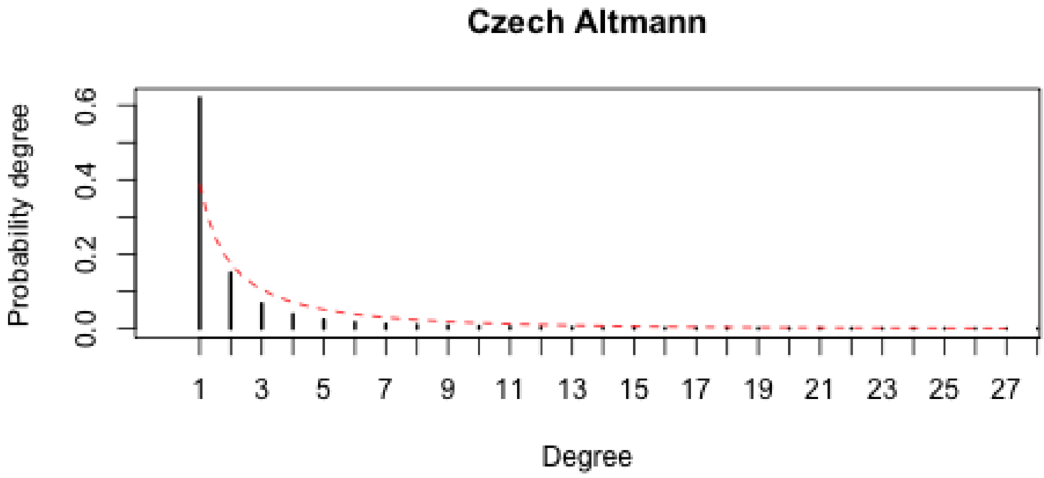

Basque Altmann
Probability degree
0.0
0.2
0.4
0.6
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree

Catalan Altmann

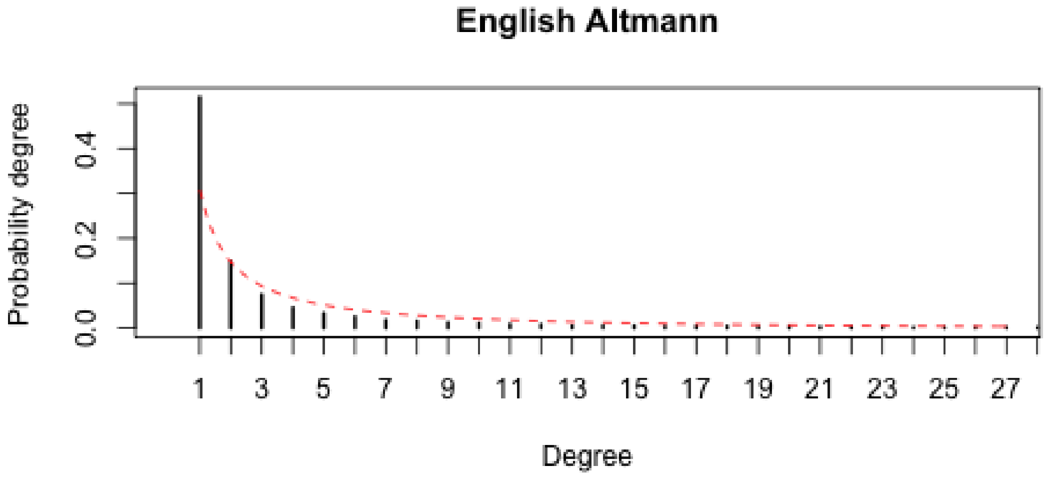


Chinese Altmann

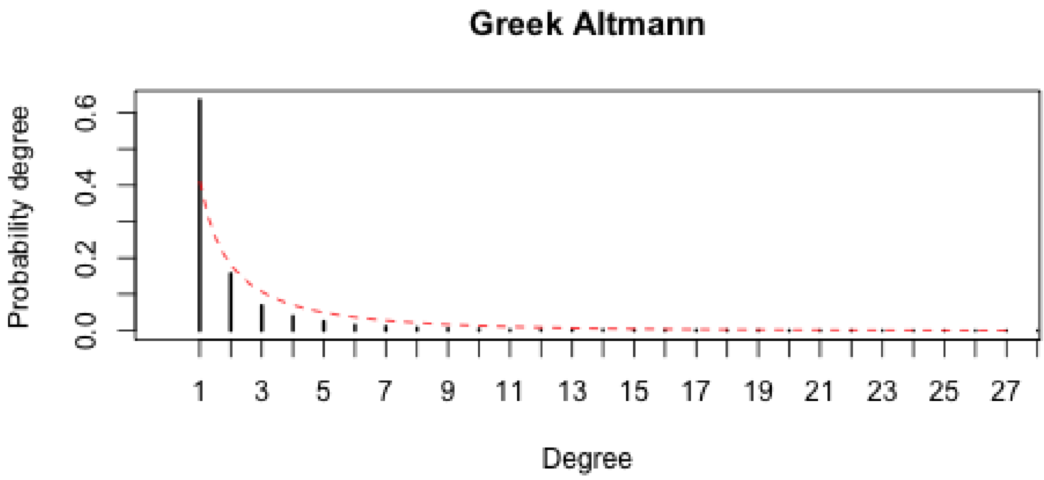


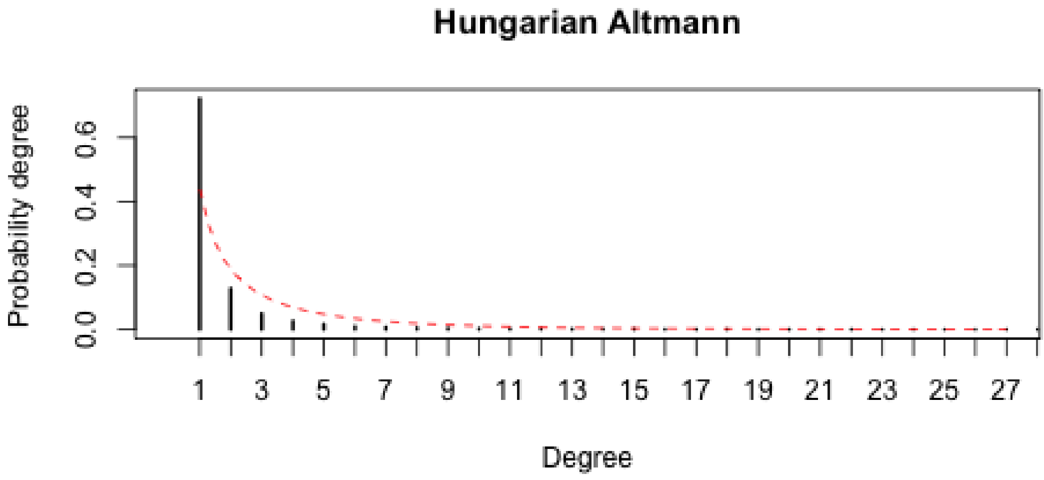


Czech Altmann



English Altmann


Greek Altmann
Probability degree
0.0
0.2
0.4
0.6
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree


Hungarian Altmann
Probability degree
0.0
0.2
0.4
0.6
1
3
5
7
9
11
13
15
17
19
21
23
25
27
Degree

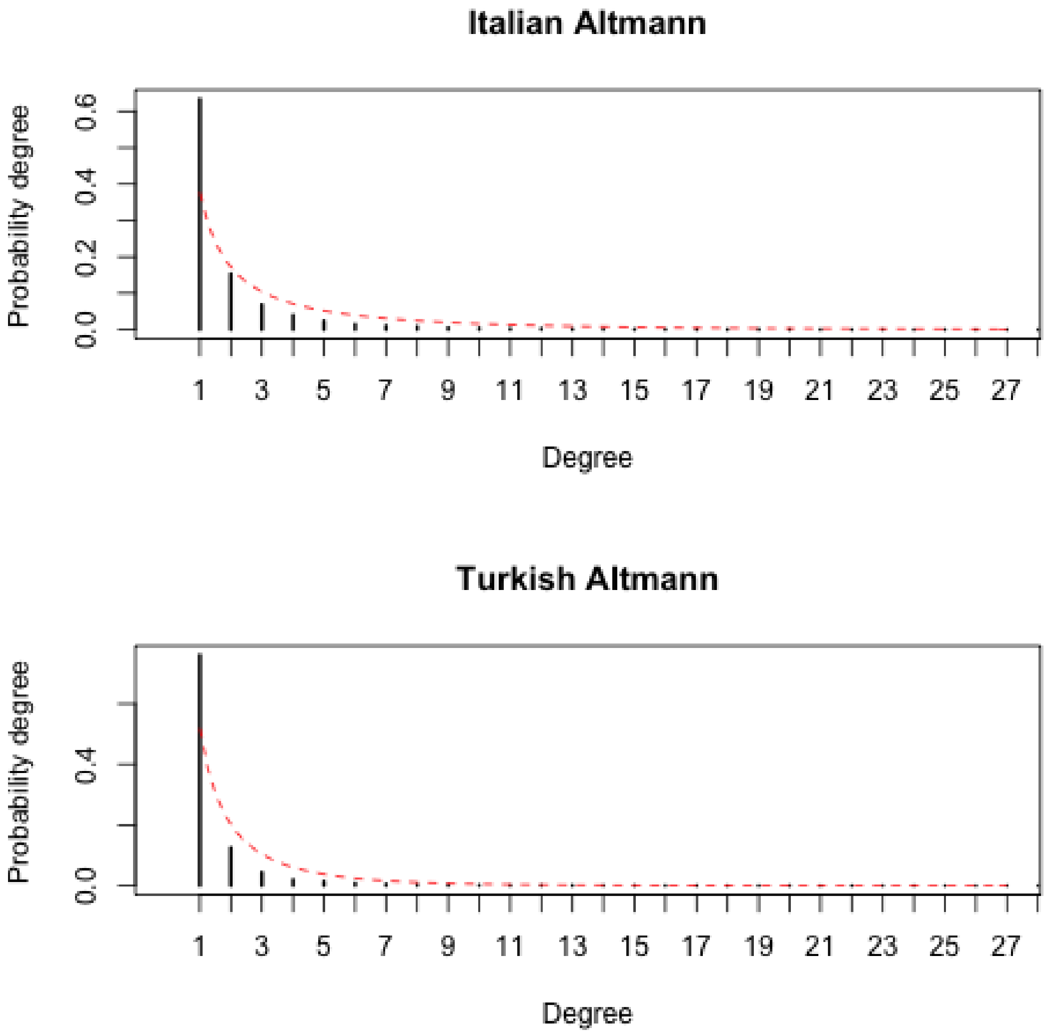


Figure 1: example caption

# 3 Discussion

When optimising model 3 we noticed that the value of $k_{max}$ stays its starting value and does not change, while manually changing the starting value of $k_{max}$ and thus the value obtained by optimising, significantly improved the AIC-value and the RSS-value. For the english-in-degree-sequence reduces changing $k_{max}$ manually from max degree to 778 the RSS by over 30%.

On the plots we see that the Altmann distribution and the (right-truncated)-zeta distribution fit the probabilities from the data well. We prefer the Altmann distribution over the the zeta-distributions because the Altmann distribution has a lower AIC.

## 3.1 Conclusion

The best model is the Altmann model, the zeta function with respect to the previous functions works better because it behaves good on the tails, the geometric distribution, even if shifted to 1, 'loses' too much probability on the tails. Finally as was expected, the difference in AIC has changed as soon as you introduce a better model.

Because the Altmann distribution also gives a reasonable fit on the plot, we think that using the altmann distribution is an appropriate way to model the data.

# 4 Methods

When using the “L-BFGS-B” optimisation method for the right truncated zeta model (model 3), we often encountered errors. We decided to change the optimisation method to the conjugate gradients method “CG”. Here we couldn't set any lower bound, but that didn't seem to give any problem for our data.

We noticed our geometric distribution gave a weight to the value 0. So we decided to make a corrected geometric distribution, by shifting the data for this distribution by 1. We called this distribution geom_corrected.

Calculating the mle-function for the Altmann distribution didn't really work out either. So we calculated the optimised parameters 'manually' with the optim-function, which is used inside the mle-function.