Lab 2: Analysis of the degree distribution

Michele Gentili and Simon Van den Eynde

October 12, 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 displaced geometric distribution
5. a displaced poisson distribution

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

At last we considered of another might be better than the previous selected models.

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

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%.

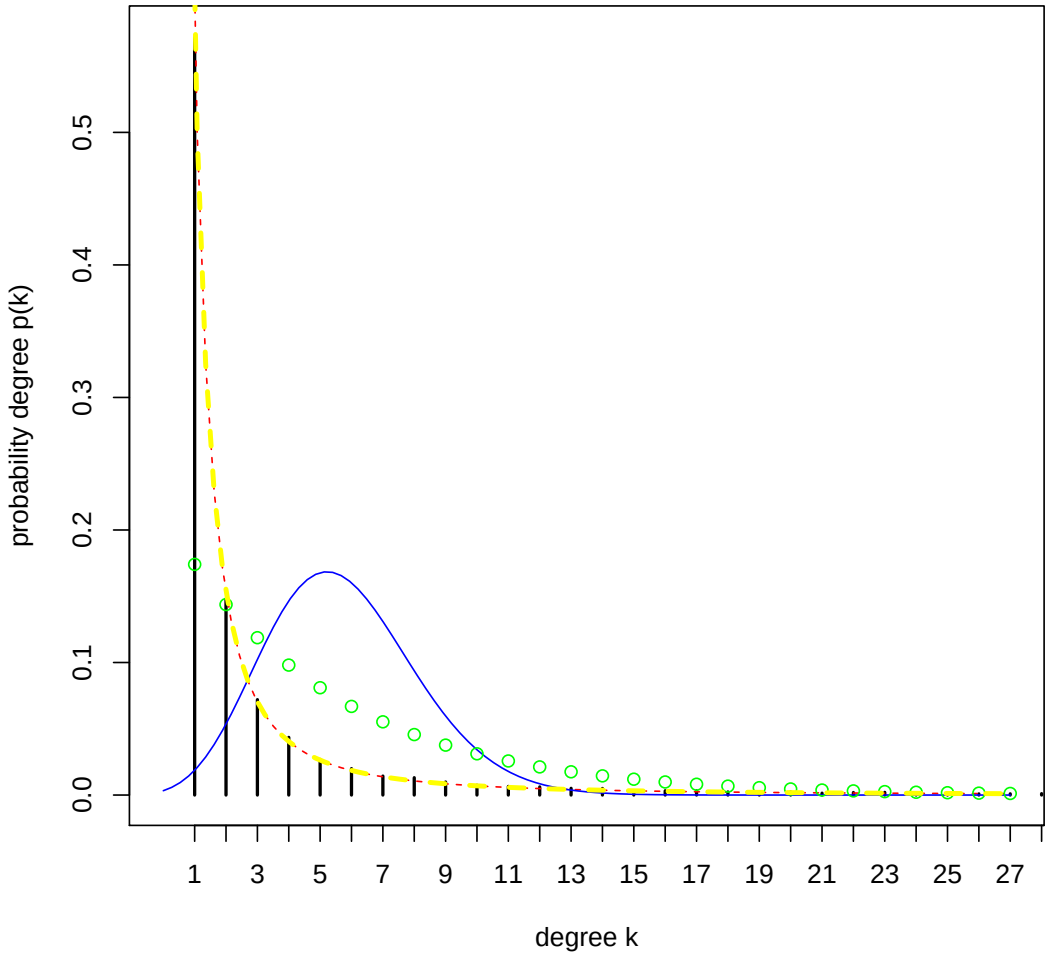

Figure 1: example caption

3.1 Conclusion

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.