Complex and Social Networks
Assignment 2 - Analysis of the degree distribution

Daniel Benedí García
José Ángel Vicente Porres

January, 2024

1 Introduction

This study initiates an investigation into the analysis of degree distributions within the global syntactic dependency networks of various languages. These networks serve as depictions of linguistic structures, where individual vertices are associated with lexical units, specifically words, while edges signify syntactic relationships, ascertained through their occurrence in dependency treebanks. Our research is primarily directed towards the examination of in-degree distributions, which provide valuable perspectives on the prevalence and structural arrangement of incoming syntactic dependencies within these intricate linguistic networks.

The primary aim of this study is to infer the underlying rules governing the observed degree distributions. To achieve this objective, the central focus will involve a comparative analysis between the empirical in-degree distributions and a predefined set of mathematical models. To ascertain the best possible fit to the observed data, we will employ the Maximum Likelihood Estimation (MLE) technique, which will aid in determining which model(s) most accurately capture the distribution patterns observed in the linguistic networks under investigation. This analytical approach will facilitate a deeper understanding of the organizational principles governing syntactic dependencies within diverse languages.

2 Results

In computational linguistics, various mathematical models are employed to represent the distribution patterns of in-degree within syntactic graphs for different languages. Commonly analyzed distributions include displaced Poisson, displaced geometric, zeta, truncated zeta, and Altmann distribution (also known

as Menzerath's law) [2].

Each distribution is characterized by specific parameters that require adjustment to achieve an optimal fit with the given dataset. The methodology employed for parameter estimation is maximum likelihood estimation, as elaborated in Section 5. To assess and compare the goodness of fit for different distributions, we utilize the Akaike information criterion (AIC) [1]. Herein, we define AIC_{best} as the minimum AIC across all models. Consequently, for each model m, we calculate $\Delta AIC_m = AIC_m - AIC_{best}$, aiming to identify the model that best approximates our data. In Table 1, we present the ΔAIC values for each model in every language, where values closer to 0 indicate a stronger alignment with our optimal approximation.

	Model					
Language	D. Poisson	D. Geom	Zeta	Truncated Z	Altmann	AIC_{best}
Arabic	222316.204	24198.794	12.753	10.106	0.000	61853.678
Basque	38422.596	8364.030	0.152	0.066	0.000	27332.538
Catalan	885719.670	61915.772	47.994	34.746	0.000	126776.094
Chinese	591632.099	48863.560	199.429	185.451	0.000	132238.223
Czech	883983.366	91126.664	36.121	27.817	0.000	204789.467
English	718960.209	45955.536	258.402	215.189	0.000	120243.434
Greek	145963.408	17133.537	3.129	0.000	4.797	36275.129
Hungarian	433637.814	53469.625	0.000	1.724	1.972	81358.519
Italian	234004.631	22877.764	0.355	0.000	1.940	39279.960
Turkish	171839.447	24615.351	0.000	1.977	1.997	39935.050

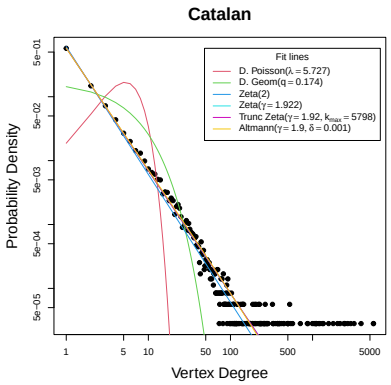
Table 1: AIC difference of the models per language

Table 1 presents a compelling argument, demonstrating that in the majority of cases, Altmann's distribution emerges as the optimal choice due to its consistently lowest AIC values. This assertion is further supported by a visual examination of the fit depicted in Figure 1a, where the Altmann distribution closely aligns with the observed data, substantiating its efficacy as an approximation.

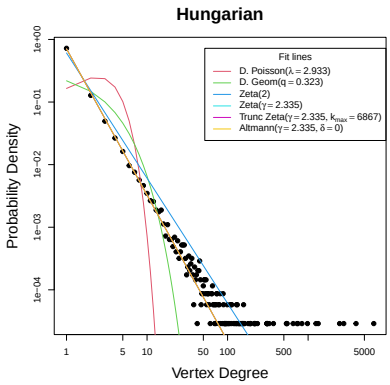
Turning our attention to the Hungarian language dataset, as depicted in Figure 1b, the ΔAIC criterion favors the selection of the zeta distribution over the truncated zeta or Altmann distributions. Although both truncated zeta and Altmann distributions exhibit commendably close ΔAIC values, the consideration of AIC's definition elucidates that these distributions had lower likelihoods due to their additional parameters. This outcome stems from AIC's penalty mechanism, wherein each extra parameter incurs a penalty of 2, and both truncated zeta and Altmann distributions possess one extra parameter compared to the zeta distribution.

By adopting the zeta distribution with $\gamma = 2$ as the baseline, a graphical comparison in Figure 1a and 1b unequivocally underscores the inadequacy of the displaced Poisson and displaced geometric distributions as approximations for

our data. Conversely, the fitted zeta distribution, the truncated variant, and the Altmann distribution stand out as highly effective approximations, closely matching the observed data in stark contrast to the zeta distribution with $\gamma = 2$. This behaviour can be observed in all the figures of appendix C.



(a) Comparison of distributions for Catalan



(b) Comparison of distributions for Hungarian

Figure 1: Distributions fitted for some example languages

3 Discussion

Log-log plots and precision of the empirical distributions

In the analysis of the provided data sets, a consistent pattern emerges across all plots. For instances characterized by low degrees, the data points exhibit a distinctive arrangement, forming an exceedingly narrow and nearly linear alignment. Nevertheless, as one progresses along the distribution towards higher degrees, a transformation occurs, leading to the gradual broadening and decreased precision of this linear alignment. Eventually, at the extreme right end of the distribution, the data points adopt a nearly horizontal orientation.

It is imperative to acknowledge the utilization of a log-log scale, where each unit of grid displacement results in exponential scale expansion. Natural plots have been omitted from presentation, as they fail to yield informative insights. However, it is important to realize that our findings are predicated upon a power-law distribution, necessitating careful consideration in the analysis of the obtained results.

In the examination of higher-degree points, an apparent trend emerges wherein the precision of their distribution appears to diminish progressively. This phenomenon can be attributed to the exponential reduction in the population of nodes. Consequently, even minute statistical uncertainties exert a substantial

influence on the accuracy of individual points.

Around the terminal segment of the plot, we have to take into account the fundamental characteristics of empirical distributions. These distributions exhibit a limitation in precision, not surpassing a granularity of $1/N$, wherein N denotes the sample size. The horizontal line depicted in the plot is result of this threshold, signifying not an actual proclivity of the distribution to persist in a horizontal trajectory, but rather the limitations imposed by small sample sizes for achieving a more refined portrayal of the underlying distribution.

Fitted lines

While examining the graphical representations (appendix C), one may initially be inclined to hypothesize that the observed curves exhibit a sub-optimal fit to the distribution in the context of higher degrees. This apparent discrepancy can be attributed, once again, to the inherent imprecision of the empirical distribution as it approaches the upper extremities of the spectrum. We have to take into account, however, that both the likelihood function and the Maximum Likelihood Estimation (MLE) method do not succumb to this limitation in precision. These statistical techniques use direct information from the sample dataset, making predictions with greater accuracy and reliability. The parameters of the fitted distributions can be observed in table 2.

	Model						
	D. Poisson	D. Geom	Zeta	Truncated Z		Altmann	
Language	λ	q	γ_1	γ_2	k_{max}	γ_3	δ
Arabic	3.217	0.298	2.104	2.103	2361	2.086	0.001
Basque	1.831	0.459	2.358	2.357	605	2.347	0.002
Catalan	5.727	0.174	1.922	1.920	5798	1.900	0.001
Chinese	5.173	0.192	1.886	1.884	8027	1.835	0.003
Czech	3.891	0.252	2.051	2.050	4963	2.041	0.001
English	6.850	0.146	1.800	1.795	4774	1.740	0.003
Greek	3.407	0.284	2.132	2.130	1135	2.132	0.000
Hungarian	2.933	0.323	2.335	2.335	6867	2.335	0.000
Italian	4.165	0.236	2.108	2.107	2812	2.108	0.000
Turkish	2.000	0.432	2.543	2.543	7039	2.543	0.000

Table 2: Values of fitted parameters per distribution and language

4 Methods

Degree distribution, understanding the plots

The initial phase of this research project focused on ensuring the precise representation of the dataset. An overview of the utilized data is provided in Table 3. Following the importation of the data points sequence, the generation of a

frequency table, and the visualization of the outcomes, it became apparent that the resulting plots did not faithfully reflect the data. Specifically, certain nodes exhibited degrees exceeding a thousand, yet were inadequately represented in the generated plots. This discrepancy was attributed to the data factorization process conducted using the R command `table`. To rectify this issue, we proceeded to transform the data into a `data.set` format, allowing for better manipulation and accurate visual representations. The distribution plots are presented in the figures of Appendix A and B.

Language	N	Maximum degree	M/N	N/M
Arabic	21065	2249	3.351	0.298418
Basque	11868	576	2.180	0.458649
Catalan	35524	5522	5.745	0.174056
Chinese	35563	7645	5.202	0.192219
Czech	66014	4727	3.972	0.251752
English	29172	4547	6.857	0.145830
Greek	12704	1081	3.524	0.283774
Hungarian	34600	6540	3.098	0.322827
Italian	13433	2678	4.231	0.236376
Turkish	20403	6704	2.313	0.432395

Table 3: Summary of dataset

Zeta distribution and right truncation

Due to a lack of precision, there is a tail at the end of the plots in the empirical distribution. Even though they do not affect the MLE method for fitting lines, we do not have any kind of information about how the data behaves in this area.

For this reason we propose the implementation of the right truncated zeta, by not taking into account the end points of the distribution.

$$TruncZ(k) = \begin{cases} \frac{1}{\sum_{x=1}^{k_{max}} x^{-\gamma}} k^{-\gamma} & \text{if } k \leq k_{max} \\ 0 & \text{otherwise} \end{cases}$$

However, MLE methods are not able to discern which is the optimal bound for this kind of models, because the likelihood function is monotonic on the k_max parameter. Therefore, we opted to slowly (1% max degree) increase k_max until the $-2\mathcal{L}$ converges in the sixth decimal place. The first value assigned to k_max was the maximum degree because otherwise, we would have had points with probability 0 and therefore a likelihood of $log(0)$, which would have been a problem. Another possibility to could have been to assign a fixed value.

5 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) involves the process of determining the parameters of an assumed probability distribution based on observed data. It entails maximizing a likelihood function such that, within the context of the assumed statistical model, the observed data has the highest probability. It is defined as follow:

$$\mathcal{L}(\text{params}|\mathbf{x}, \text{model}) = \sum_{i=1} \log p(x_i; \text{params})$$

In the problem statement, the likelihood functions where given for some probability functions.

	Probability function		Likelihood function
Displaced Poisson	$p(k) = \lambda^k \frac{e^{-\lambda}}{k!(1-e^{-\lambda})}$		$\mathcal{L}(\lambda) = M \log \lambda - N(\lambda + \log(1 - e^{-\lambda})) - C$
Displaced Geometric	$p(k) = (1-q)^k q$		$\mathcal{L}(q) = (M - N)\log(1-q) + N \log q$
Zeta	$p(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}$		$\mathcal{L}(\gamma) = -\gamma M' - N \log \zeta(\gamma)$
Truncated Zeta	$p(k) = \begin{cases} \frac{1}{\sum_{x=1}^{k_{max}} x^{-\gamma}} k^{-\gamma} & \text{if } k \leq k_{max} \\ 0 & \text{otherwise} \end{cases}$		$\mathcal{L}(\gamma, k_m ax) = -\gamma M' - N \log \sum_{x=1}^{k_m ax} x^{-\gamma}$

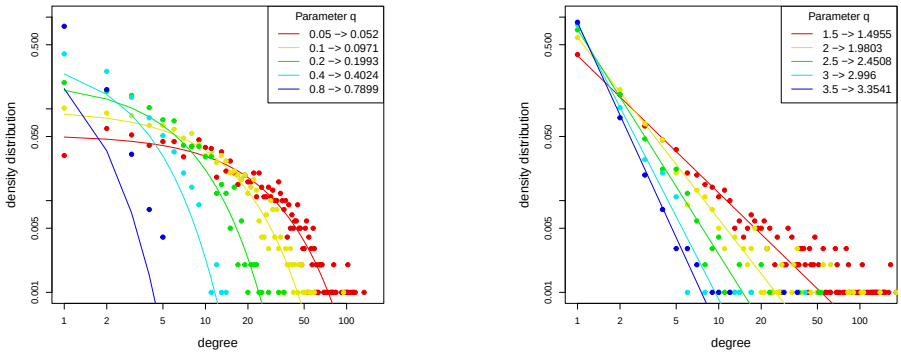
Moreover, we used the Altmann distribution $p(k) = k^{-\gamma} e^{-\delta k} \frac{1}{\sum_{x=1}^{N} x^{-\gamma} e^{-\delta x}}$, for which we had to derive its likelihood function:

$$\begin{aligned}
\mathcal{L}(\gamma, \delta) &= \sum_{i=1} \log p(k_i) \\
&= \sum_{i=1} \log \left(k_i^{-\gamma} e^{-\delta k_i} \frac{1}{\sum_{x=1}^{N} x^{-\gamma} e^{-\delta x}} \right) \\
&= \sum_{i=1} \log(k_i^{-\gamma}) + \log(e^{-\delta k_i}) - \log(\sum_{x=1}^{N} x^{-\gamma} e^{-\delta x}) \\
&= \sum_{i=1} -\gamma \log k_i - \delta k_i - \log(d) \\
&= -\gamma M' - \delta M - N \log(d)
\end{aligned}$$

In order to maximize this likelihood we used the `mle` method of the R package `stats4`. We configured it to use Limited-memory Broyden–Fletcher–Goldfarb–Shannon bounded (L-BFGS-B). We opted to use L-BFGS-B because t allowed us to establish bounds which is not possible with other methods such as BFGS, CG (more fragile than the BFGS methods), SANN (only allows one dimensional parameters so we could not use it for Altmann) and showed a good performance.

Validation

The hardest part of this kind of models, is to make sure that the methods used are in fact an effective representation of the data sample displayed. In this kind of cases, the simplest way is to get a sample known for following the distribution and study the fitting of the model. Here, in 2 we present the lines obtained at fitting the geometric and zeta MLE models over their respective empirical distributions for different parameters



(a) Geometric distribution

(b) Zeta distribution

Figure 2: Validations fit for different distributions and parameters

We can observe, that in most cases, the model fit quite well the empirical distributions. However, it seems to fail in the geometric when the parameter tends to 1.

For this project, as we have seen in the discussion, the models that better fit our real samples are the ones similar to the zeta distribution. As we can see in 2b, this kind of models fit well in our experimental range.

References

[1] Hirotogu Akaike. "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Selected Papers of Hirotugu Akaike*. Ed. by Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa. New York, NY: Springer New York, 1998, pp. 199–213. ISBN: 978-1-4612-1694-0. DOI: 10.1007/978-1-4612-1694-0_15. URL: https://doi.org/10.1007/978-1-4612-1694-0_15.

[2] Gabriel Altmann. "Prolegomena to Menzerath's law". In: *Glottometrika* 2.2 (1980), pp. 1–10.

A Figures of the spectrum of each language

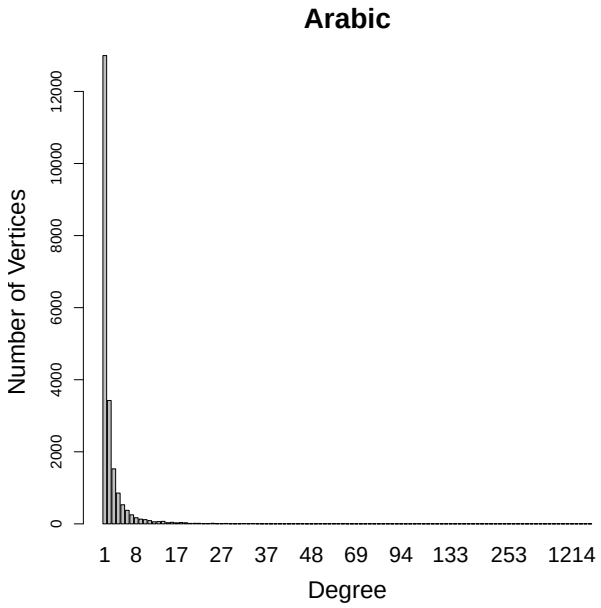


Figure 3: Degree distribution spectrum of Arabic

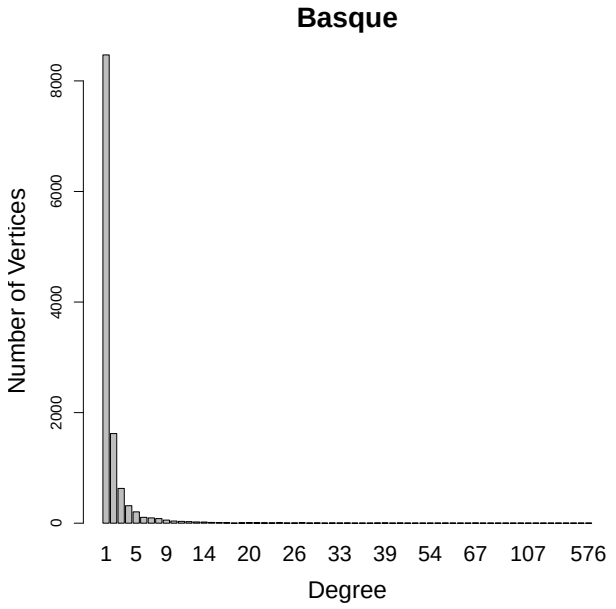


Figure 4: Degree distribution spectrum of Basque

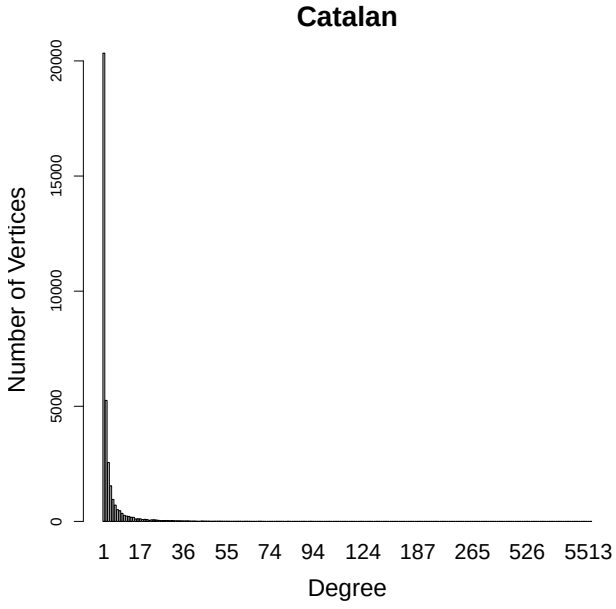


Figure 5: Degree distribution spectrum of Catalan

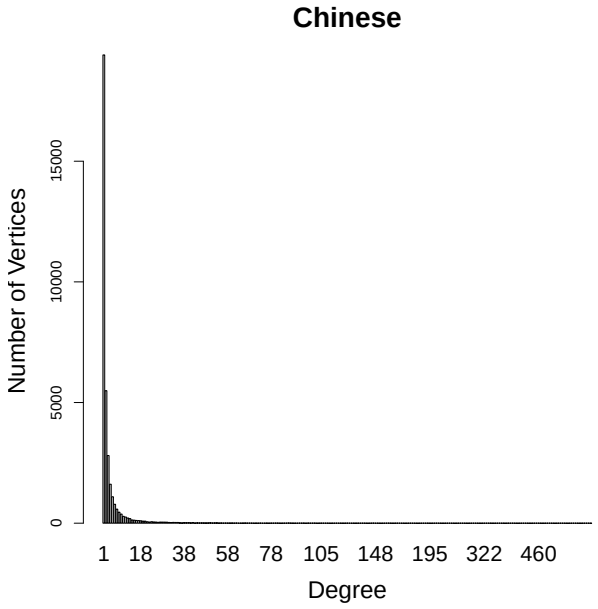


Figure 6: Degree distribution spectrum of Chinese

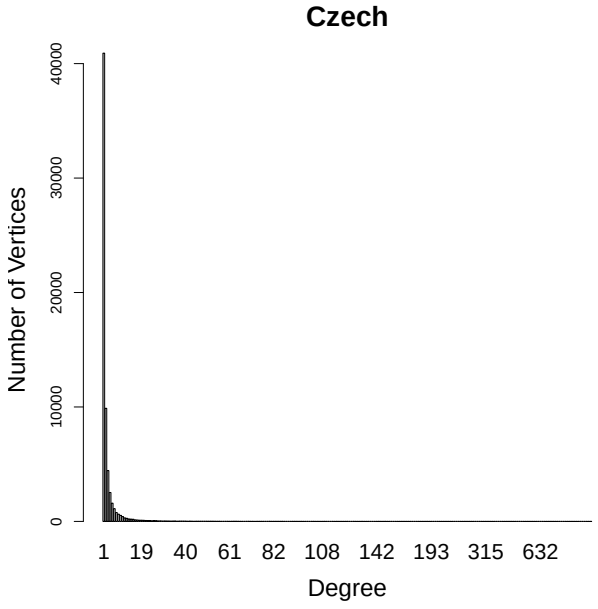


Figure 7: Degree distribution spectrum of Czech

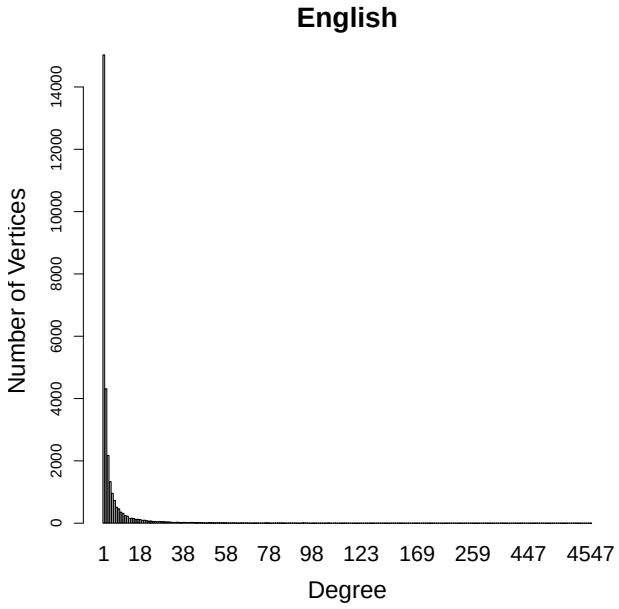


Figure 8: Degree distribution spectrum of English

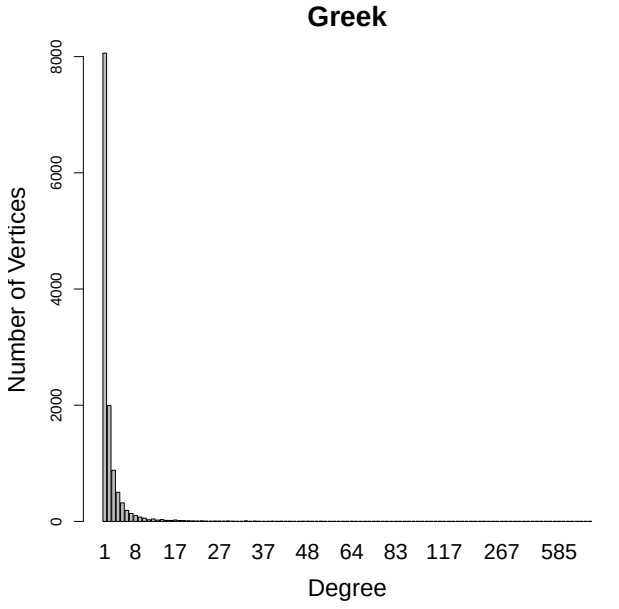


Figure 9: Degree distribution spectrum of Greek

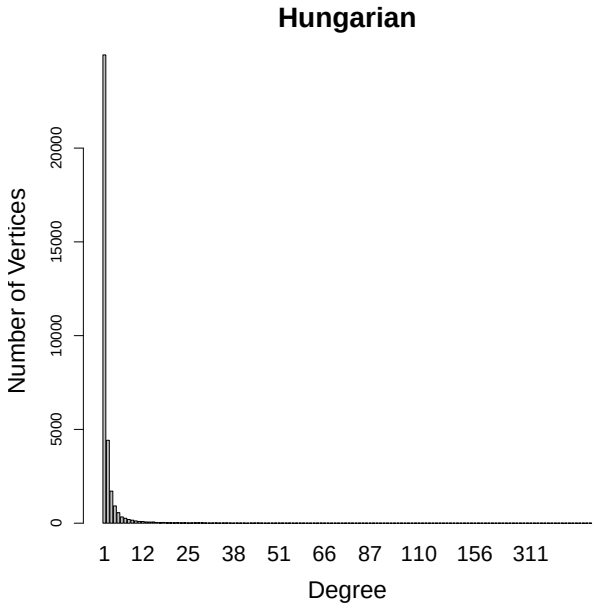


Figure 10: Degree distribution spectrum of Hungarian

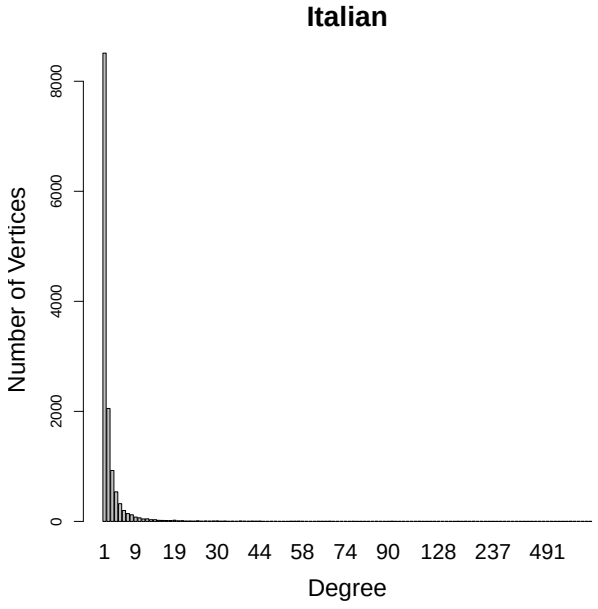


Figure 11: Degree distribution spectrum of Italian

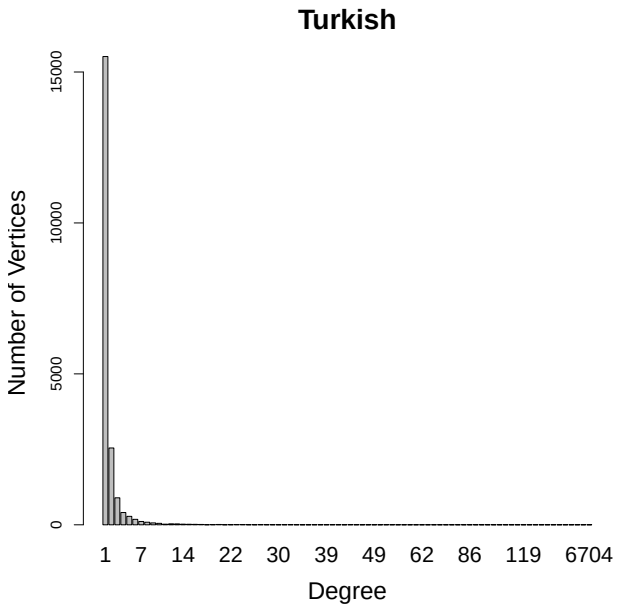


Figure 12: Degree distribution spectrum of Turkish

B Figures of the spectrum of each language using log-log

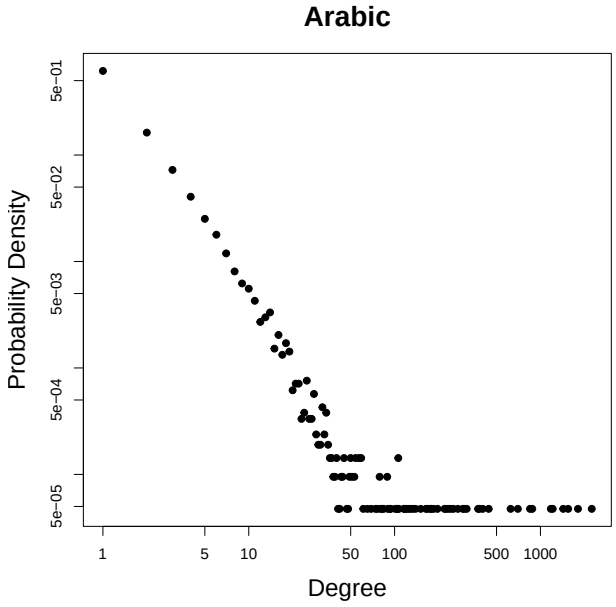


Figure 13: Degree distribution spectrum of Arabic

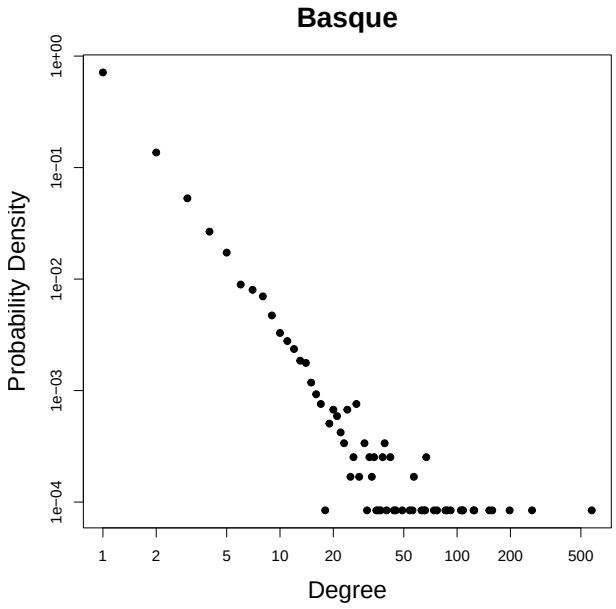


Figure 14: Degree distribution spectrum of Basque

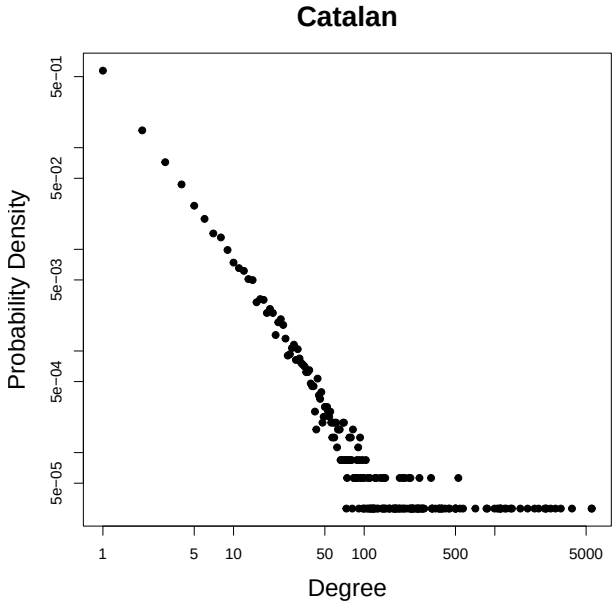


Figure 15: Degree distribution spectrum of Catalan

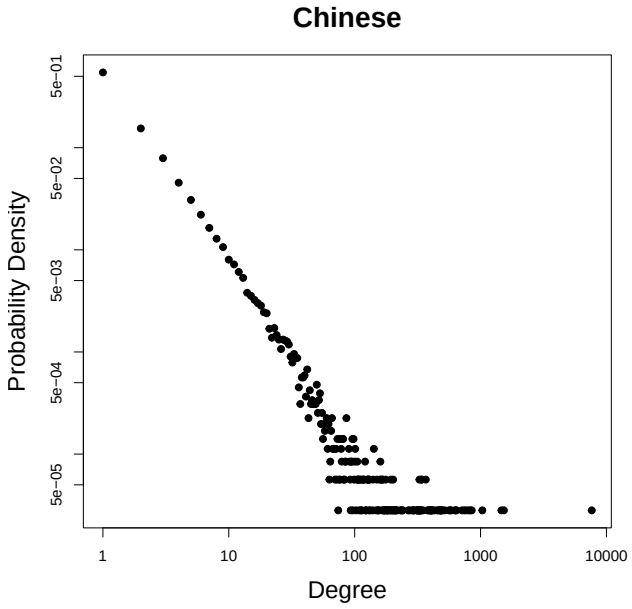


Figure 16: Degree distribution spectrum of Chinese

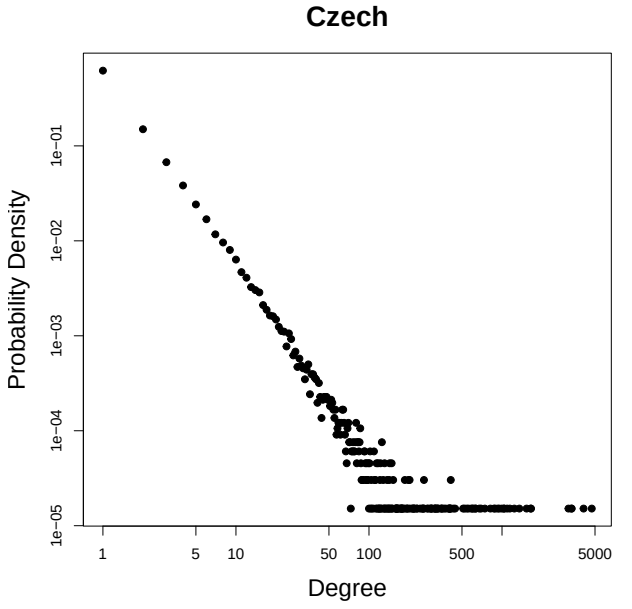


Figure 17: Degree distribution spectrum of Czech

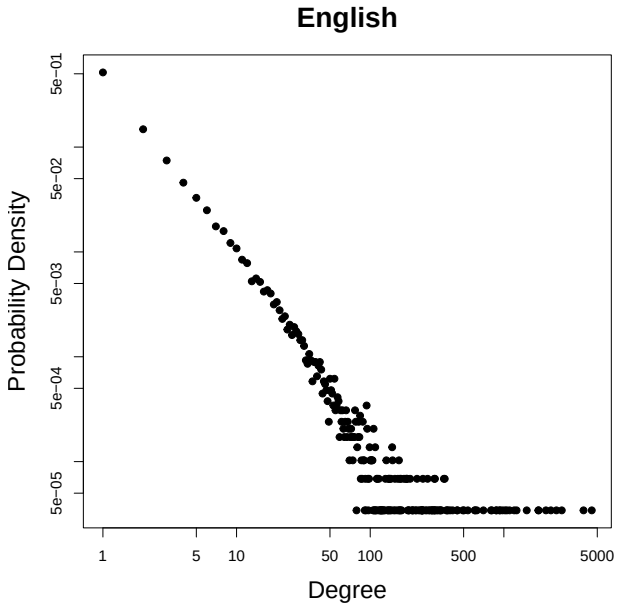


Figure 18: Degree distribution spectrum of English

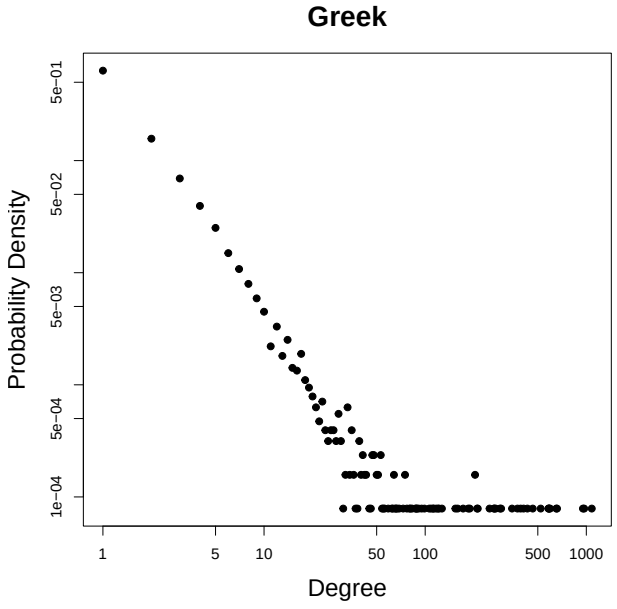


Figure 19: Degree distribution spectrum of Greek

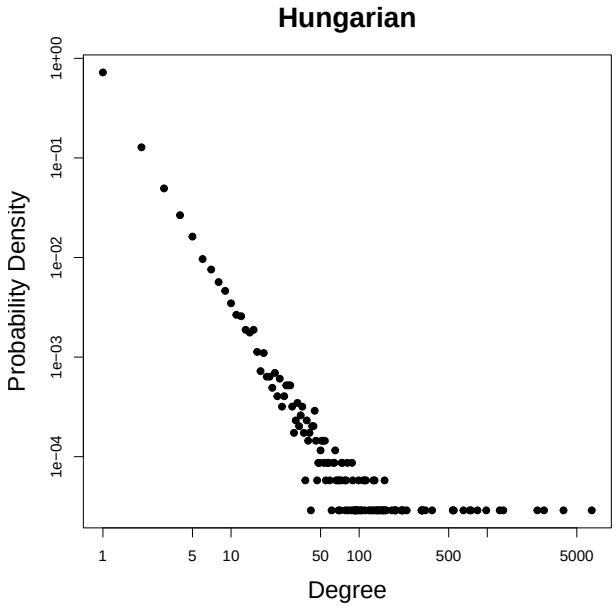


Figure 20: Degree distribution spectrum of Hungarian

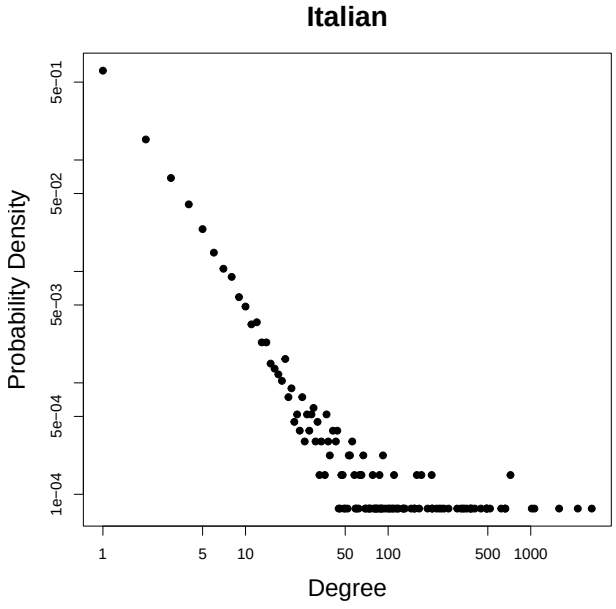


Figure 21: Degree distribution spectrum of Italian

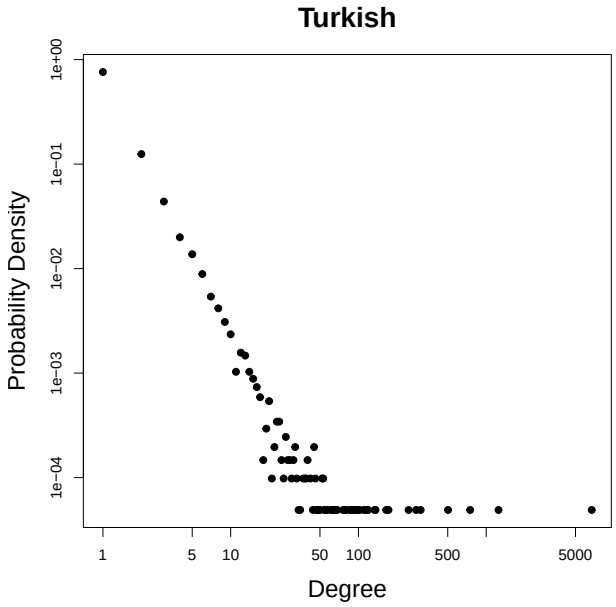


Figure 22: Degree distribution spectrum of Turkish

C Figures with all the models fitted for each language

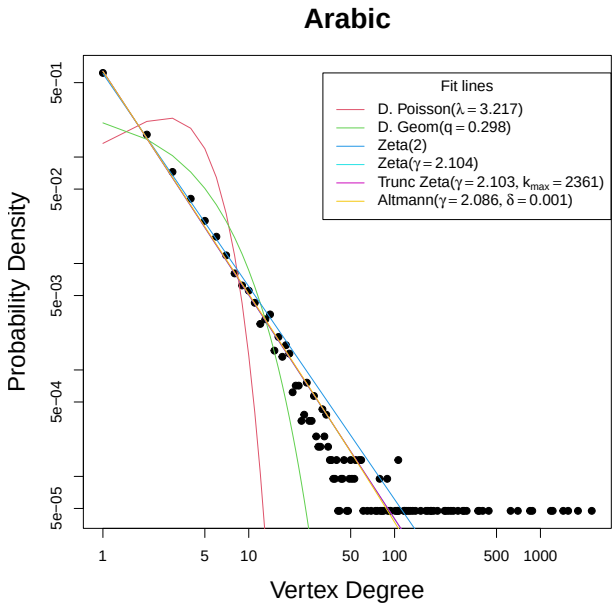


Figure 23: Fitted probability distributions for Arabic

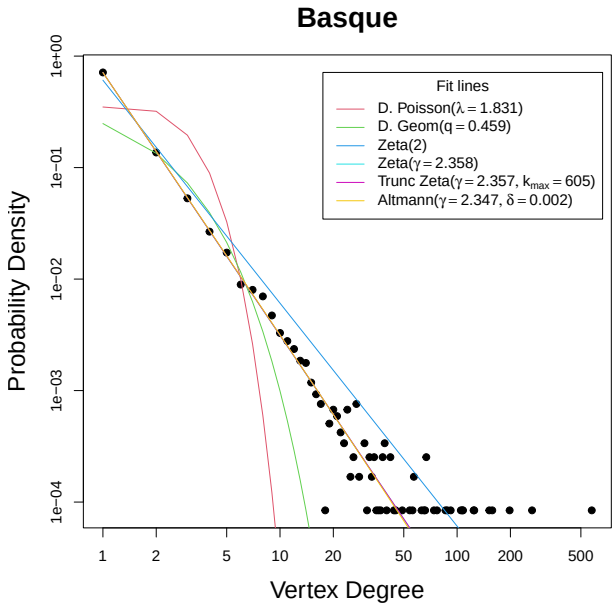


Figure 24: Fitted probability distributions for Basque

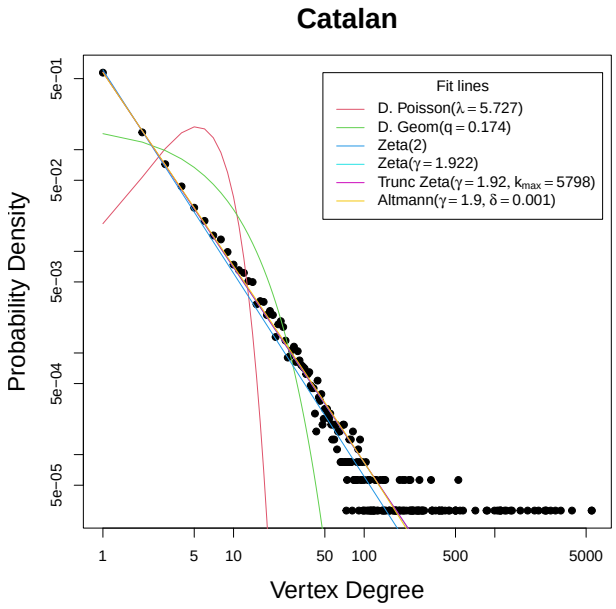


Figure 25: Fitted probability distributions for Catalan

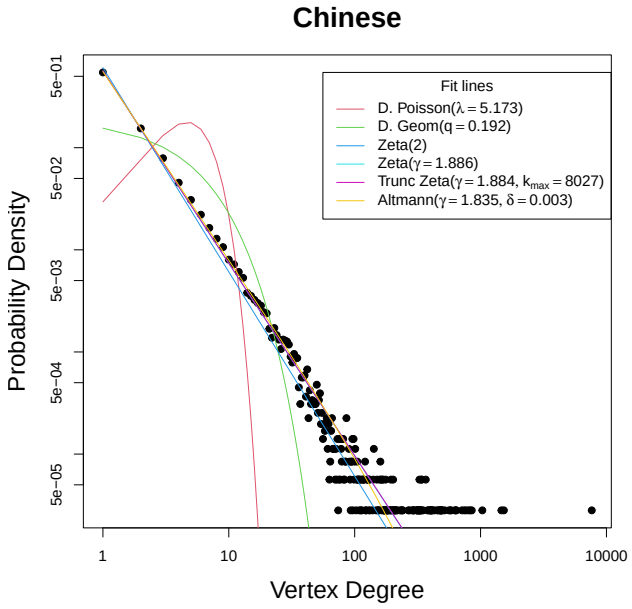


Figure 26: Fitted probability distributions for Chinese

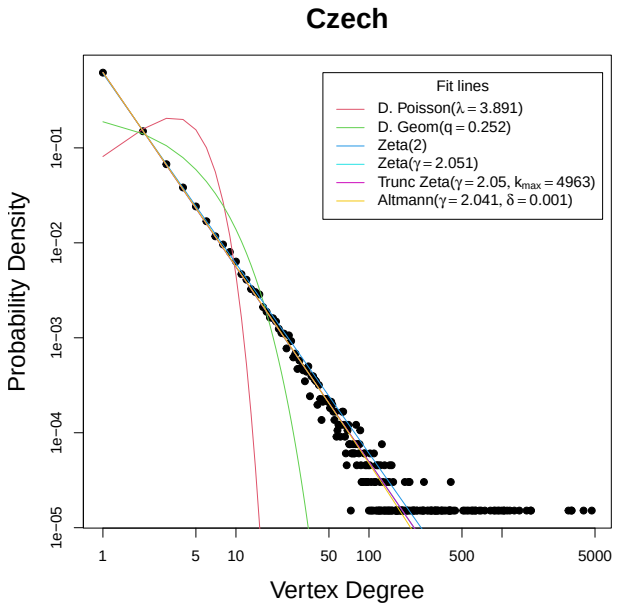


Figure 27: Fitted probability distributions for Czech

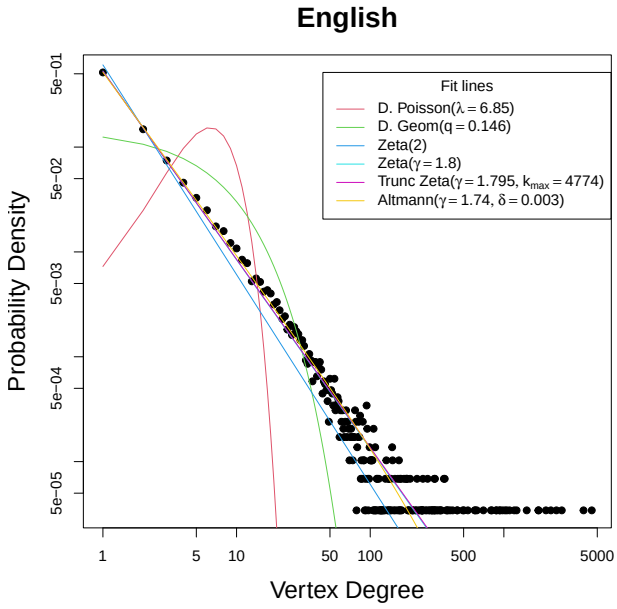


Figure 28: Fitted probability distributions for English

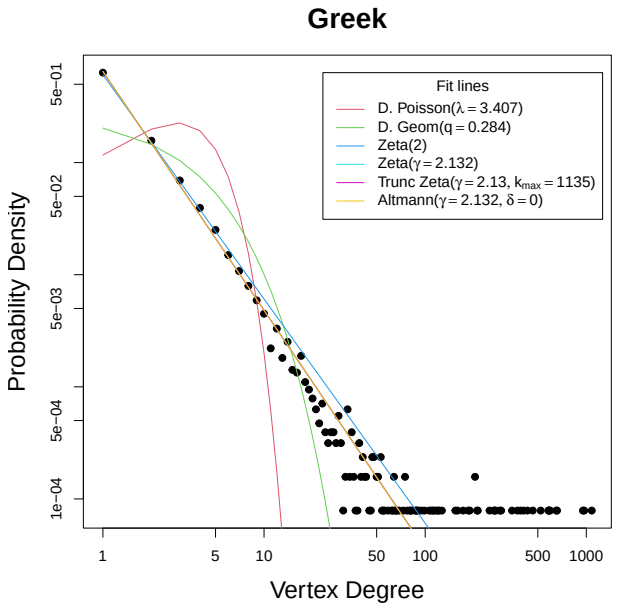


Figure 29: Fitted probability distributions for Greek

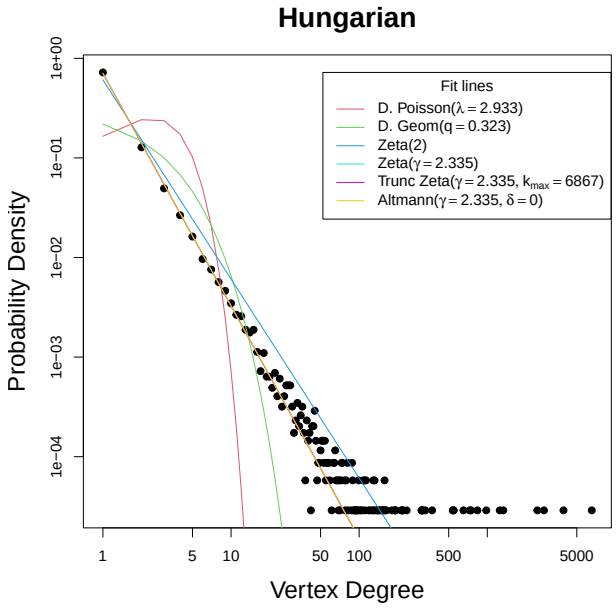


Figure 30: Fitted probability distributions for Hungarian

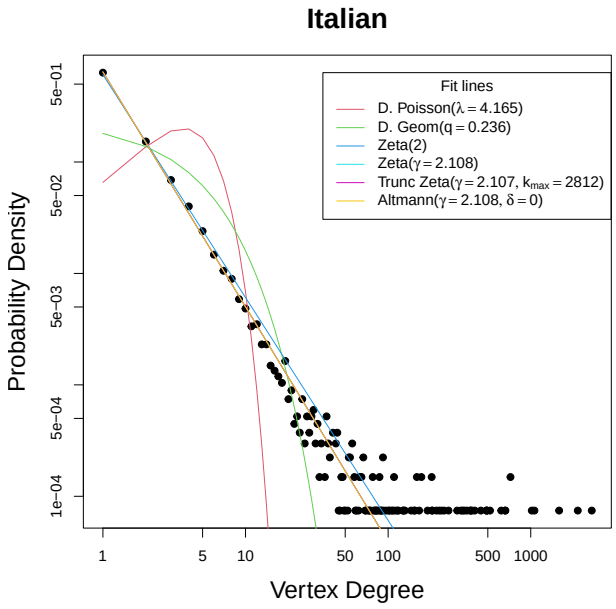


Figure 31: Fitted probability distributions for Italian

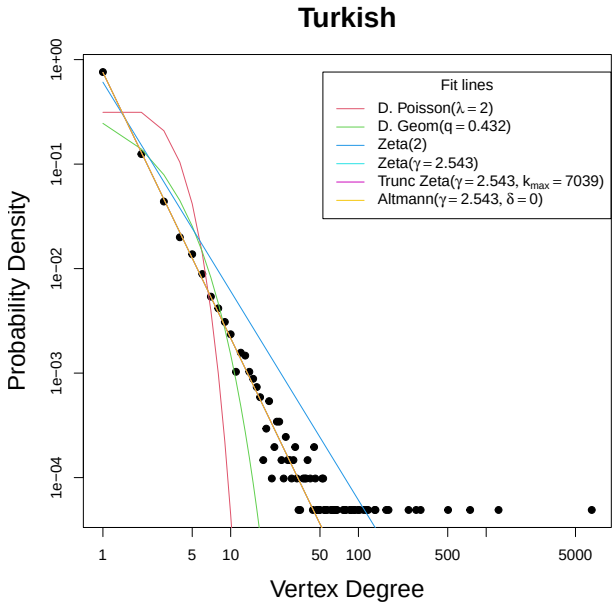


Figure 32: Fitted probability distributions for Turkish