CSN Lab2 Report

Raphaël Vignon
Ali Arabyarmohammadi

October 2022

1 Introduction

Degree distribution models are essential tools for assessing and understanding the form and nature of social networks, and they can also be helpful for the development of efficient graph algorithms. We will compare multiple distributions and look at each parameter's likelihood and number of parameters to identify a distribution that suits the data in this project.
The goal is to figure out if all of these languages have the same kind of word link distribution and have the same distribution parameters.
We'll be working on out-degree distributions for ten different languages and consider several different degree distributions listed below.

- Poisson distribution
- Geometric distribution
- Zeta distribution
- Zeta distribution with $\gamma = 2$
- Right truncated distribution

In the end, we will investigate the consequences of adding a new probability distribution called the **Altmann function** that is able to lead to a better fit than the best model we examined before.
In the first step, we will also produce a table using some elementary properties of the out-degree sequence for each language which will be needed for the rest of our analysis.
Parameters included in the Table 1 are listed below.

- ***Language*** **:** Names of the 10 different languages.
- $\boldsymbol{N}$: Number of nodes.
- ***Maximum Degree*** : Greatest number in the degree sequence.
- $\frac{M}{N}$: Mean degree where M is the sum of degrees.
- $\frac{N}{M}$: Reverse of the Mean degree.

Language	$\boldsymbol{N}$	**Maximum Degree**	$\frac{M}{N}$	$\frac{N}{M}$
Arabic	15678	4896	4.502424	0.2221026
Basque	6188	2097	4.181642	0.2391405
Catalan	24727	6622	8.253933	0.1211544
Chinese	23946	7537	7.726259	0.1294287
Czech	41912	12671	6.256394	0.1598365
English	17775	7040	11.25406	0.08885678
Greek	9280	2737	4.824138	0.2072909
Hungarian	25534	1020	4.197462	0.2382392
Italian	12285	1671	4.625885	0.2161748
Turkish	15287	4488	3.086675	0.3239732

Table 1: Summary of the properties of the out-degree sequences.

Then, to obtain a visual sense of the degree distribution, we'll look at bar plots for the English language. The following plots show that nodes with a low degree are more frequent than those with a higher degree.

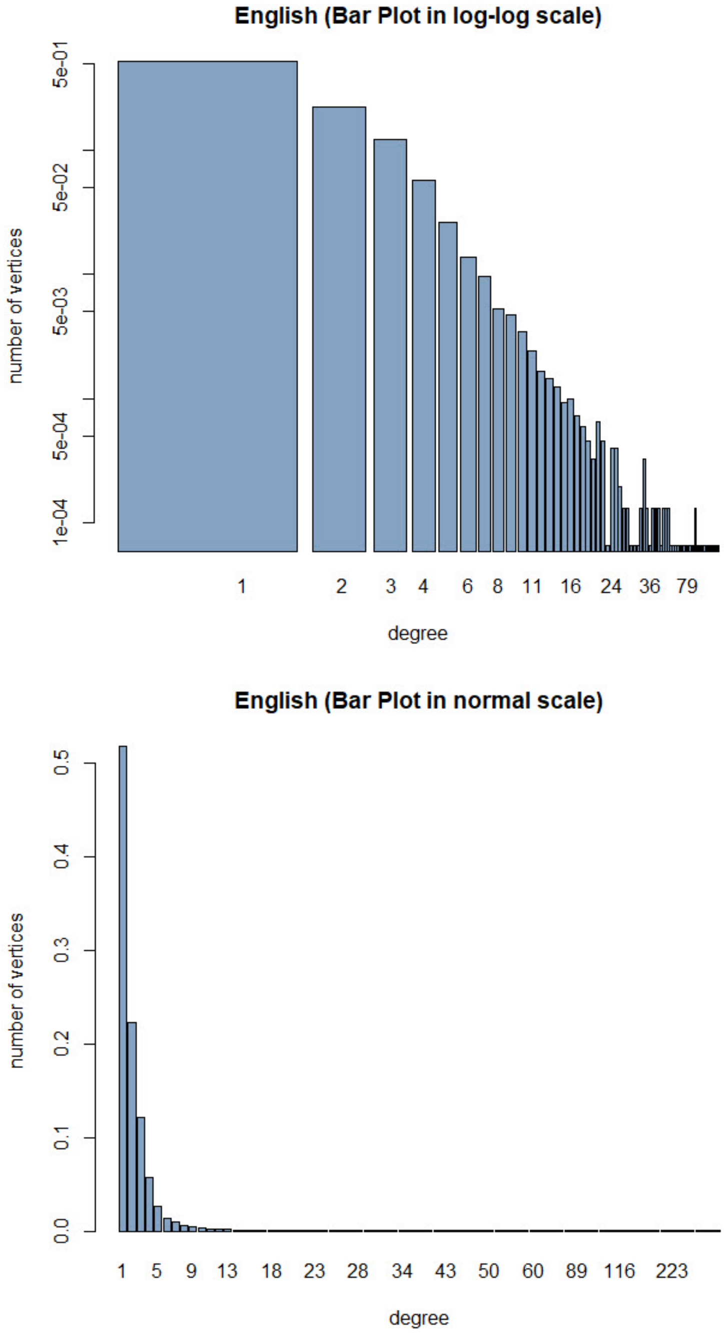

Figure 1: Degree distribution of English language in normal and log-log scale

2 Results

Firstly, we will compare the results without considering the Altman distribution, and then we will investigate the results including it.

2.1 Results without Altmann

In the table 2 we show the ΔAIC for different languages. The objective of this table is to see which is the best model for each of the languages and be able to see how good are each of the other models compared to the best fit.
We can now go further and retrieve the log Likelihood for each model and compute the AIC for the real cases after obtaining the parameters and validating our approaches. The best AIC for each Language can then be subtracted from the AIC for the other ways to get the delta AIC table.
Table 2 represents the outcome.

	Model				
Language	**Zeta ($\gamma = 2$)**	**Zeta**	**Zeta RT**	**Poisson**	**Geometric**
Arabic	792	7	0	203627	9831
Basque	82	1	0	67106	5469
Catalan	7767	94	0	541699	14165
Chinese	4365	42	0	604803	23775
Czech	6036	36	0	824167	30602
English	7733	135	0	646683	14345
Greek	1257	20	0	90513	1964
Hungarian	1762	12	0	164540	8065
Italian	1580	21	0	95428	1881
Turkish	21	0	1	166481	11597

Table 2: Table showing the ΔAIC for each language/distribution combination

With the exception of the **Turkish** language, where the Zeta distribution appears to work better, it is obvious from this table that the approach that best suits these out-degree sequences is the Right-Truncated Zeta distribution.

2.2 Results with Altmann

In the following section we add the Altman function to see if it provides a better model and which impact it will have on our result. In the table 3 we also show the ΔAIC for different languages.

	Model					
Language	**Zeta ($\gamma = 2$)**	**Zeta**	**Zeta RT**	**Poisson**	**Geometric**	**Altmann**
Arabic	1540	756	748	204375	10579	0
Basque	185	104	103	67209	5572	0
Catalan	10991	3317	3223	544922	17388	0
Chinese	5692	1369	1327	606131	25103	0
Czech	8907	2907	2871	827038	33473	0
English	9923	2325	2190	648872	16535	0
Greek	2578	1342	1322	91835	3286	0
Hungarian	4200	2450	2438	166978	10503	0
Italian	3476	1917	1896	97324	3777	0
Turkish	137	116	117	166597	11713	0

Table 3: Table showing the ΔAIC for each language/distribution combination

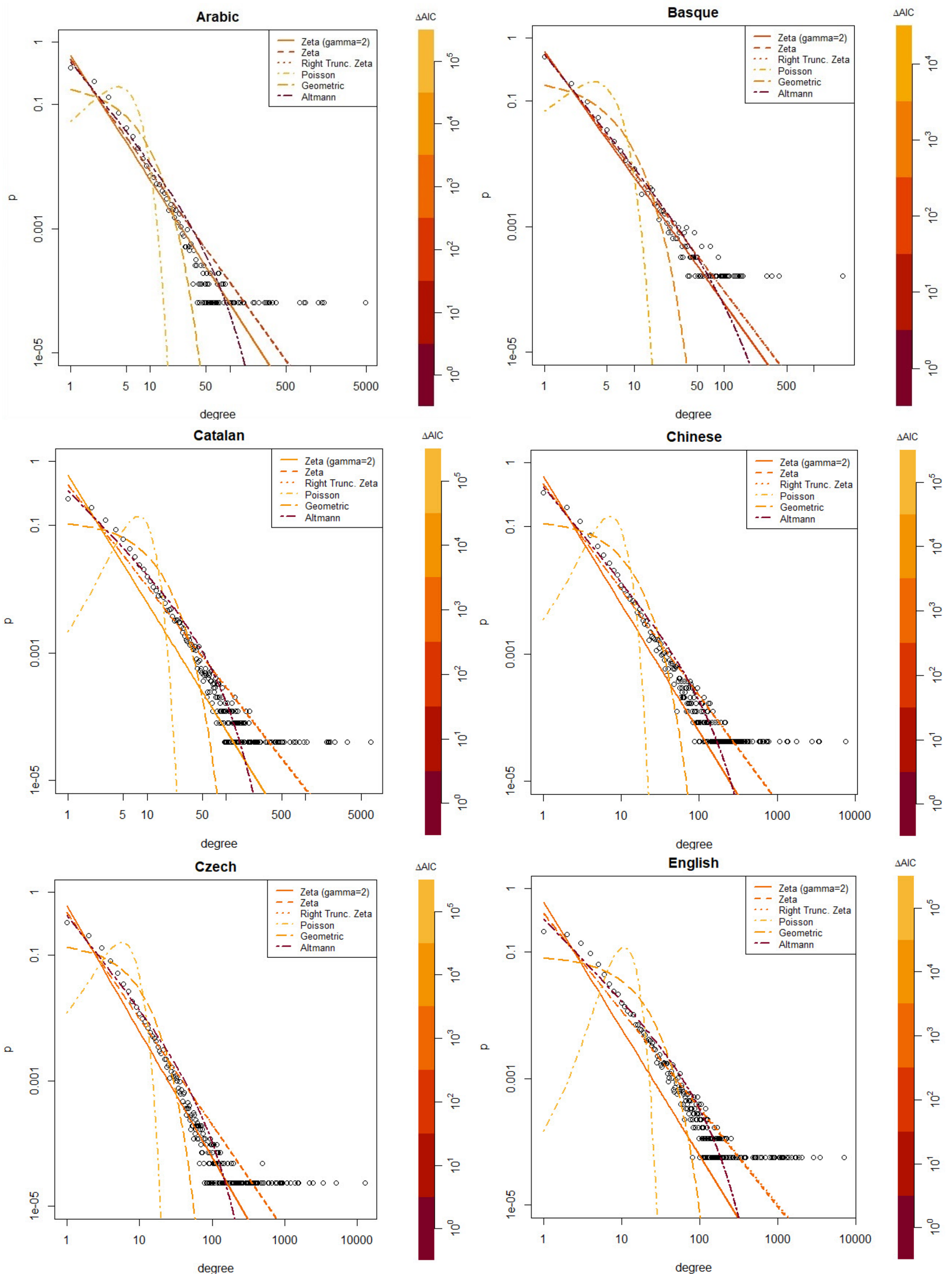

Figure 2: Plots showing the degree distribution of the word interaction graphs for 10 different languages and fits by different probability distributions in a log-log scale. The colors of the distributions show how well it fits the data compared to the other distributions using the ΔAIC.

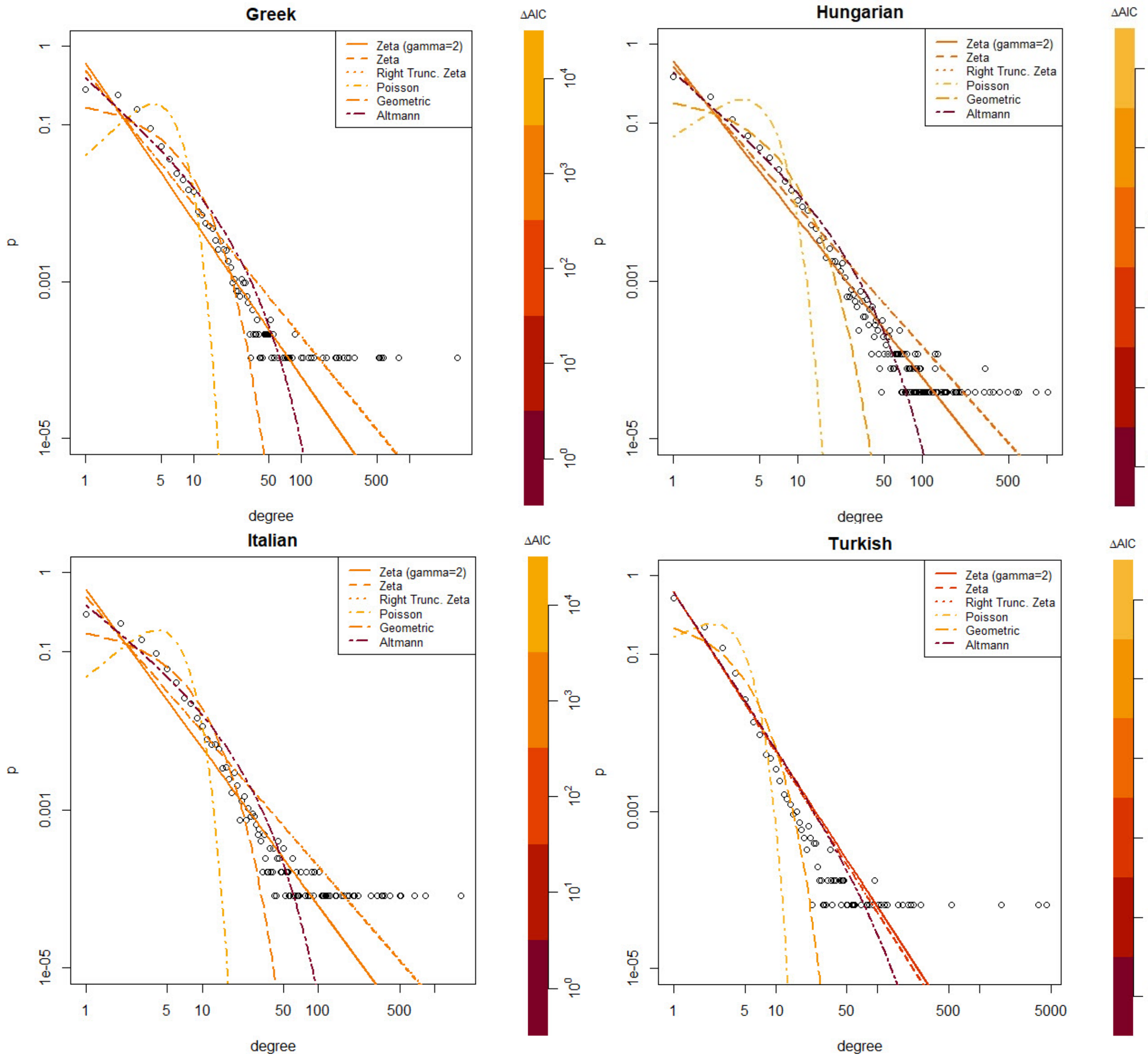

Figure 3: Continuation of figure 2 for other languages

Figures 2 and 3 show the degree distribution on the original graph, and different theoretical distributions plotted on top of them. The colors of the plot indicate how good the fit is (in terms of AIC). The objective of these plots is to visualize the different distributions in a single view, how good they are, and how they fit the graph data.

	Zeta	**Zeta RT**		**Poisson**	**Geometric**	**Altmann**	
Language	γ	γ	k_{max}	λ	q	γ	δ
Arabic	1.797628	1.795492	15678	4.449833	0.2221026	1.554883	0.02239932
Basque	1.88715	1.884904	6188	4.113253	0.2391405	1.763406	0.01055607
Catalan	1.590978	1.583235	24727	8.25178	0.1211544	1.24891	0.01920106
Chinese	1.662662	1.658106	23946	7.722839	0.1294287	1.466829	0.009919987
Czech	1.690866	1.688268	41912	6.244246	0.1598365	1.439122	0.016869
English	1.545278	1.53242	17775	11.25392	0.08885679	1.255782	0.01146553
Greek	1.699111	1.692925	9280	4.783788	0.2072909	1.195903	0.05288204
Hungarian	1.76932	1.767426	25534	4.129952	0.2382392	1.352568	0.04749585
Italian	1.704723	1.699731	12285	4.57837	0.2161748	1.156094	0.06149305
Turkish	2.042634	2.04231	15287	2.920239	0.3239732	1.949688	0.0105813

Table 4: Table showing the maximum likelihood estimations of the parameters of each distribution for each language.

The table 4 shows the most likely parameters for each of the distributions. The Altmann function now has new parameters, γ and δ.
The objective of this table is to be able to compare the parameters across distributions and languages.

Language	**Data**	**Zeta($\gamma = 2$)**	**Zeta**	**Zeta RT**	**Poisson**	**Geometric**	**Altmann**
Arabic	0.00102	0.00303	0.00969	0.0095	0	0.222	0
Basque	0.00064	0.00303	0.00579	0.0055	0	0.239	0
Catalan	0.00186	0.00303	0.03193	0.0314	0	0.121	0
Chinese	0.00300	0.00303	0.02110	0.0207	0	0.129	0
Czech	0.00150	0.00303	0.01793	0.0177	0	0.159	0
English	0.00376	0.00303	0.04159	0.0408	0	0.088	0
Greek	0.00140	0.00303	0.01709	0.0164	0	0.207	0
Hungarian	0.00054	0.00303	0.01140	0.0112	0	0.238	0
Italian	0.00113	0.00303	0.01655	0.0160	0	0.216	0
Turkish	0.00045	0.00303	0.00237	0.0023	0	0.323	0

Table 5: Table showing the complementary cumulative distribution ($P(k > 200)$) for each language and each distribution and also the percentage of nodes with higher than degree 200 on the original graph for comparison.

The table 5 shows the percentage of nodes with degrees higher than 200 from the original graph and for each of the theoretical distributions. The idea of this table is to see how well each distribution predicts the number of nodes in the tail.

3 Discussion

First of all, as can be seen in figure 2 and table 3 the Altmann distribution gives the best fit to the data. This distribution follows the shape of the data on the plot quite well but does not follow it on the tail. The second-best distributions are the Zeta and Right-truncated Zeta. The truncated version is always slightly superior to the original.

These two distributions are linear on the log-log plot, and that is not what the original data show, but they look like a good approximation of the distribution that generated the data. Next, we have the geometric and the fixed Gamma distribution.

The worst of all distributions is the Poisson. That does not even have the same shape as the data, and their ΔAIC are orders of magnitude higher than the rest. For some languages like Arabic, Hungarian, Italian and Turkish, the data points on the plot shows the percentage of nodes with degree 2 is higher than nodes with degree 1.

None of the distribution show this kind of behavior.

As already discussed, the Altman distribution seems to fail to predict the long tail of the distribution in the figure2. To understand better, we can look into table 5 and see that Altmann and Poison distributions have a probability of 0 for nodes with degrees higher than 200, but we have some nodes with higher degrees in our data. From this perspective, the better models are the Zeta family, which can predict long tails.

On the other hand, the Geometric distribution indicates too many nodes with a high degree compared to the input data.

Looking at table 4 we can see that γ is almost the same for both Zeta and Right-truncated Zeta. We can see also that k_{max} is always the same as the maximum degree of the graphs from table 1. This may be a problem because it is unlikely that the model that generated the data has this characteristic. One option to avoid this problem is to set the lower limit of the k_{max} parameter to the total number of nodes on the graph.

Another observation that we can make is that for this example, the number of parameters has little importance in calculating the AIC because the ΔAIC, as well as the likelihood, are orders of magnitude bigger than the number of parameters. The number of parameters may play a more important role if it is related to the number of data points as in K-Nearest Neighbours (KNN) regression.

Finally, we can note that adding a function that is better than the other (i.e., Altmann function) will impact the result of the table showing the ΔAIC. However, we can still see if models are better than others; it won't change most of our interpretations. For example, in the table 2we can observe that Zeta is better than Zeta RT for Turkish since its value is 0. When we look at the table3, we can still see that the ΔAIC value for Zeta is still lower than the one for Zeta RT.

4 Methods

To estimate the maximum likelihood of the right-truncated Zeta distribution, we had challenges of setting the lower bound on the variables for the "L-BFGS-B" method, and by some search and try and error, finally, we figured out how to set the right bounds.

In order to compare the tail of the distribution, which can not be clearly seen on the plot because of the scale, we added table 5, that shows the complementary cumulative distribution ($P(k > k^*)$) for $k^* = 200$. We choose 200 because it is the limit in most languages where the data starts to be more sparse in terms of the number of degrees without nodes. Using this table, we were able to see which distribution fitted better the tail of the data.