Complex Social Networks - Lab 2

Afonso Ramos
Lukas Ramroth

October 14, 2019

1 Introduction

The objective of this second lab project is to practice on information theoretic model selection for the degree distribution of global syntactic dependency networks. To achieve this we will be analysing several degree distributions with six theoretic models and based on the obtained results select the best one that fits the distribution.

Three different sequences were given to the students:

- Undirected degree distributions;
- In-degree distribution sequences;
- Out-degree distribution sequences.

From these we've selected the out-degree distribution for the provided 10 different languages with the following theoretic models:

- Poisson distribution (γ parameter)
- Geometric distribution (q parameter)
- Zeta distribution (γ parameter)
- Zeta distribution ($\gamma = 2$ parameter)
- Right truncated distribution (γ and k_{max} parameters)
- Altmann distribution (δ and γ parameters)

2 Property Summary

For starters, we decided to compute the different properties of the out-degree sequences as recommended by the provided guide, and, as we have learnt throughout the development of this task, we have decided to include two extra parameters which were useful to obtain the likely hoods of the several languages.

In the following table we can then find:

- **N**: Number of nodes;

- **M**: Sum of degrees;
- **Maximum Degree**: Largest out-degree;
- **M/N**: Mean degree;
- **N/M**: Inverse of the mean degree;
- **M'**: Sum of degree logarithms;
- **C**: Sum of logarithm of degree factorials.

Language	N	Max Degree	M	M/N	N/M	M'	C
Arabic	15678	4896	70589	4.50242	0.2221	12530.41	170079.88
Basque	6188	2097	25876	4.18164	0.2391	4231.38	56296.61
Catalan	24727	6622	204095	8.25393	0.1212	29926.06	565816.20
Chinese	23946	7537	185013	7.72626	0.1294	24832.10	555080.18
Czech	41912	12671	262218	0.15984	6.2564	41038.66	730668.60
English	17775	7040	200041	0.08886	11.2540	23919.12	660363.15
Greek	9280	2737	44768	0.20729	4.8241	8938.33	92879.20
Hungarian	25534	1020	107178	0.23824	4.1975	21493.72	184050.32
Italian	12285	1671	56829	0.21617	4.6259	11701.85	106778.81
Turkish	15287	4488	47186	0.32397	3.0867	8162.51	113935.58

Table 1: Summary Table of out-degree sequences for all 10 languages.

To better understand the degree distribution we can observe from the following graph, which was done by following the guide's recommendations, that the nodes with small out-degree are more frequent than nodes with high out-degree.

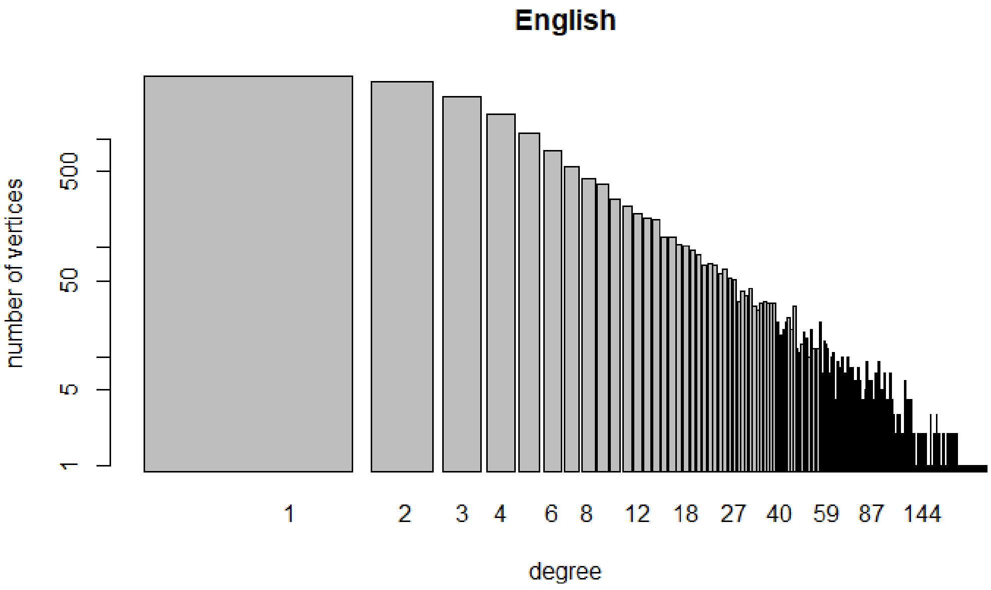


Figure 1: English log-log out-degree sequence plot.

3 Results (without Altmann)

With all the necessary parameters obtained and calculated we can now proceed to determining which ones minimise the negative logarithmic likelihood function.

As a means of optimising the search for the best parameters we chose to start with the default parameters, which acted as our best initial assumption, which consisted of:

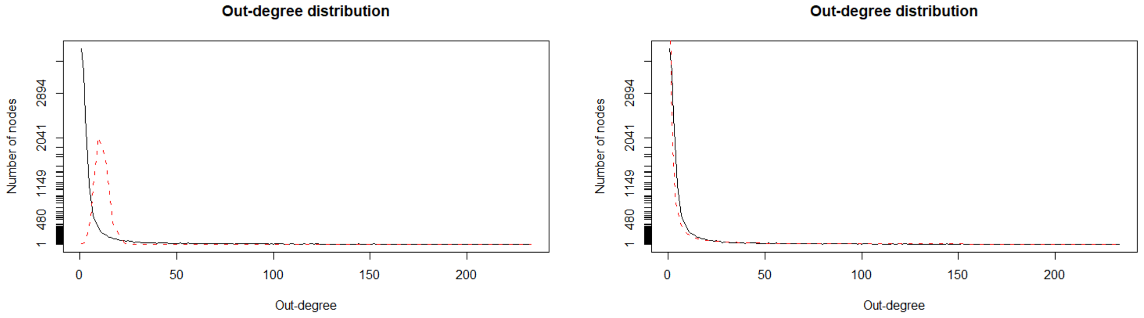
- λ_0 = M/N
- q_0 = N/M
- $\gamma_0 = 2$
- $k_{max,0}$ = N

In the end, with all the parameters and methods' results, we can compute the AIC for all the cases and, by subtracting the best AIC of each language, obtain the ΔAIC for each of the languages.

Language	**Poisson**	**Geometric**	**Zeta**	**Zeta (γ=2)**	**R. Trunc. Zeta**
Arabic	$2.036 * 10^5$	$9.829 * 10^3$	$7.924 * 10^2$	7.494	0
Basque	67105.701	5467.462	82.421	1.455	0
Catalan	541699.024	14163.356	7767.499	93.627	0
Chinese	604803.167	23773.076	4364.603	41.521	0
Czech	824167.355	30600.486	6035.890	36.052	0
English	646682.542	14343.474	7732.677	134.967	0
Greek	90513.117	1962.453	1256.834	20.398	0
Hungarian	164540.132	8063.318	1762.370	12.137	0
Italian	95428.171	1878.663	1580.394	20.940	0
Turkish	$1.665 * 10^5$	$1.159 * 10^4$	$2.110 * 10^1$	0	1.208

Table 2: ΔAIC for the 5 models applied to all 10 languages.

From this table we can clearly observe that the method better fits these out-degree sequences is almost always the Right-Truncated Zeta distribution, with the exception of the Turkish language, where the Zeta distribution seems to fit better.



(a) English with Poisson method.

(b) English with Zeta method.

Figure 2: Comparison of plots showing Poisson and Zeta methods for English.

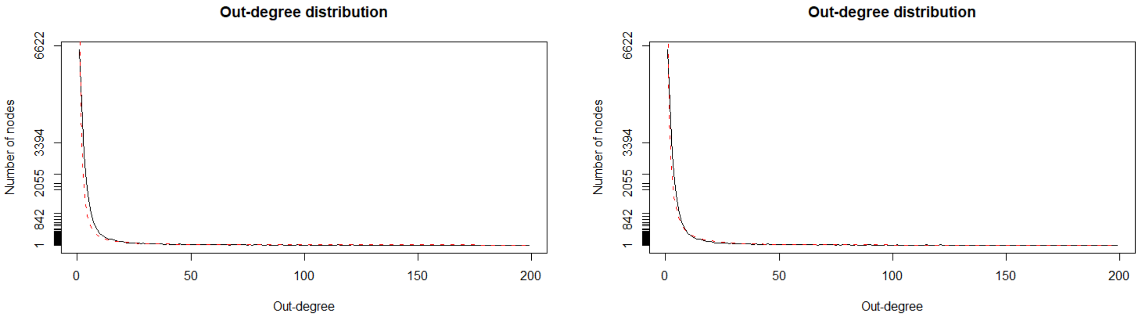
4 Results (with Altmann)

We then take into account the Altmann model to our set of distributions, again, as a means of identifying the best fit to out data. To to this two parameters were added in order to calculate it: γ & δ.

Language	Poisson	Geometric	Zeta	Zeta(γ=2)	RT Zeta	Altmann
Arabic	204373	10574.6	1538.41	753.546	746.052	0
Basque	67206.6	5568.40	183.357	102.392	100.937	0
Catalan	544920	17384.4	10988.5	3314.67	3221.04	0
Chinese	606128	25098.7	5690.09	1367.00	1325.48	0
Czech	827036	33469.4	8904.84	2905.00	2868.95	0
English	648870	16531.4	9920.60	2322.89	2187.93	0
Greek	91832.8	3282.09	2576.47	1340.03	1319.64	0
Hungarian	166975	10499.0	4198.03	2447.79	2435.66	0
Italian	97322.1	3772.55	3474.28	1914.83	1893.89	0
Turkish	166595	11708.8	135.343	114.235	115.443	0

Table 3: ΔAIC for the 6 models (Altmann included) applied to all 10 languages.

As we can clearly observe, Altmann is now the overall best method for all our data, independently of the language. Let's try and compare Catalan with the Right Truncated method (the previously best method for our data) with the Altmann method.



(a) Catalan with Right Truncated Zeta method.

(b) Catalan with Altmann method.

Figure 3: Comparison of plots showing Poisson and Zeta methods for English.

As expected, Altmann is clearly a better fit, however one needs to look very closely, which is expected, as we are comparing the best method with the second best.

5 Discussion

Table 1 displays some important parameters of all the out-degree distributions. Analysing this table we find that Czech has very high maximum degree that is almost always the double of every other language, while Greek and Basque have a very small amount of nodes than the rest. English's average is also quite high when compared to the other languages.

Table 2 is the ΔAIC table where we can observe that Right Truncated Zeta was chosen in almost all cases with the exception of Turkish. The Zeta distribution was also a very good fit coming in a very close second for Arabic and Basque. On another note, Poisson and Geometric distributions should not be used to model out-degree distributions based on the obtained values.

Table 3 is the ΔAIC with the Altmann function included. γ & δ are the new parameters used for the Altmann function and we can see that this method was chosen as the best fit for every language without any exceptions. It's worth noting that Zeta was a rather close second for both Turkish and Basque. With Figure 3 we tried to demonstrate the differences between the Altmann function and the Right Truncated method (previous best), where the first provides a slightly better fit.

6 Methods

We developed three different scripts, each one with its own function: $summary_table.R$, which printed a summary table for all the languages, *solution.R*, which calculates the best models for all the cases (including calcuation of parameters, likelyhoods, solver use, and altmann implementation which required a little more of work), and *plot.R*, which plots the provided language and model.