CSN - Second Lab

Kymry Burwell, Laura Cebollero

October 7th, 2018

Introduction

The aim of this lab project is to analyze a degree distribution and select a theoretic model that best fits it. There are three sequences of which we can work on:

1. Undirected degree sequence.
2. In-degree sequence.
3. Out-degree sequence.

In our case, we have chosen to work with the out-degree one for 10 different languages.

The distributions we will be testing are the following:

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

The first step on the analysis is to compute different metrics for each language, such as the length of the sequence (N) and the maximum degree, among others.

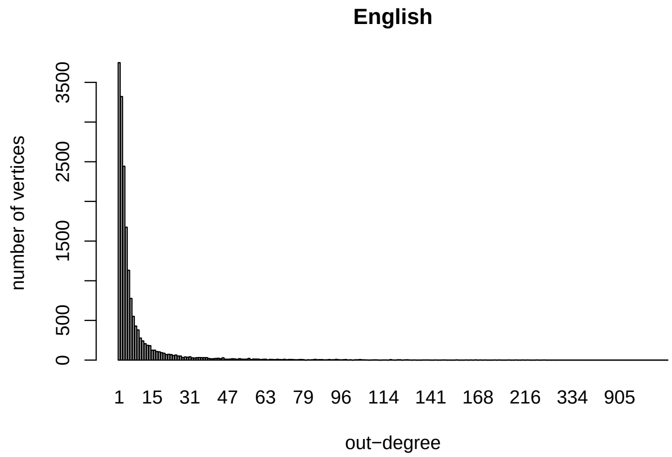
Additionally, to make our computations easier we have added a couple of metrics that we were not required in Table 1, which are MP and C.

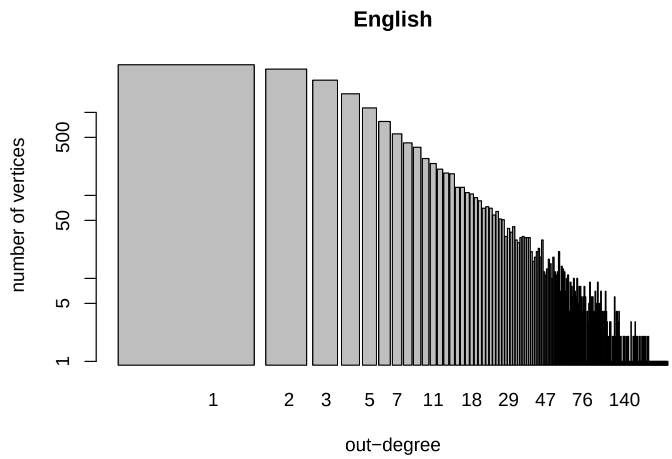
The resulting table is the following:

Language	N	M	Maximum Degree	M/N	N/M	MP	C
Arabic	15678	70589	4896	4.502424	0.2221026	12530.413	165907.83
Basque	6188	25876	2097	4.181642	0.2391405	4231.383	54154.09
Catalan	24727	204095	6622	8.253933	0.1211544	29926.062	561322.53
Chinese	23946	185013	7537	7.726259	0.1294287	24832.108	549519.06
Czech	41912	262218	12671	6.256394	0.1598365	41038.656	721024.15
English	17775	200041	7040	11.254065	0.0888568	23919.120	657764.54
Greek	9280	44768	2737	4.824138	0.2072909	8938.332	91074.93
Hungarian	25534	107178	1020	4.197462	0.2382392	21493.722	177186.08
Italian	12285	56829	1671	4.625885	0.2161748	11701.853	104228.03
Turkish	15287	47186	4488	3.086675	0.3239732	8162.505	108443.77

In the table above, N represents the number of nodes in the network, M is the sum of degrees of all nodes, Maximum Degree is the highest degree in the give language, M/N is the average degree, N/M is the inverse of the average degree, MP is the sum of the log of degrees, and C is the the following $\sum_{i=1}^{N} \sum_{j=2}^{k_i} log(j)$.

Next, we will look at a few bar plots for the English langauge to get a visual idea of the degree distribution.





We can see from the above plots that nodes with small out-degree are more common than nodes with high out-degree.

Results

Having computed the basic metrics, we now proceed to compute the most likely parameters for the various distributions. To do this, we are trying to find the parameters that minimize the minus log-likelihood function. To help expedite the process, we begin with default parameters, which act as our best initial guess. Theses consist of the following:

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

Language	lambda	q	gamma 1	gamma 2	k_max
Arabic	4.449833	0.2221026	1.797628	2	1.792754
Basque	4.113253	0.2391405	1.887150	2	1.881876
Catalan	8.251780	0.1211544	1.590979	2	1.575434
Chinese	7.722839	0.1294287	1.662662	2	1.653665
Czech	6.244246	0.1598365	1.690866	2	1.685455
English	11.253920	0.0888568	1.545278	2	1.524973
Greek	4.783788	0.2072909	1.699111	2	1.685881
Hungarian	4.129952	0.2382392	1.769320	2	1.752150
Italian	4.578370	0.2161748	1.704723	2	1.687240

Language	lambda	q	gamma 1	gamma 2	k_max
Turkish	2.920239	0.3239732	2.042634	2	2.041608

In order to verify the methods we used to compute the minus log-likelihood parameters, we will test them on contrived data where the distribution is known a priori. We have 8 different data sets that were created using geometric and zeta distributions of varying parameters.

Geometric test

For the different instances of the geometric test table, we have computed first the coefficients table:

Test	Lambda	q	gamma 1	gamma 2	k_max
probability_0.05	19.242000	0.0519696	1.332493	2	1.001000
probability_0.1	10.297653	0.0971062	1.419459	2	1.062752
probability_0.2	4.983631	0.1992826	1.571629	2	1.200408
probability_0.4	2.213289	0.4024145	1.891404	2	1.544710
probability_0.8	1.000001	0.7898894	2.999125	2	2.769765

And the delta AIC one:

Test	Poisson	Geometric	Zeta Gamma 2	Zeta	RT Zeta
probability_0.05	11537.5153	0	1334.4464	3084.1451	513.3898
probability_0.1	5180.8823	0	957.3043	1941.3070	455.6337
probability_0.2	1376.4305	0	686.5640	1042.7039	322.5124
probability_0.4	254.8481	0	349.1629	361.4936	157.5391
probability_0.8	209.3423	0	56.2003	353.5856	21.9838

Zeta test

Now we are going to apply the same tests with a prior known.

For the case of the coefficients table we get the results below:

Test	Lambda	q	gamma 1	gamma 2	k_max
exponent_2	5.044273	0.1969667	1.980298	2	1.966280
exponent_2.5	1.367588	0.5449591	2.450770	2	2.431334
exponent_3	1.000001	0.7326007	2.996023	2	2.990149
exponent_3.5	1.000001	0.7861635	3.354107	2	3.351363
And for the del	ta AIC:				

Test	Poisson	Geometric	Zeta Gamma 2	Zeta	RT Zeta
exponent_2	16112.4546	1695.9256	3.9722601	4.325777	0
exponent_2.5	1484.8257	412.1626	3.2846050	105.332119	0
exponent_3	652.7395	224.5615	0.4967527	296.812569	0
exponent_3.5	800.1710	275.9646	0.1566700	412.827364	0

Test conclusions

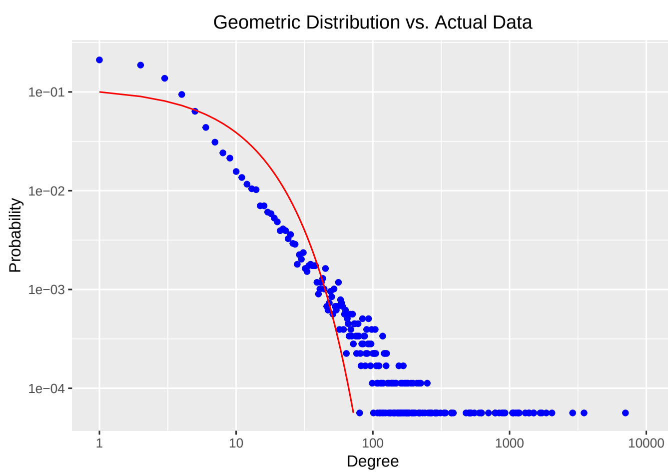
Having obtained these parameters, we can now proceed to obtain the -2 log Likelihood for each method and compute the AIC.

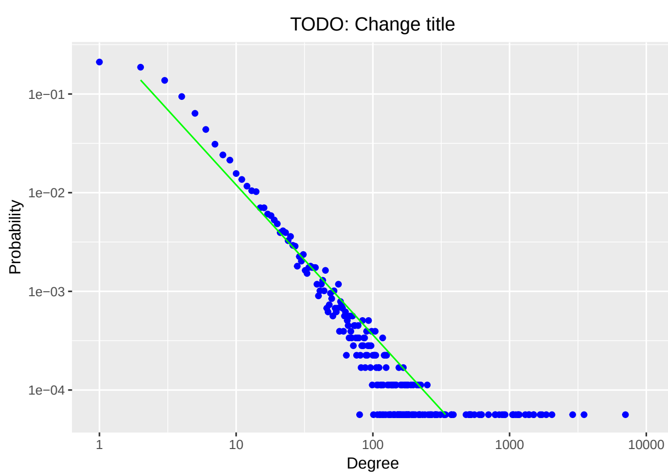
Once computed, we can produce the delta table by substracting the best AIC of each Language from the other methods' AIC. The resulting table is the following:

Language	Poisson	Geometric	Zeta gamma = 2	Zeta	RT Zeta
Arabic	195299.90	9845.304	24.290144	811.15054	0
Basque	62828.34	5475.134	9.127080	92.09307	0
Catalan	532832.91	14284.587	214.857445	7890.72998	0
Chinese	593734.76	23826.904	95.348947	4420.43102	0
Czech	804930.45	30652.481	88.047080	6089.88546	0
English	641583.86	14442.010	233.502463	7833.21273	0
Greek	86938.16	1996.018	53.963015	1292.39877	0
Hungarian	150977.08	8228.737	177.556870	1929.78923	0
Italian	90403.46	1955.514	97.791005	1659.24507	0
Turkish	155500.29	11597.395	2.854096	25.96224	0

We can see that, in the case of the out-degree sequence, the method what works best is the right-truncated zeta one.

Discussion





Methods