Complex and Social Networks: Lab session 4

Model selection for k^2

Amalia & Egon

Contents

Introduction

In this session, we are going to practice on the fit of a non-linear function to data using collections of syntactic dependency trees from different languages. In a syntactic dependency trees, the vertices are the words (tokens) of a sentence and links indicate syntactic dependencies between words [Ferrer-i-Cancho, 2013].

We will investigate the scaling of $\left\langle k^2 \right\rangle$ as a function of n, where $\left\langle k^2 \right\rangle$ is defined as the degree 2nd moment.

Data preparation

In order to start our analysis, we ensure that the validity of $\left\langle k^2 \right\rangle$ holds, where $\left\langle k^2 \right\rangle$ should satisfy

$$4 - 6/n \leq \left\langle k^2 \right\rangle \leq n - 1$$

and

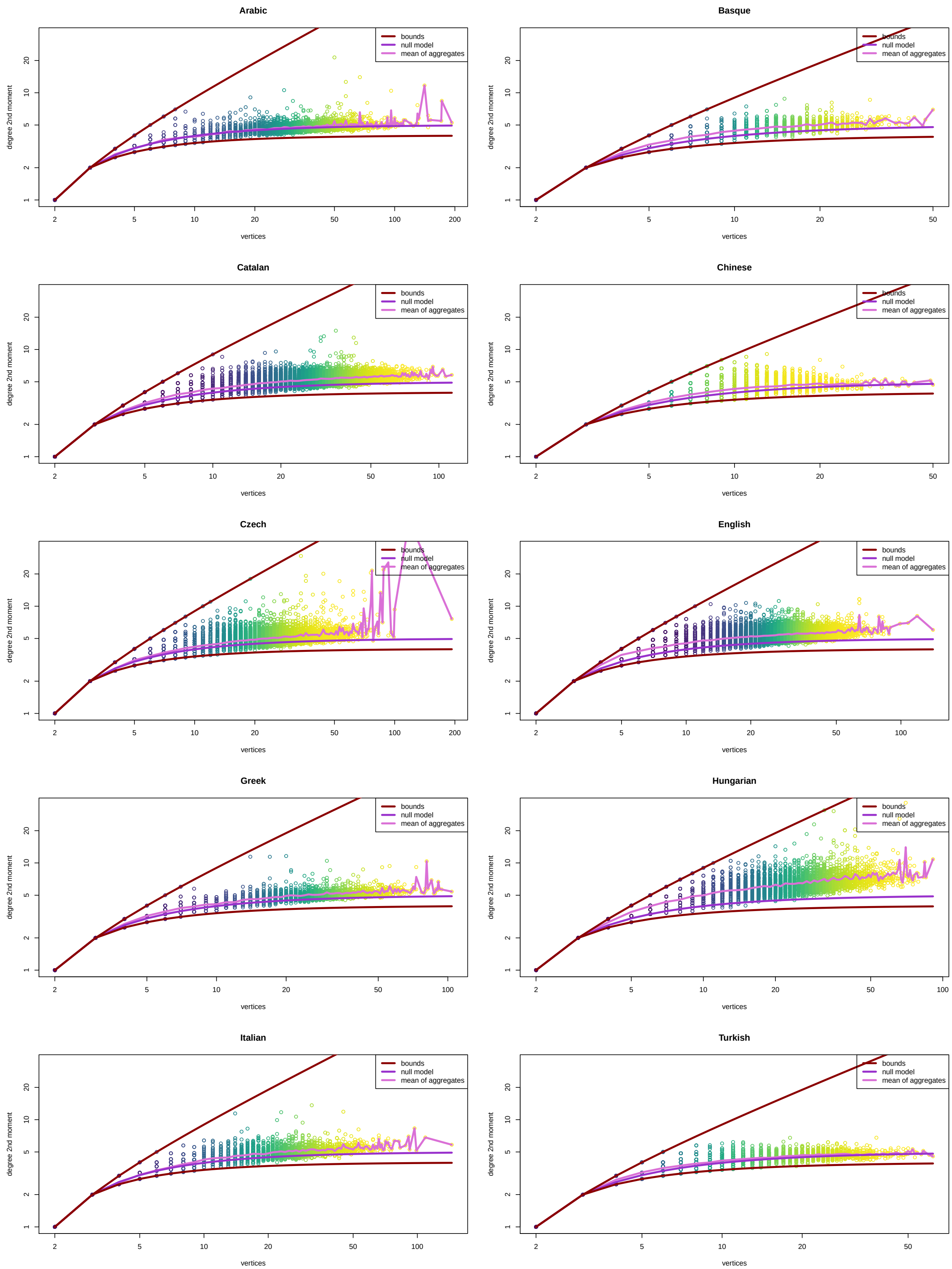
$$\frac{n}{8(n-1)} \left\langle k^2 \right\rangle + \frac{1}{2} \leq \langle d \rangle \leq n - 1$$

We set as an a acceptance threshold of e^{-5}. Then we produce a table that summarizes the properties of the syntactic dependency trees for all the languages.

```
## [1] "0 error detected"
```

	N	μ_n	σ_n	μ_k	σ_k
Arabic	4108	26.957644	20.649242	4.160443	1.2747542
Basque	2933	11.335493	6.528244	4.143336	1.0891315
Catalan	15053	25.571713	13.618702	4.961791	0.8237621
Chinese	54238	6.248884	3.310410	3.218085	1.0661125
Czech	25037	16.427647	10.721571	4.292722	1.2987953
English	18779	24.046222	11.223216	5.170150	0.8022908
Greek	2951	22.820400	14.381896	4.599747	1.0700644
Hungarian	6424	21.659869	12.566434	5.955709	1.7058038
Italian	4144	18.406612	13.345733	4.340449	1.1705960
Turkish	6030	11.101658	8.281824	3.759063	0.9341041

To have a glance of our data set and to check also visually the bounds, we plot below the preliminary visualizations of all the languages.


Arabic
Basque
Catalan
Chinese
Czech
English
Greek
Hungarian
Italian
Turkish
bounds
null model
mean of aggregates
degree 2nd moment
vertices

Results

Following, we present the results of our analysis. Firstly, we present the plots of the original data with the best fitted model and null model. Following, in Table 1 we present residual standard error of each model and in Table 2 the AIC of reach model. Finally, in Table 3 we present the AIC differences with respect to the best AIC in our ensemble of models.

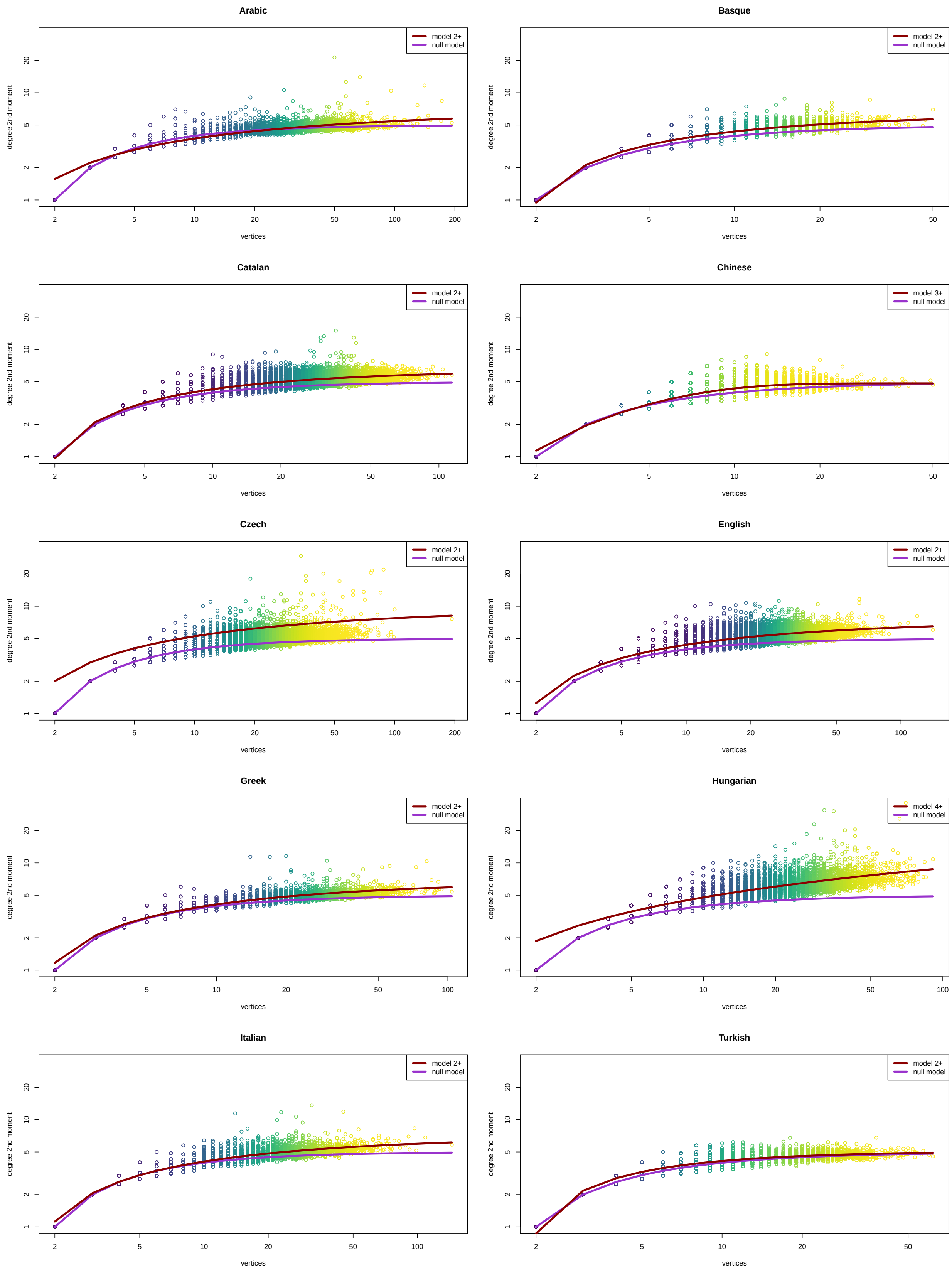

Arabic
model 2+
null model
degree 2nd moment
vertices
Basque
model 2+
null model
degree 2nd moment
vertices
Catalan
model 2+
null model
degree 2nd moment
vertices
Chinese
model 3+
null model
degree 2nd moment
vertices
Czech
model 2+
null model
degree 2nd moment
vertices
English
model 2+
null model
degree 2nd moment
vertices
Greek
model 2+
null model
degree 2nd moment
vertices
Hungarian
model 4+
null model
degree 2nd moment
vertices
Italian
model 2+
null model
degree 2nd moment
vertices
Turkish
model 2+
null model
degree 2nd moment
vertices

Table 1: residual standard error of each model

	0	1	2	3	4	1+	2+	3+	4+	5
Arabic	4.102689	1.184978	0.7214354	0.7945320	0.8244184	0.7324911	0.7032236	0.7925453	0.7101210	0.7242936
Basque	3.684651	1.110297	0.4761478	0.6996916	0.5063874	0.5311356	0.2894630	0.3356299	0.3961709	0.3906088
Catalan	3.986825	1.170825	0.3889786	0.6054668	0.5561838	0.4391683	0.1946655	0.5976964	0.3293115	0.2968101
Chinese	3.734550	1.108562	0.4622285	0.6547797	0.5885212	0.5092122	0.1736408	0.1645445	0.3984868	0.2802140
Czech	3.767383	6.338515	6.3528770	6.5038957	6.6111555	6.3712810	6.8502309	6.6211785	6.5820075	6.3898819
English	3.926751	1.290084	0.5339175	0.7337273	0.6453667	0.5742069	0.4448171	0.5264377	0.4908780	0.5111986
Greek	3.965243	1.215584	0.6555272	0.7983149	0.7348555	0.6830178	0.5937077	0.6379252	0.6245618	0.6306236
Hungarian	3.714361	1.619771	0.9724685	1.1991892	0.9285518	1.0219834	1.0193215	0.9731546	0.9228295	0.9538962
Italian	3.961409	1.162587	0.5001893	0.7095665	0.5896824	0.5397196	0.4059884	0.4707072	0.4537302	0.4674913
Turkish	3.882096	1.055345	0.3919239	0.5622000	0.5877715	0.4315421	0.1139717	0.1433716	0.3422016	0.2454410

Table 2: Akaike information criterion of each model

	0	1	2	3	4	1+	2+	3+	4+	5
Arabic	681.3395	384.2749	266.16540	289.32792	297.20256	269.81540	261.00781	289.70577	262.37162	268.093095
Basque	230.7416	130.9674	60.81138	93.14394	65.02066	69.99166	19.94116	32.37161	45.36524	45.114200
Catalan	539.9713	305.7110	95.12272	180.07799	162.79290	118.42352	-36.81437	178.57122	63.15084	44.173080
Chinese	231.8715	130.8361	58.31919	87.57131	77.64678	66.45086	-22.98607	-27.50593	45.85485	17.213364
Czech	485.1762	577.7369	579.11787	583.25274	585.14894	579.62700	593.35455	587.36897	585.35391	581.110832
English	492.4682	297.5560	143.26765	199.21736	175.65061	156.07136	112.10466	141.75533	128.47561	136.585372
Greek	482.9990	280.6302	175.39555	209.29045	194.06157	182.46150	159.32852	171.68395	167.07255	169.703934
Hungarian	444.4455	310.9919	229.32029	263.26954	220.85290	237.36567	237.91131	230.40268	220.83259	227.164600
Italian	477.2415	269.8238	127.42496	186.86845	154.42451	140.35571	92.92212	118.06708	110.85272	116.901643
Turkish	318.3858	170.8911	58.94067	100.07073	104.16857	69.91858	-80.90859	-54.74683	43.47457	6.541522

Table 3: AIC differences with respect to the best AIC in our ensemble of models

	0	1	2	3	4	1+	2+	3+	4+	5
Arabic	420.3317	123.26712	5.157588	28.32010	36.1947405	8.807583	0.000000	28.69795	1.363807	7.085280
Basque	210.8005	111.02626	40.870229	73.20278	45.0795070	50.050505	0.000000	12.43046	25.424086	25.173044
Catalan	576.7857	342.52536	131.937091	216.89236	199.6072673	155.237883	0.000000	215.38559	99.965205	80.987446
Chinese	259.3775	158.34202	85.825119	115.07724	105.1527107	93.956789	4.519857	0.00000	73.360776	44.719294
Czech	0.0000	92.56073	93.941714	98.07658	99.9727857	94.450842	108.178389	102.19281	100.177749	95.934673
English	380.3635	185.45131	31.162985	87.11270	63.5459523	43.966704	0.000000	29.65067	16.370950	24.480711
Greek	323.6705	121.30168	16.067030	49.96192	34.7330510	23.132977	0.000000	12.35543	7.744026	10.375412
Hungarian	223.6129	90.15935	8.487695	42.43695	0.0203093	16.533073	17.078718	9.57009	0.000000	6.332008
Italian	384.3194	176.90165	34.502845	93.94634	61.5023956	47.433596	0.000000	25.14496	17.930606	23.979525
Turkish	399.2944	251.79965	139.849257	180.97932	185.0771587	150.827172	0.000000	26.16176	124.383159	87.450112

Table 4: best fittings for model parameters

	Arabic	Basque	Catalan	Chinese	Czech	English	Greek	Hungarian	Italian	Turkish
1: b	0.4512250	0.6149676	0.5038438	0.5795037	0.6251585	0.5266968	0.5127181	0.6187676	0.5113627	0.5310229
2: a	2.5482871	2.2257613	2.7693416	2.4003487	0.3483891	2.7259454	2.5404635	2.5069573	2.5168473	2.5461158
2: b	0.1678985	0.2583998	0.1750107	0.2070931	0.7713861	0.1947691	0.1964578	0.2835524	0.1992913	0.1753552
3: a	4.0746292	3.6407175	4.3609231	3.6085510	4.4334094	4.5199560	4.1381843	4.9388527	4.2358559	3.7217363
3: c	0.0028679	0.0115088	0.0037733	0.0086448	0.0085207	0.0041893	0.0047291	0.0075865	0.0042247	0.0059642
4: a	1.2485380	1.6226295	1.4229568	1.4855724	1.9505059	1.5207777	1.4288740	1.9700695	1.4319624	1.3757218
1+: b	0.3112882	0.4502544	0.3399191	0.3829720	0.6387276	0.3702604	0.3620231	0.4930659	0.3632876	0.3366462
1+: d	2.1466016	1.9386527	2.4601615	2.0009873	-0.4264037	2.4965765	2.2315574	2.6411404	2.2240029	2.0980963
2+: a	-6.8112951	-8.3572640	-8.1901516	-8.3497294	-10.0000000	-8.4283702	-7.8553808	-10.0000000	-8.0272883	-7.9387197
2+: b	-0.3125482	-0.5945588	-0.5577472	-0.9243988	-0.3225141	-0.4274324	-0.4355463	-0.3554523	-0.4195528	-0.9056871
2+: d	7.0553826	6.4776542	6.5272167	5.2105319	10.0000000	7.5154740	6.9801687	9.9508749	7.1230184	5.1060520
3+: a	10.0000000	-5.6587970	10.0000000	-6.0885807	-8.5462102	-4.5530482	-4.2641004	-6.3768787	-4.4592920	-5.5083800
3+: c	0.0013548	-0.1720884	0.0018908	-0.2506649	-0.0220420	-0.0858583	-0.0861897	-0.0537432	-0.0896434	-0.2255580
3+: d	-5.9946825	5.4037967	-5.7124884	4.8272838	10.0000000	6.0718439	5.6760633	8.2435009	5.6833995	4.7803803
4+: a	0.7894978	1.2271915	0.9151143	0.9466820	2.9955377	1.0464711	0.9831902	1.8019057	1.0052245	0.7989028
4+: b	1.8755846	1.2318294	1.9570060	1.6784053	-3.9669514	1.7962037	1.6757035	0.6208072	1.6038420	1.9441683
5: a	2.4911974	1.5033763	1.9561945	1.4076547	0.2914703	2.2517722	1.9450465	1.9088502	1.9993473	1.6125379
5: b	0.1769061	0.4932613	0.3291229	0.5351842	0.8238923	0.2789184	0.3175517	0.4064232	0.3008405	0.4293627
5: c	-0.0001918	-0.0128132	-0.0043521	-0.0185104	-0.0005634	-0.0023779	-0.0036447	-0.0037940	-0.0029018	-0.0116954

Discussion

The best model seems to be, in most cases, the model 2+, $f(n) = an^b + d$. The fit seems to be good, both visually and in terms of AIC, we have usually a better fit than the fit of the null model.

We have some exceptions: The best model for the Czech seems to be the null model, but we think that is not significant because seems very hard in general to find good fits for it, maybe due to the high numbers of heavy outliers.

The best model for Hungarian is 4+, closely followed by model 4. These types of models seems to really capture it, in fact, we have a really different fit in respect to the null model.

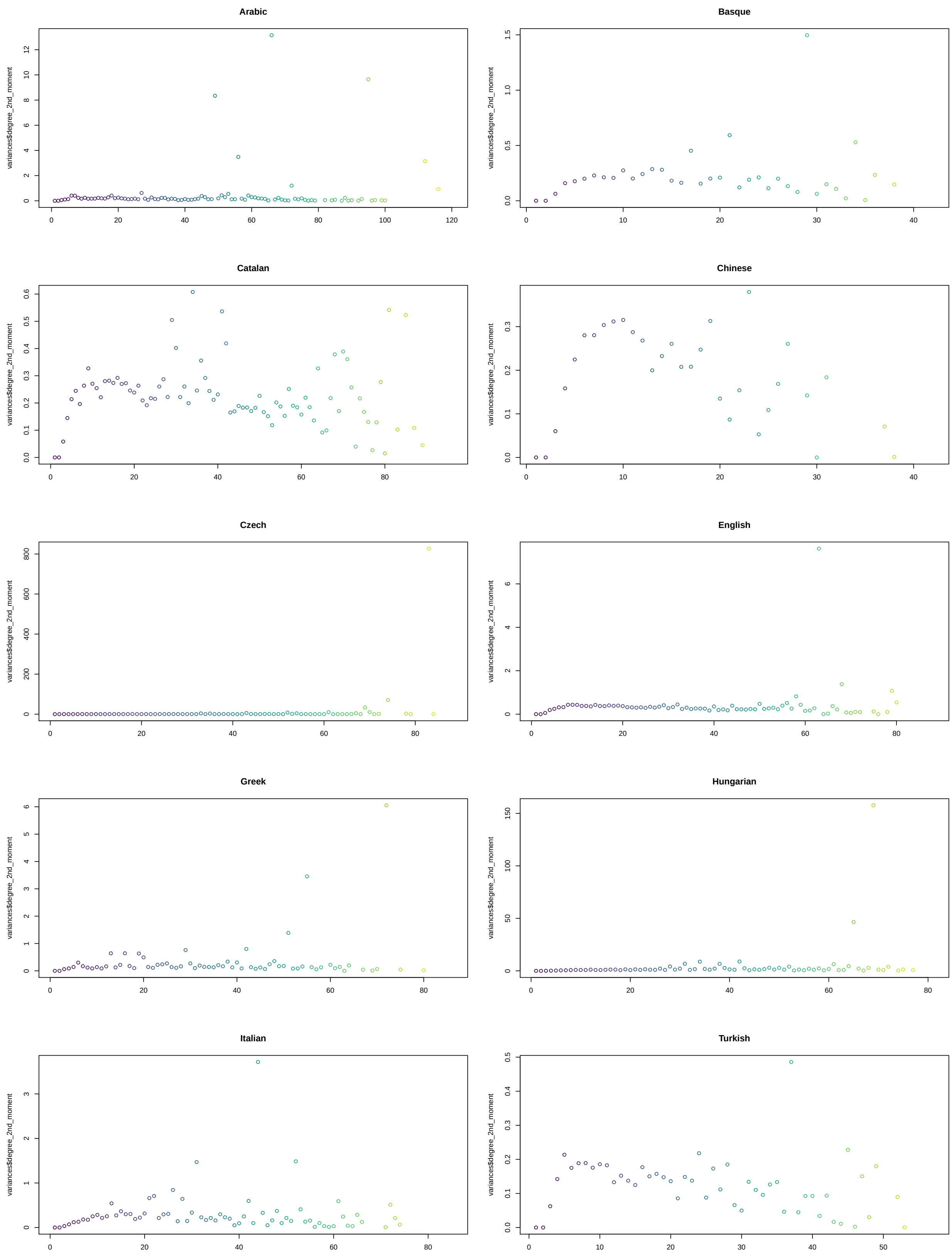
The best model for Chinese is 3+, but 2+ performs quite good as well.

To conclude, we can say that the 2+ model seems to outperform all the others method; even when is not the best, it performs well.

With these results it is very interesting to see that the linguistics networks that we are studying are explained indeed by a power law for the majority of the cases. We see how even for these cases that we study within our lab session show interesting perspective we can have by studying languages from a network perspective and how this approach can serve in understanding their universal properties.

Methods

Homoschedasticity and choice of aggregated data



As we see from the plots above, the points of the degree second moment show a big variability as the number of vertices grow. As well as a digitization of the points, since for the same number of vertices we see the different number for the degree, which result to what we see above as vertical lines of points. Additionally to these plots, we want to check if the assumption of homoscedasticity holds. Following, we see that for each language the variance of points as a function of the number of vertices. For all cases we see that there is no homogeneity of the variance. For this reasons we decided to proceed our analysis using the aggregated version of the data. Since we will take the average of all point for a specific number of vertices, it will serves us in having a more homogeneous version of the data.

fitting the models

We fit all the models explained in the task (since we have two times model 4, we called the model in the advanced section "model 5").

We used the nls(..) function. For model with just one or two parameters, we discovered that is usually not important to give very precise starting point (unless in some cases where some specific starting point caused the failure of the algorithm), because the fit is very easy and the tedious search for better starting points didn't seem to reward us in terms of better algorithm (we usually don't need a lot of iterations).

Things became tricky once we introduced more complicated models: to fit 3 or more parameters seemed a completely different task for our poor optimizer.

We tried to change the algorithm of nls(..) by setting it to 'port' that let us impose some bounds on the variable, but without a good starting point it only solved the problem by squishing one of the 3 parameters to the bound and then solving for the others.

The 'heuristic' that we used to get some results is this: we found good parameters for the Catalan data set that seems overall easier to work with, and then we set that values as starting points for other languages, letting the bounds be an interval containing them, not too small to constrain too much, but not to big to made the algorithm fail.

However, this doesn't seem to help for the model 5+; 4 parameters are really difficult to recover. We tried to set the nls algorithm to stop after some iterations, but the results were not satisfactory, so we decided to not include that.