# Homework 2
# Degree Distribution Analysis

Juan Pablo Royo Sales & Francesc Roy Campderrós

Universitat Politècnica de Catalunya

October 17, 2020

## Contents

# 1 Introduction

In this homework we have analyzed Incomming Degree Language Set (Incomming Degree) 5 which contains the distribution over different natural languages for the incomming connection words frequency.

The goal of this work is to analyzed according to what have been discussing in class, what is the best model that fits (or explain), this distribution over the whole set of languages, taking into consideration only **five** different models: Displaced Poisson Model (Displaced Poisson), Displaced Geometric Model (Displaced Geometric), Zeta Model with fixed Gamma with value 2 (Zeta-Gamma-2), Zeta Model (Zeta) and Right-Truncated Zeta Model (Right-Truncated Zeta)

Additionally we are going to compare these results using Akaike Information Criterion (Akaike) in order to determine what model fits best for each language.

At the end of the work the we are going to run the distribution against Altmann Model (Altmann), that has been pointed out during the lectures to fit better.

This analysis work is going to be done in the following way:

- **Results** section: We are going to show the results obtained for each language according to the previous statements.
- **Discussion and Methods** section: In this section we are going to conclude and give some explanaition based on the results.
- **Conclusions** section: Finally we are going to give our opinion about the results obtained and what we have learnt.

# 2 Results

Lets divide the results for each language, but first lets see the Summary of the data that we are analysing.

## 2.1 Distributions

| Language | N | Maximum degree | M/N | N/M |
|---|---|---|---|---|
| Arabic | 21065 | 2249 | 3.35100878234037 | 0.298417600475995 |
| Basque | 11868 | 576 | 2.18031681833502 | 0.458648941103725 |
| Catalan | 35524 | 5522 | 5.74527080283752 | 0.174056199318945 |
| Chinese | 35563 | 7645 | 5.20240137221269 | 0.192218925156611 |
| Czech | 66014 | 4727 | 3.97215742115309 | 0.25175235872442 |
| English | 29172 | 4547 | 6.8572946661182 | 0.14583010482851 |
| Greek | 12704 | 1081 | 3.52392947103275 | 0.283774124374553 |
| Hungarian | 34600 | 6540 | 3.09763005780347 | 0.322827445931068 |
| Italian | 13433 | 2678 | 4.23055162659123 | 0.236375794048813 |
| Turkish | 20403 | 6704 | 2.31269911287556 | 0.432395201966685 |

Table 1: Language List - Paramters

## 2.2 Fitting Model Results

Lets remind the model classification in the Appendix section C

| | 1 | 2 | 4 | 5 |
|---|---|---|---|---|
| Language | $\lambda$ | $q$ | $\gamma$ | $k_{max}$ |
| Arabic | 3.216673 | 0.2984176 | 2.104113 | 24065 |
| Basque | 1.830868 | 0.4586489 | 2.358005 | 14868 |
| Catalan | 5.726551 | 0.1740562 | 1.922218 | 38524 |
| Chinese | 5.172914 | 0.1922189 | 1.885594 | 38563 |
| Czech | 3.891029 | 0.2517524 | 2.050873 | 69014 |
| English | 6.850030 | 0.1458301 | 1.799843 | 32172 |
| Greek | 3.407164 | 0.2837741 | 2.131930 | 15704 |
| Hungarian | 2.932666 | 0.3228275 | 2.335141 | 37600 |
| Italian | 4.164842 | 0.2363758 | 2.108419 | 16433 |
| Turkish | 1.999577 | 0.4323952 | 2.543344 | 23403 |

Table 2: Model Fitting by Language

### 2.3 AIC Results

| Language | AIC 1 | AIC 2 | AIC 3 | AIC 4 | AIC 5 |
|---|---|---|---|---|---|
| Arabic | 240318.35 | 24186.041 | 172.3717 | **0.0000000** | 1.643912 |
| Basque | 50164.36 | 8363.878 | 845.1732 | **0.0000000** | 1.973784 |
| Catalan | 913866.77 | 61868.421 | 212.5898 | 0.6420711 | **0.000000** |
| Chinese | 618335.67 | 48666.087 | 491.2543 | 1.9569369 | **0.000000** |
| Czech | 940668.86 | 91090.542 | 138.4353 | **0.0000000** | 1.354231 |
| English | 739544.11 | 45704.816 | 1432.2985 | 7.6821378 | **0.000000** |
| Greek | 157133.81 | 17130.408 | 160.6789 | **0.0000000** | 1.740389 |
| Hungarian | 468252.20 | 53469.625 | 2220.2829 | **0.0000000** | 1.971684 |
| Italian | 245803.03 | 22877.409 | 117.8825 | **0.0000000** | 1.668910 |
| Turkish | 193345.03 | 24615.351 | 2740.4657 | **0.0000000** | 1.996789 |

Table 3: AIC Results

### 2.4 Altmann Results

| Language | $\gamma$ | $\delta$ |
|---|---|---|
| Arabic | 2.086211 | 0.0014726478 |
| Basque | 2.346528 | 0.0017016417 |
| Catalan | 1.907315 | 0.0008281837 |
| Chinese | 1.834734 | 0.0028665707 |
| Czech | 2.035468 | 0.0010169897 |
| English | 1.740599 | 0.0025428652 |
| Greek | 2.131722 | 0.0000010000 |
| Hungarian | 2.335120 | 0.0000010000 |
| Italian | 2.108182 | 0.0000010000 |
| Turkish | 2.543335 | 0.0000010000 |

Table 4: Altmann Results

## 3 Discussion and Methods

### 3.1 Considerations

It is known that all the likelihood functions has been provided and the only thing that need to be adjust it to negate them. On the other hand there it is important to point out that the **lower** and **upper** bound parameters needs to be adjust for each case. The idea here is to explain those consideration that we have taken into account for each Model.

### 3.1.1 Zeta Model

In this case as we can see $\gamma = 1.0000001$ as lower bound and $\gamma = 2$ as an started point

```
fit_with_zeta <- function(N, M_P){
  minus_log_likelihood_zeta <- function(gamma) {
    N * log(zeta(gamma)) + gamma * M_P
  }
  mle( minus_log_likelihood_zeta
     , start = list(gamma = 2)
     , method = "L-BFGS-B"
     , lower = c(1.0000001))
}
```

Listing 1: Extracted from source Solution.R

### 3.1.2 Zeta Truncated

One of the important setup on this model was start value of $k_{max}$. Since it is known that $k \gg N$ we cannot put as a started value of $k_{max} = N$ because on some cases, particularly in **English**, this cannot be optimized. Therefore we have chosen $k_{max} = N + 3000$ as a started value which is $\gg N$.

```
fit_with_zeta_truncated <- function(N, M_P, MAX){
  minus_log_likelihood_right_truncated_zeta <-
  ↪  function(gamma,k_max) {
    harmonic=0
    for (i in 1:k_max){
      harmonic=harmonic+ 1/(i^(gamma))
    }
    N * log(harmonic) + gamma * M_P
  }
  mle( minus_log_likelihood_right_truncated_zeta
     , start = list(gamma = 2, k_max= N+3000 )
     , method = "L-BFGS-B"
     , lower = c(1.0000001,MAX)
  )
}
```

Listing 2: Extracted from source Solution.R

### 3.1.3 Displaced Poisson and Geometric

On those cases we have done the regular setup but taking into consideration that we are calculating $C$ value from the sample.

```
calculate_c <- function(x,n){
  C= 0
  for (i in 1:n){
    for (j in 2:x[i]){
      C= C + log(j)
    }
  }
  C
}
```

Listing 3: Extracted from source Solution.R

```
fit_with_displaced_poisson <- function(N, M, C){
  minus_log_likelihood_displaced_poisson <- function(lamda) {
    -(M*log(lamda)) + N*(lamda + log(1-exp(-lamda))) + C
  }
  mle( minus_log_likelihood_displaced_poisson
     , start = list(lamda = M/N)
     , method = "L-BFGS-B"
     , lower = c(0.0000001)
  )
}
```

Listing 4: Extracted from source Solution.R

## 3.2 General Analysis

As we can see in the table 3 the best models are between Zeta and Right-Truncated Zeta because are the ones in which the $\Delta = AIC - AIC_{best}$ is the smallest. As long as the $AIC$ value approximates to the best possible value that difference approximates to 0, indicating that the prediction is fitting the model better for all languages in general.

If we have to pick only **one** model which fits better for all languages, according to the Akaike results this should be Zeta. This is pretty obvious since in almost all languages $\Delta = 0$. For the rest of the languages which are not 0,

like **Catalan, Chinese and English**, it is still an small difference and it can be explain that it is also a good model.

In these cases as we can see in the distribution table 1 $M/N$ the average degree frequency is higher for those 3 languages. In those cases since it is clear that we have more degrees frequencies in few terms the distribution is not straightforward enough to be fitted by a Zeta model and we might be needing more parameters to fit better the model. And that the case of Right-Truncated Zeta.

Taking into consideration that we are analyzing a Incomming Degree $\{w|(u,w) \in E\}$, this means all the words that have edges from some other words connecting to this. Although we cannot give an intuitive explanation for *Chinese*, for the case of the *Catalan and English* it is clear that those words are connectors and/or prepositions in the language like *in, at, of, etc* for **English** or *a, amb, des, dalt, etc* for **Catalan**. In that case it is straightforward to see for someone who speak those languages that there are a lot of this kinds of words with high incoming degree, turning the **power law** like distribution, not so clear on the distribution tail. This would explain that we need extra parameters for fitting the model best when the degree is higher.

On the other hand as well as Zeta is the best function to fit the model, we can appreciate that Displaced Poisson is the worst for fitting them. This could be explain because *Poisson* distribution approximates well when $p$ probability is low and $m$ the number of elements is big. As we are working with an reduce number of terms of the languages, in this case Incomming Degree, the distribution is not fitting accordingly.

## 3.3 Altmann Model

One of the topic proposal in the homework was to use an Altmann function which it is known that approximates better to the model. In order to do that

we need first to derive **Log Likelihood** function.

$$
\begin{aligned}
Ł &= \sum_{i=o}^{N} \log\left[c k_i^{-\gamma} e^{-\delta k_i}\right] && \text{(1a)}\\
&= \sum_{i=o}^{N} \log(c) + \log k_i^{-\gamma} + (-\delta k_i) log(e) && \text{(1b)}\\
&= \sum_{i=o}^{N} \log(c) + -\gamma \log(k_i) + (-\delta k_i) && \text{(1c)}\\
&= N \log(c) - \gamma M' - \delta M && \text{(1d)}\\
& && \text{(1e)}
\end{aligned}
$$

As we can see in the table of results 4 $\gamma$ parameters fits well regarding Zeta which is the one that explain better the model.

# 4 Conclusions

In conclusion we can say that we have learnt how to pick a good Model for fitting a real network with the tools provided. Although that and taking into consideration that we have Akaike for measuring the accuracy of the model, we can notice the big difference between the different models such as Displaced Poisson and Zeta for example. This leads to think to the fact that without proper measuring tools like Akaike it is really difficult to determine what is the best model.

On the other hand, the difference is quite strong which indicates that real networks are difficult to predict because of its particular behavior, as we have explained in the previous section for example the case of **English** or **Catalan** where some specific details in the language can lead to a not straightforward **Power Law like** distribution.

To sum up we think it has been a great exercise to put in practice the theoretical concepts we have seen in class, but with a real case.

# A Source Code Considerations

If you want to run `./Solution.R` script to compare the results in this technical report with the sample, you need to take into consideration the following:

- **stats4** package is required to be installed in your **R** Studio
- **VGAM** package is required to be installed in your **R** Studio
- Unzip `./in-degree_sequences.tar.gz` in the same Root folder of the script. The script is looking for `./data/LANG_in_degree_sequence.txt` files to be placed in a folder in the same place that the script is running. Sometimes you need to setup as a working directory in your **R** Studio the folder of the script.
- Place `./list_in.txt` in the same root folder of the script. Same as previous item.

# B Listing with Languages Files

```
language file
Arabic ./data/Arabic_in-degree_sequence.txt
Basque ./data/Basque_in-degree_sequence.txt
Catalan ./data/Catalan_in-degree_sequence.txt
Chinese ./data/Chinese_in-degree_sequence.txt
Czech ./data/Czech_in-degree_sequence.txt
English ./data/English_in-degree_sequence.txt
Greek ./data/Greek_in-degree_sequence.txt
Hungarian ./data/Hungarian_in-degree_sequence.txt
Italian ./data/Italian_in-degree_sequence.txt
Turkish ./data/Turkish_in-degree_sequence.txt
```

Listing 5: Extracted from source list_in.txt

# C Model Classification

| Model | Function |
|---|---|
| 1 | Displaced Poisson |
| 2 | Displaced geometric |
| 3 | Zeta with $\gamma = 2$ |
| 4 | Zeta |
| 5 | Right-truncated zeta |

Table 5: Model Classification

# D Plotting

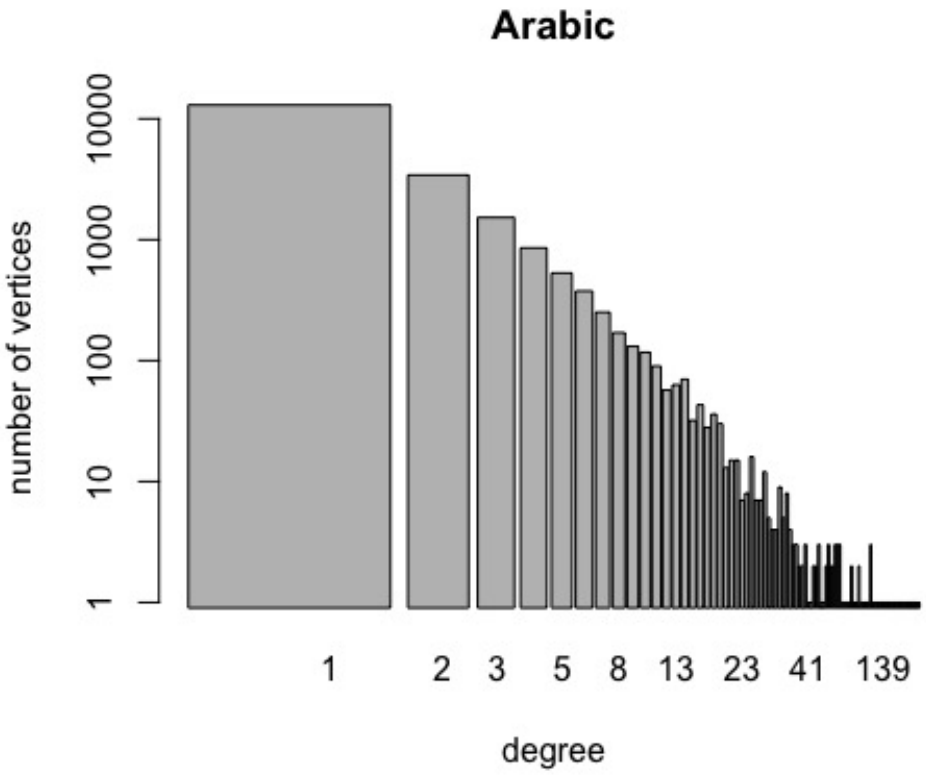


Figure 1: Arabic In Degree

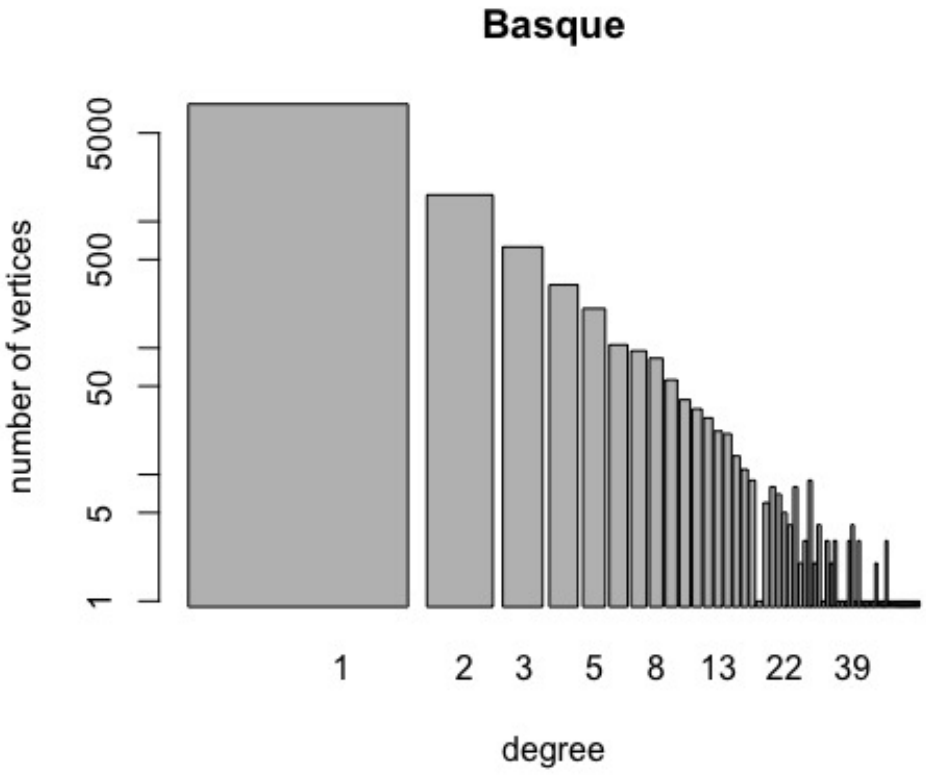


Figure 2: Basque In Degree

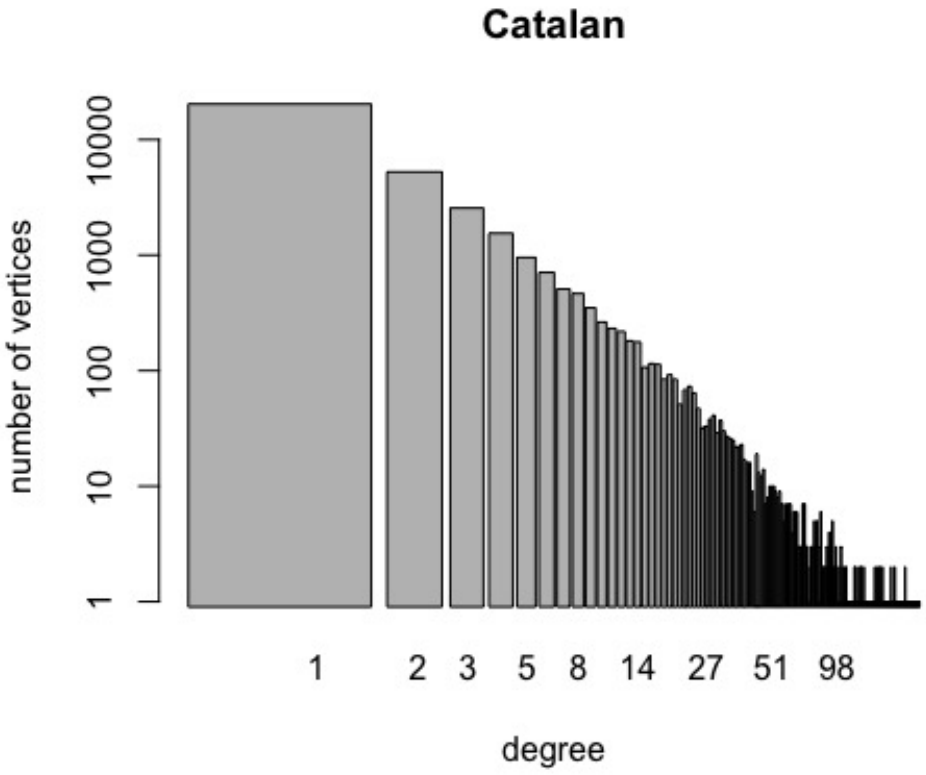


Figure 3: Catalan In Degree

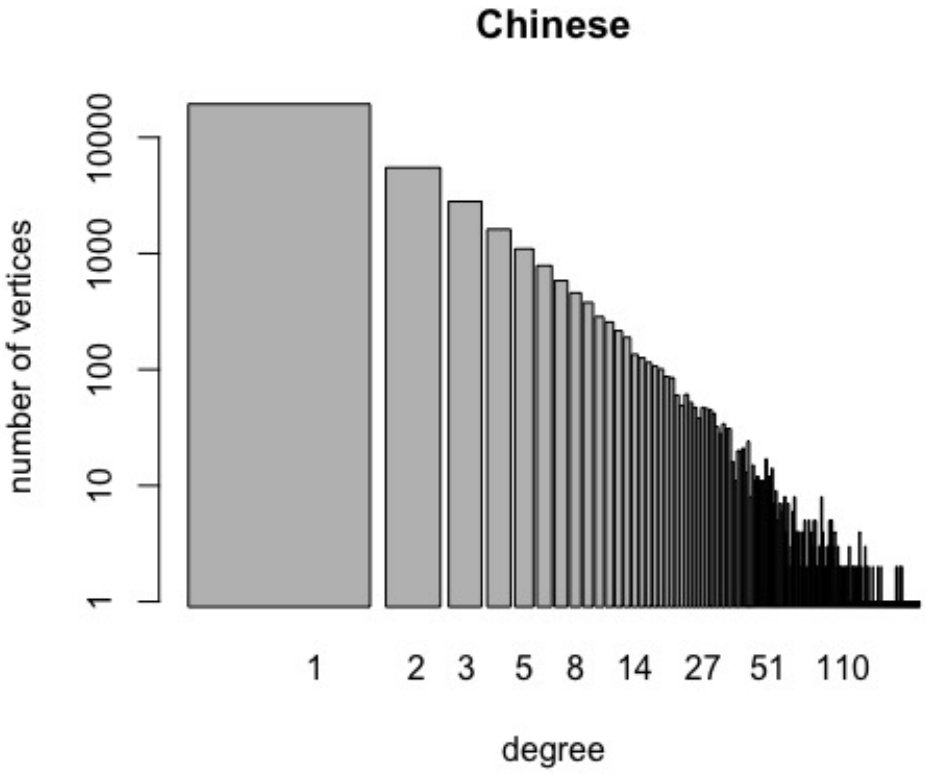


Figure 4: Chinese In Degree

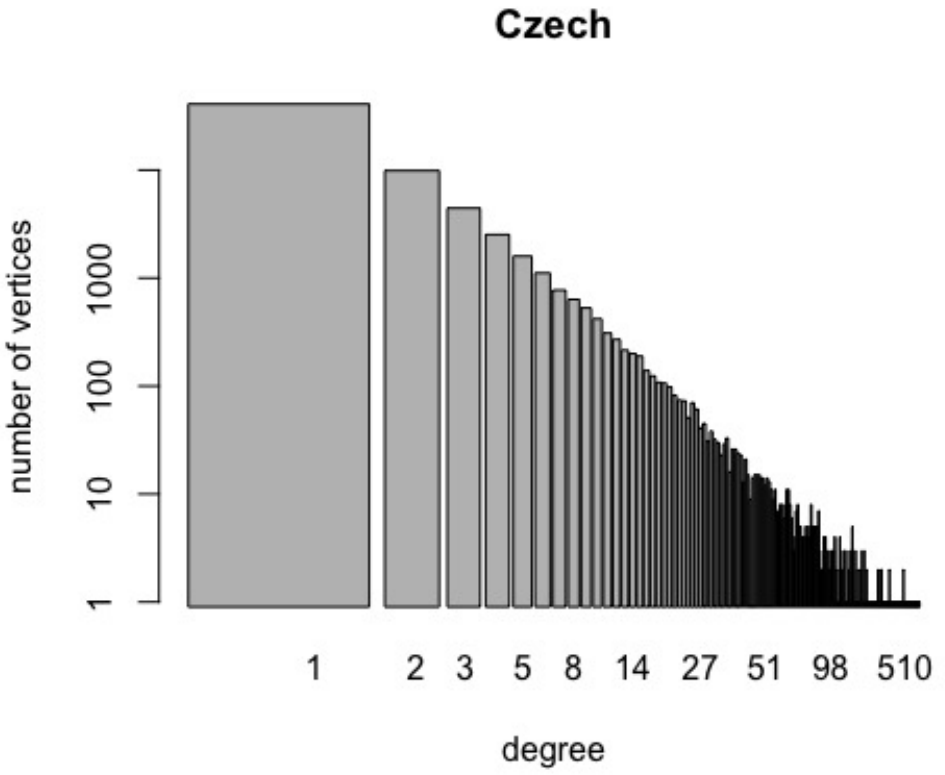


Figure 5: Czech In Degree

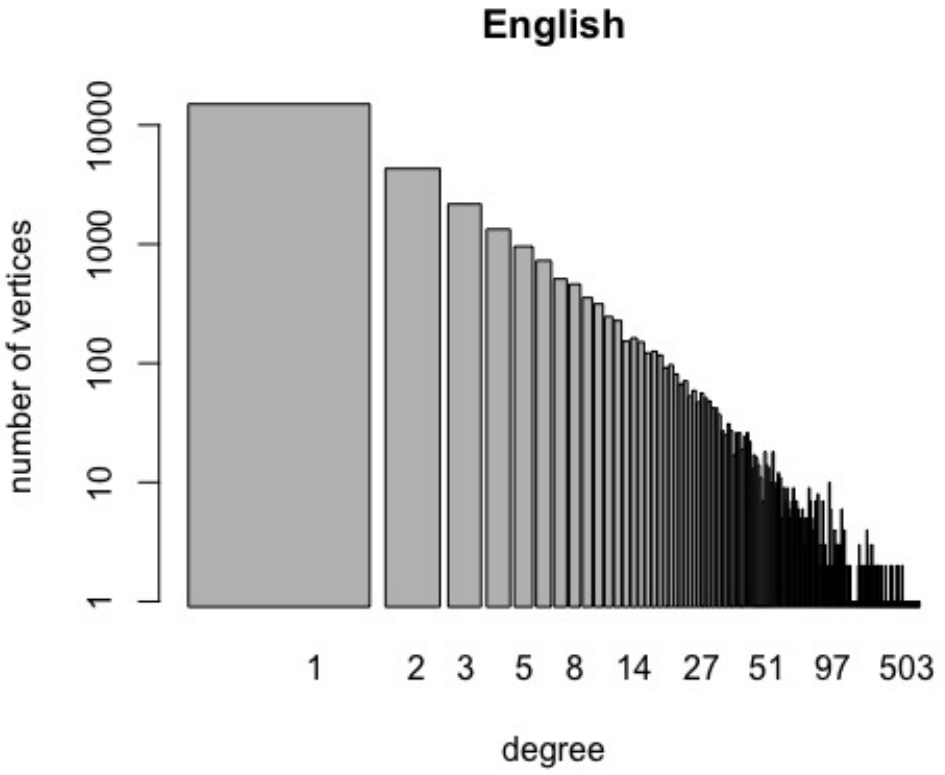


Figure 6: English In Degree

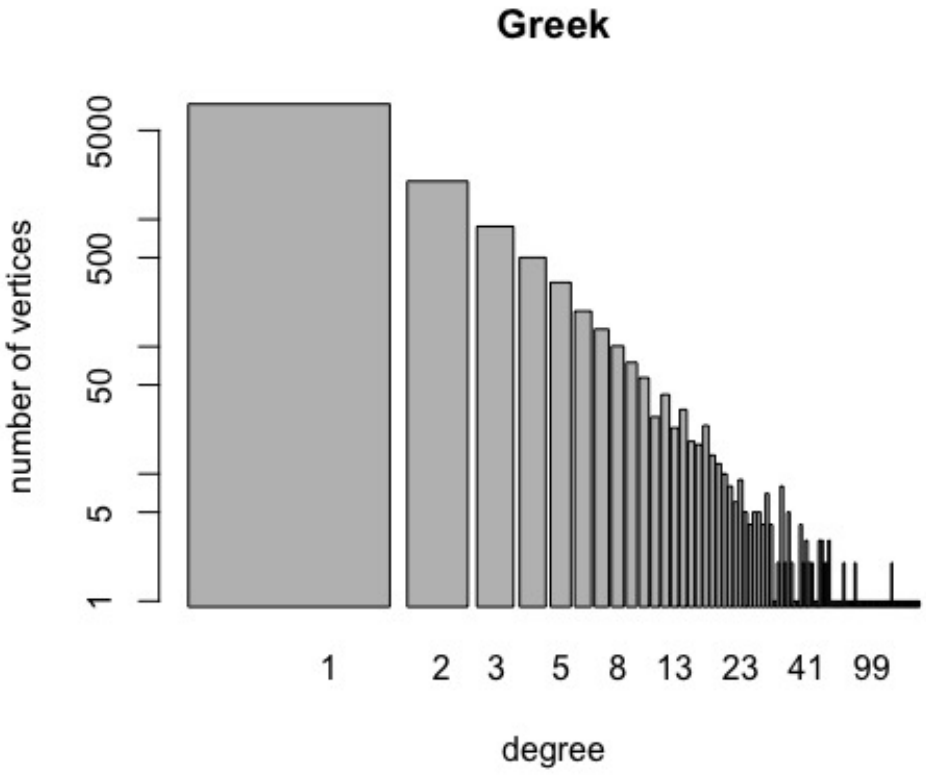


Figure 7: Greek In Degree

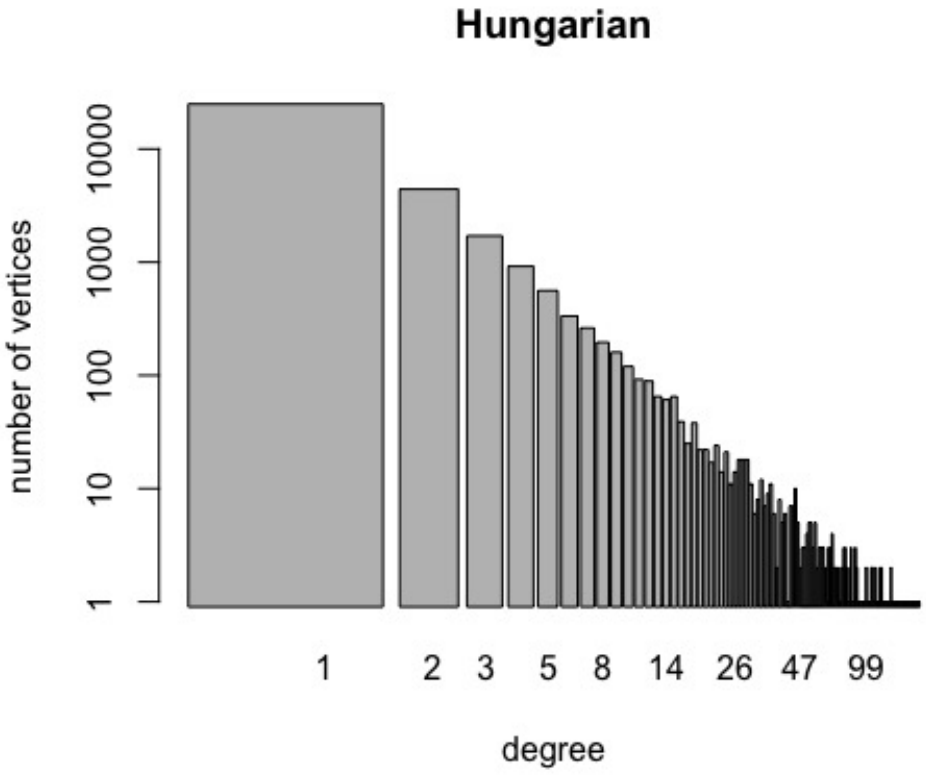


Figure 8: Hungarian In Degree

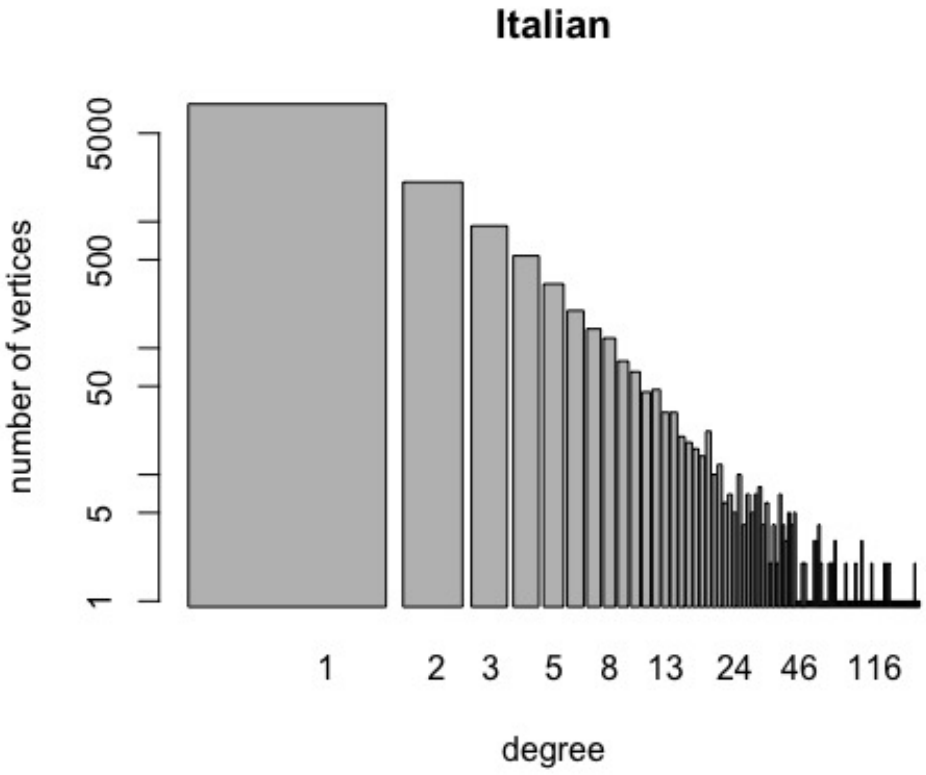


Figure 9: Italian In Degree

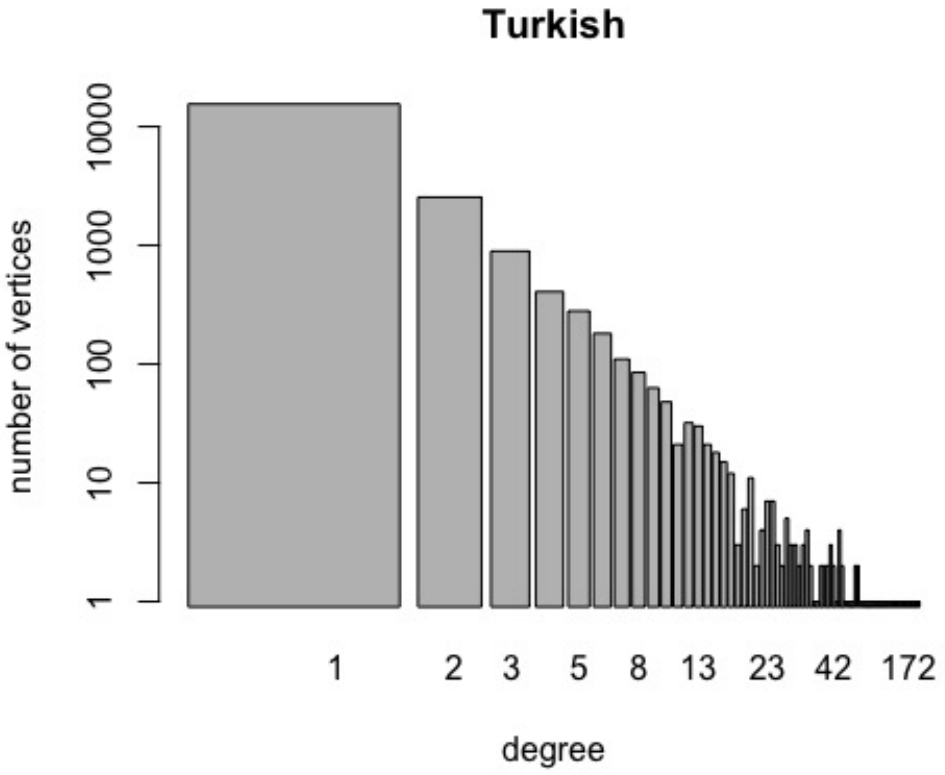


Figure 10: Turkish In Degree

# E Source code

In the source code there are 3 folders with code: