

# Fitting degree distribution

## Fitting degree distribution

The purpose of this section is to use more analytical methods to try and evaluate which distribution best represents the frequency of different orders of patents.

Because these methods are computationally expensive only a subset of the data will be analysed, choosing the years 1978, 1992, 2002, and 2012. All but the latter are the same years chosen by valverde in his plot and the latter is to represent the new data available at an appropriate gap (~10 years).

We can use the “powerLaw” package to fit different models to the in-degree distribution. This package is based on the Newman paper “Power-law distributions in empirical data” (2009). It includes several methods, \* A maximisation of likelihood search to optimise parameters of 4 different discrete models to fit the data: power law, lognormal, exponential and poisson. \* We have done this for each of the 4 discrete distributions and shown it in a graph. \* A likelihood ratio test to compare how two different models fit the data. \* We compare log-normal to power law and vice versa. \* A bootstrapping method to evaluate a p value for the data being from the distribution. \* Currently computationally too expensive ## Methods

Using a bootstrapping method to obtain a p value for distributions

- The paper says that for a given accuracy in the p value the number of simulations required scales  $\frac{1}{4}\eta^{-1/2}$ .
- The paper also approximates a p value cut off of about 0.1 to disregard the hypothesis of the tested distribution.
- To get an accuracy of 0.01 we need 2500 simulations, this is very computationally expensive so no results yet. To find which is a better fit

Likelihood ratio to compare distributions.

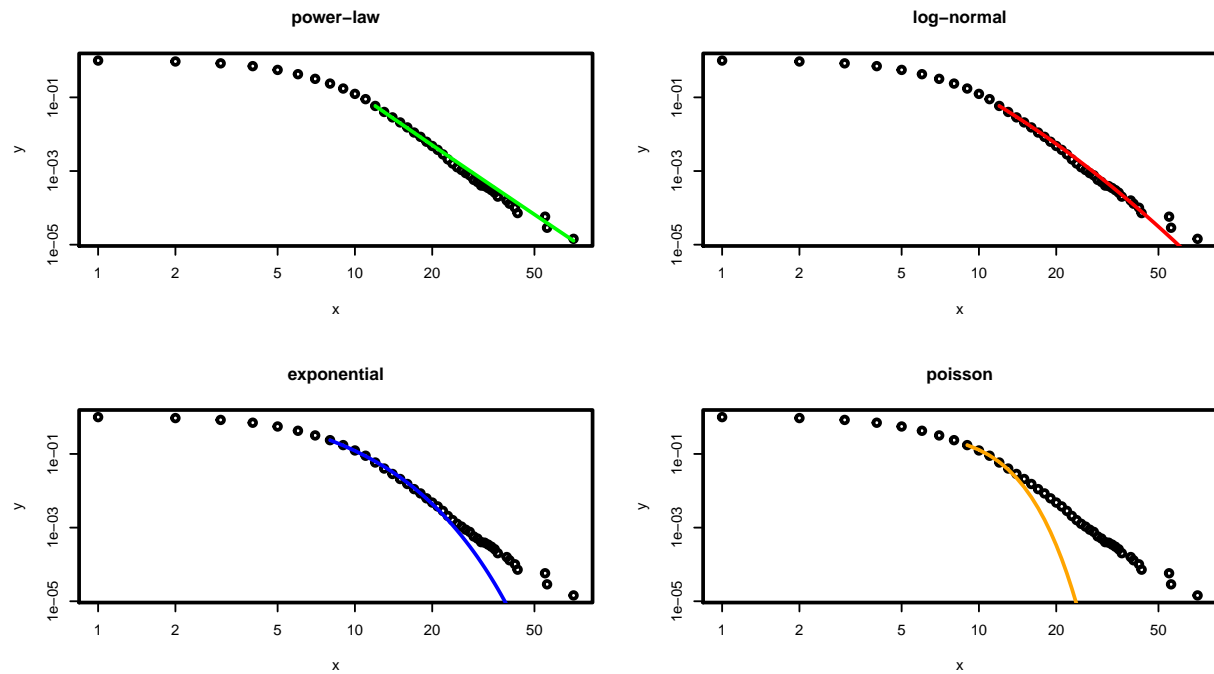
- one sided p value is the upper limit on getting that small a log-likelihood ratio if the first distribution ( $m_{pl}$ ) is true.
- two sided p value is the probability of getting a log-likelihood ratio which deviates that much from zero in either direction if the two distributions are equally good.
- Test statistic is the sample average of the log-likelihood ratio standardised by an estimated standard deviation.

## Results

We can see from the graph that exponential and poisson distributions do not fit the data well. Power-law distributions, being perfectly straight on a log-log plot are unable to bend to the dip in the tail and log-normal shows the closest fit while not necessarily describing the tail perfectly in all years.

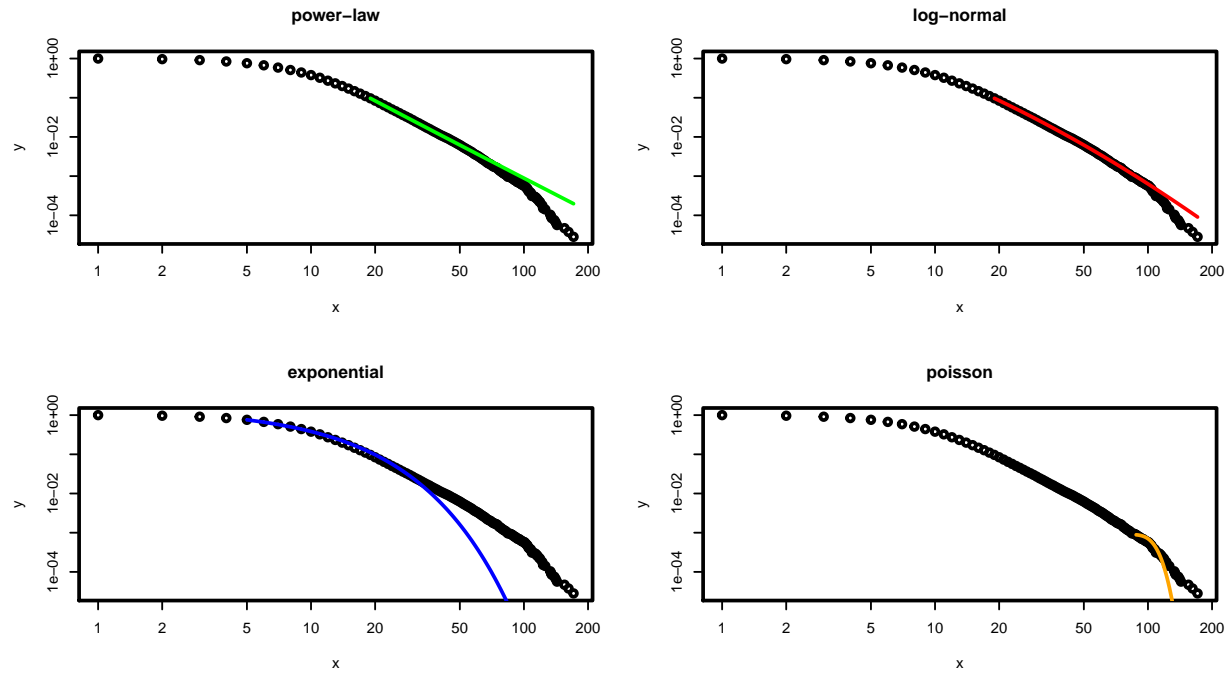
For the likelihood ratio test we reaffirm what we saw in the graphs where the power-law is optimised for a specific point in the graph where it is straightest but as you move away from that point it performs substantially worse, exemplified by the behaviour at high orders (in the tail). Furthermore the difference between power law and log-normal increases over time as the network grows larger.

1978



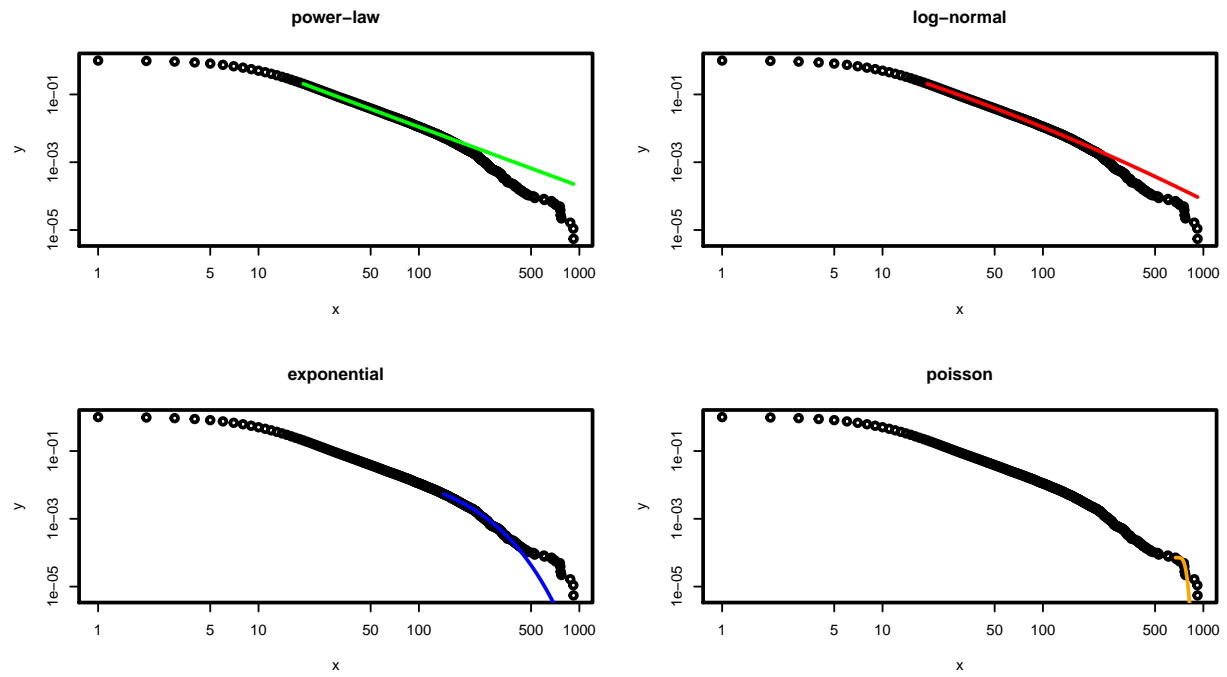
```
## $test_statistic
## [1] -1.66488
##
## $p_one_sided
## [1] 0.04796831
##
## $p_two_sided
## [1] 0.09593661
##
## $test_statistic
## [1] 1.66488
##
## $p_one_sided
## [1] 0.9520317
##
## $p_two_sided
## [1] 0.09593661
```

1992



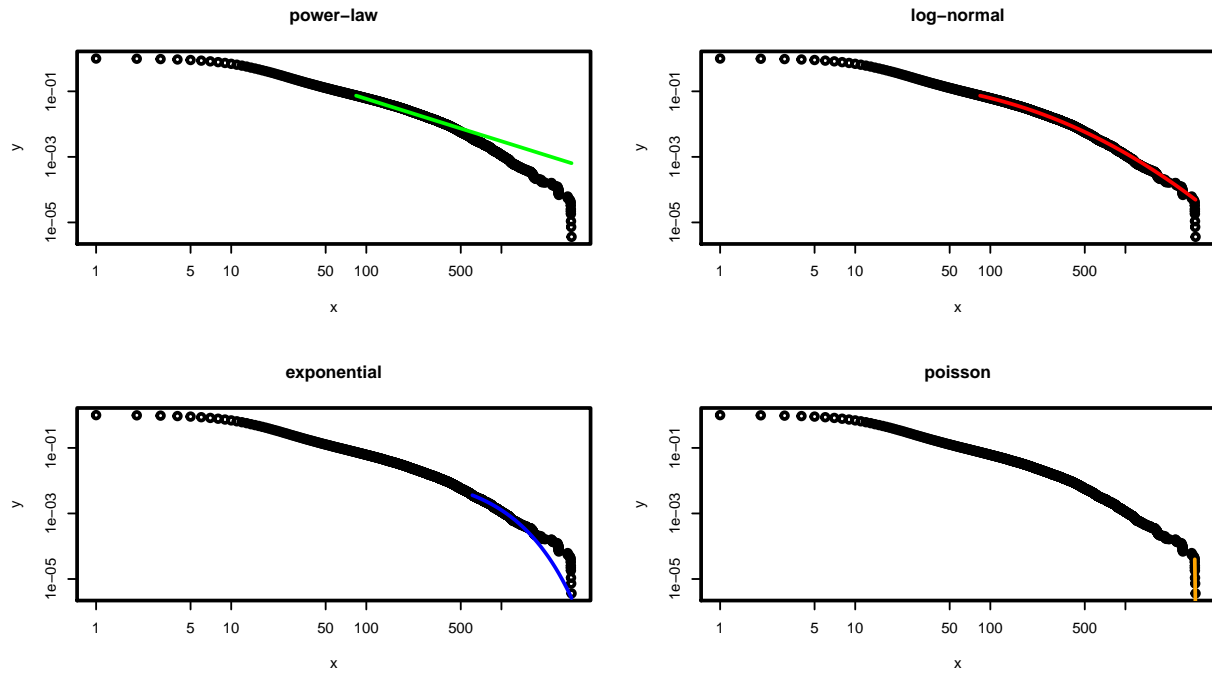
```
## $test_statistic
## [1] -3.635451
##
## $p_one_sided
## [1] 0.0001387472
##
## $p_two_sided
## [1] 0.0002774944
##
## $test_statistic
## [1] 3.635451
##
## $p_one_sided
## [1] 0.9998613
##
## $p_two_sided
## [1] 0.0002774944
```

2002



```
## $test_statistic
## [1] -6.342428
##
## $p_one_sided
## [1] 1.130857e-10
##
## $p_two_sided
## [1] 2.261714e-10
##
## $test_statistic
## [1] 6.342428
##
## $p_one_sided
## [1] 1
##
## $p_two_sided
## [1] 2.261715e-10
```

2012



```
## $test_statistic
## [1] -19.5905
##
## $p_one_sided
## [1] 9.318451e-86
##
## $p_two_sided
## [1] 1.86369e-85
##
## $test_statistic
## [1] 19.5905
##
## $p_one_sided
## [1] 1
##
## $p_two_sided
## [1] 0
```