# Comparing Cited by Examiners to Cited by Other

In this notebook we compare the distribution of citations by which group of people cited them. Cited by examiner is exactly that whereas "Cited by other"" indicates those cited in a protest, by an attorney or agent not acting in a representative capacity but on behalf of a single inventor, and by the applicant.

### Data

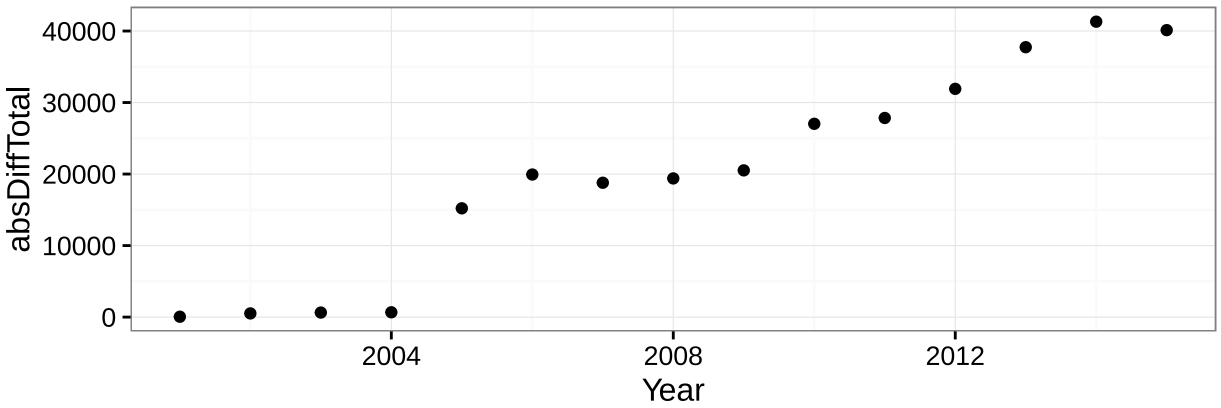
In order to analyse these two groups we first used map reduce functions in mongodb to group the citations by "CitedBy" and "Patent" and count the totals. Another mapreduce added this information to the patent collection before a final mapreduce aggregated frequencies of these orders each year.

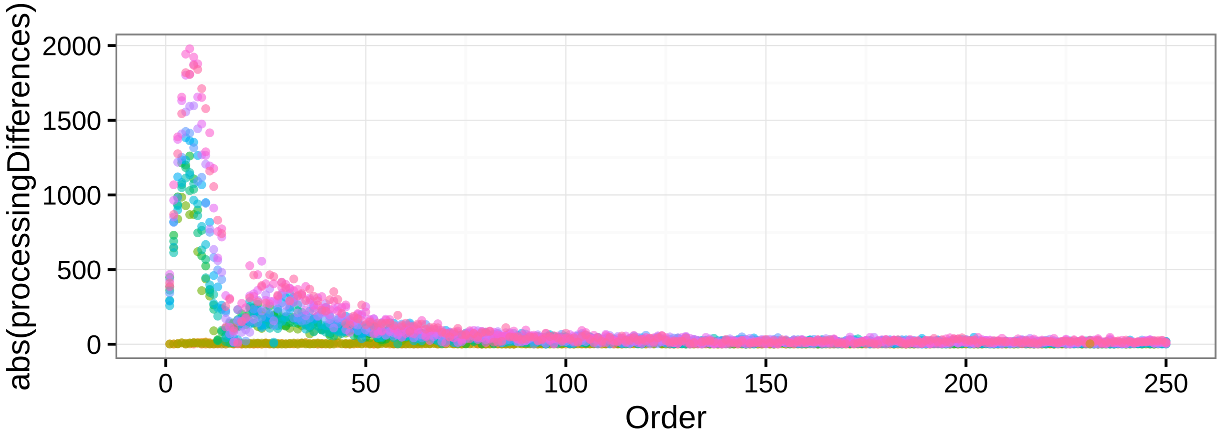
Below we can see how the data has been formatted by thesee map reduces, each Year/Order combination has a row and the frequencies of patents which that order of citations for each category is recorded (Examiner, Other, Total, Total2). Total indicates the total when previously calculating order (while originally processing the data) and Total2 is the equivalent through map reduce.

```
## 'data.frame':    16935 obs. of  6 variables:
##  $ Year    : num  2001 2001 2001 2001 2001 ...
##  $ Order   : num  0 1 2 3 4 5 6 7 8 9 ...
##  $ Examiner: num  16029 17212 19515 20835 19972 ...
##  $ Other   : num  62899 11728 10743 10670 9862 ...
##  $ Total   : num  2112 4977 7608 10350 12013 ...
##  $ Total2  : num  2112 4977 7608 10350 12013 ...
```

There are some differences (processingDifferences) between the order calculated through the citations using map reduce (Total2) and the order calculated while processing the data (Total).

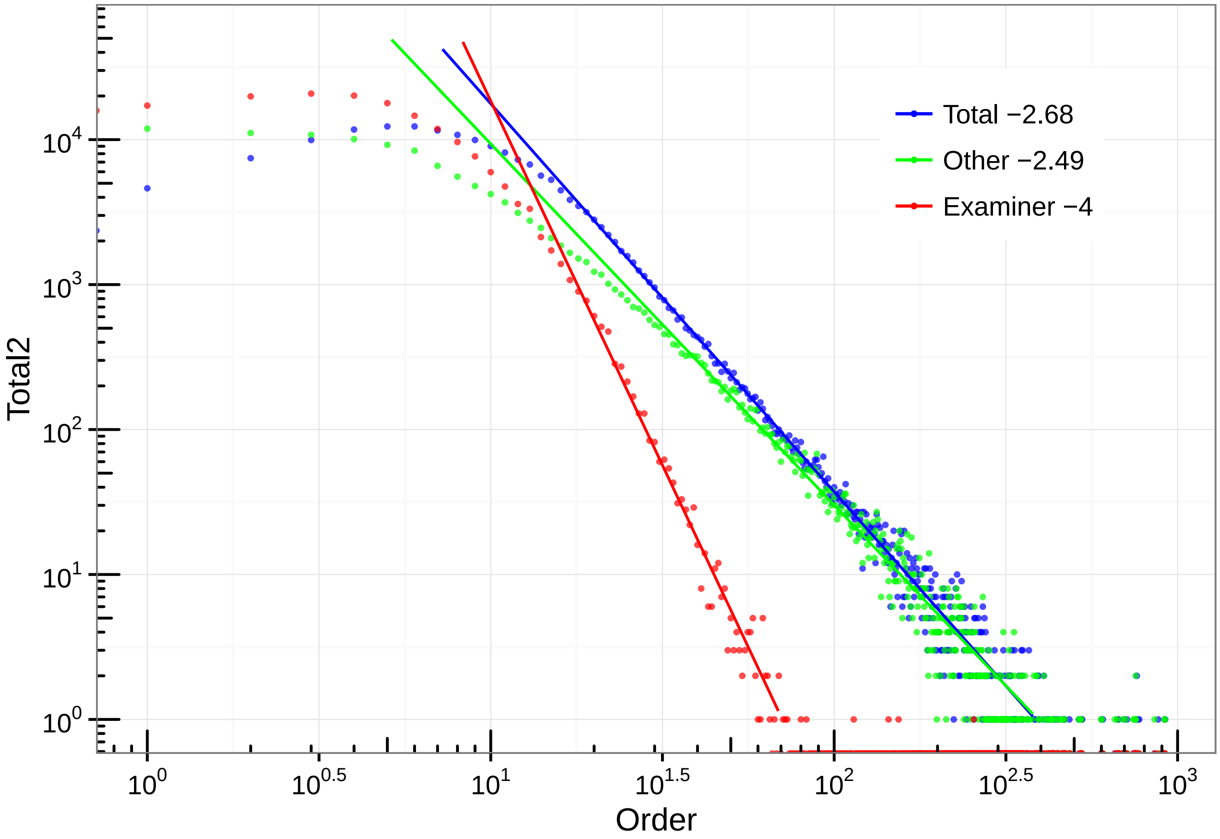
The average error is 0, which means all the citations each year are being processed in both methods. We can see that the errors largely alternate between positive and negative, namely that the differences may be one additional citation on one patent meaning one less on the adjacent patent. The errors are only present in the xml not the sgml years 2005 onwards and while increase over time this is likely in line with increasing number of patents. In 2015 there were r orderFrequencies %>% filter(Year == 2015) %>% select(Total) %>% sum()‘ total citations so about 10% are effected.





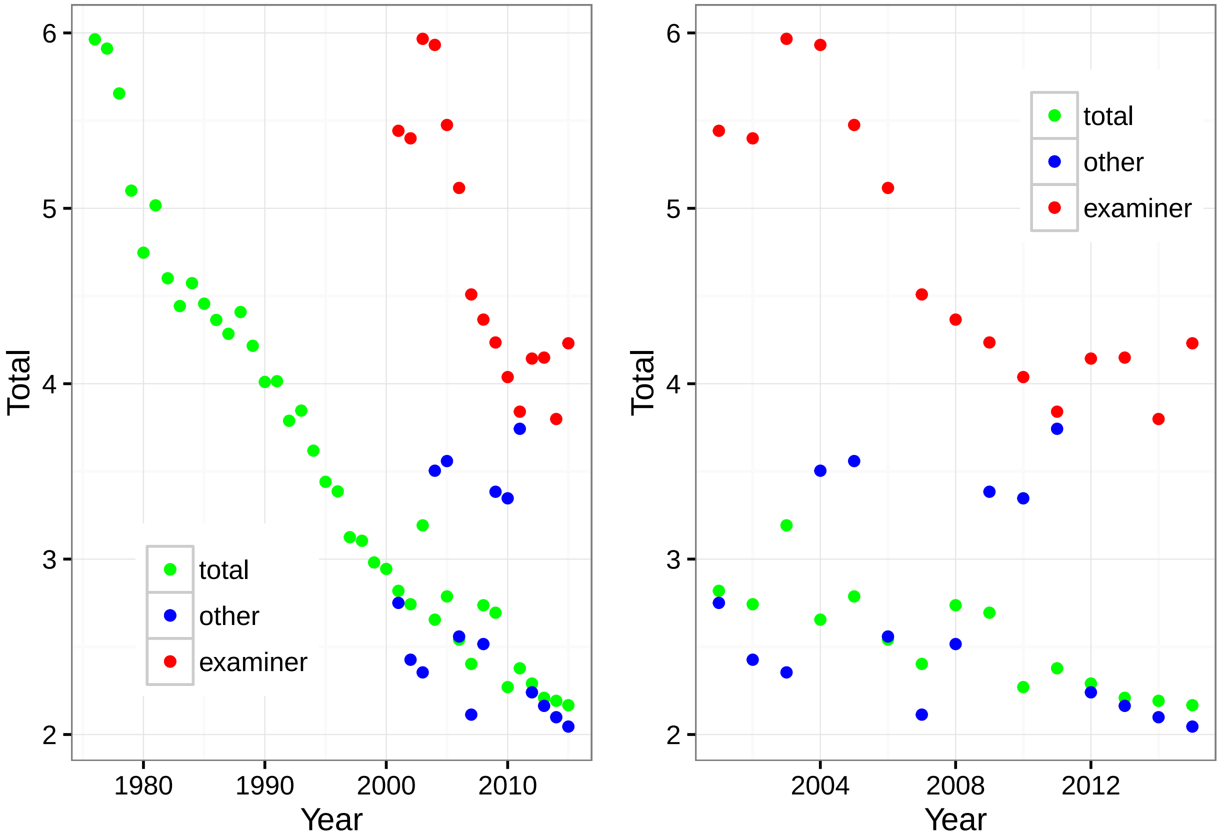
```
## processingDifferences
## Min.   :-1979
## 1st Qu.:    0
## Median :    1
## Mean   :    0
## 3rd Qu.:    4
## Max.   :  556
```

### Order Distribution

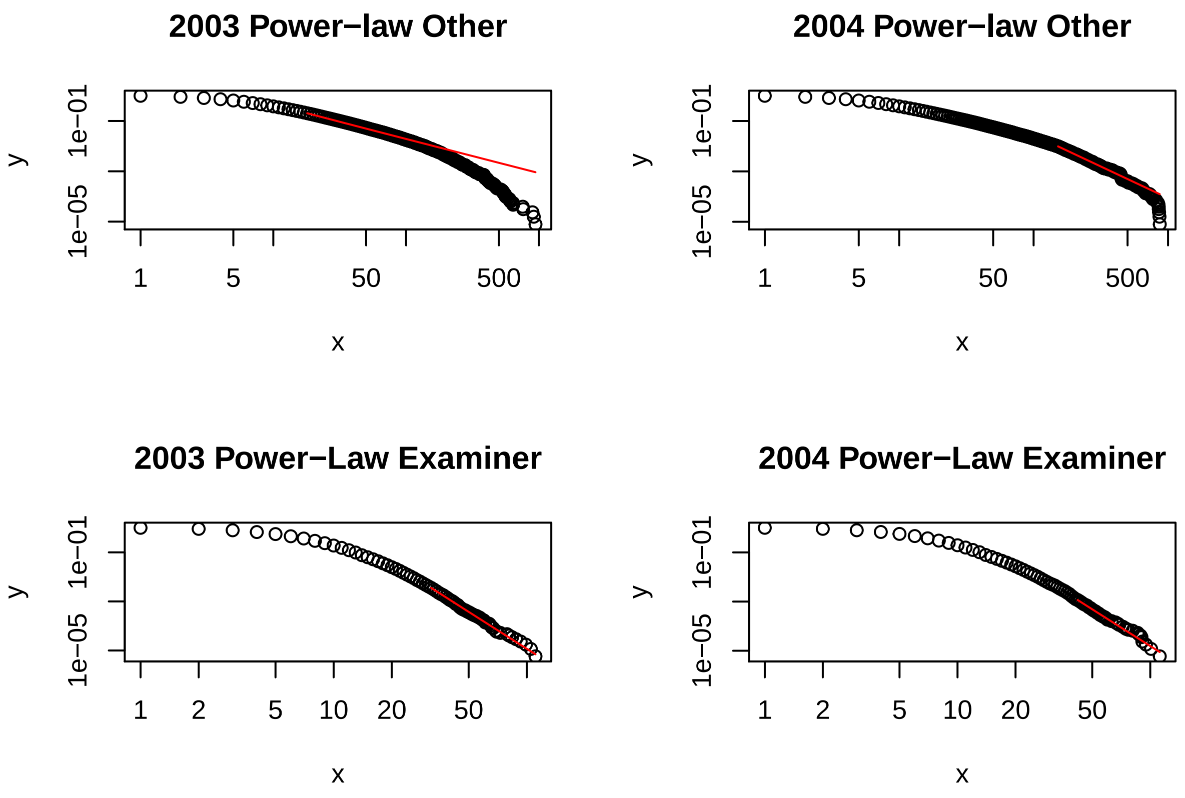


* Examiner has a smaller range of values (steeper gradient), goes to about 100 rather than 1000. * Examiner has larger frequencies at low orders. * Examiner seems to be a little better fit for a power law.

### Change in Power-Law gradient over time



* The gradients of examiner and other vary because it is clearly not a power law but we are trying to fit it to one. * When the other is fitted to the start of the curve it appears to loosely follow the total. * This is not the case with the examiner plot, which show much stronger resemblence to power laws. * While the examiner gradients vary it could be said to loosely follow the shape of the total. * An interesting conclusion here is that the total is dominated by the other which doesn't follow a power law in the years we have the data for, so the total also is probably not followint a power law for these earlier years.



2003 Power−law Other

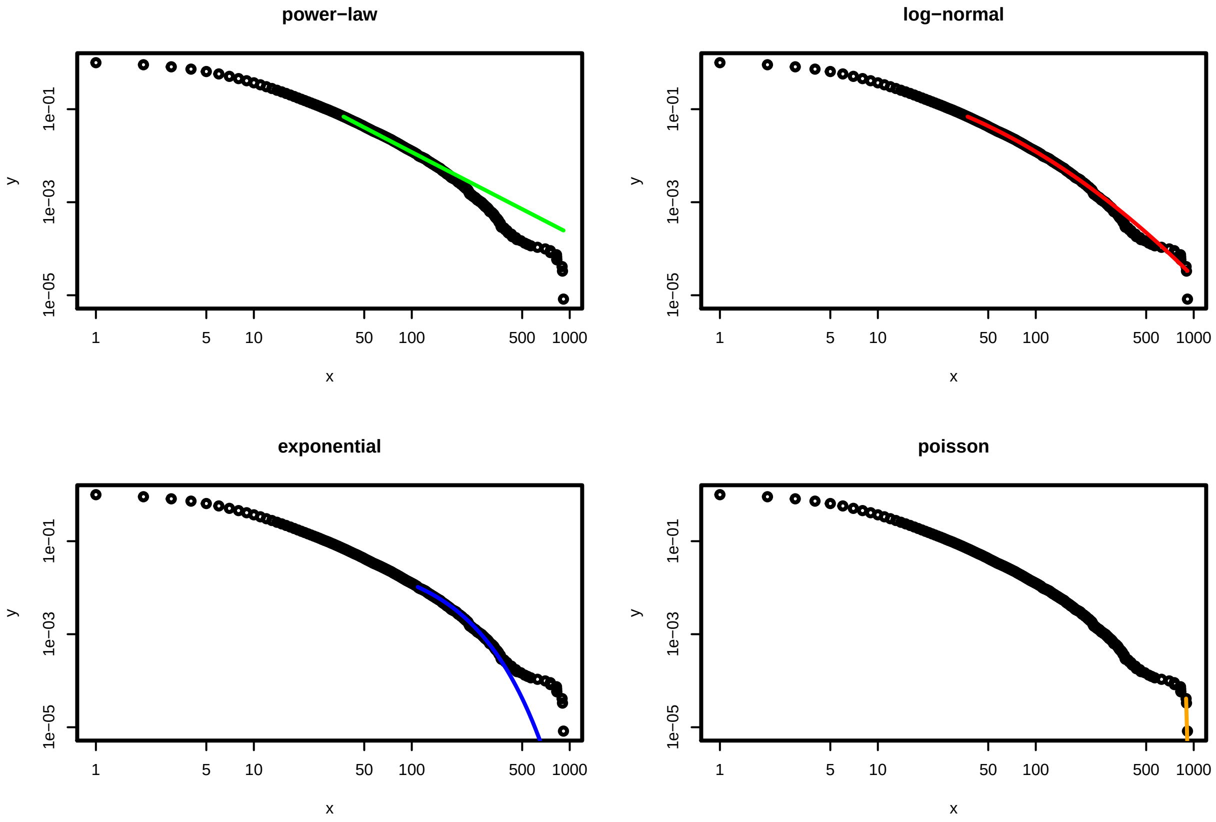
2004 Power−law Other

2003 Power−Law Examiner

2004 Power−Law Examiner

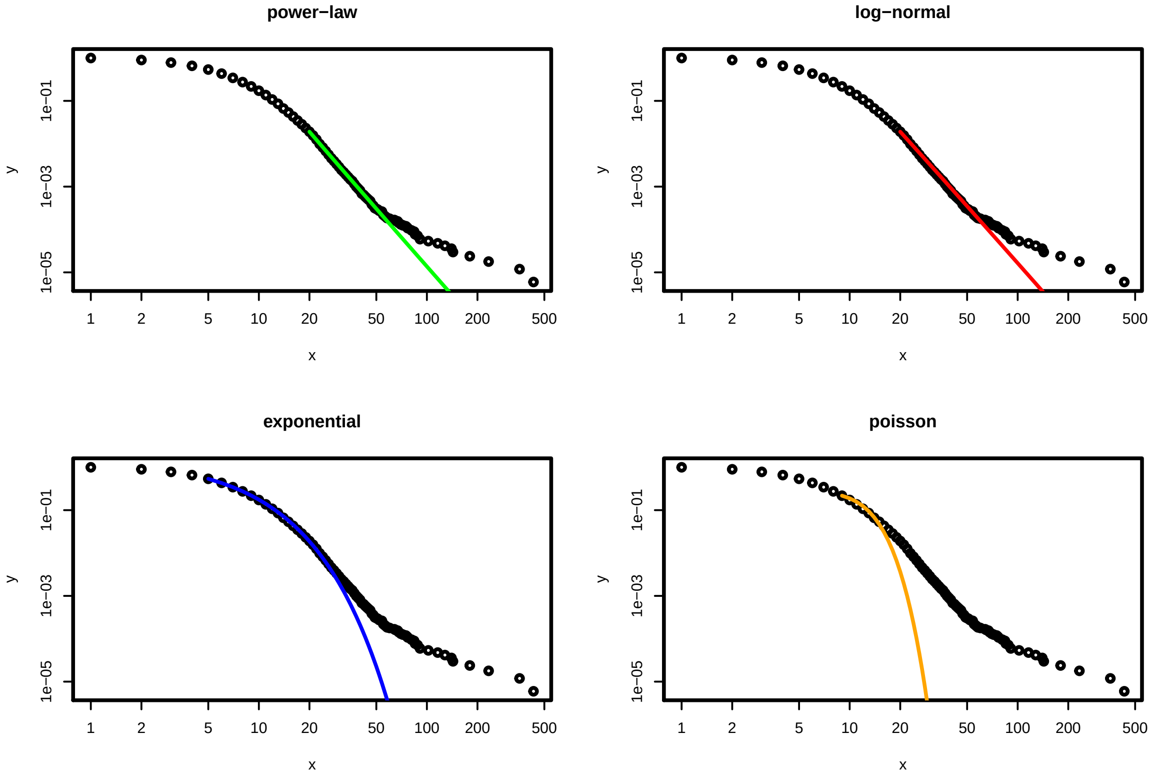
## Power-law vs. log-normal

In this section we compare how well power law fits perform relative to log-normal fits of the data. We first consider a sample year and use the "poweRlaw" package to fit power-law, log-normal, exponential and poisson distributions to the "Examiner" and "Other" distributions seperately.



```
## [1] "Upper Limit on probability given power law is true 6.45351593433556e-15"
## [1] "Upper Limit on probability given log-normal is true 0.999999999999994"
## [1] "Upper Limit on probability given both are true 1.28785870856518e-14"
```

- While none of the distributions catch the shape of the tail the log-normal seems to most closely relate to the distribution with poisson clearly being false. Exponential capturing the start of the distribution and power-law capturing the center of the distribution.
- This therefore could both be a power-law with exponential cut-off (e.g. extended power-law) or a log-


power−law

log−normal

exponential

poisson

normal.

```
## [1] "Upper Limit on probability given power law is true 0.961351748874968"
## [1] "Upper Limit on probability given log-normal is true 0.0386482511250321"
## [1] "Upper Limit on probability given both are true 0.0772965022500642"
```

- The log0normal here is able to capture the entire distribution including most of the variation in the tail.
- Exponential and poisson don't fit any part of the distribution to a reasonable degree (even the start like before)
- Power-law looks more dubious as no part of the distribution looks truely straight on log-log scale.
- Conclusion: This clearly looks like a log-normal distribution.

## Scatterplot between orderOther and orderExaminer