# MAG - 'Computer networks'

Jan Overgoor

```
to build from command-line, run with:
R -e "rmarkdown::render('mag_networks.Rmd', output_file='mag_networks.pdf')"
```

- total number of papers after 1950 (=nodes): 813948
- total number of references: 6277994
- total number of references to known nodes (=edges): 3730543
- share of references to known nodes: 0.5942253
- total number of papers with at least one out-link: 475490
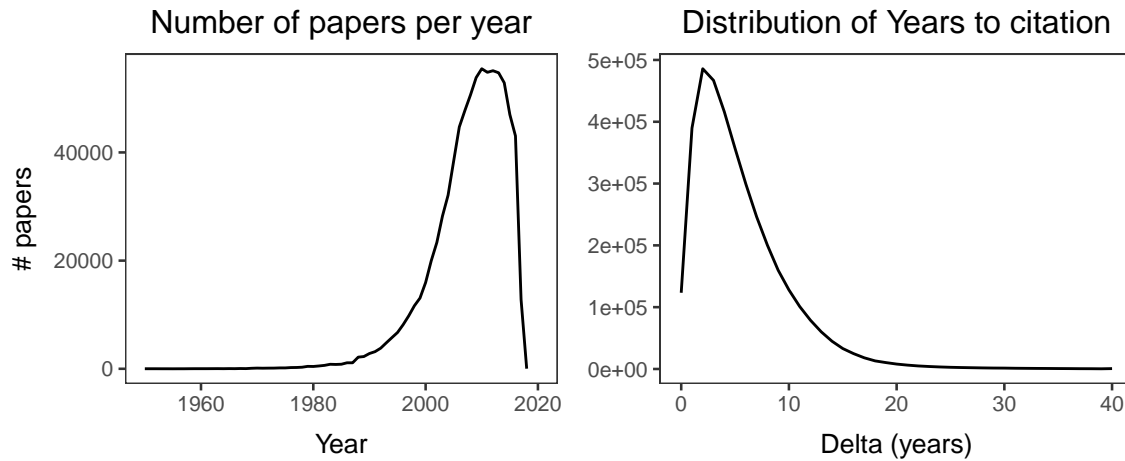
- Left: number of papers per year, linear increase since 2000, drop for recent years
- Right: distribution of years between publication and getting cited. Most citations happen within 2-3 years of publication.

Next, we compare the stated number of citations to the amount we can actually find in the data.
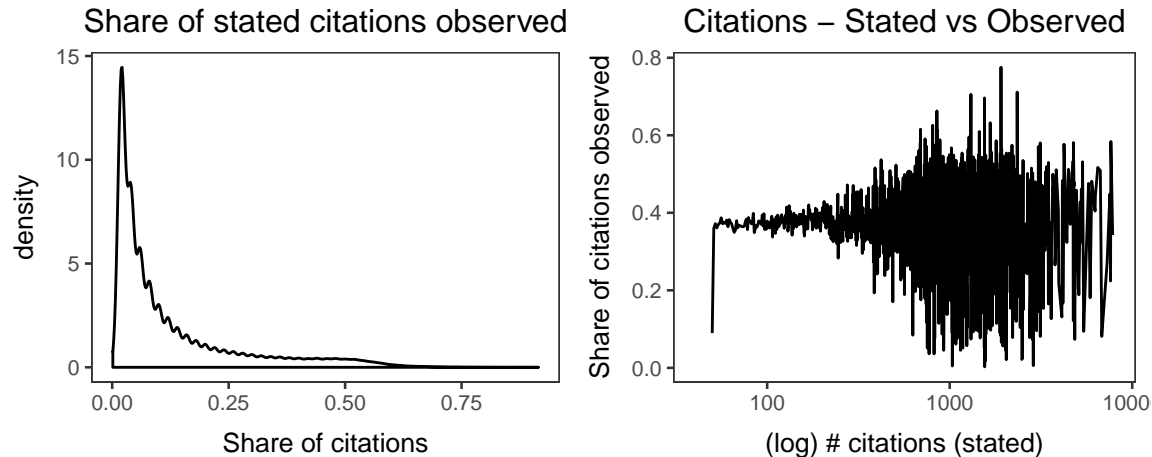
- Left: distribution of "share of stated observations observed", for mostly papers this is <25%. Since I look at the graph filtered by field of study, citing papers might not be included. The whole graph is hard to work with, so this is what we got.
- Right: per "stated number of citations", what is the average "share of citations observed"? Very stable by x, with a much higher variance for the highly cited papers (as there are fewer of them).
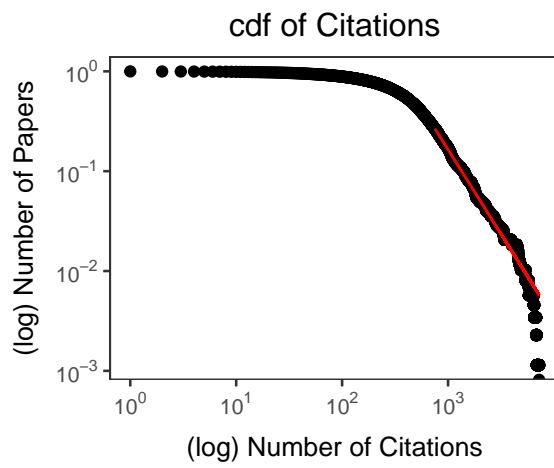
1

Are the degree distributions similar for the stated and observed citation counts?
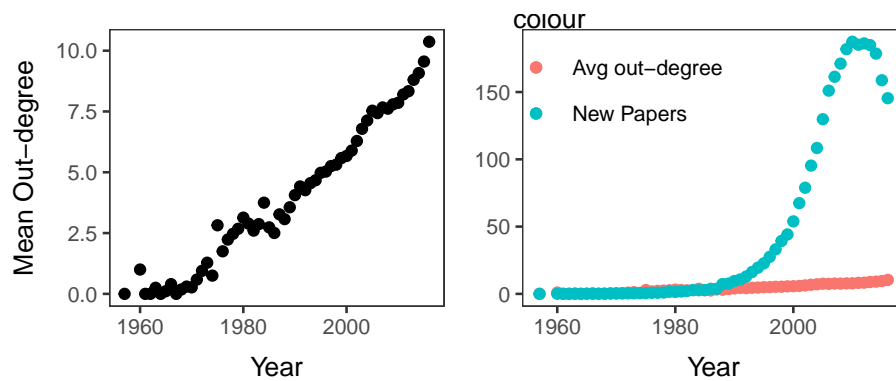
Degree Distribution
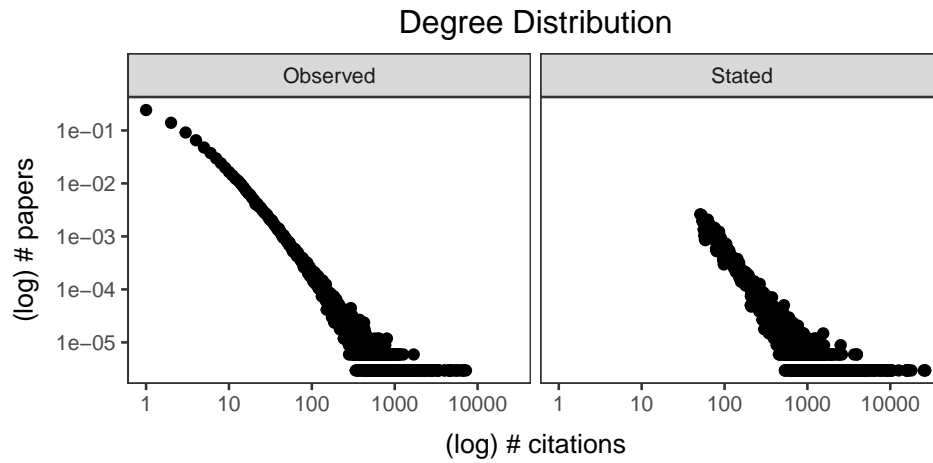
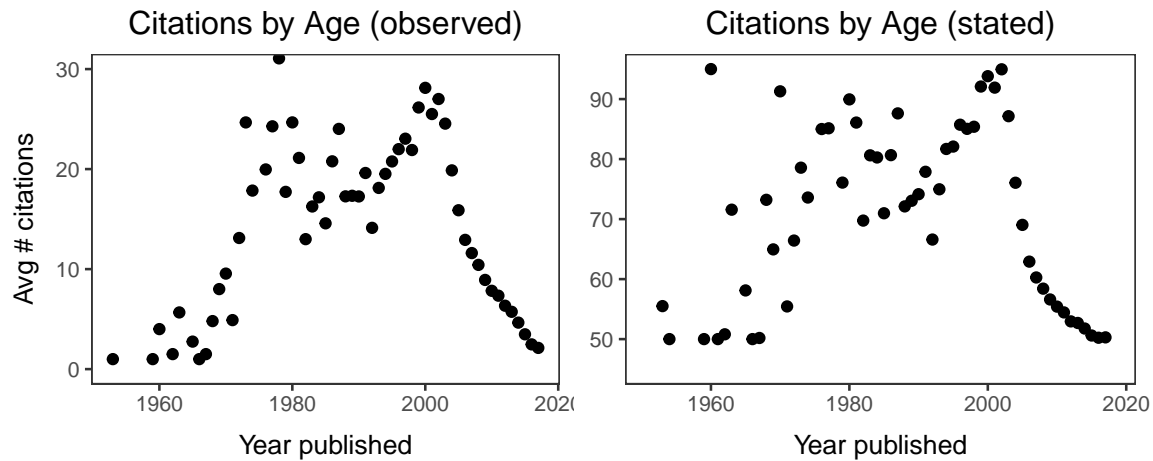Yes, and this shows the censoring at 50 very clearly.

Here is the cdf and Clauset-Shalizi-Newman powerlaw fit:

```
## [1] "plfit:   alpha=2.691   xmin=753"
```

cdf of Citations

Out-degree over time (global average is 7.78):

- Left: average number of citations by year of publishing (as observed). linear increase until 2000 (newer papers more cited), but then drops off
- Right: same, but as stated. The trend is the same, but the numbers are inflated by about 50.

Here is the distribution of papers/author:

```
## [1] "plfit:  alpha=2.391  xmin=88"
```
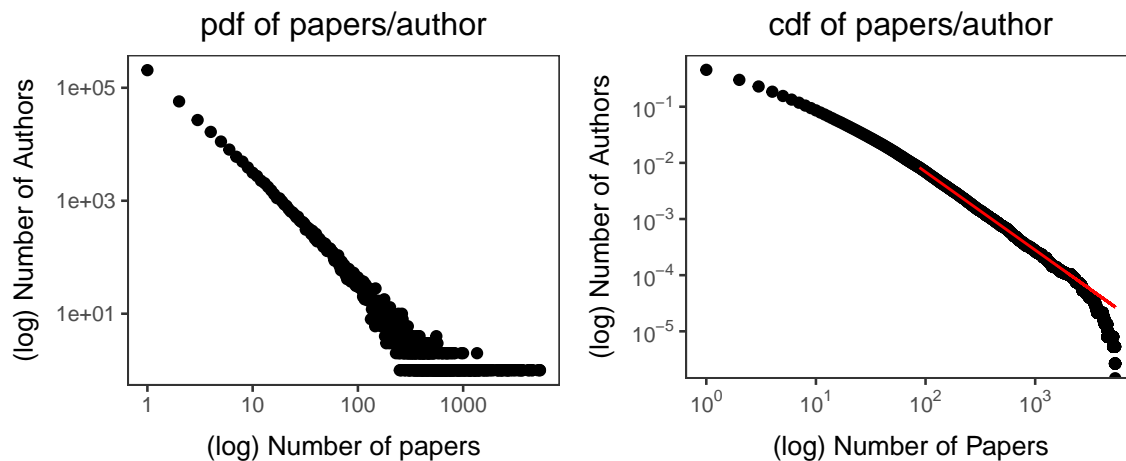

Very heavy-tailed as well.

What are the top keywords?

keywords	n
quality of service	43272
protocols	42369
wireless sensor networks	39752
telecommunication traffic	35324
internet	35221
routing	33588
wireless communication	27664
wireless network	27393
wireless sensor network	26506
ad hoc networks	24194

Model

Data construction process:

- sample 10000 citations from after 2011
- for each actual citation, sample 24 non-cited papers (from before publication date)
- for each of the (paper,option) pairs, compute features (n citations, years since, has same author)

```
## 
## ================================================================================
##                                                  y
##                         (1)        (2)        (3)        (4)        (5)        (6)
## --------------------------------------------------------------------------------
## log Citations          0.746***   0.795***   1.104***   1.096***   1.095***   0.868***
##                        (0.008)    (0.008)    (0.011)    (0.011)    (0.012)    (0.011)
## 
## Has degree             1.386***   1.404***   1.979***   1.890***   1.873***   1.405***
##                        (0.042)    (0.046)    (0.049)    (0.049)    (0.052)    (0.047)
## 
## Has same author                   5.306***   4.913***   4.842***   4.654***   5.207***
##                                   (0.085)    (0.097)    (0.096)    (0.106)    (0.088)
## 
## log Age                                      -1.627***  -1.552***  -1.510***
##                                              (0.023)    (0.024)    (0.026)
## 
## Max papers by author                                     0.076***   0.064***
##                                                         (0.008)    (0.008)
## 
## # same keywords                                                     1.541***
##                                                                    (0.028)
## 
## +10 years                                                                     -2.890***
##                                                                               (0.098)
## 
## +10 years*log Citations                                                        0.303***
##                                                                               (0.023)
## 
## --------------------------------------------------------------------------------
## Observations           10,046     10,046     10,046     10,046     10,046     10,046
## Log Likelihood        -19,394.290 -16,952.770 -13,670.760 -13,638.840 -11,650.330 -15,413.050
## ================================================================================
## Note:                                                      *p<0.1; **p<0.05; ***p<0.01
## [1] "Train accuracy  &   0.432 & 0.502 & 0.588 & 0.588 & 0.651 & 0.545  \\"
## [1] "Test accuracy   &   0.409 & 0.475 & 0.55 & 0.552 & 0.617 & 0.513  \\"
```

4