MAG - 'Computer networks'

overgoor

Contents


```
# cleaner theme
my_theme <- function(base_size=10) {
  # Set the base size
  theme_bw(base_size=base_size) +
    theme(
      # Center title
      plot.title = element_text(hjust = 0.5),
      # Make the background white
      panel.background=element_rect(fill='white', colour='white'),
      panel.grid.major=element_blank(),
      panel.grid.minor=element_blank(),
      # Minimize margins
      plot.margin=unit(c(0.2, 0.2, 0.2, 0.2), "cm"),
      panel.margin=unit(0.25, "lines"),
      # Tiny space between axis labels and tick labels
      axis.title.x=element_text(margin=ggplot2::margin(t=6.0)),
      axis.title.y=element_text(margin=ggplot2::margin(r=6.0)),
      # Simplify the legend
      legend.key=element_blank(),
      legend.title=element_blank(),
      legend.background=element_rect(fill='transparent')
    )
}
```

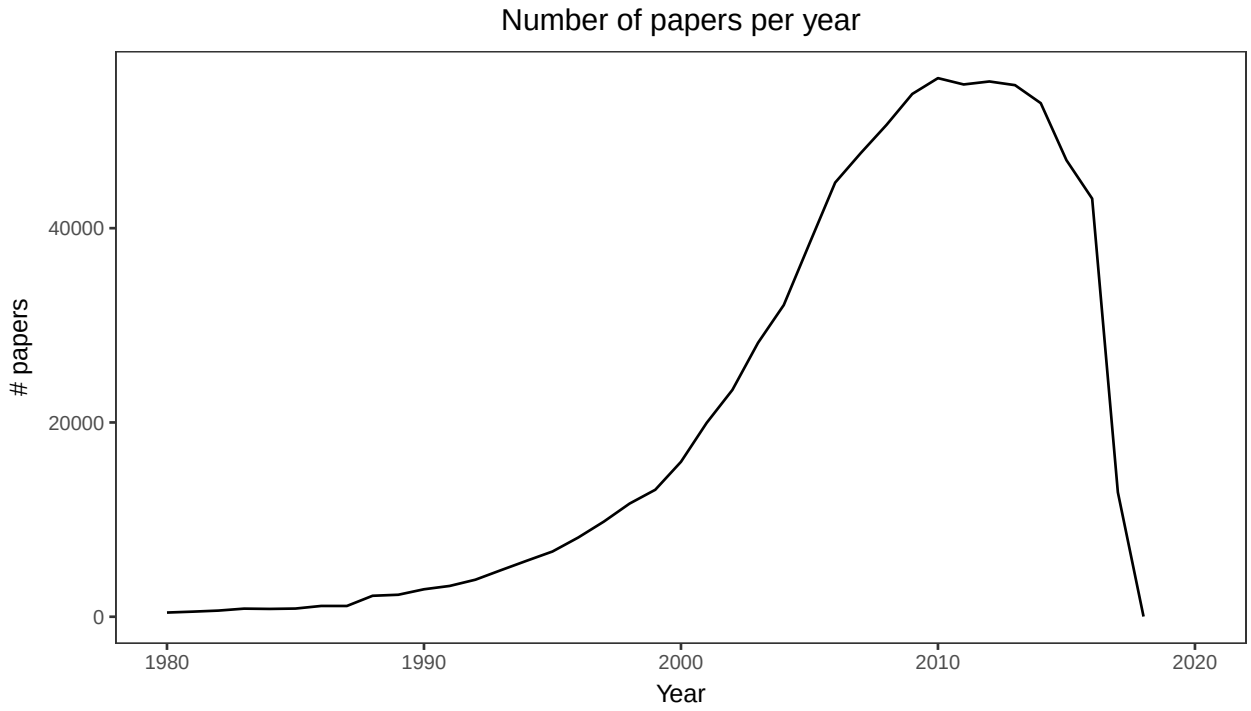
Read Data

```
d_mat = read_csv("~/choosing_to_grow/data_academic/processed/mag_net.txt", col_types='cccddcc')
# explode citations into edges
edges = d_mat %>%
  separate_rows(references, sep=',') %>%
  select(id, cites=references, year_cited=year) %>%
  left_join(d_mat %>% select(cites=id, year_published=year))
```

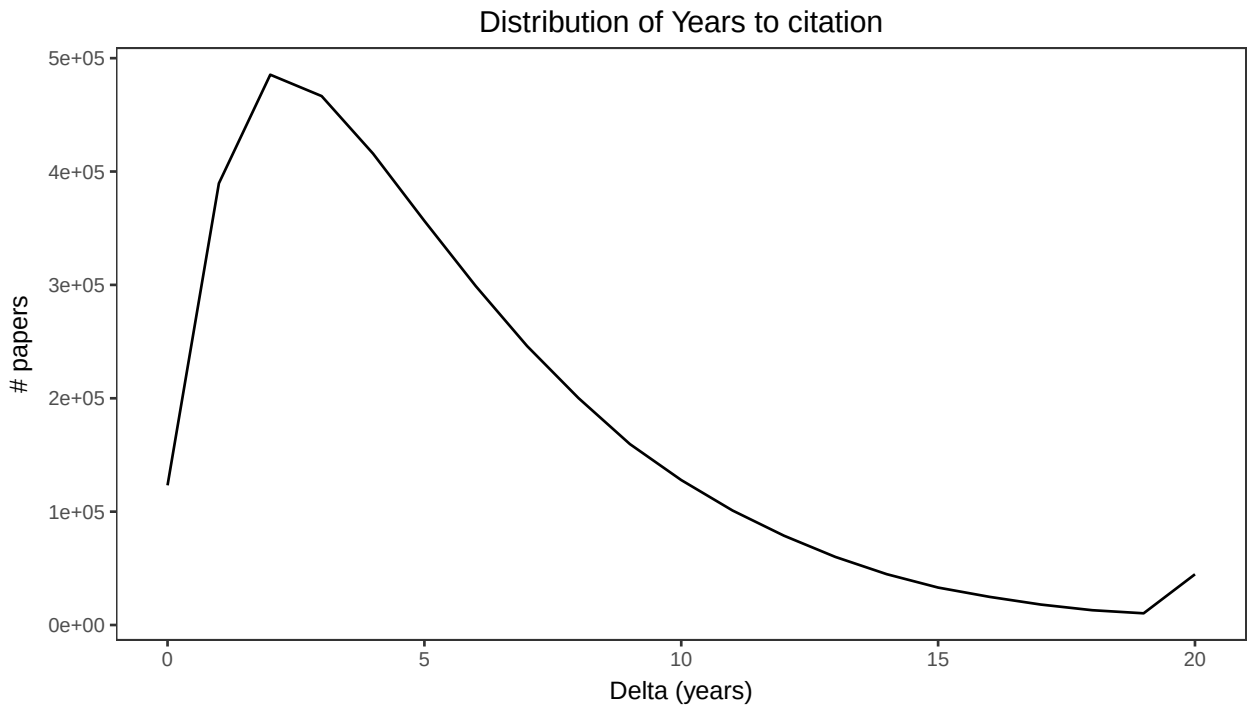
Plots

Role of Year

```
# distribution of when papers are published
d_mat %>%
  group_by(year) %>% summarize(n=n()) %>%
  ggplot(aes(year, n)) + geom_line() +
    scale_x_continuous("Year", limits=c(1980,2020)) +
    scale_y_continuous("# papers") +
    ggtitle("Number of papers per year") +
    my_theme()
```



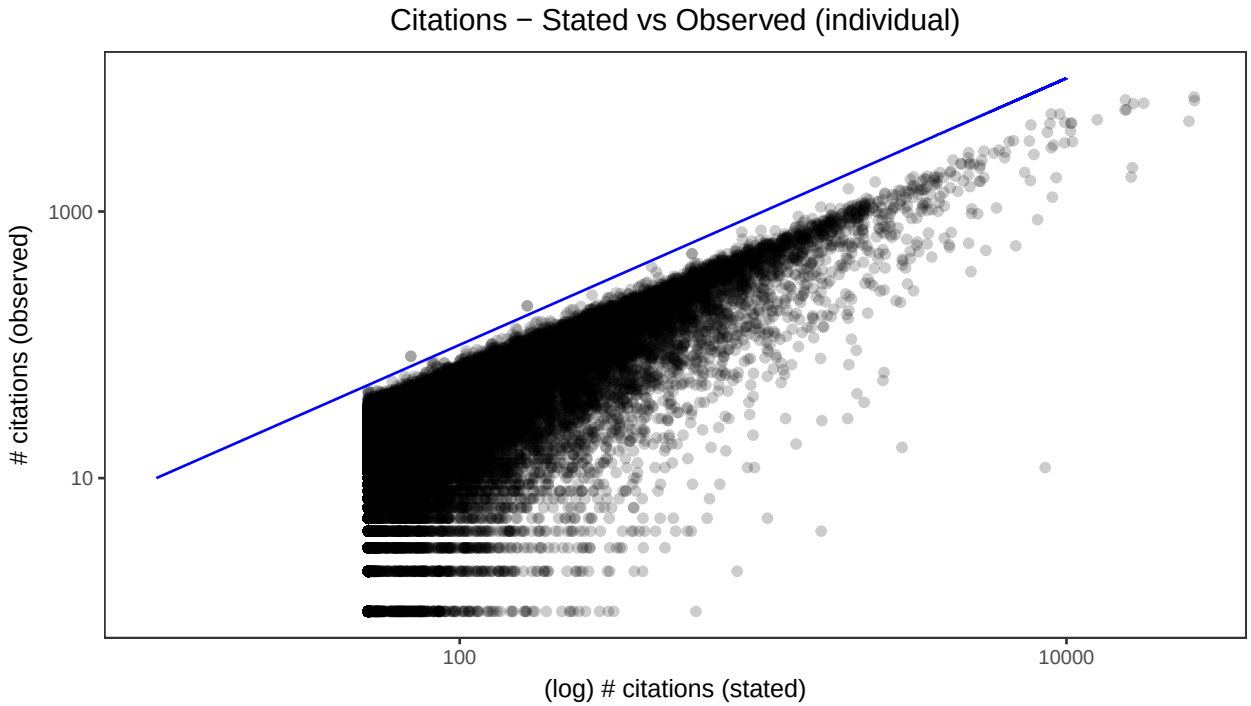

```
# distribution of time between a papers publishing year and year of citation
edges %>%
  mutate(delta=year_cited-year_published) %>%
  mutate(delta=ifelse(delta < 0, NA, ifelse(delta > 20, 20, delta))) %>%
  group_by(delta) %>% summarize(n=n()) %>%
  filter(!is.na(delta)) %>%
  ggplot(aes(delta, n)) + geom_line() +
    scale_x_continuous("Delta (years)") +
    scale_y_continuous("# papers") +
    ggtitle("Distribution of Years to citation") +
    my_theme()
```



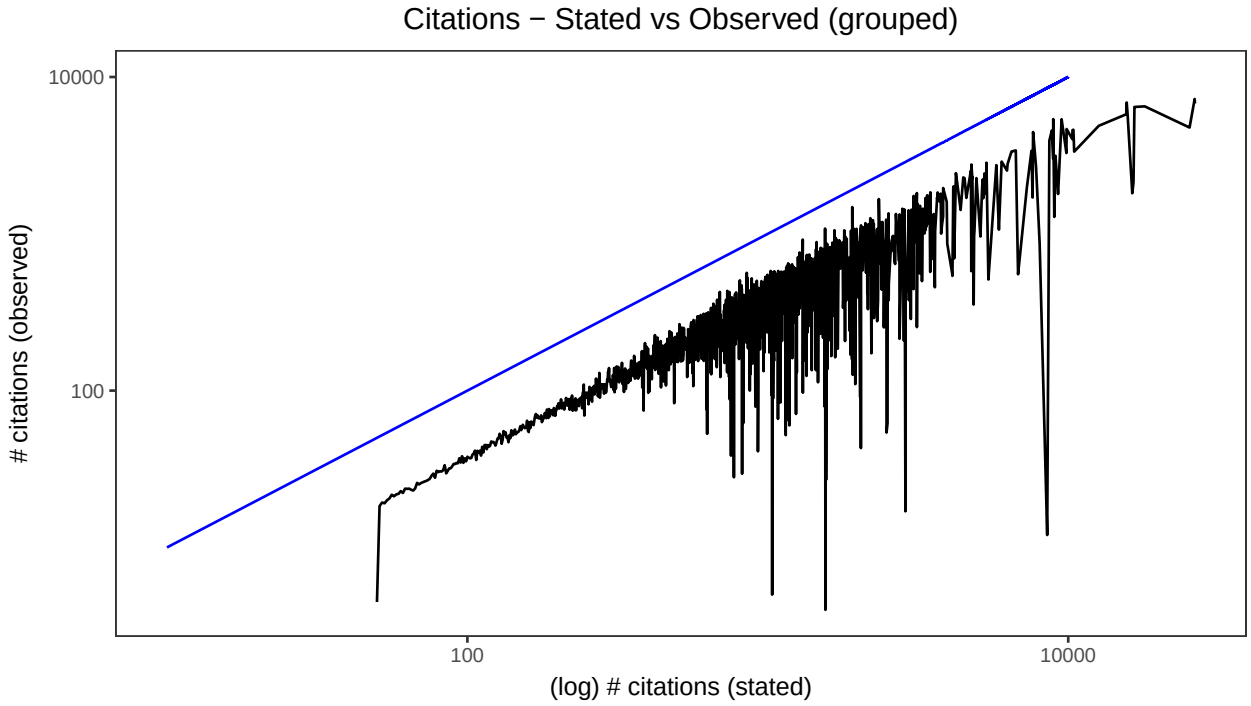
Share of citations observed

```
# how many of the citations in "n_citation" do we actually observe in the data?
DF = inner_join(d_mat,
           edges %>% group_by(cites) %>% summarize(n_citation_data=n()),
           by=c('id'='cites')) %>%
  select(id, n_citation, n_citation_data) %>%
  mutate(p_citation_data=n_citation_data/n_citation) %>%
  filter(!is.na(id))
```

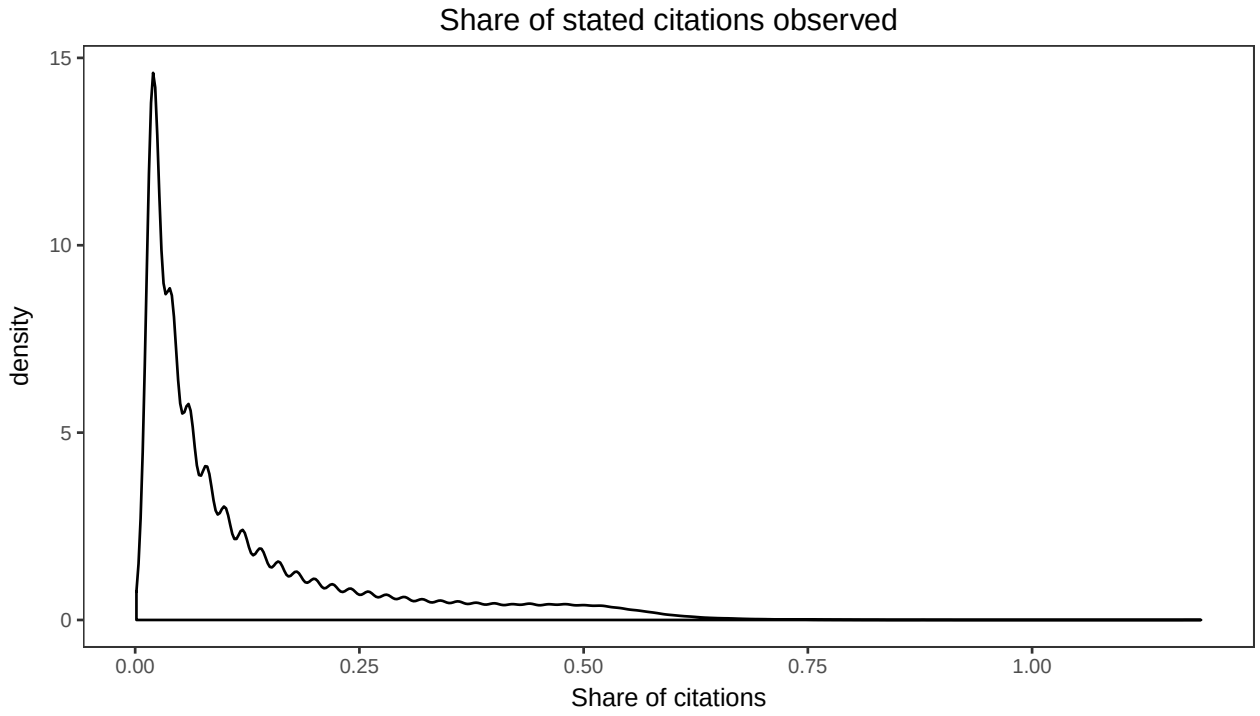
```
# scatter
ggplot(DF, aes(n_citation, n_citation_data)) + geom_point(alpha=0.2) +
  scale_x_log10("(log) # citations (stated)") +
  scale_y_log10("# citations (observed)") +
  geom_line(data=data.frame(x=10:10000), aes(x, x), color='blue') +
  ggtitle("Citations - Stated vs Observed (individual)") +
  my_theme()
```




```
# grouped
DF %>% group_by(n_citation) %>% summarize(stat=mean(n_citation_data)) %>%
  ggplot(aes(n_citation, stat)) + geom_line() + #geom_point(alpha=0.2) +
    scale_x_log10("(log) # citations (stated)") +
    scale_y_log10("# citations (observed)") +
    geom_line(data=data.frame(x=10:10000), aes(x, x), color='blue') +
    ggtitle("Citations - Stated vs Observed (grouped)") +
    my_theme()
```

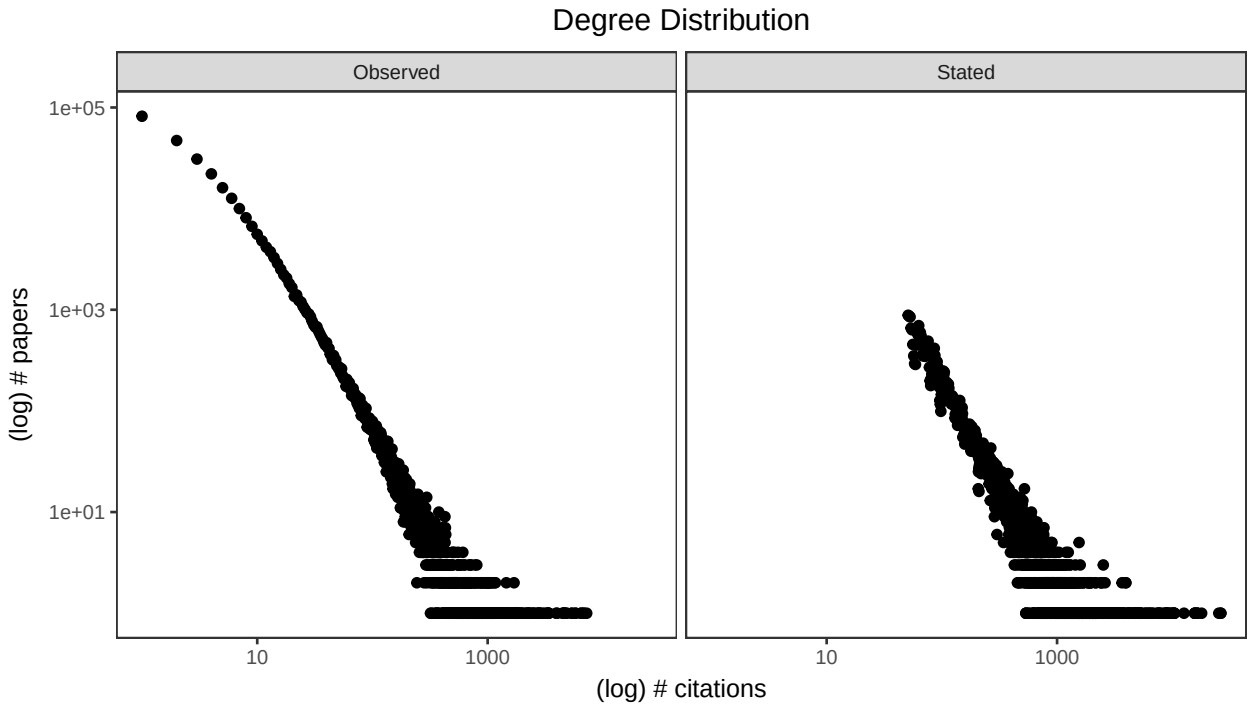


```
# density
ggplot(DF, aes(p_citation_data)) + geom_density() +
  scale_x_continuous("Share of citations") +
  ggtitle("Share of stated citations observed") +
  my_theme()
```



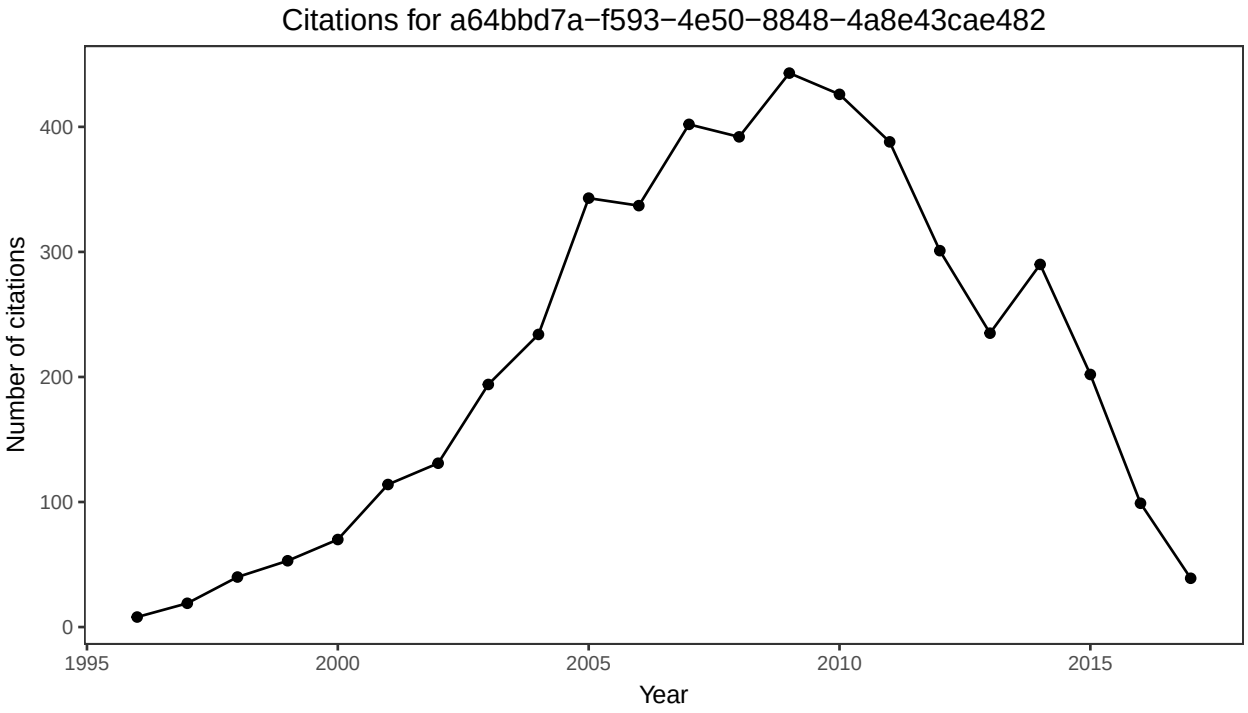
Degree distribution

```
# degree distribution
rbind(
    DF %>% mutate(stat=n_citation) %>% filter(stat > 50) %>% group_by(stat) %>% summarize(n=n()) %>% mu
    DF %>% mutate(stat=n_citation_data) %>% group_by(stat) %>% summarize(n=n()) %>% mutate(g='Observed'
  ) %>%
  ggplot(aes(stat, n)) + geom_point() +
    scale_x_log10("(log) # citations") +
    scale_y_log10("(log) # papers") +
    ggtitle("Degree Distribution") +
    facet_wrap(~g) +
    my_theme()
```



Individual paper's timeline

```
# timeline of citations
edges %>%
  filter(cites=='a64bbd7a-f593-4e50-8848-4a8e43cae482') %>%
  group_by(year_cited) %>% summarize(n=n()) %>%
  arrange(year_cited) %>%
  ggplot(aes(year_cited, n)) + geom_point() + geom_line() +
    ggtitle("Citations for a64bbd7a-f593-4e50-8848-4a8e43cae482") +
    xlab("Year") + ylab("Number of citations") +
    my_theme()
```


Citations for a64bbd7a−f593−4e50−8848−4a8e43cae482
Number of citations
0
100
200
300
400
1995
2000
2005
2010
2015
Year