# Basic analysis

Jinxi_Hu-48528608

2025-11-09

```
library(readr)
library(igraph)
library(RColorBrewer)
set.seed(48528608)
```

```
# this file is for do some most basic analysis to the data we collected.
nodes <- read.csv("data/nodes.csv")

# head of data
head(nodes)
```

```
##   local_id                       paper_id
## 1    P0001 https://openalex.org/W4365143687
## 2    P0002 https://openalex.org/W4205164650
## 3    P0003 https://openalex.org/W4295951577
## 4    P0004 https://openalex.org/W2947423323
## 5    P0005 https://openalex.org/W3042276730
## 6    P0006 https://openalex.org/W3180959755
##
## 1
## 2
## 3
## 4                                                                                                                    I
## 5
## 6 Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) fo
##   year       first_author             institution country
## 1 2023       Michael Moor     Stanford University      US
## 2 2022   Pranav Rajpurkar      Harvard University      US
## 3 2022   Julián N. Acosta         Yale University      US
## 4 2019 Philippe Schwaller   Ibm Research - Zurich      CH
## 5 2020     Jessica Morley    University Of Oxford      GB
## 6 2021     Gary S Collins John Radcliffe Hospital      GB
##                         venue                                                subtopic
## 1                      Nature                       Machine Learning in Healthcare
## 2           Nature Medicine Artificial Intelligence in Healthcare and Education
## 3           Nature Medicine Artificial Intelligence in Healthcare and Education
## 4       ACS Central Science                 Machine Learning in Materials Science
## 5 Social Science & Medicine Artificial Intelligence in Healthcare and Education
## 6                   BMJ Open Artificial Intelligence in Healthcare and Education
##   citations references n_authors author_share
## 1      1155         50         7   0.14285714
```

```
## 2      1931        127         4   0.25000000
## 3       810        180         4   0.25000000
## 4       722         40         7   0.14285714
## 5       671        174         7   0.14285714
## 6       667         33        13   0.07692308
```

```r
# row and colum number
dim(nodes)
```

```
## [1] 60152    13
```

```r
# colum names
names(nodes)
```

```
##  [1] "local_id"     "paper_id"     "title"        "year"         "first_author"
##  [6] "institution"  "country"      "venue"        "subtopic"     "citations"
## [11] "references"   "n_authors"    "author_share"
```

```r
# type of colums
str(nodes)
```

```
## 'data.frame':    60152 obs. of  13 variables:
##  $ local_id    : chr  "P0001" "P0002" "P0003" "P0004" ...
##  $ paper_id    : chr  "https://openalex.org/W4365143687" "https://openalex.org/W4205164650" "https:/
##  $ title       : chr  "Foundation models for generalist medical artificial intelligence" "AI in healt
##  $ year        : int  2023 2022 2022 2019 2020 2021 2016 2020 2020 2024 ...
##  $ first_author: chr  "Michael Moor" "Pranav Rajpurkar" "Julián N. Acosta" "Philippe Schwaller" ...
##  $ institution : chr  "Stanford University" "Harvard University" "Yale University" "Ibm Research - Zu
##  $ country     : chr  "US" "US" "US" "CH" ...
##  $ venue       : chr  "Nature" "Nature Medicine" "Nature Medicine" "ACS Central Science" ...
##  $ subtopic    : chr  "Machine Learning in Healthcare" "Artificial Intelligence in Healthcare and Edu
##  $ citations   : int  1155 1931 810 722 671 667 351 445 762 833 ...
##  $ references  : int  50 127 180 40 174 33 22 42 60 99 ...
##  $ n_authors   : int  7 4 4 7 7 13 4 9 44 34 ...
##  $ author_share: num  0.143 0.25 0.25 0.143 0.143 ...
```

```r
# summary
summary(nodes)
```

```
##    local_id           paper_id            title                year
##  Length:60152       Length:60152       Length:60152       Min.   :2015
##  Class :character   Class :character   Class :character   1st Qu.:2022
##  Mode  :character   Mode  :character   Mode  :character   Median :2023
##                                                           Mean   :2023
##                                                           3rd Qu.:2024
##                                                           Max.   :2025
##                                                           NA's   :57542
##  first_author       institution          country            venue
##  Length:60152       Length:60152       Length:60152       Length:60152
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
##
##
##
##
##    subtopic           citations         references      n_authors
##  Length:60152       Min.   :   0.00   Min.   :  0.0   Min.   :  1.00
##  Class :character   1st Qu.:   1.00   1st Qu.: 10.0   1st Qu.:  3.00
##  Mode  :character   Median :   6.00   Median : 27.0   Median :  5.00
##                     Mean   :  34.77   Mean   : 36.2   Mean   :  7.65
##                     3rd Qu.:  25.00   3rd Qu.: 48.0   3rd Qu.:  9.00
##                     Max.   :2383.00   Max.   :629.0   Max.   :100.00
##                     NA's   :57542     NA's   :57542   NA's   :57542
##   author_share
##  Min.   :0.01
##  1st Qu.:0.11
##  Median :0.20
##  Mean   :0.26
##  3rd Qu.:0.33
##  Max.   :1.00
##  NA's   :57542
```

```
# read edge
edges = read.csv("data/edges.csv")
# number of cititations
dim(edges)
```

```
## [1] 94493     2
```

```
# form graph
graph <- graph_from_data_frame(edges, nodes, directed = FALSE)
# remove the parallel edges and self loops
graph <- simplify(graph, remove.multiple = TRUE, remove.loops = TRUE)
# nodes in graph
vcount(graph)
```

```
## [1] 60152
```

```
# edges in graph
ecount(graph)
```

```
## [1] 94458
```

```
# plot graph (color by university)
institutions <- unique(V(graph)$institution)
palette <- brewer.pal(min(length(institutions), 8), "Set2")
color_map <- setNames(rep(palette, length.out = length(institutions)), institutions)
V(graph)$color <- color_map[V(graph)$institution]

plot(graph, vertex.size=2, edge.size=0.1, vertex.color=V(graph)$color,
     main="Overall network (colored by university)", vertex.label=NA)
```

## Overall network (colored by university)