hw3

Jiahui Xu

2026-02-01

Q2: Dynamic content or document object model changes are two ways that can break a pipeline. In the former case, websites could update how they deliver data such as java script renddering. In the latter case, the specific content could change.

Detect broken pipeline: Tes whether certain elements are found on the sample page; Use logging and warning. Make the analysis replicable: save raw date with timestamps; document data provenance (describe the process of data scraping and the tiem of collection).

Q3: Citation over time (10 professors from psu)

```
# load packages
library(rvest)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(scholar)
library(stringr) # text patterns
library(tibble)

# Hard-code Google Scholar IDs
# -----------------------------------------------------------------------------
matt_scholar_id   <- "yPbxmSwAAAAJ"
sona_scholar_id   <- "Cuz1fTcAAAAJ"
derek_scholar_id  <- "9c6_ChYAAAAJ"
bumba_scholar_id <- "6sS40fEAAAAJ"
micheal_scholar_id <- "SrGrUPsAAAAJ"
boliang_scholar_id <- "Vav4qHMAAAAJ"
joe_scholar_id <- "DV5ECYgAAAAJ"
doug_scholar_id <- "16ceDW4AAAAJ"
cyanne_scholar_id <-'IMUIrJMAAAAJ'
xun_scholar_id <- 'w18ZmkEAAAAJ'
```

```r
# pull profiles
# ------------------------------------------------------------------------------
library(scholar)
matt_profile <- get_profile(matt_scholar_id)
sona_profile <- get_profile(sona_scholar_id)
derek_profile <- get_profile(derek_scholar_id)
bumba_profile <- get_profile(bumba_scholar_id)
micheal_profile <- get_profile(micheal_scholar_id)
boliang_profile <- get_profile(boliang_scholar_id)
joe_profile <- get_profile(joe_scholar_id)
doug_profile <- get_profile(doug_scholar_id)
cyanne_profile <- get_profile(cyanne_scholar_id)
xun_profile <- get_profile(xun_scholar_id)

cat("\n------------------------------\n") # "\n" start a new line;
```

```
##
## ------------------------------
```

```r
cat("Google Scholar Profile Summaries\n")
```

```
## Google Scholar Profile Summaries
```

```r
cat("------------------------------\n")
```

```
## ------------------------------
```

```r
cat("\n", matt_profile$name, "\n", sep = "") # sep = "" don't insert additional separator
```

```
##
## Matt Golder
```

```r
print(matt_profile)
```

```
## $id
## [1] "yPbxmSwAAAAJ"
##
## $name
## [1] "Matt Golder"
##
## $affiliation
## [1] "Professor of Political Science, Pennsylvania State University"
##
## $total_cites
## [1] 16122
##
## $h_index
## [1] 28
##
## $i10_index
## [1] 32
##
## $fields
## [1] "Political science"    "comparative politics" "methodology"
## [4] "European politics"
##
## $homepage
## [1] "http://mattgolder.com/"
## 
## $coauthors
##  [1] "William Roberts Clark"  "Sona Golder"            "Thomas Brambor"        
##  [4] "Charles Crabtree"       "David A Siegel"         "Nils-Christian Bormann"
##  [7] "William D. Berry"       "Daniel Milton"          "Garrett Glasgow"       
## [10] "Indridi H. Indridason"  "Thomas Gschwend"        "Ben Gaskins"           
## [13] "Benjamin Ferland"       "Michael Gilligan"       "Leonard Wantchekon"    
## [16] "Kostanca Dhima"         "Ray Block Jr."          "Sort by citations"     
## [19] "Sort by year"           "Sort by title"         
## 
## $available
## [1] 3
## 
## $not_available
## [1] 0
```

```r
cat("\n", sona_profile$name, "\n", sep = "")
```

```
## 
## Sona Golder
```

```r
print(sona_profile)
```

```
## $id
## [1] "Cuz1fTcAAAAJ"
## 
## $name
## [1] "Sona Golder"
## 
## $affiliation
## [1] "Professor of Political Science, Pennsylvania State University"
## 
## $total_cites
## [1] 2810
## 
## $h_index
## [1] 21
## 
## $i10_index
## [1] 24
## 
## $fields
## [1] "comparative politics" "institutions"         "government formation"
## [4] "methodology"         
## 
## $homepage
## [1] "http://sonagolder.com/"
## 
## $coauthors
##  [1] "Matt Golder"             "William Roberts Clark"  
##  [3] "Van der Straeten Karine" "Laura B. Stephenson"    
##  [5] "André Blais"             "Jean-Francois Laslier"  
##  [7] "Damien Bol"              "Philipp Harfst"         
##  [9] "Garrett Glasgow"         "David A Siegel"
## [11] "Mark Souva"              "Gary Uzonyi"
## [13] "Thomas Gschwend"         "IGNACIO LAGO"
## [15] "Elisabeth Gidengil"      "Allan Sikk"
## [17] "Raimondas Ibenskas"      "Charles Crabtree"
## [19] "Kostanca Dhima"          "Paul Poast"
##
## $available
## [1] 6
##
## $not_available
## [1] 0
```

```r
cat("\n", derek_profile$name, "\n", sep = "")
```

```
##
## Derek Kreager
```

```r
print(derek_profile)
```

```
## $id
## [1] "9c6_ChYAAAAJ"
##
## $name
## [1] "Derek Kreager"
##
## $affiliation
## [1] "Professor of Criminology and Sociology, Penn State"
##
## $total_cites
## [1] 5573
##
## $h_index
## [1] 33
##
## $i10_index
## [1] 54
##
## $fields
## [1] "Incarceration"          "Delinquency"            "Social Networks"
## [4] "Romantic Relationships"
##
## $homepage
## character(0)
##
## $coauthors
##  [1] "Dana L. Haynie"    "David R. Schaefer" "Ross L. Matsueda"
##  [4] "Jeremy Staff"      "Jacob T.N. Young"  "Martin Bouchard"
##  [7] "Sara Wakefield"    "Eva S. Lefkowitz"  "Diane Felmlee"
## [10] "Elena A. Erosheva" "Mark Feinberg"     "René Veenstra"
## [13] "Rose Wesche"       "Mo Yu"             "Duane Alwin"
## [16] "James Moody"       "Kelly Rulison"     "Corey Whichard"
## [19] "Richard Felson"    "Wayne Osgood"
##
## $available
## [1] 44
```

```
##
## $not_available
## [1] 2
```

```r
cat("\n", bumba_profile$name, "\n", sep = "")
```

```
##
## Bumba Mukherjee
```

```r
print(bumba_profile)
```

```
## $id
## [1] "6sS40fEAAAAJ"
##
## $name
## [1] "Bumba Mukherjee"
##
## $affiliation
## [1] "Professor, Political Science, Penn State University"
##
## $total_cites
## [1] 4606
##
## $h_index
## [1] 26
##
## $i10_index
## [1] 50
##
## $fields
## [1] "International Political Economy" "Civil Conflict"
## [3] "Statistical Methodology"
##
## $homepage
## [1] "https://sites.psu.edu/bumbamukherjee/"
##
## $coauthors
##  [1] "Benjamin E. Bagozzi" "Vineeta Yadav"       "Will H. Moore"
##  [4] "David Leblang"       "Ore Koren"           "Sergio Béjar"
##  [7] "helen v. milner"     "Daniel W. Hill, Jr." "Minnie Minhyung Joo"
## [10] "David Andrew Singer" "Brandon Bolte"       "Nguyen Khoi Huynh"
## [13] "Nathan Jensen"       "Xun Cao"             "Hugh Ward"
## [16] "Quan Li"             "Justin Esarey"       "Nicolás Schmidt"
## [19] "Anna Harvey"         "Alexandra Guisinger"
##
## $available
## [1] 2
##
## $not_available
## [1] 1
```

```r
cat("\n", micheal_profile$name, "\n", sep = "")
```

```
##
## Michael J. Nelson
```

```r
print(micheal_profile)
```

```
## $id
## [1] "SrGrUPsAAAAJ"
##
## $name
## [1] "Michael J. Nelson"
##
## $affiliation
## [1] "Pennsylvania State University, Department of Political Science"
##
## $total_cites
## [1] 2104
##
## $h_index
## [1] 24
##
## $i10_index
## [1] 43
##
## $fields
## [1] "Judicial Politics"       "Judicial Behavior"
## [3] "Empirical Legal Studies" "State Politics"
##
## $homepage
## [1] "http://www.mjnelson.org/"
##
## $coauthors
##  [1] "James L. Gibson"      "Amanda Driscoll"      "Rachael K. Hinkle"
##  [4] "Morgan Hazelton"      "Charles Crabtree"     "Christina L. Boyd"
##  [7] "Lee Epstein"          "Christopher Witko"    "Patrick Tucker"
## [10] "Ian Ostrander"        "Alicia Uribe-McGuire" "Jay N Krehbiel"
## [13] "Andrew D. Martin"     "Ethan D. Boldt"       "Markus Neumann"
## [16] "Zachary D. Baumann"   "Michael Burnham"      "Sangyeon Kim"
## [19] "Erin Heidt-Forsythe"  "Morrgan Herlihy"
##
## $available
## [1] 14
##
## $not_available
## [1] 0
```

```r
cat("\n", boliang_profile$name, "\n", sep = "")
```

```
##
## Boliang Zhu
```

```r
print(boliang_profile)
```

```
## $id
## [1] "Vav4qHMAAAAJ"
##
## $name
## [1] "Boliang Zhu"
##
```

```
## $affiliation
## [1] "Pennsylvania State University"
##
## $total_cites
## [1] 825
##
## $h_index
## [1] 11
##
## $i10_index
## [1] 12
##
## $fields
## [1] "Political Economy"         "Political Science"
## [3] "Foreign Direct Investment" "Corruption"
## [5] "Chinese Politics"
##
## $homepage
## [1] "http://sites.psu.edu/bxz14/"
##
## $coauthors
## [1] "Pablo M. Pinto"     "Xiaojun Li"          "Joseph Wright"
## [4] "Bruce A. Desmarais" "Sort by citations"  "Sort by year"
## [7] "Sort by title"      "About Scholar"      "Search help"
##
## $available
## [1] 3
##
## $not_available
## [1] 0
```

```
cat("\n", joe_profile$name, "\n", sep = "")
```

```
##
## Joseph Wright
```

```
print(joe_profile)
```

```
## $id
## [1] "DV5ECYgAAAAJ"
##
## $name
## [1] "Joseph Wright"
##
## $affiliation
## [1] "Pennsylvania State University"
##
## $total_cites
## [1] 9699
##
## $h_index
## [1] 38
##
## $i10_index
## [1] 52
```

```
##
## $fields
## [1] "Authoritarianism"     "Comparative Politics" "Democratization"
## [4] "Foreign Aid"          "Political Science"
##
## $homepage
## [1] "http://sites.psu.edu/wright/"
##
## $coauthors
##  [1] "Erica Frantz"             "Abel Escribà-Folch"
##  [3] "Barbara Geddes"           "Simone Dietrich"
##  [5] "Covadonga Meseguer"       "David B. Carter"
##  [7] "Heather Boushey"          "Deniz Aksoy"
##  [9] "John Chin"                "George Derpanopoulos"
## [11] "Wonjun Song"              "Matthew S. Winters"
## [13] "Jia Li"                   "Xu Xu"
## [15] "Daehee Bak"               "Boliang Zhu"
## [17] "Matthew Charles Wilson"   "Margaret (Molly) Ariotti"
## [19] "Sophia McClennen"         "Elizabeth Stein"
##
## $available
## [1] 8
##
## $not_available
## [1] 0
```

```r
cat("\n", doug_profile$name, "\n", sep = "")
```

```
##
## Douglas Lemke
```

```r
print(doug_profile)
```

```
## $id
## [1] "16ceDW4AAAAJ"
##
## $name
## [1] "Douglas Lemke"
##
## $affiliation
## [1] "Professor of Political Science, Pennsylvania State University"
##
## $total_cites
## [1] 6491
##
## $h_index
## [1] 30
##
## $i10_index
## [1] 35
##
## $fields
## [1] "conflict"    "development"
##
## $homepage
## character(0)
##
## $coauthors
##  [1] "Jacek Kugler"        "William Reed"        "David E. Cunningham"
##  [4] "patrick regan"       "Alex Braithwaite"    "Jeff Carter"
##  [7] "Charles Crabtree"    "Allan C. Stam"       "Michael R. Kenwick"
## [10] "Brian Efird"         "Sort by citations"   "Sort by year"
## [13] "Sort by title"       "About Scholar"       "Search help"
##
## $available
## [1] 3
##
## $not_available
## [1] 0
```

```
cat("\n", cyanne_profile$name, "\n", sep = "")
```

```
##
## Cyanne E. Loyle
```

```
print(cyanne_profile)
```

```
## $id
## [1] "IMUIrJMAAAAJ"
##
## $name
## [1] "Cyanne E. Loyle"
##
## $affiliation
## [1] "Professor, Pennsylvania State University & PRIO"
##
## $total_cites
## [1] 1618
##
## $h_index
## [1] 19
##
## $i10_index
## [1] 26
##
## $fields
## [1] "Conflict Studies"     "Human Rights"          "Transitional Justice"
##
## $homepage
## [1] "http://www.cyanneloyle.com/"
##
## $coauthors
##  [1] "[christian davenport]"         "Helga Malmin Binningsbø"
##  [3] "Kathleen Gallagher Cunningham" "Danielle F. Jung"
##  [5] "Benjamin Appel"                "Reyko Huang"
##  [7] "Jon Elster"                    "Scott Gates"
##  [9] "Christopher Michael Sullivan"  "Samuel E. Bestvater"
## [11] "Jessica Maves Braithwaite"     "R Joseph Huddleston"
## [13] "Anjanette (Angie) Raymond"     "Federica Carugati"
## [15] "Michael A. Rubin"              "Scott Shackelford"
## [17] "Abbey Stemler"                 "Jessica Steinberg"
## [19] "Haley Swedlund"                "Ilayda B. Onder"
##
## $available
## [1] 2
##
## $not_available
## [1] 3
```

```
cat("\n", xun_profile$name, "\n", sep = "")
```

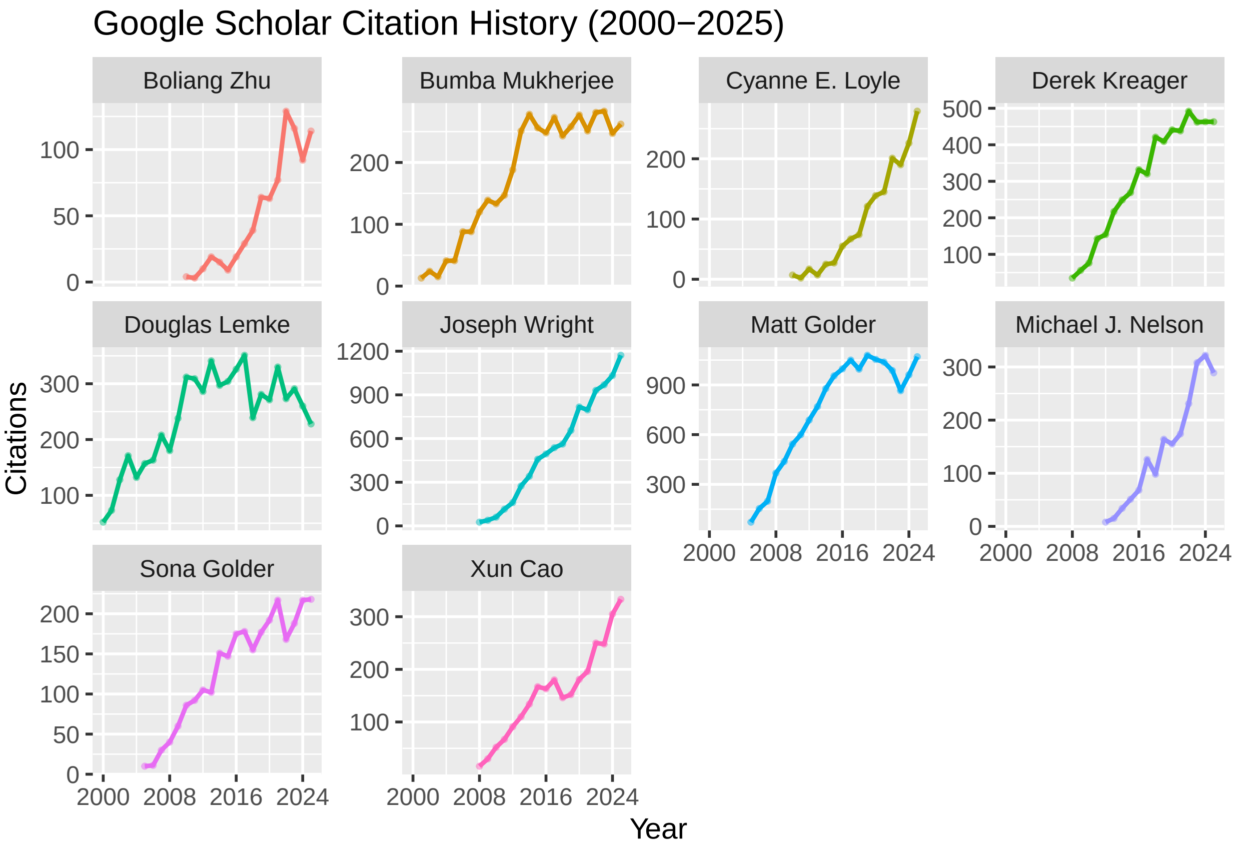
```
##
## Xun Cao
```

```
print(xun_profile)
```

```
## $id
## [1] "w18ZmkEAAAAJ"
##
## $name
## [1] "Xun Cao"
##
## $affiliation
## [1] "Penn State University"
##
## $total_cites
## [1] 2883
##
## $h_index
## [1] 25
##
## $i10_index
## [1] 31
##
## $fields
## [1] "political economy"      "climate change"         "environment and energy"
## [4] "conflicts"              "political geography"
##
## $homepage
## [1] "http://sites.psu.edu/xuncao/research"
##
## $coauthors
##  [1] "Aseem Prakash"           "Michael D. Ward"
##  [3] "Hugh Ward"               "Chuyu Liu"
##  [5] "Genia Kostka"            "Adam Luedtke"
##  [7] "Christian Breunig"       "Kristin M. Bakke"
##  [9] "James Piazza"            "Theodora-Ismene Gizelis"
## [11] "Brian Greenhill"         "helen v. milner"
## [13] "Bumba Mukherjee"         "Anthony A. Pezzola"
## [15] "Amanda Fidalgo"          "Andrew N. Kleit"
## [17] "Sort by citations"       "Sort by year"
## [19] "Sort by title"           "About Scholar"
##
## $available
## [1] 4
```

```
##
## $not_available
## [1] 1
```

```r
# pull citation
#----------------------------------
matt_ct <- get_citation_history(matt_scholar_id) %>%
  mutate(name = matt_profile$name)
sona_ct <- get_citation_history(sona_scholar_id) %>%
  mutate(name = sona_profile$name)
derek_ct <- get_citation_history(derek_scholar_id) %>%
  mutate(name = derek_profile$name)
bumba_ct <- get_citation_history(bumba_scholar_id) %>%
  mutate(name = bumba_profile$name)
micheal_ct <- get_citation_history(micheal_scholar_id) %>%
  mutate(name = micheal_profile$name)
boliang_ct <- get_citation_history(boliang_scholar_id) %>%
  mutate(name = boliang_profile$name)
joe_ct <- get_citation_history(joe_scholar_id) %>%
  mutate(name = joe_profile$name)
doug_ct <- get_citation_history(doug_scholar_id) %>%
  mutate(name = doug_profile$name)
cyanne_ct <- get_citation_history(cyanne_scholar_id) %>%
  mutate(name = cyanne_profile$name)
xun_ct <- get_citation_history(xun_scholar_id) %>%
  mutate(name = xun_profile$name)

# create df
citation_df <- bind_rows(matt_ct, sona_ct, derek_ct, bumba_ct, micheal_ct,
                         boliang_ct, joe_ct, doug_ct, cyanne_ct, xun_ct) %>%
  filter(year >= 2000, year <= 2025) # choose between 2000-2025

# plot citation over year
ggplot(citation_df, aes(x = year, y = cites, color = name)) +
  geom_line(linewidth=0.8) +
  geom_point(size = 0.6, alpha = 0.5) +
  facet_wrap(~ name, scales = "free_y", ncol = 4) +
  scale_x_continuous(
    breaks = seq(2000, 2025, by = 8)
  ) +
  labs(
    title = "Google Scholar Citation History (2000-2025)",
    x = "Year",
    y = "Citations",
    color = "Faculty"
  ) + theme(legend.position = "none")
```



Q4: How works of these professors overlap using the most common publication keywords

```
# create a list of scholar id
ids <- c(matt_scholar_id, sona_scholar_id, derek_scholar_id,
         bumba_scholar_id,micheal_scholar_id,boliang_scholar_id,
         joe_scholar_id,doug_scholar_id,cyanne_scholar_id,xun_scholar_id)

# pull publications
publications <- function(id){
  profile <- get_profile(id) #
  pubs <- get_publications(id)

  pubs %>%
    transmute(
      name  = profile$name,
      title = title
    ) # create a new df
}
library(purrr)
all_pubs <- map_dfr(ids, publications)

# cleansing title
all_pubs <- all_pubs %>% mutate(
  clean_title = title %>%
      str_to_lower() %>% # lowercase
      str_replace_all("[[:punct:]]", "") %>% #delete punctuation
      str_squish() # delete additional blank space
)
```

```
# check overlap
pair_overlap <- all_pubs %>%
  distinct(name, clean_title) %>%
  inner_join(
    all_pubs,
    by = "clean_title"
  ) %>%
  filter(name.x < name.y) %>%
  count(name.x, name.y, name = "n_shared") %>%
  arrange(desc(n_shared))
```

```
## Warning in inner_join(., all_pubs, by = "clean_title"): Detected an unexpected many-to-many relations
## i Row 9 of `x` matches multiple rows in `y`.
## i Row 9 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
print(pair_overlap)
```

```
##             name.x        name.y n_shared
## 1     Matt Golder   Sona Golder       13
## 2     Boliang Zhu Joseph Wright        1
## 3 Bumba Mukherjee       Xun Cao        1
```

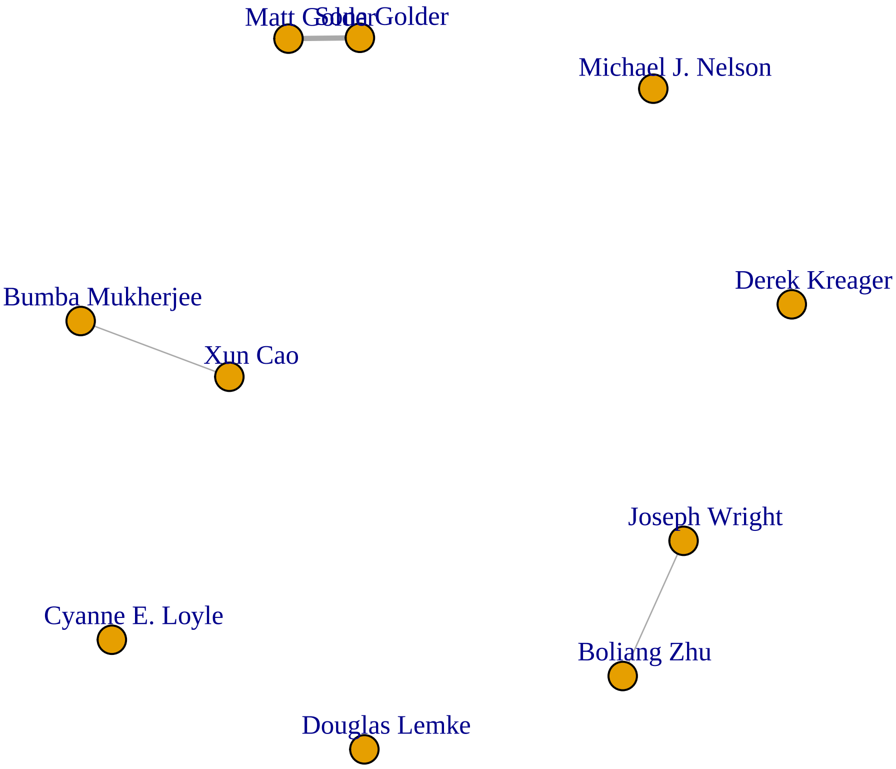
```
# name list  nodes)
name_list <- all_pubs %>%
  distinct(name) %>%
  pull(name)
nodes <- tibble(name = name_list)

# all pairs
all_pairs <- tibble(from = name_list) %>%
  tidyr::crossing(to = name_list) %>%
  filter(from < to)

#create edges
edges_full <- all_pairs %>%
  left_join(
    pair_overlap %>% transmute(from = name.x, to = name.y, weight = n_shared),
    by = c("from", "to")
  ) %>%
  mutate(weight = tidyr::replace_na(weight, 0))

#plot
library(igraph)
```

```
##
## Attaching package: 'igraph'

## The following objects are masked from 'package:purrr':
##
##     compose, simplify

## The following object is masked from 'package:tibble':
##
##     as_data_frame
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

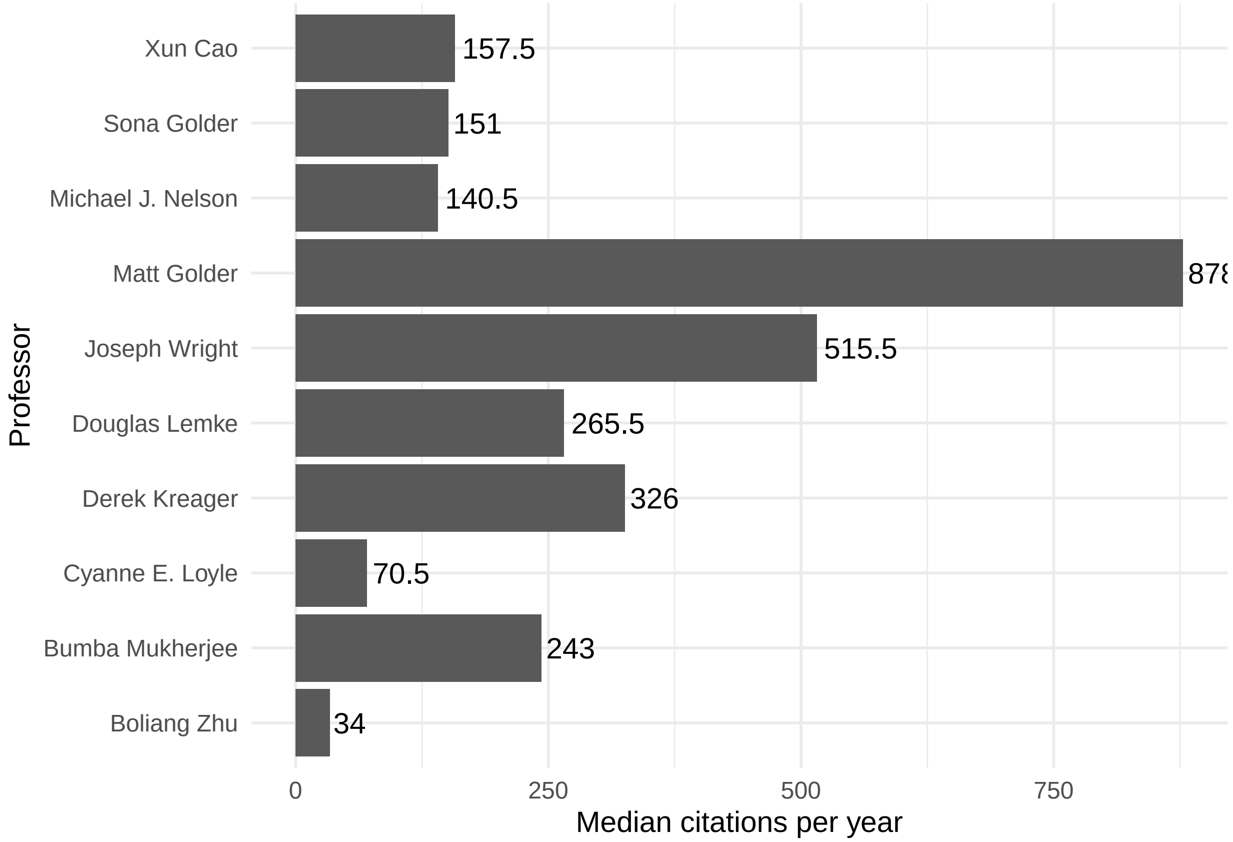
```
## The following object is masked from 'package:base':
##
##     union
```

```r
edges_nonzero <- edges_full %>% filter(weight > 0)
g <- graph_from_data_frame(
  d = edges_nonzero,
  directed = FALSE,
  vertices = nodes
)

plot(
  g,
  vertex.size = 8,
  vertex.label.cex = 0.8,
  vertex.label.dist = 1.2,
  edge.width = log(E(g)$weight + 1)
)
```



I visualize coauthorship among these ten professors using a network plot, where nodes represent professors and edge width reflects the frequency of coauthored publications. The network shows three coauthor relationships: Bumba and Xun have one shared publication, Joseph and Boliang have one shared publication, and Mat and Sonna have thirteen shared publications, which is reflected by the thicker edge connecting them.

Q5: median citation count (per year)



I calculate median citations based on observed years only. Matt has the highest median citation (878), followed by Joe (515.5) and Derek Kreager (326). A middle tier includes Douglas Lemke (265.5), Bumba Mukherjee (243), Xun Cao (157.5), Sona Golder (151), and Michael J. Nelson (140.5). By contrast, Cyanne E. Loyle (70.5) and Boliang Zhu (34) have comparatively lower median citation counts per year.