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)

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

```r
# 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)
```

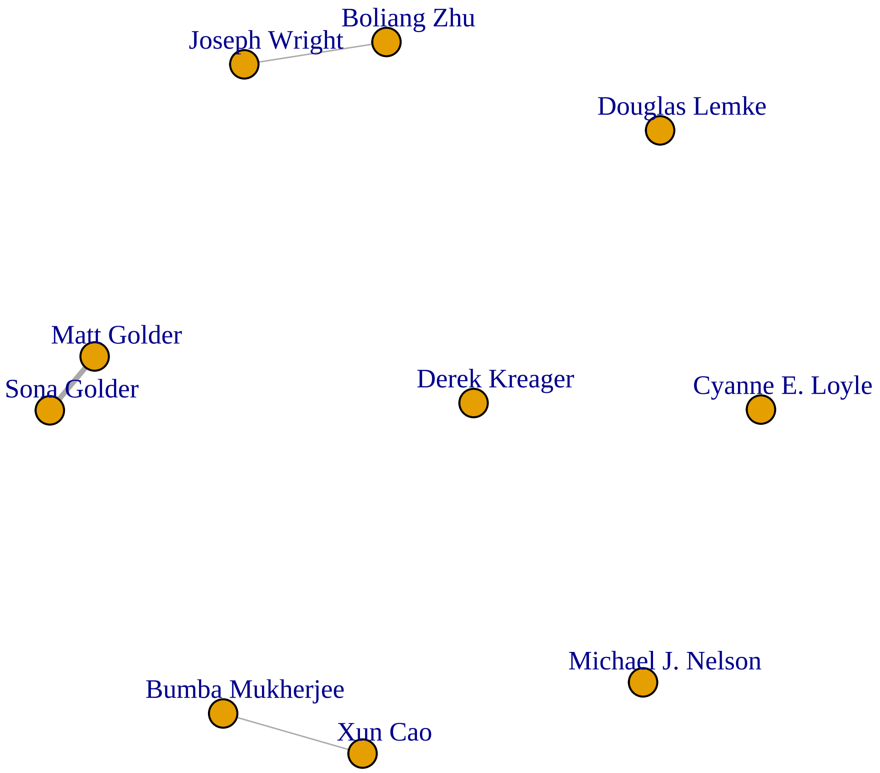
```
##            name.x        name.y n_shared
## 1     Matt Golder   Sona Golder       13
## 2     Boliang Zhu Joseph Wright        1
## 3 Bumba Mukherjee       Xun Cao        1
```

```
# 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':
```

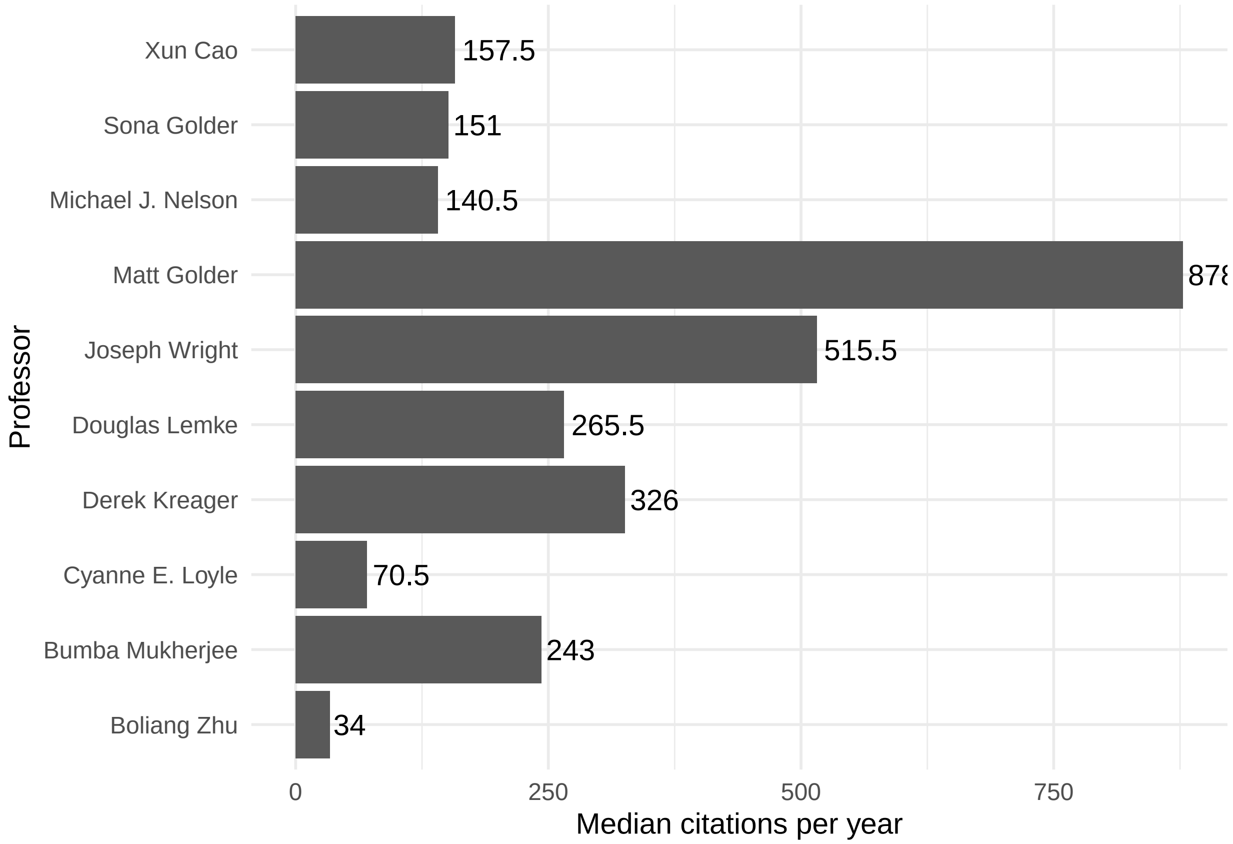
```
##
##     decompose, spectrum

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

```
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.