

Basic analysis and Network statistics

Jinxi_Hu-48528608, Samarth_Grover-38220463

2025-11-09

set uup

```
library(readr)
library(igraph)
library(RColorBrewer)
library(ggplot2)
library(reshape2)
library(scales)
set.seed(48528608)
```

```
# Basic Analysis (Jinxi Hu)
# this file is for do some most basic analysis to the data we collected.
nodes <- read.csv("data/nodes.csv")
# only keep the node with title
nodes_clean <- subset(nodes, !(is.na(title) | trimws(title) == ""))

# head of data
head(nodes_clean)
```

```
##   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
## 5
## 6 Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) f
```

year	first_author	institution	country
2023	Michael Moor	Stanford University	US
2022	Pranav Rajpurkar	Harvard University	US
2022	Julián N. Acosta	Yale University	US
2019	Philippe Schwaller	Ibm Research - Zurich	CH
2020	Jessica Morley	University Of Oxford	GB
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

# row and column number
dim(nodes_clean)

## [1] 2610 13

# column names
names(nodes_clean)

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

# type of columns
str(nodes_clean)

## 'data.frame': 2610 obs. of 13 variables:
## $ local_id : chr "P0001" "P0002" "P0003" "P0004" ...
## $ paper_id : chr "https://openalex.org/W4365143687" "https://openalex.org/W4205164650" "https://openalex.org/W4205164650" ...
## $ title   : chr "Foundation models for generalist medical artificial intelligence" "AI in healthcare and medicine" ...
## $ 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 - Zurich" ...
## $ 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 Education" ...
## $ citations  : int 1155 1931 810 722 671 351 445 762 833 ...
## $ 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 ...

# summary
summary(nodes_clean)

##      local_id      paper_id      title        year
##  Length:2610      Length:2610      Length:2610      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
##   first_author      institution       country          venue
##   Length:2610        Length:2610        Length:2610        Length:2610
##   Class :character   Class :character   Class :character   Class :character
##   Mode  :character   Mode  :character   Mode  :character   Mode  :character
##                                         Mean  :2023
##                                         3rd Qu.:2024
##                                         Max.  :2025
##   subtopic           citations        references      n_authors
##   Length:2610        Min.   : 0.00   Min.   : 0.0   Min.   : 1.000
##   Class :character   1st Qu.: 1.00   1st Qu.: 10.0  1st Qu.: 3.000
##   Mode  :character   Median : 6.00   Median : 27.0  Median : 5.000
##                                         Mean   :34.77   Mean   :36.2   Mean   : 7.654
##                                         3rd Qu.:25.00   3rd Qu.:48.0  3rd Qu.: 9.000
##                                         Max.   :2383.00  Max.   :629.0  Max.   :100.000
##   author_share
##   Min.   :0.0100
##   1st Qu.:0.1111
##   Median :0.2000
##   Mean   :0.2622
##   3rd Qu.:0.3333
##   Max.   :1.0000

```

```

# read edge
edges = read.csv("data/edges.csv")
# only keep valid edges
edges_clean <- subset(edges, source %in% nodes_clean$local_id &
                      target %in% nodes_clean$local_id)
# number of citations
dim(edges_clean)

```

```
## [1] 3757     2
```

```

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

```

```
## [1] 2610
```

```

# edges in graph
ecount(graph)

```

```
## [1] 3722
```

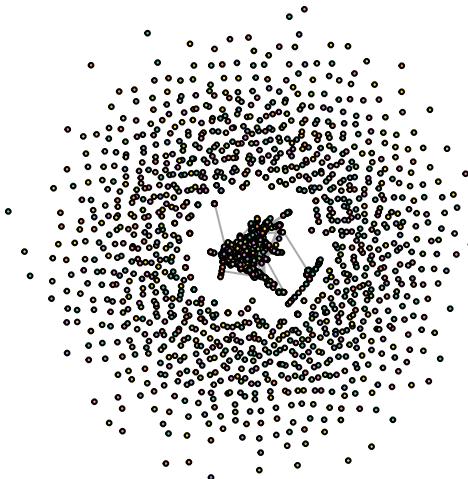
```

# 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)



```

# Initial Network Statistics (Samarth Grover)
cat("==> NETWORK STATISTICS ==>\n")

## ==> NETWORK STATISTICS ==>

cat("Total nodes:", vcount(graph), "\n")

## Total nodes: 2610

cat("Total edges (citations):", ecount(graph), "\n")

## Total edges (citations): 3722

```

```

cat("Network density:", edge_density(graph), "\n")

## Network density: 0.00109318

cat("Average degree:", mean(degree(graph, mode="all")), "\n")

## Average degree: 2.852107

cat("Average in-degree (citations received):",
    mean(degree(graph, mode="in")), "\n")

## Average in-degree (citations received): 2.852107

cat("Average out-degree (citations made):",
    mean(degree(graph, mode="out"))), "\n\n")

## Average out-degree (citations made): 2.852107

# Analyze isolated components
cat("==> COMPONENT ANALYSIS ==>\n")

## ==> COMPONENT ANALYSIS ==>

components <- components(graph, mode="weak")
cat("Number of weakly connected components:", components$no, "\n")

## Number of weakly connected components: 1086

cat("Size of largest component:", max(components$csize), "\n")

## Size of largest component: 1433

cat("Proportion of nodes in largest component:",
    round(max(components$csize) / vcount(graph) * 100, 2), "%\n\n")

## Proportion of nodes in largest component: 54.9 %

# Count isolated nodes
isolated_count <- sum(degree(graph, mode="all") == 0)
cat("Number of isolated nodes (no citations in or out):",
    isolated_count, "\n")

## Number of isolated nodes (no citations in or out): 1043

```

```

cat("Proportion of isolated nodes:",
    round(isolated_count / vcount(graph) * 100, 2), "%\n")

## Proportion of isolated nodes: 39.96 %

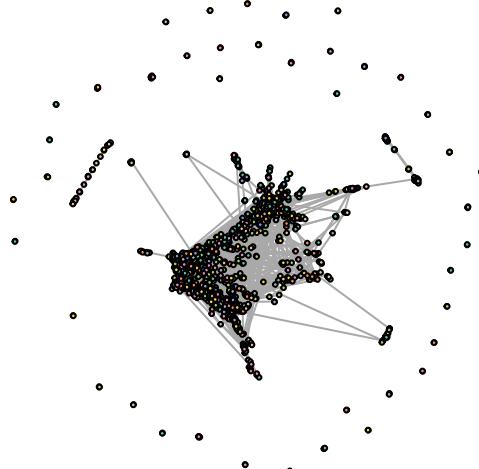
# Remove isolated nodes for cleaner visualization
non_isolated <- V(graph)[degree(graph, mode="all") > 0]
graph_connected <- induced_subgraph(graph, non_isolated)

# plot graph without isolated nodes (color by university)
institutions_connected <- unique(V(graph_connected)$institution)
palette_connected <- brewer.pal(min(length(institutions_connected), 8), "Set2")
color_map_connected <- setNames(rep(palette_connected,
                                      length.out = length(institutions_connected)),
                                 institutions_connected)
V(graph_connected)$color <- color_map_connected[V(graph_connected)$institution]

plot(graph_connected, vertex.size=2, edge.size=0.1,
      vertex.color=V(graph_connected)$color,
      main="Network without isolated nodes (colored by university)",
      vertex.label=NA)

```

Network without isolated nodes (colored by university)



```

library(igraph)
nodes2 <- read.csv("data/nodes_connected.csv")

```

```

edges2 <- read.csv("data/edges_connected.csv")

head(nodes2)

##   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
## 5
## 6 Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) f
##   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

dim(nodes2)

## [1] 1582 13

names(nodes2)

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

```

```
str(nodes2)
```

```
## 'data.frame': 1582 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 heal...  
## $ 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 - Z...  
## $ 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 ...
```

```
summary(nodes2)
```

```
##      local_id          paper_id          title           year  
##  Length:1582    Length:1582    Length:1582    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  
##  
##      first_author        institution        country          venue  
##  Length:1582    Length:1582    Length:1582    Length:1582  
##  Class :character  Class :character  Class :character  Class :character  
##  Mode  :character  Mode  :character  Mode  :character  Mode  :character  
##  
##  
##      subtopic         citations       references      n_authors  
##  Length:1582    Min.   : 0.00  Min.   : 0.00  Min.   : 1.000  
##  Class :character  1st Qu.: 2.00  1st Qu.:18.00  1st Qu.: 3.000  
##  Mode  :character  Median :12.00  Median :33.00  Median : 6.000  
##                                         Mean   :50.88  Mean   :45.12  Mean   : 8.419  
##                                         3rd Qu.:42.00  3rd Qu.:56.00  3rd Qu.:10.000  
##                                         Max.   :2383.00  Max.   :629.00  Max.   :100.000  
##  
##      author_share  
##  Min.   :0.0100  
##  1st Qu.:0.1000  
##  Median :0.1667  
##  Mean   :0.2322  
##  3rd Qu.:0.3333  
##  Max.   :1.0000
```

```
graph2 <- graph_from_data_frame(edges2, vertices = nodes2, directed = TRUE)  
graph2 <- simplify(graph2, remove.multiple = TRUE, remove.loops = TRUE)
```

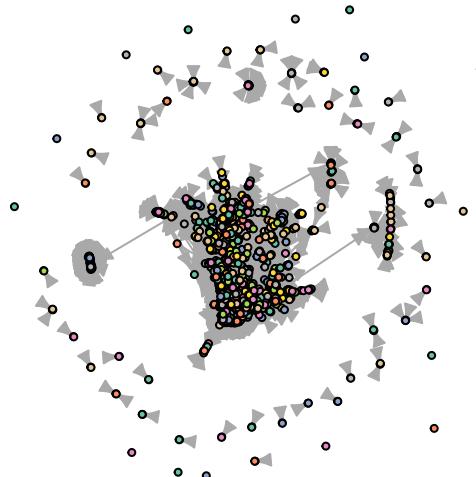
```

# plot graph (color by university)
institutions <- unique(V(graph2)$institution)
palette <- brewer.pal(min(length(institutions), 8), "Set2")
color_map <- setNames(rep(palette, length.out = length(institutions)),
                      institutions)
V(graph2)$color <- color_map[V(graph2)$institution]

plot(graph2, vertex.size=3, edge.size=1, vertex.color=V(graph2)$color,
      main="Overall network (colored by university)",
      vertex.label=NA, edge.arrow.size=0.5)

```

Overall network (colored by university)



```
cat("Total nodes:", vcount(graph2), "\n")
```

```
## Total nodes: 1582
```

```
cat("Total edges (citations):", ecount(graph2), "\n")
```

```
## Total edges (citations): 3730
```

```
cat("Network density:", edge_density(graph2), "\n")
```

```
## Network density: 0.001491319
```

```

cat("Average degree:", mean(degree(graph2, mode="all")), "\n")

## Average degree: 4.71555

cat("Average in-degree (citations received):",
    mean(degree(graph2, mode="in")), "\n")

## Average in-degree (citations received): 2.357775

cat("Average out-degree (citations made):",
    mean(degree(graph2, mode="out")), "\n\n")

## Average out-degree (citations made): 2.357775

# Component Analysis
cat("== COMPONENT ANALYSIS ==\n")

## == COMPONENT ANALYSIS ==

components <- components(graph2, mode="weak")
cat("Number of weakly connected components:", components$no, "\n")

## Number of weakly connected components: 58

cat("Size of largest component:", max(components$csize), "\n")

## Size of largest component: 1433

cat("Proportion of nodes in largest component:",
    round(max(components$csize) / vcount(graph2) * 100, 2), "%\n\n")

## Proportion of nodes in largest component: 90.58 %

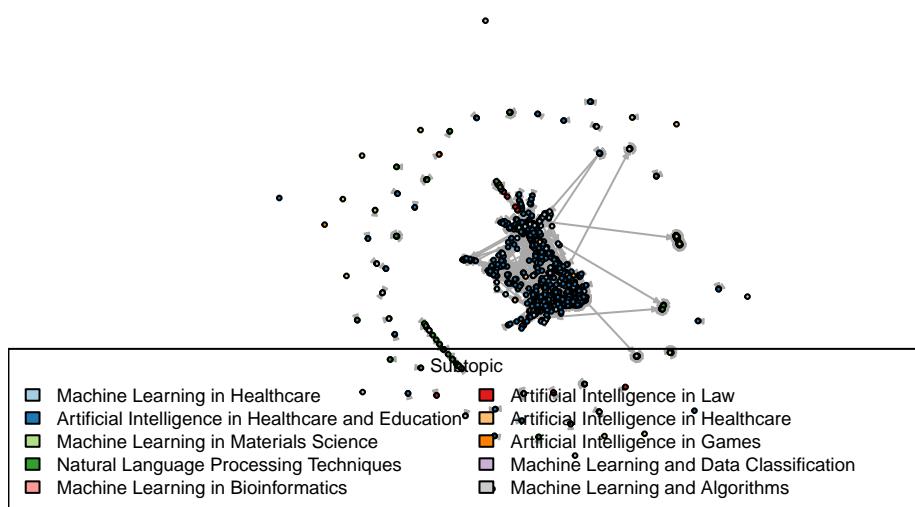
top_subtopics <- names(sort(table(V(graph2)$subtopic),
    decreasing = TRUE)[1:10])
V(graph2)$subtopic_group <- ifelse(V(graph2)$subtopic %in% top_subtopics,
    V(graph2)$subtopic, "Other")

# Create color palette
subtopic_groups <- unique(V(graph2)$subtopic_group)
n_groups <- length(subtopic_groups)
palette_sub <- c(brewer.pal(min(n_groups-1, 11), "Paired"), "gray80")
color_map_sub <- setNames(palette_sub[1:n_groups], subtopic_groups)
V(graph2)$color <- color_map_sub[V(graph2)$subtopic_group]

# Plot
plot(graph2, vertex.size=2, edge.arrow.size=0.2, vertex.color=V(graph2)$color,
    main="Citation Network (colored by subtopic)", vertex.label=NA)
legend("bottomleft", legend=names(color_map_sub), fill=color_map_sub,
    cex=0.6, title="Subtopic", ncol=2)

```

Citation Network (colored by subtopic)



```
# Calculate all centrality metrics
V(graph2)$in_degree <- degree(graph2, mode="in")
V(graph2)$out_degree <- degree(graph2, mode="out")
V(graph2)$total_degree <- degree(graph2, mode="all")

# PageRank (influence - highly cited by other highly cited papers)
V(graph2)$pagerank <- page_rank(graph2, directed = TRUE)$vector

# Betweenness (bridging papers, potentially unifying ideas)
V(graph2)$betweenness <- betweenness(graph2, directed = TRUE, normalized = TRUE)

# Closeness (foundational papers)
V(graph2)$closeness_out <- closeness(graph2, mode = "out", normalized = TRUE)
V(graph2)$closeness_in <- closeness(graph2, mode = "in", normalized = TRUE)

# Eigenvector centrality (alternative influence measure)
V(graph2)$eigenvector <- eigen_centrality(graph2, directed = TRUE)$vector
```

Analyze the isolated nodes (Jinxi Hu)

```
# Load isolated nodes data
nodes_isolated <- read.csv("data/nodes_isolated.csv")

cat("== ISOLATED NODES ANALYSIS ==\n")
```

```

## === ISOLATED NODES ANALYSIS ===

cat("Total isolated nodes:", nrow(nodes_isolated), "\n")

## Total isolated nodes: 1028

cat("Isolated nodes have no citations in or out\n\n")

## Isolated nodes have no citations in or out

# Basic statistics for isolated nodes
cat("Year range for isolated papers:", min(nodes_isolated$year, na.rm=TRUE), "-",
    max(nodes_isolated$year, na.rm=TRUE), "\n")

## Year range for isolated papers: 2015 - 2025

cat("Average citations for isolated papers:",
    round(mean(nodes_isolated$citations, na.rm=TRUE), 2), "\n")

## Average citations for isolated papers: 9.97

cat("Average references for isolated papers:",
    round(mean(nodes_isolated$references, na.rm=TRUE), 2), "\n\n")

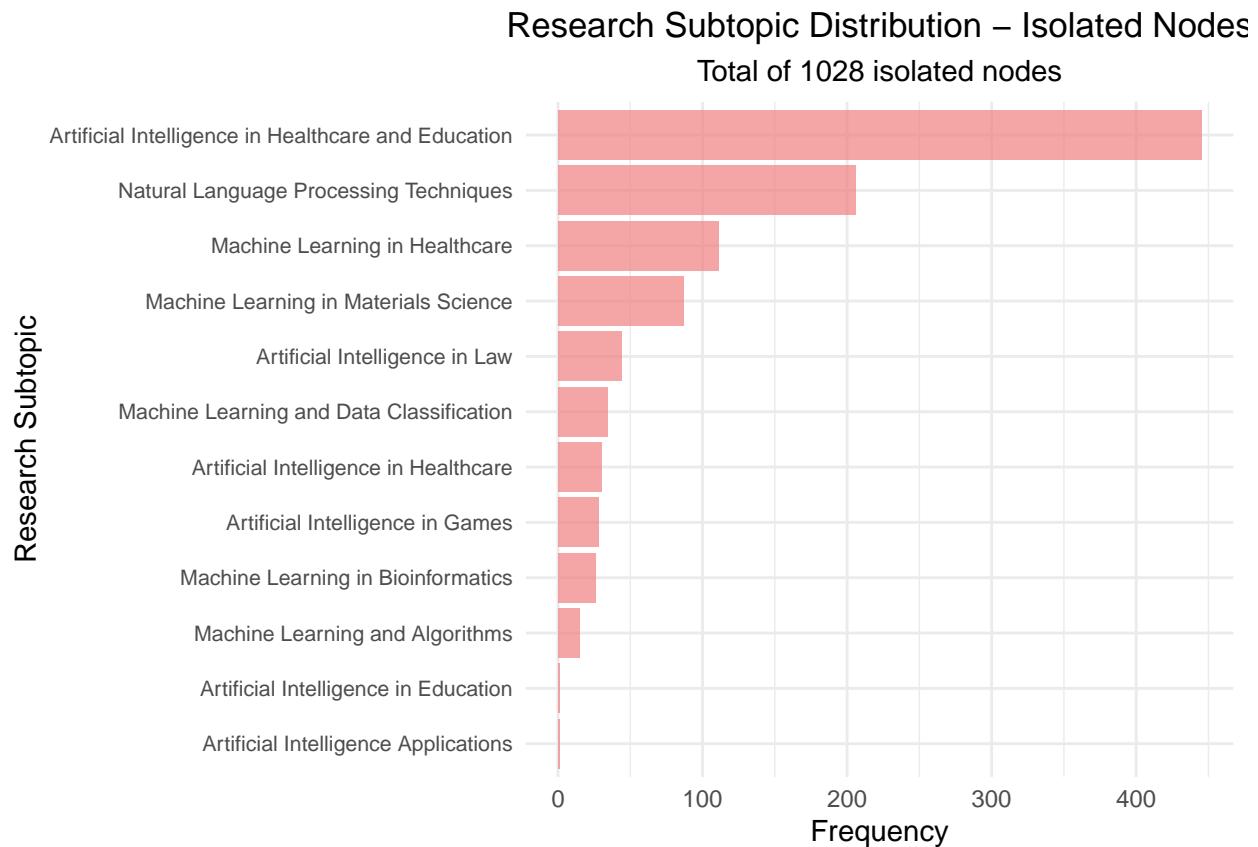
## Average references for isolated papers: 22.48

# Frequency analysis by research subtopic for isolated nodes

# Count frequency of subtopics in isolated nodes
subtopic_freq_isolated <- table(nodes_isolated$subtopic)
subtopic_df_isolated <- data.frame(
  subtopic = names(subtopic_freq_isolated),
  frequency = as.numeric(subtopic_freq_isolated))
subtopic_df_isolated <- subtopic_df_isolated[order(
  subtopic_df_isolated$frequency, decreasing = TRUE), ]

# Create histogram for subtopics (isolated nodes)
ggplot(subtopic_df_isolated, aes(x = reorder(subtopic, frequency),
                                 y = frequency)) +
  geom_bar(stat = "identity", fill = "lightcoral", alpha = 0.7) +
  coord_flip() +
  labs(title = "Research Subtopic Distribution - Isolated Nodes",
       subtitle = paste("Total of", nrow(nodes_isolated), "isolated nodes"),
       x = "Research Subtopic",
       y = "Frequency") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 8),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))

```



```
# Print top subtopics for isolated nodes
cat("Top 10 research subtopics in isolated nodes:\n")
```

```
## Top 10 research subtopics in isolated nodes:
```

```
print(head(subtopic_df_isolated, 10))
```

	subtopic	frequency
## 5	Artificial Intelligence in Healthcare and Education	445
## 12	Natural Language Processing Techniques	206
## 10	Machine Learning in Healthcare	111
## 11	Machine Learning in Materials Science	87
## 6	Artificial Intelligence in Law	44
## 8	Machine Learning and Data Classification	34
## 4	Artificial Intelligence in Healthcare	30
## 3	Artificial Intelligence in Games	28
## 9	Machine Learning in Bioinformatics	26
## 7	Machine Learning and Algorithms	15

```
# Frequency analysis by institution for isolated nodes
```

```
# Count frequency of institutions in isolated nodes
institution_freq_isolated <- table(nodes_isolated$institution)
institution_df_isolated <- data.frame(
```

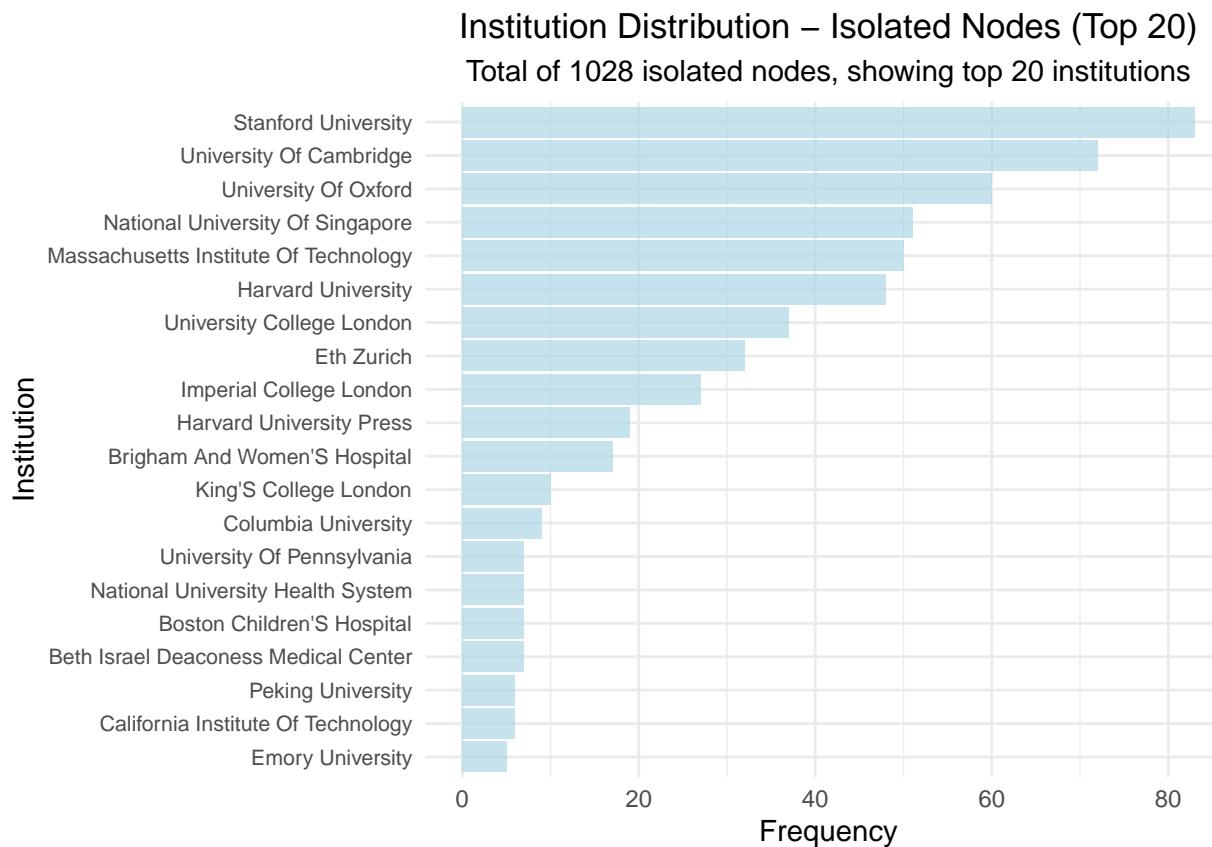
```

institution = names(institution_freq_isolated),
frequency = as.numeric(institution_freq_isolated)
)
institution_df_isolated <- institution_df_isolated[order(
  institution_df_isolated$frequency, decreasing = TRUE), ]

# Show only top 20 institutions for better visualization
top_institutions_isolated <- head(institution_df_isolated, 20)

# Create histogram for institutions (isolated nodes)
ggplot(top_institutions_isolated, aes(x = reorder(institution, frequency),
                                       y = frequency)) +
  geom_bar(stat = "identity", fill = "lightblue", alpha = 0.7) +
  coord_flip() +
  labs(title = "Institution Distribution – Isolated Nodes (Top 20)",
       subtitle = paste("Total of", nrow(nodes_isolated),
                     "isolated nodes, showing top 20 institutions"),
       x = "Institution",
       y = "Frequency") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 8),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))

```



```

# Print top institutions for isolated nodes
cat("\nTop 15 institutions with isolated nodes:\n")

## 
## Top 15 institutions with isolated nodes:

print(head(institution_df_isolated, 15))

##                                     institution frequency
## 258                      Stanford University      83
## 308                  University Of Cambridge      72
## 332                  University Of Oxford       60
## 206 National University Of Singapore       51
## 173 Massachusetts Institute Of Technology     50
## 104                  Harvard University      48
## 291          University College London      37
## 80                     Eth Zurich            32
## 120          Imperial College London      27
## 105          Harvard University Press      19
## 33          Brigham And Women'S Hospital     17
## 146          King'S College London       10
## 63             Columbia University            9
## 26 Beth Israel Deaconess Medical Center       7
## 29 Boston Children'S Hospital                7

# Compare isolated vs connected nodes

# For connected nodes
subtopic_freq_connected <- table(V(graph2)$subtopic)
institution_freq_connected <- table(V(graph2)$institution)

# Create comparison data frame for subtopics
subtopic_comparison <- data.frame(
  subtopic = unique(c(names(subtopic_freq_isolated),
                      names(subtopic_freq_connected))),
  isolated = 0,
  connected = 0
)

# Fill in frequencies
for(i in 1:nrow(subtopic_comparison)) {
  topic <- subtopic_comparison$subtopic[i]
  subtopic_comparison$isolated[i] <- ifelse(
    topic %in% names(subtopic_freq_isolated),
    subtopic_freq_isolated[topic], 0)
  subtopic_comparison$connected[i] <- ifelse(
    topic %in% names(subtopic_freq_connected),
    subtopic_freq_connected[topic], 0)
}

# Calculate proportions
subtopic_comparison$total <- subtopic_comparison$isolated + subtopic_comparison$connected

```

```

subtopic_comparison$isolated_prop <- subtopic_comparison$  

  isolated / subtopic_comparison$total  

subtopic_comparison <- subtopic_comparison[order(  

  subtopic_comparison$isolated_prop, decreasing = TRUE), ]  

cat("\n==== COMPARISON: ISOLATED vs CONNECTED NODES ===\n")  

##  

## === COMPARISON: ISOLATED vs CONNECTED NODES ===  

cat("Research subtopics with highest isolation rates:\n")  

## Research subtopics with highest isolation rates:  

print(head(subtopic_comparison[subtopic_comparison$total >= 5,  

  c("subtopic", "isolated", "connected",  

  "isolated_prop")], 10))  

##  

##          subtopic isolated connected  

## 7      Machine Learning and Algorithms      15       1  

## 8      Machine Learning and Data Classification    34       4  

## 3      Artificial Intelligence in Games      28       4  

## 9      Machine Learning in Bioinformatics     26       7  

## 12     Natural Language Processing Techniques   206      78  

## 6      Artificial Intelligence in Law        44      23  

## 4      Artificial Intelligence in Healthcare   30      18  

## 11     Machine Learning in Materials Science   87      74  

## 10     Machine Learning in Healthcare        111     167  

## 5  Artificial Intelligence in Healthcare and Education  445     1206  

##  isolated_prop  

## 7      0.9375000  

## 8      0.8947368  

## 3      0.8750000  

## 9      0.7878788  

## 12     0.7253521  

## 6      0.6567164  

## 4      0.6250000  

## 11     0.5403727  

## 10     0.3992806  

## 5      0.2695336  

# Visualize comparison of isolated vs connected by subtopic  

# Prepare data for comparison plot (top 15 subtopics by total count)
top_subtopics_total <- head(subtopic_comparison[order(subtopic_comparison$total,  

  decreasing = TRUE), ], 15)
comparison_melted <- melt(top_subtopics_total[,  

  c("subtopic", "isolated", "connected")],  

  id.vars = "subtopic",  

  variable.name = "type",  

  value.name = "count")

```

```

# Create stacked bar chart
ggplot(comparison_melted, aes(x = reorder(subtopic, count), y = count,
                               fill = type)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_fill_manual(values = c("isolated" = "lightcoral",
                               "connected" = "lightblue"),
                    labels = c("Isolated Nodes", "Connected Nodes")) +
  labs(title = "Research Subtopic Distribution Comparison",
       Isolated vs Connected Nodes",
       subtitle = "Top 15 research subtopics ordered by total paper count",
       x = "Research Subtopic",
       y = "Number of Papers",
       fill = "Node Type") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 8),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = "bottom")

```

