Understanding the Academic World

By: Berkan Teber
Supervised By: Kamer Kaya, Hüsnü Yenigün

3 Questions to Answer

- Which papers are the most influential in their field?
- Which researchers are more successful?
- Which scientific areas are more promising?

3 Tasks to Accomplish

- Evaluate papers — citation count vs pagerank
- Evaluate authors — h, g, h’ indexes, h_{rwr}
- Evaluate fields — community detection, statistics

Datasets

- 2 datasets:
  - AMiner dataset — 154 million papers
  - Microsoft Academic Graph — 166 million papers
  - 64 million one-to-one matching between them

Data Preparation

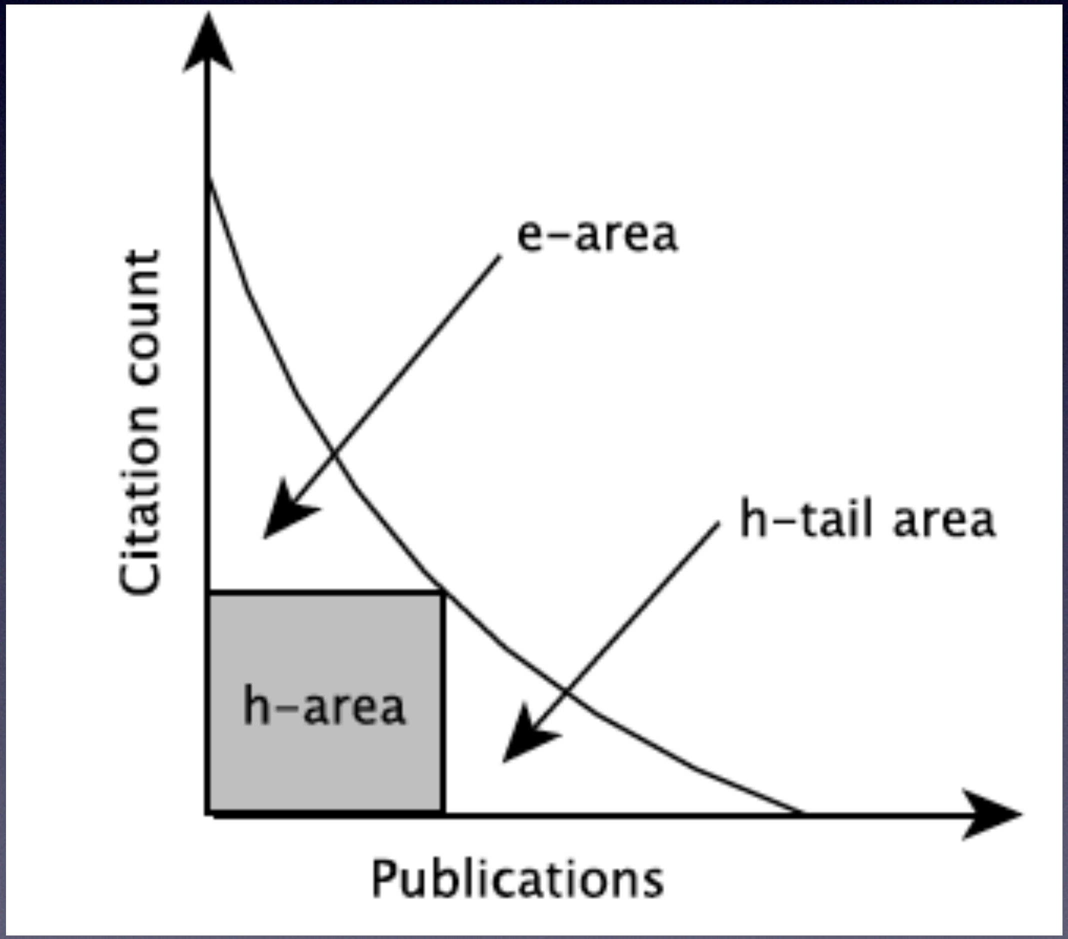
- Filtering necessary attributes — year, authors & affiliations, references, keywords, fields of study
- Constructing CRS (compressed row storage) formatted graphs
  - paper-cites-paper
  - paper-citedby-paper
  - author-writes-paper
  - author-collaborates-author
  - author-cites-author

Evaluating Papers

- Looking at citation counts
  - most current methods use this approach — h-index, g-index, h’-index
- PageRank (based on Random Walk w/ Restarts)
  - a newer approach in literature
  - implemented by another team

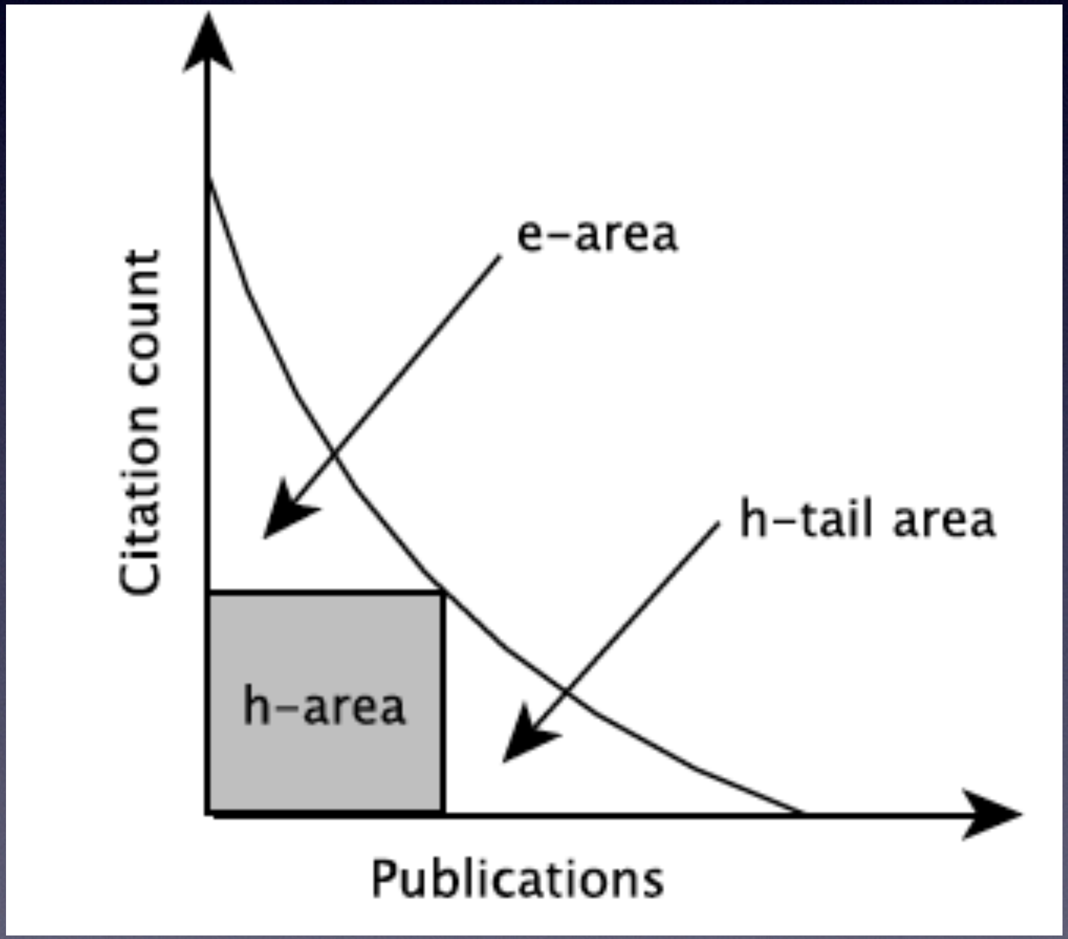
Evaluating Authors

- h-index
  - a researcher has an h-index of h if h of his papers have at least h citations and all of the remaining papers have less than h citations
  - uses only the h-area



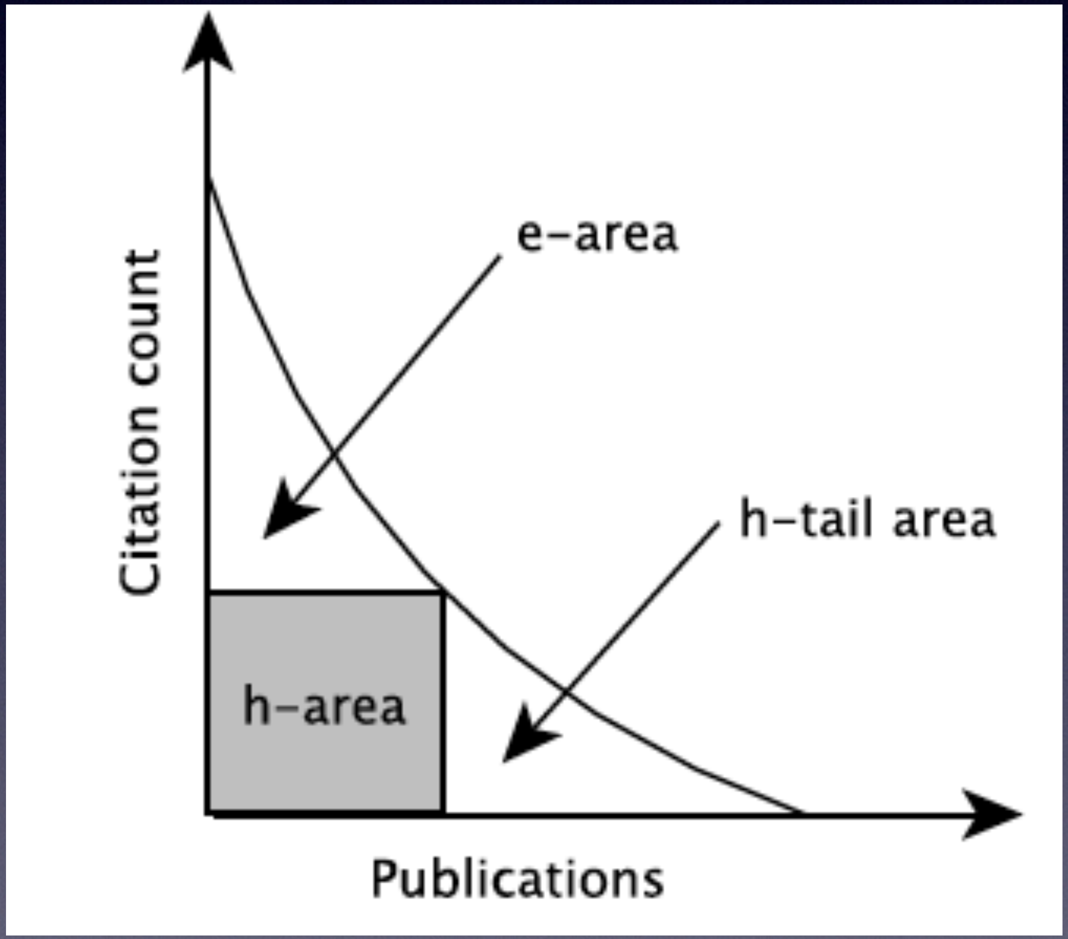
Evaluating Authors

- g-index
  - a researcher has an g-index of g if top g papers have at least g^2 citations in total and top (g+1) papers have less than $(g+1)^2$ citations
  - uses both h-area and e-area



Evaluating Authors

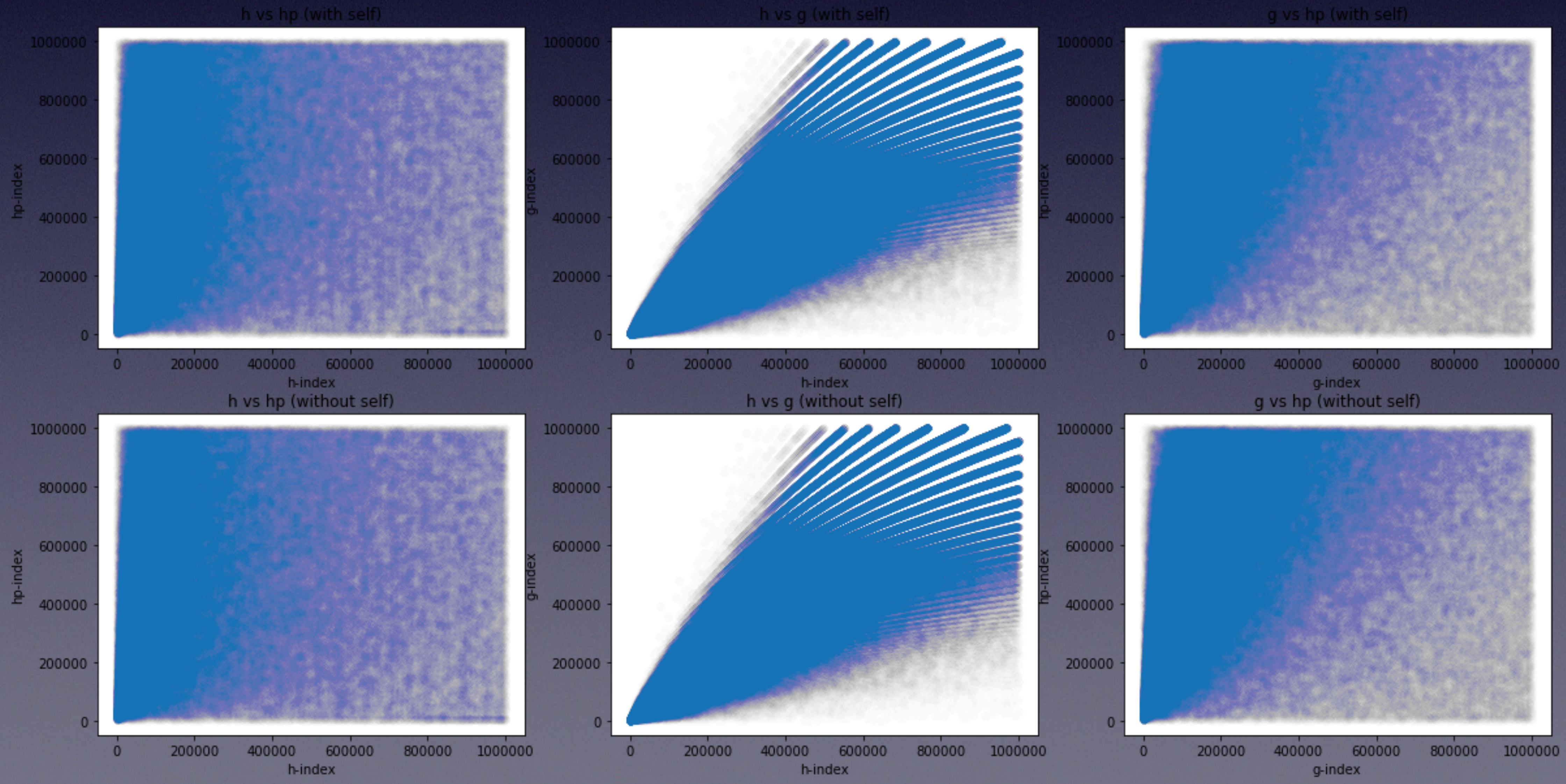
- h'-index
  - a researcher has an h'-index h' = eh/t where h is his h-index, and e and t are the square roots of e-area and h-tail area respectively
  - uses all three areas



Comparing Indexes

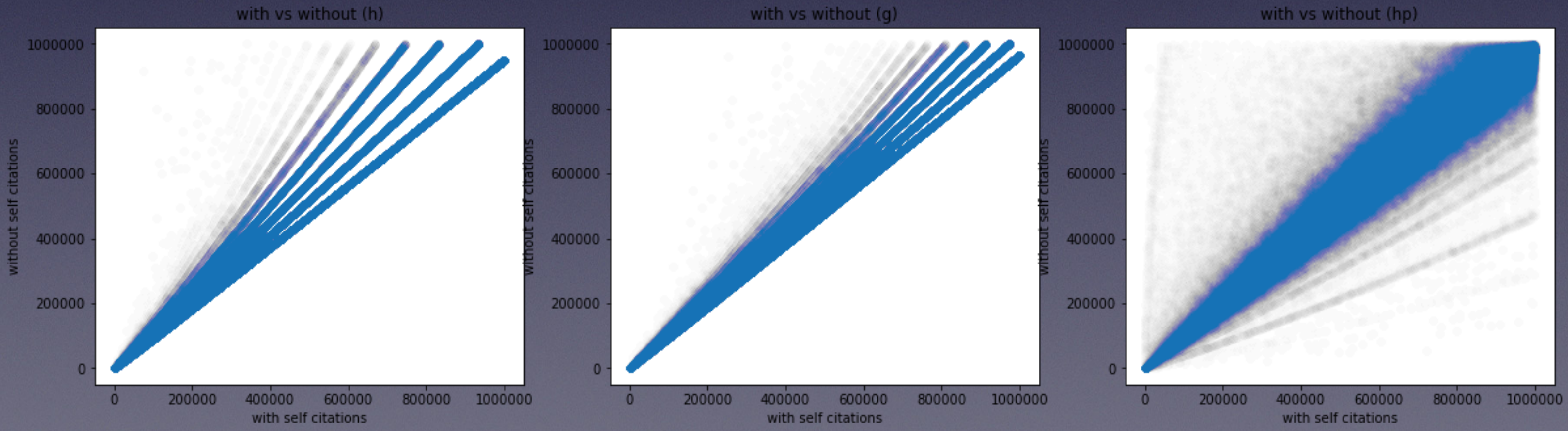
- There are no correlations between h’-index and other indexes.

- However, there is a correlation between h-index and g-index.



Comparing Indexes

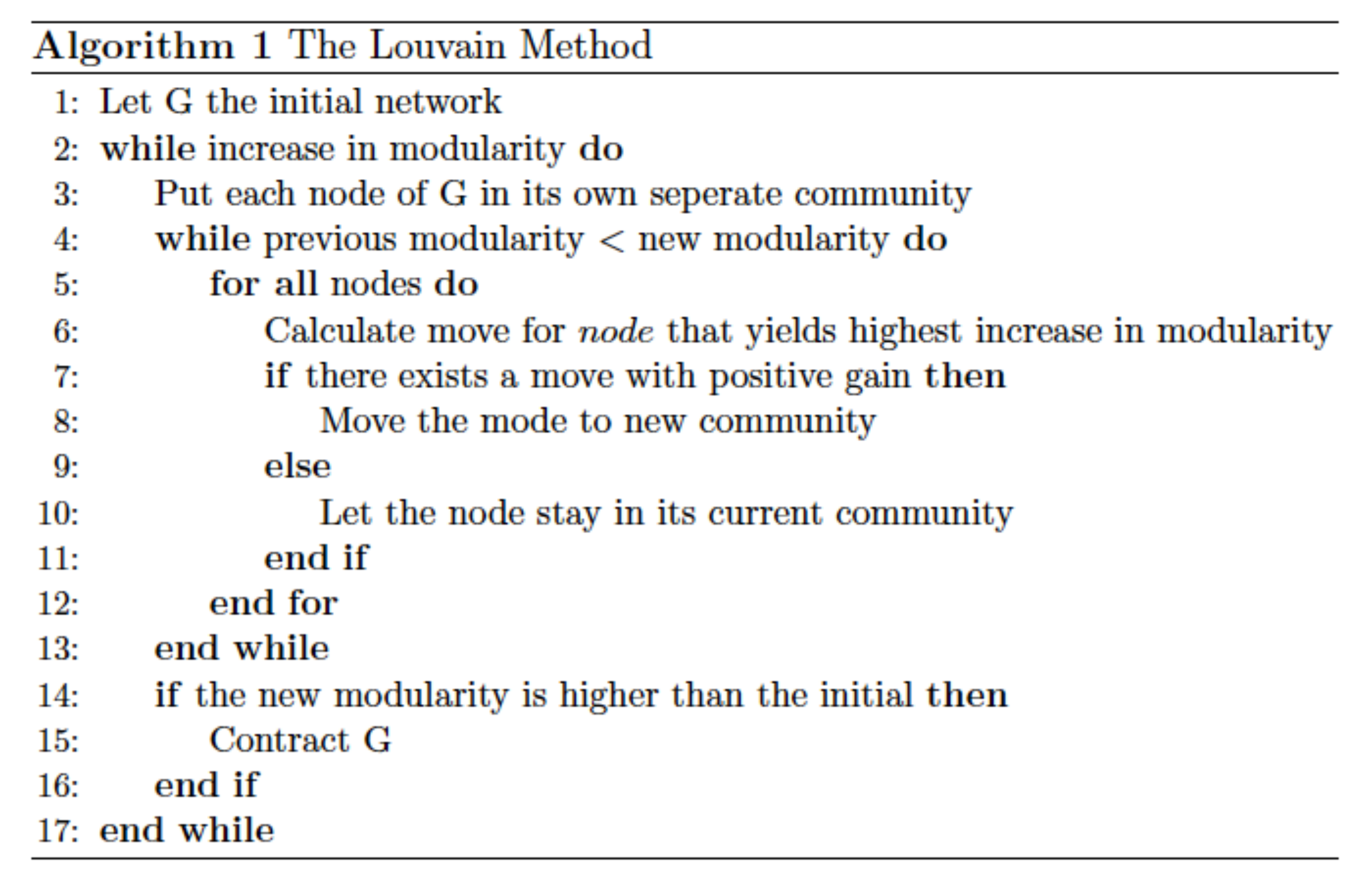
- The effect of self-citations is the most in h’-index.
- This effect is the least in g-index.



Evaluating Fields

- Field identification
    - Community detection — Louvain method
    - Frequency analysis on keywords
    - Word clouds
- Comparison of fields with h-index

Community Detection — Louvain Method



Algorithm 1 The Louvain Method

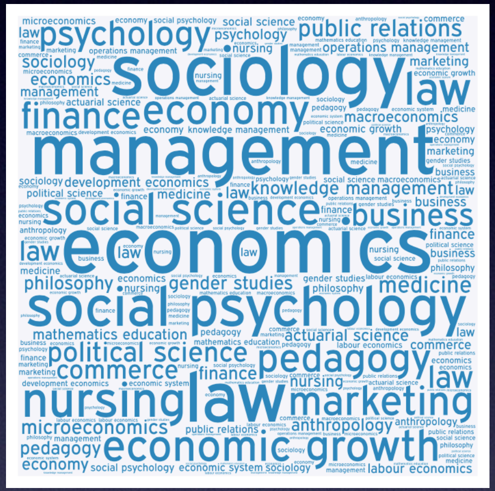
```
Let G the initial network
while increase in modularity do
    Put each node of G in its own seperate community
    while previous modularity $<$ new modularity do
        for all nodes do
            Calculate move for node that yields highest increase in modularity
            if there exists a move with positive gain then
                Move the mode to new community
            else
                Let the node stay in its current community
            end if
        end for
    end while
    if the new modularity is higher than the initial then
        Contract G
    end if
end while
```

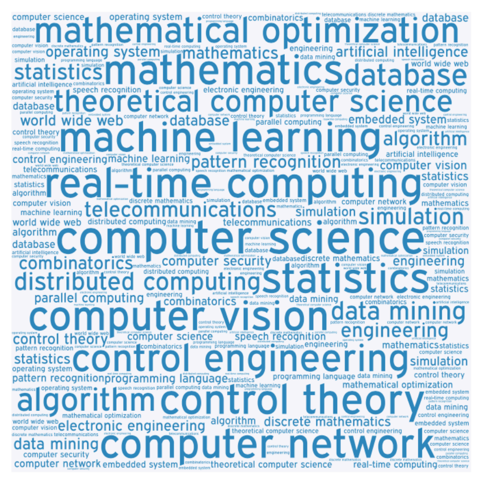
(Herman Moyner Lund, 2017)

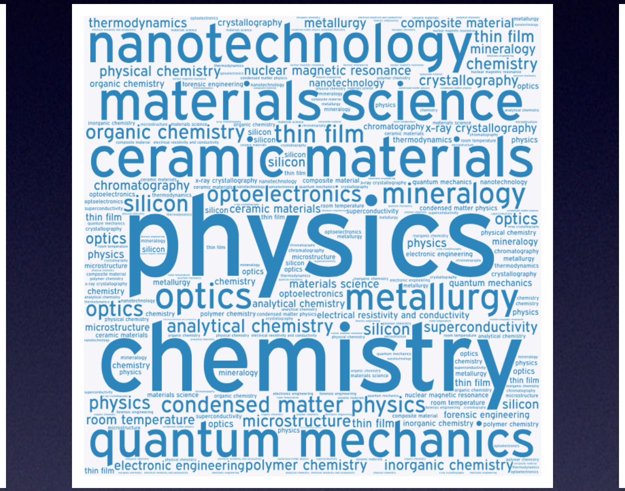
Frequency Analysis of Keywords

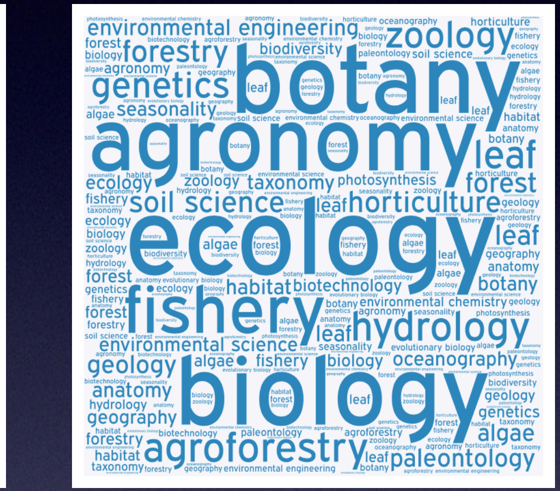
- score = $(\text{freq})^2$ x expected
    - where freq is the frequency of keywords within the community and expected is the expected frequency obtained by the general distribution and the community size

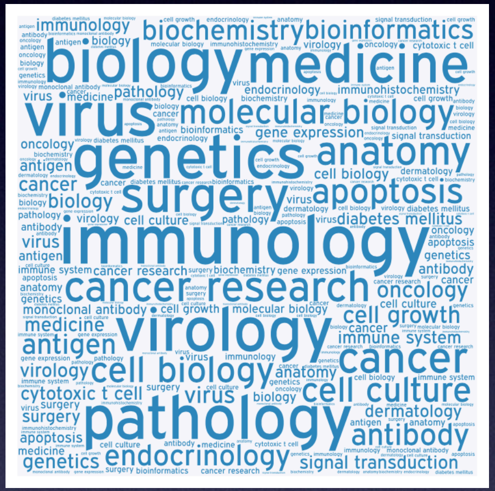
Word Clouds

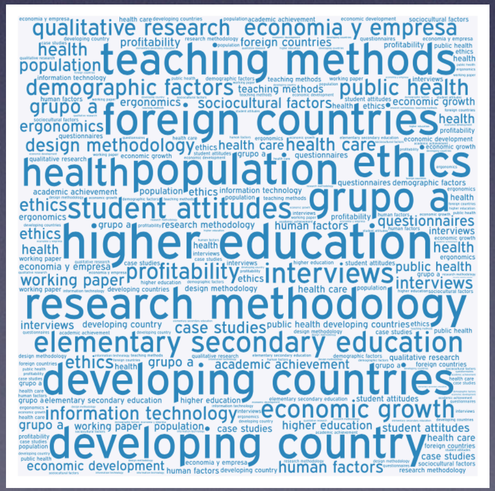

psychology
public relations
sociology
law
finance
economy
management
social science
business
economics
social psychology
political science
pedagogy
commerce
nursing
marketing
economic growth

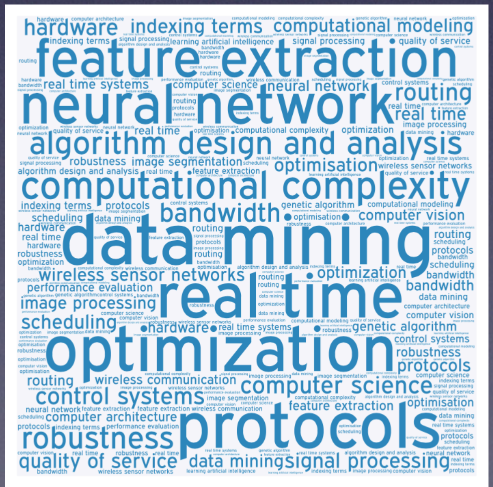

mathematical optimization
mathematics
statistics
database
theoretical computer science
machine learning
real-time computing
computer science
distributed computing
computer vision
control engineering
algorithm
control theory
computer network

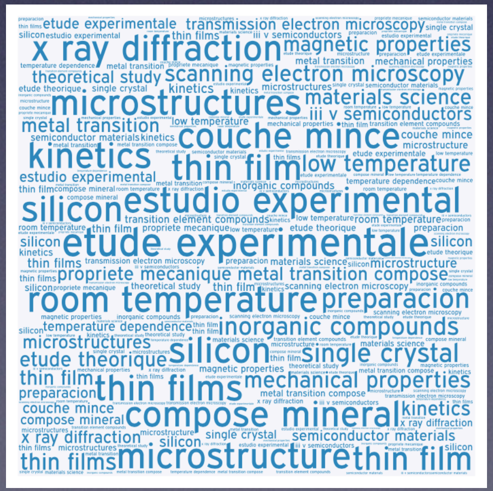

nanotechnology
materials science
ceramic
materials
physics
optoelectronics
mineralogy
optics
metallurgy
chemistry
condensed matter physics
quantum mechanics

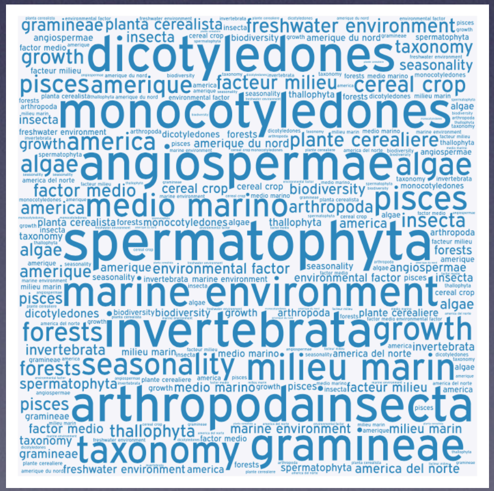

environmental engineering
forestry
zoology
genetics
botany
agronomy
leaf
soil science
horticulture
ecology
fishery
hydrology
biology
agroforestry
paleontology

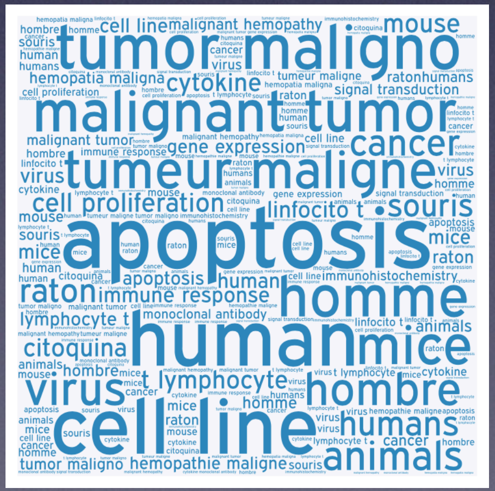

biochemistry
bioinformatics
biology
medicine
virus
molecular biology
genetics
anatomy
surgery
apoptosis
immunology
cancer research
oncology
virology
cell biology
cancer
cell culture
pathology
antibody
endocrinology
signal transduction


qualitative research
economía y empresa
teaching methods
demographic factors
public health
foreign countries
healthpopulation
ethics
student attitudes
grupo a
higher education
interviews
research methodology
elementary secondary education
developing countries
economic growth
developing country


indexing terms
computational modeling
feature extraction
neural network
routing
algorithm design and analysis
computational complexity
bandwidth
data mining
real time
optimization
control systems
computer science
robustness
protocols
quality of service
signal processing


x ray diffraction
magnetic properties
electron microscopy
microstructures
materials science
couche mince
kinetics
thin film
low temperature
silicon
estudio experimental
etude experimentale
room temperature
preparacion
inorganic compounds
single crystal
thin films
mechanical properties
compose mineral
microstructure
thin film


gramineae
freshwater environment
dicotyledones
taxonomy
pisces
amerique
facteur milieu
cereal crop
monocotyledones
america
plante cerealiere
angiospermae
algae
medio marino
pisces
estudio experimental
spermatophyta
marine environment
invertebrata
growth
seasonality
milieu marin
arthropoda
insecta
taxonomy
gramineae

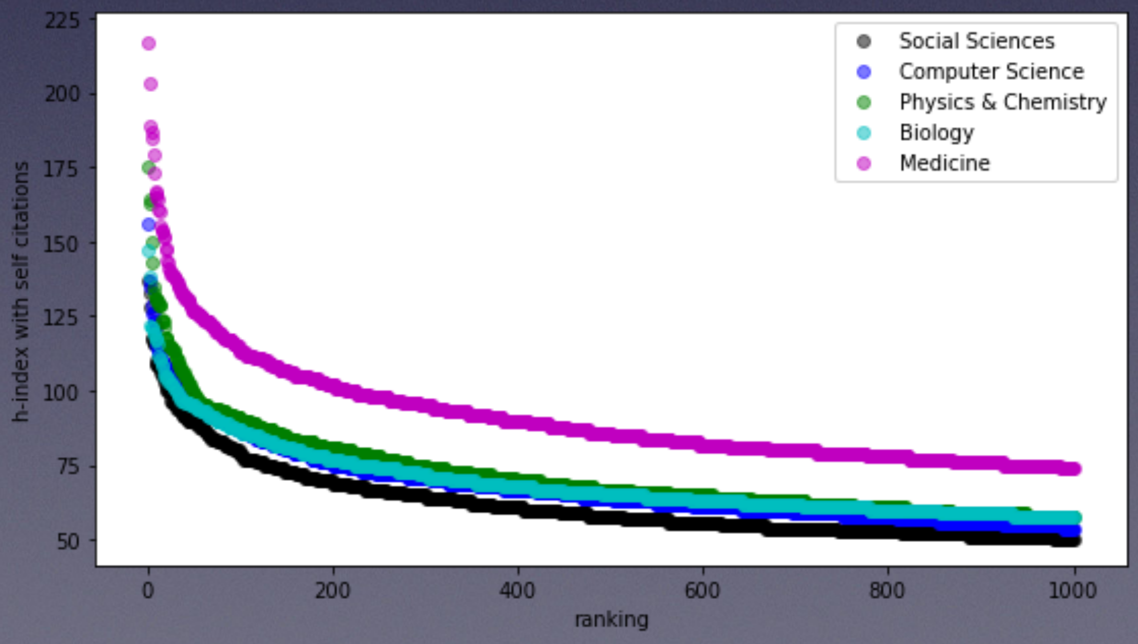

tumor maligno
mouse
malignant tumor
cancer
tumeur maligne
virus
cell proliferation
souris
apoptosis
mice
raton
homme
immune response
human
mice
citoquina
lymphocyte
hombre
virus
humans
cell line
animals

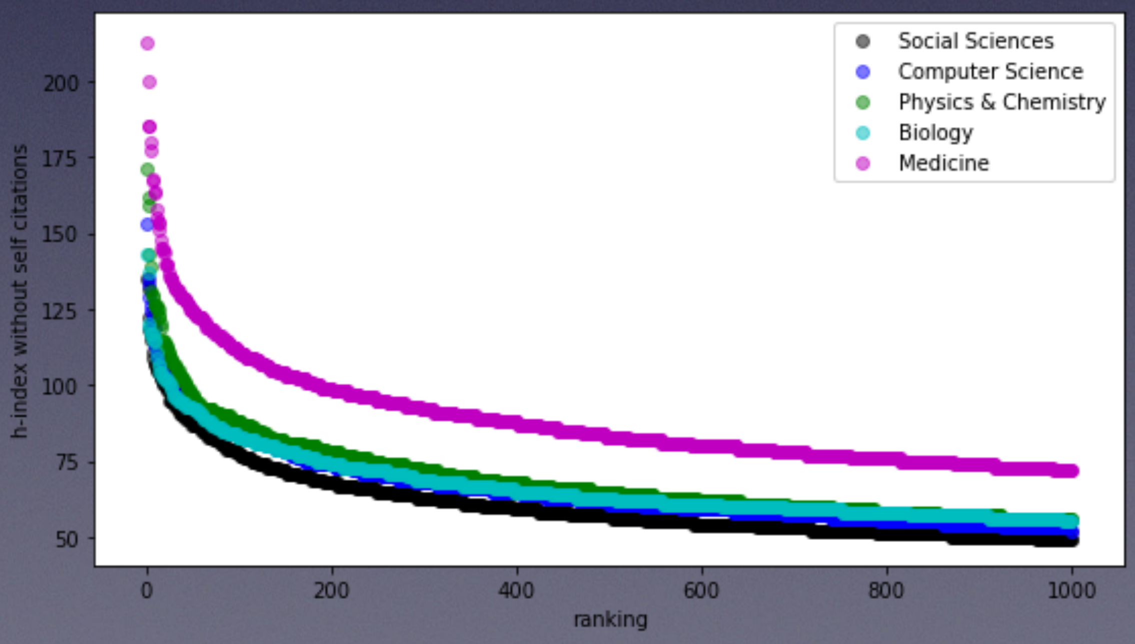
Field Identification

1. Social Sciences
2. Computer Science
3. Physics & Chemistry
4. Biology
5. Medicine

Field Comparison

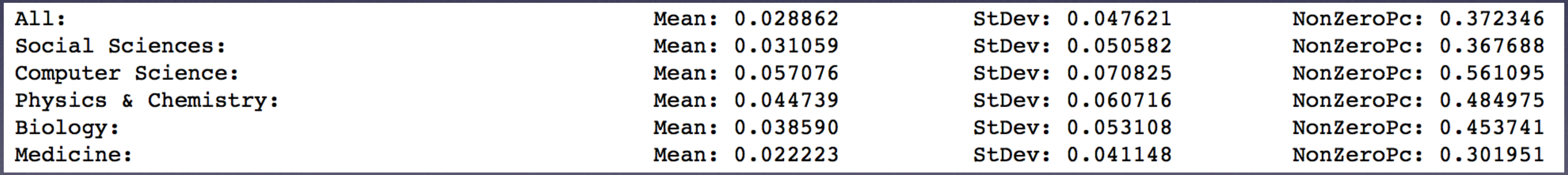
- As seen in the figure, top authors of some fields, e.g. medicine, have much higher h-indexes compared to the top authors of other fields.





Field Comparison

- As seen in the table, of the top 5 fields, self-citation has the most effect in computer science field and the least effect in medicine field.



```
All:                                              Mean: 0.028862          StDev: 0.047621          NonZeroPc: 0.372346
Social Sciences:                                  Mean: 0.031059          StDev: 0.050582          NonZeroPc: 0.367688
Computer Science:                                 Mean: 0.057076          StDev: 0.070825          NonZeroPc: 0.561095
Physics & Chemistry:                              Mean: 0.044739          StDev: 0.060716          NonZeroPc: 0.484975
Biology:                                          Mean: 0.038590          StDev: 0.053108          NonZeroPc: 0.453741
Medicine:                                         Mean: 0.022223          StDev: 0.041148          NonZeroPc: 0.301951
```

References

- Jie Tang, Jing Zahng, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. *ArnetMiner: Extraction and Mining of Academic Social Networks.* In Proceedings of the 14th ACM SIGKDD Internatiinal Conference on Knowledge Discovery and Data Mining: p. 990-998.

- Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. (2015). *An Overview of Microsoft Academic Service (MAS) and Applications.* In Proceedings of the 24th International Conference on World Wide Web: p. 243-246.

- Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre. (2008). *Fast Unfolding of Communities in Large Networks.* In Journal of Statistical Mechanics: Theory and Experiment 2008(10): P10008 (12pp).

- Herman Moyner Lund. (2017). *Community Detection in Complex Networks.* Master Thesis in Department of Informatics, University of Berlin.

References

- Hirsch, J. E. (2005). *An index to quantify an individual's research output.* In Proceedings of the National Academy of Sciences of the United States of America. 102(46): p. 16569-16572.

- Egghe, L. (2006). *Theory and practice of the g-index.* In Scientometrics. 69(1): p.131-152.

- Zhang, C. T. (2013). *The h'-Index, Effectively Improving th h-Index Based on the Citation Distribution.* In PLoS ONE. 8(4): p. e59912.

- Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web.* Technical report, Stanford InfoLab.