Assessing the scientific impact of research

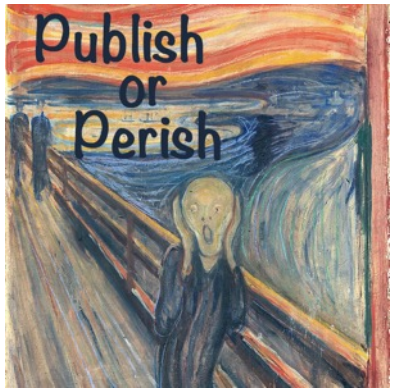
Thanasis Vergoulis

1/10/2020

ATHENA Research & Innovation Information Technologies

elixir GREECE

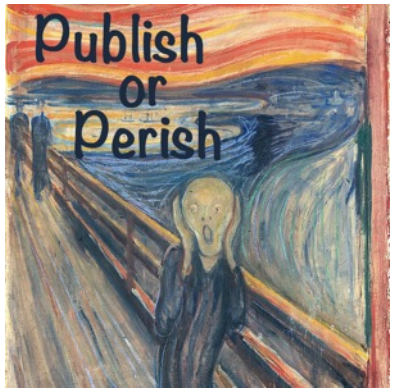
The constantly increasing size of scientific output



- **Exponential growth of scientific publications**
  - Increase in the number of researchers worldwide [1]
  - Large number of journals [2]
  - “Publish or perish” [3]

Images: Rob Curran on Unsplash, Jon Tyson on Unsplash, Wikimedia Commons (public domain)

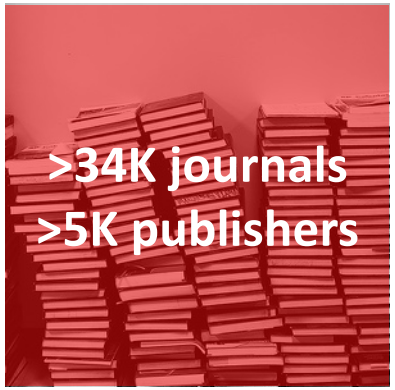
The constantly increasing size of scientific output

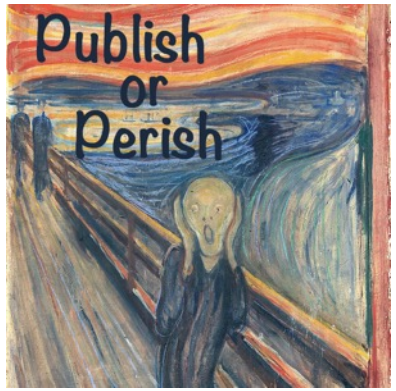



- **Exponential growth of scientific publications**
  - Increase in the number of researchers worldwide [1]
  - Large number of journals [2]
  - “Publish or perish” [3]

Images: Rob Curran on Unsplash, Jon Tyson on Unsplash, Wikimedia Commons (public domain)

The constantly increasing size of scientific output





- **Exponential growth of scientific publications**
  - Increase in the number of researchers worldwide [1]
  - Large number of journals [2]
  - “Publish or perish” [3]

Images: Rob Curran on Unsplash, Jon Tyson on Unsplash, Wikimedia Commons (public domain)

In many cases, articles contain low quality research[4,5]…

honest mistakes because of time pressure (publish or perish)

predatory publishers

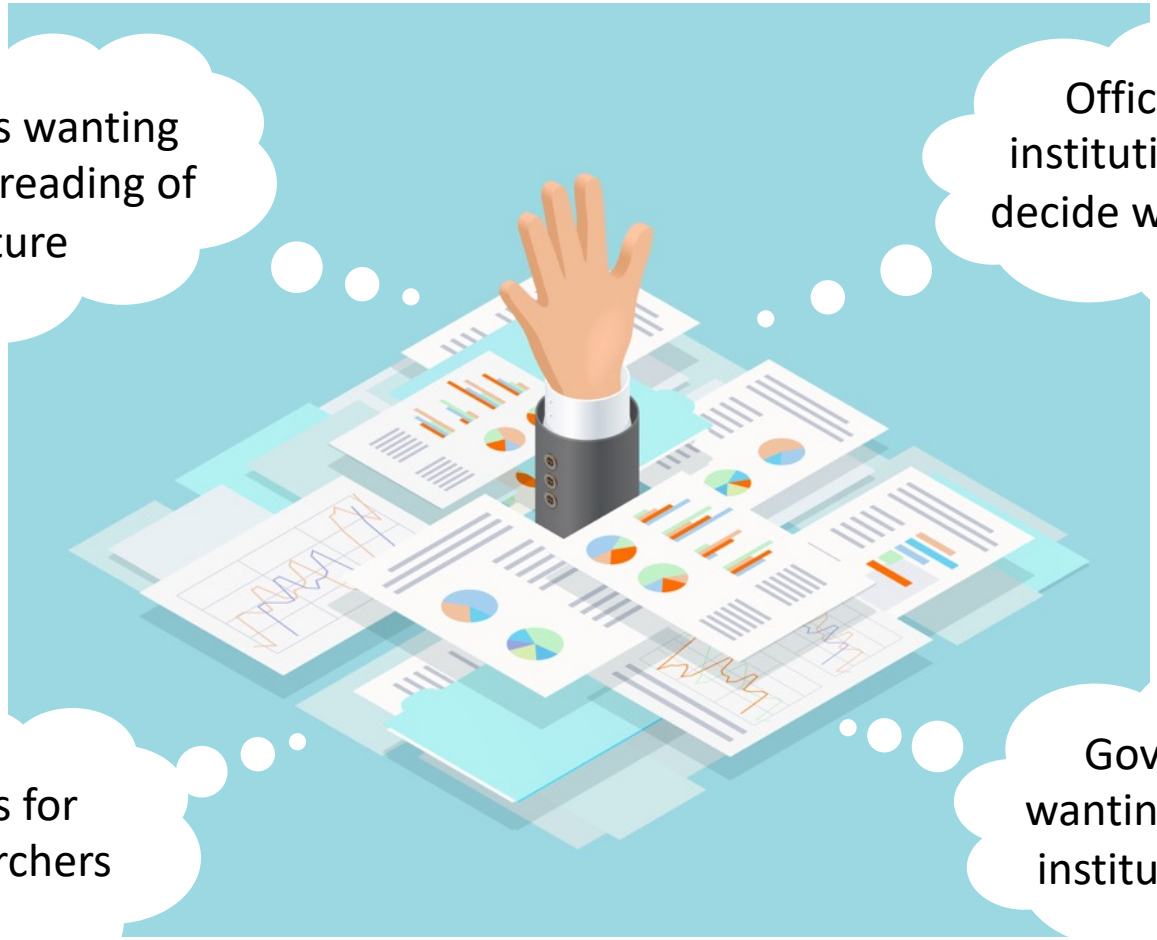
data manipulation
e.g. pi-hacking

peer-review scams
plagiarism

Assessing scientific impact is crucial for…

Researchers wanting to prioritize reading of literature

Officers in funding institutions that want to decide which field to fund



Hiring commitees for proffesors & researchers positions

Government officers wanting to assess research institutions & universities

Graphics: Designed by macrovector / Freepik

A major problem…

Useful data for reseach assessment:
- Publication content
- References (citations)
- Social media data (e.g., #tweets)
- Usage data (e.g., #downloads)

These data have been **isolated in data silos** of publishers or/and research organizations

Photo by Jeremy Olson on Unsplash

Open Science - Open Access - Science 2.0

- Ongoing change:
  - Many **open science & open access initiatives**
    - BOAI https://www.budapestopenaccessinitiative.org/
    - cOAlation S https://www.scienceeurope.org/coalition-s/
    - Scientific data should be **FAIR**: Findable, Accessible, Interoperable, Reusable
  - Most research is **publicly funded** → the results should be open

Melanie Imming, & Jon Tennant. (2018, June 8). Sticker open science: just science done right. Zenodo. (http://doi.org/10.5281/zenodo.1285575)

Scientific impact has many aspects

Many diverge indicators/metrics have been proposed in the literature.

Each captures a slightly or completely **different impact aspect**.

We will focus on **citation-based** metrics.

Designed by stories / Freepik

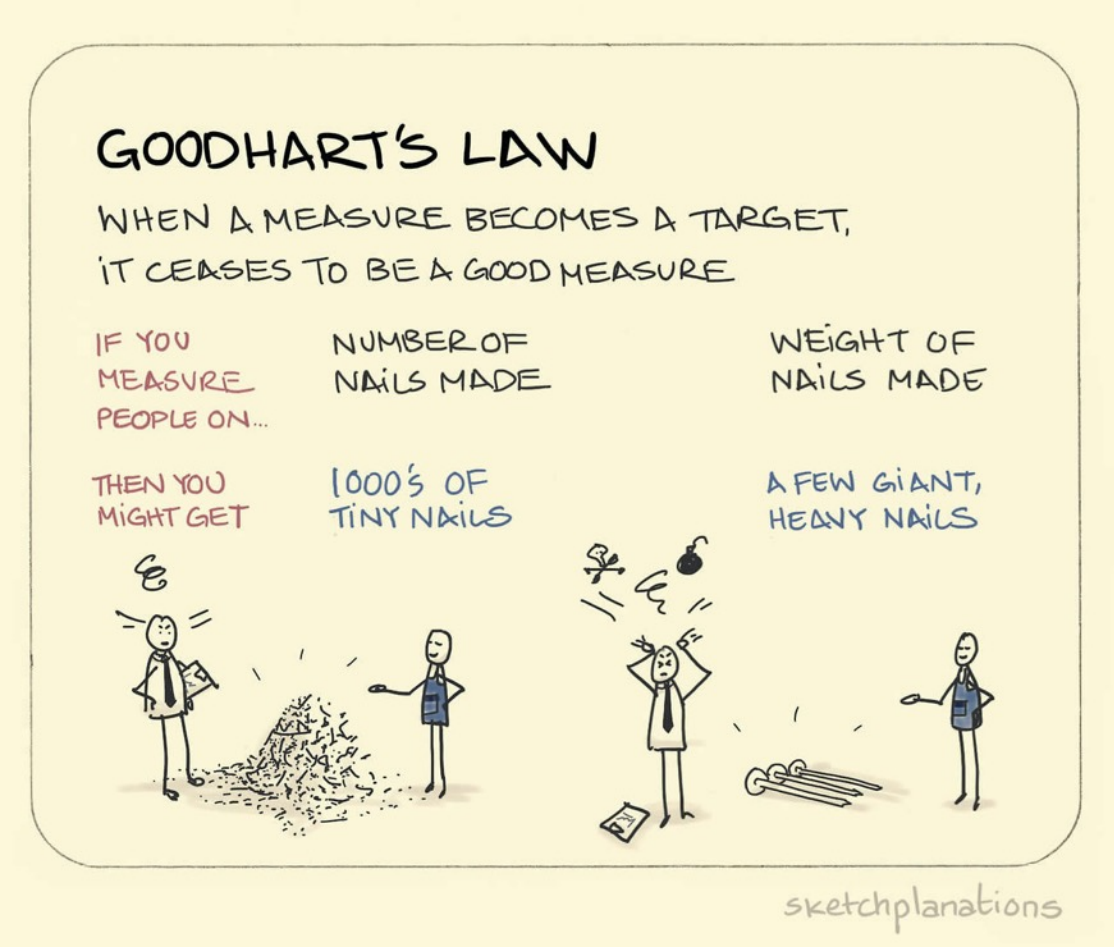
No silver bullets

- Naïve to believe that **one metric captures the "holistic impact".**

- The impact of a paper has **many different aspects** [6]
  - **Popularity**: Short-term impact (having a "hype")
  - **Influence**: Long-term impact (being "fundamental" for a discipline)
  - **Social impact**: Having hype in social media (e.g., altmetrics)

- Different aspects may be more important for different users.
  - Different algorithms **may capture better different impact aspects**.

Don’t forget Goodhart’s law!



Measuring impact

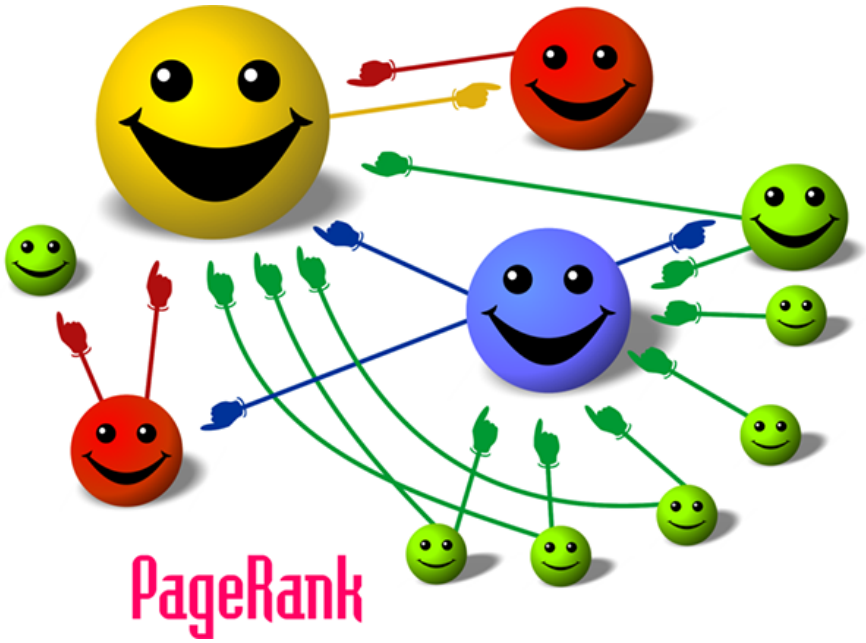
- How?
  - Use **citation count** (CC) to measure impact
    - Not all citations equally important
  - **Citation networks ~ Web** (page=paper, link=citation)
    - Use **PageRank** (PR) instead (link analysis – measures centrality
    - Considers the impact/importance of those citing a paper

$$s_i = \alpha \sum_j P_{i,j} s_j + (1-\alpha) v_i$$

$P_{i,j}$: 1/out_deg(j)

$(1-\alpha)$: random jump probability

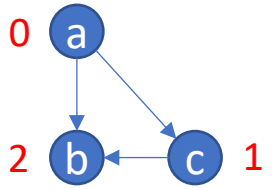
v_i: landing probability



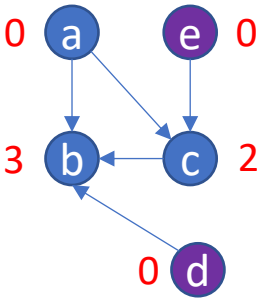
Bias against recent papers

- Centrality-based impact (CC,PR) inserts **bias against recently published papers** [7,8,9]
  - New papers have almost **zero in-degree**.
  - Citations to these papers will appear after months or even years (**citation lag**). [10,11,12,13]
    - However, these **"late" citations indicate contemporary impact**, not future impact (papers being read now)

year X

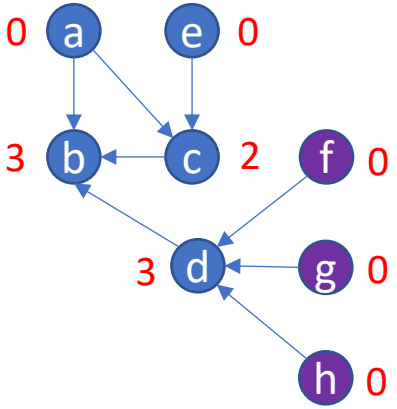


year X+1



- In this case, **network evolution is much slower than in the case of the Web** making the "bias-against-recent-nodes" problem more crucial.
  - Ranking algorithms trying to alleviate this issue have been proposed in the last years
    - Large room for improvements (as you will see)

year X+2

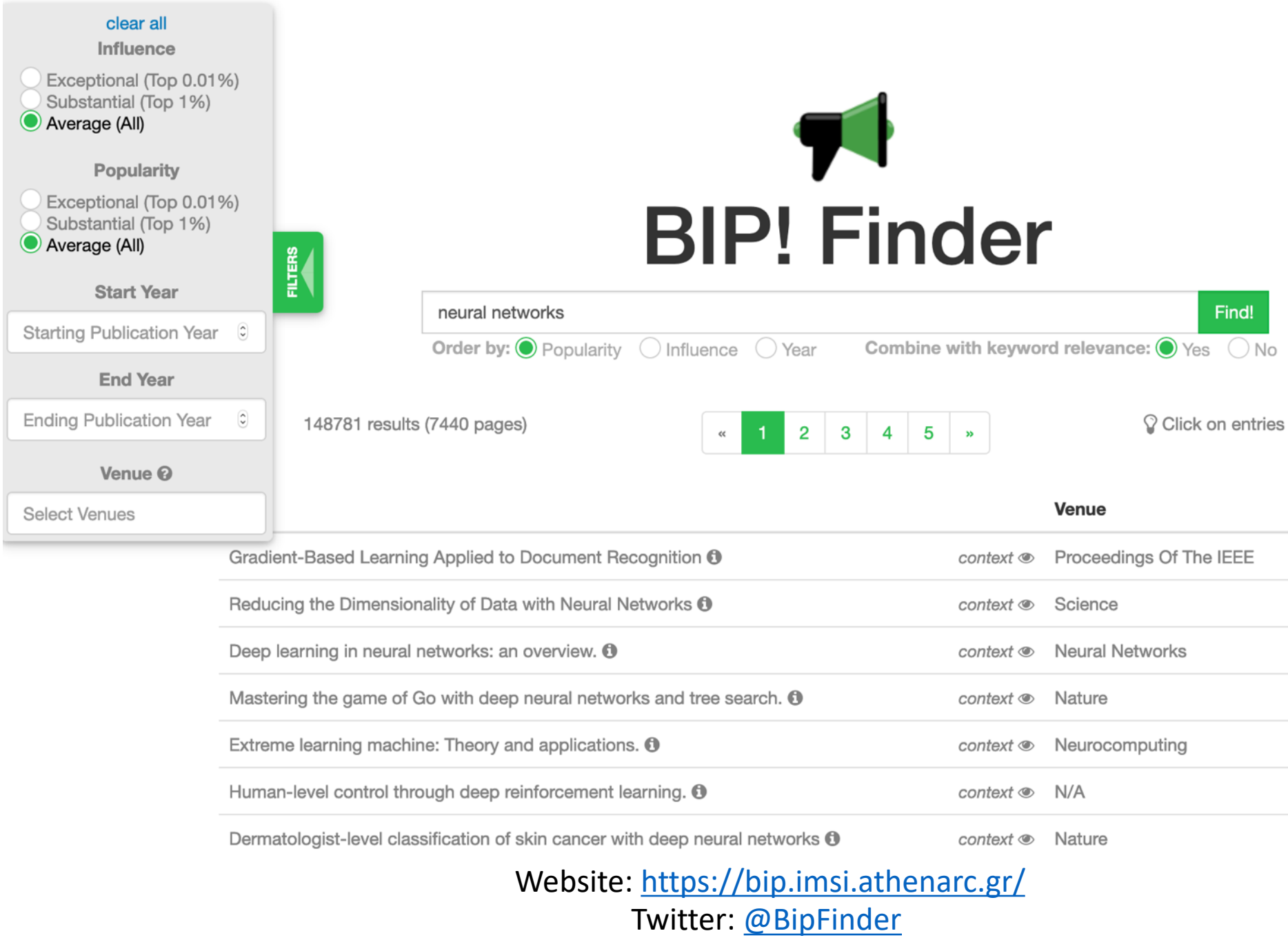


Existing approaches

23+1 methods

7 approaches

Method	Basic PR variants	Time Aware: Network Matrix	Time Aware: Landing Probability	Metadata: Venue	Metadata: Author	Multiple Networks	Ensemble	Other
Non-Linear PageRank (NPR) [50]	✓							
SPR [53]	✓							
SCEAS [41]	✓							
Focused PageRank [31]	✓							
Weighted Citation (WC) [49]		✓		✓				
Retained Adjacency Matrix (RAM) [19]		✓						
Timed PageRank [51,52]		✓		✓	✓			
Effective Contagion Matrix (ECM) [19]		✓						
NewRank (NR) [13]		✓	✓					
NTUWeightedPR [11]		✓	✓	✓	✓			
EWPR [34]		✓		✓	✓		✓	
CiteRank (CR) [45]			✓					
FutureRank (FR) [40]			✓		✓	✓		
YetRank (YR) [24]			✓	✓				
Wang et al. [47]			✓	✓	✓	✓		
PopRank [36]						✓		
MutualRank [27]						✓		
NTUTriPartite (WSDM) [17]				✓	✓	✓	✓	
NTUEnsemble [9]		✓	✓	✓	✓	✓	✓	
bletchleypark [22]		✓		✓	✓		✓	
ALEF [48]					✓		✓	
S-RCR [38]								✓
Citation Wake [29]								✓



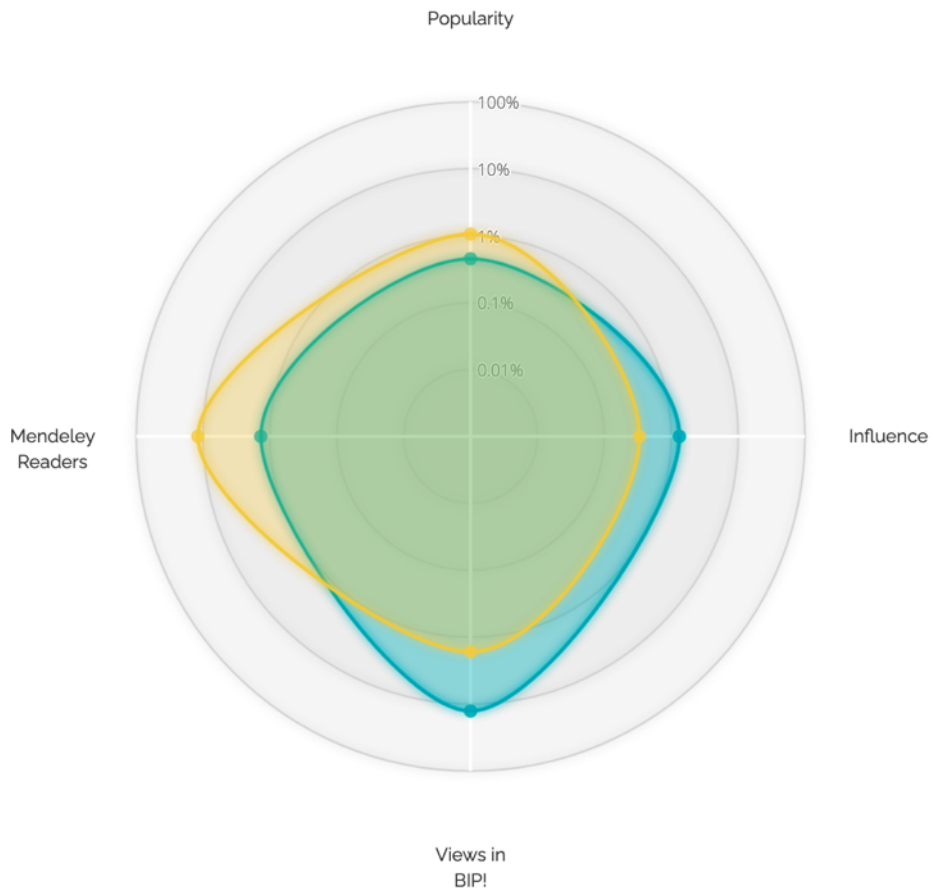
Website: https://bip.imsi.athenarc.gr/
Twitter: @BipFinder

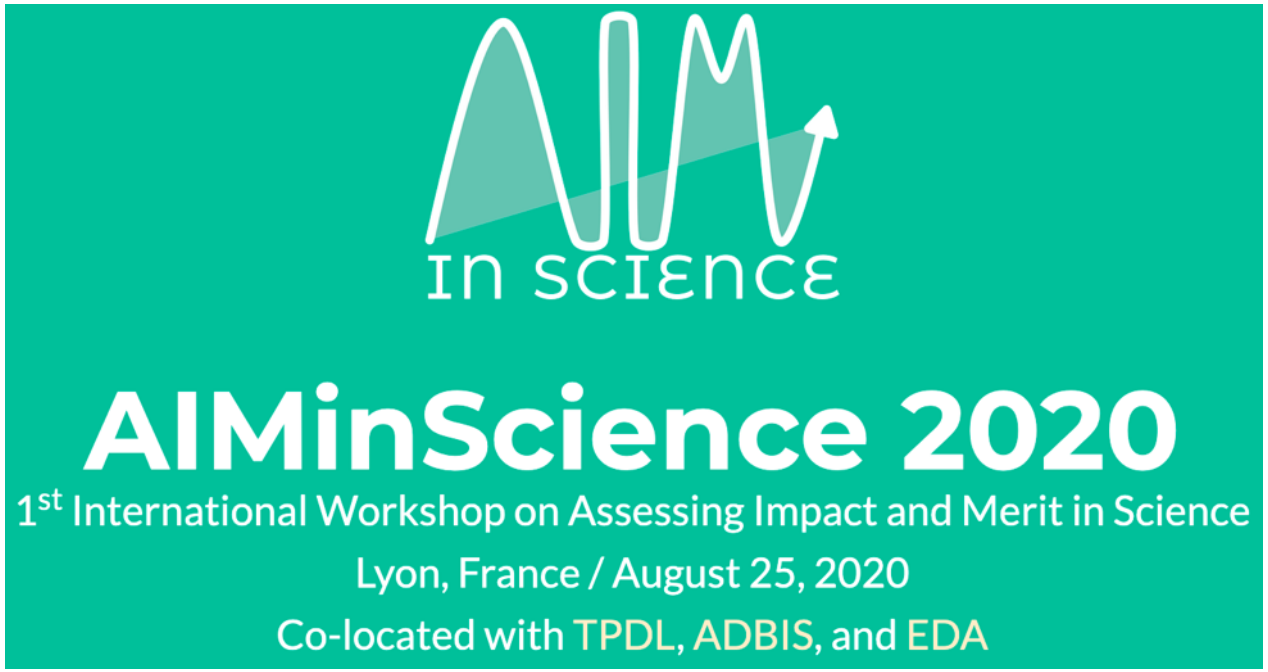
BIP! comparison

	Title	Venue	Year	Impact
●	Fast Pattern Matching in Strings	SIAM J Comput	1973	
●	BRIEF: binary robust independent elementary features	European Conference On Co (...)	2010	

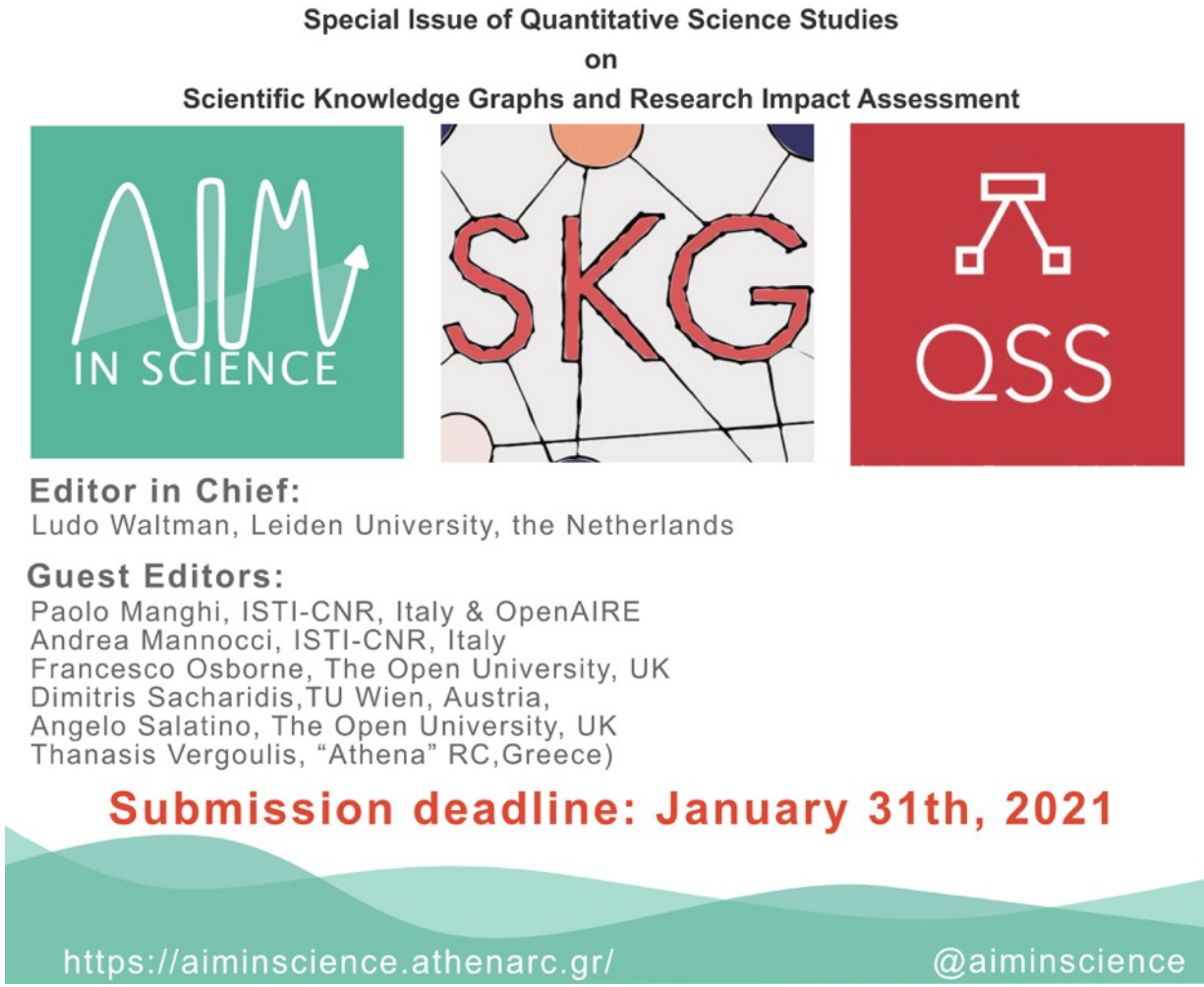
Impact aspects and other metrics

Citations per year





Website: https://aiminscience.athenarc.gr/
Twitter: @aiminscience



Call for Papers

Twitter: @vergoulis Email: vergoulis@athenarc.gr

References

[1] UNESCO Science Report: towards 2030. UNESCO Publishing, 2015

[2] SCImago, (n.d.). SJR — SCImago Journal & Country Rank [Portal]. Retrieved 19/11/2018, from http://www.scimagojr.com

[3] D. Fanelli. Do pressures to publish increase scientists' bias? an empirical support from us states data. PLOS ONE, 5(4):1–7, 04 2010.

[4] J. P. Ioannidis. Why most published research findings are false. PLoS Med, 2(8):e124, 2005.

[5] D. Sarewitz. The pressure to publish pushes down quality. Nature, 533(7602):147–147, 2016.

[6] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with googles pagerank algorithm. Journal of Informetrics, 1(1):8{15, 2007.

[7] W.-S. Hwang, S.-M. Chae, S.-W. Kim, and G. Woo. Yet another paper ranking algorithm advocating recent publications. In Proceedings of the 19th international conference on World wide web, pages 1117-1118. ACM, 2010.

[8] P. S. Yu, X. Li, and B. Liu. On the temporal dimension of search. In Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, pages 448-449. ACM, 2004.

[9] E. V. Bernstam, J. R. Herskovic, Y. Aphinyanaphongs, C. F. Aliferis, M. G. Sriram, and W. R. Hersh. Using citation data to improve retrieval from medline. Journal of the American Medical Informatics Association, 13(1):96{105, 2006.

[10] V. P. Diodato and P. Gellatly. Dictionary of Bibliometrics (Haworth Library and Information Science). Routledge, 1994.

[11] P. Groth and T. Gurney. Studying scientific discourse on the web using bibliometrics: A chemistry blogging case study. 2010.

[12] D. R. Smith. A 30-year citation analysis of bibliometric trends at the archives of environmental health, 1975{2004. Archives of environmental & occupational health, 64(sup1):43{54, 2009.

[13] J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute. A principal component analysis of 39 scientific impact measures. PloS one, 4(6):e6022, 2009.

[14] X. Bai, H. Liu, F. Zhang, Z. Ning, X. Kong, I. Lee, and F. Xia. An overview on evaluating and predicting scholarly article impact. Information, 8(3):73, 2017.