This talk bases on: Kacem, A., Flatt, J. W., & Mayr, P. (2020). Tracking self-citations in academic publishing. Scientometrics, 123(2), 1157–1165. https://doi.org/10.1007/s11192-020-03413-9

Analysing self-citations in a large bibliometric database

Philipp Mayr

ISKO UK, Challenges of scholarly communication: bibliometric transparency and impact

May 25, 2022

gesis Leibniz Institute for the Social Sciences

Agenda

- Introduction
- Related work
- Self-citation study (Kacem et al., 2020)
- Summary

This talk bases on: Kacem, A., Flatt, J. W., & Mayr, P. (2020). Tracking self-citations in academic publishing. Scientometrics, 123(2), 1157–1165. https://doi.org/10.1007/s11192-020-03413-9

Introduction

nature > news feature > article

a natureresearch jou

nature > news > article

a natureresearch

nature

nature

NEWS FEATURE · 19 AUGUST 2019

Hundreds of extreme se
revealed in new databas

Some highly cited academics seem to be
researchers warn against policing self-ci

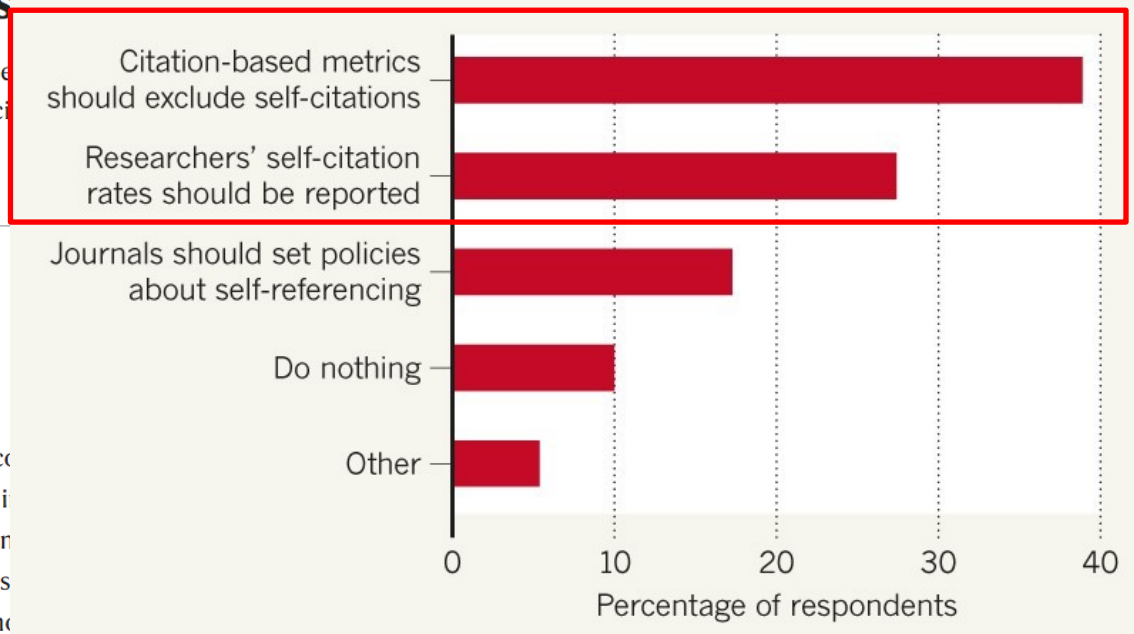
Richard Van Noorden & Dalmeet Singh Chawla

PDF version

The world's most-cited researchers, acc
eclectic bunch. Nobel laureates and emi
names, such as Sundarapandian Vaidyan
Vaidyanathan and hundreds of other res
come from their own papers, or from th

THE NUMBERS GAME

A *Nature* poll asked what (if anything) should be done to curb excessive self-citation. Respondents said that citation-based indicators are useful, but should be deployed in more nuanced and open ways.



Of 5,575 respondents, 2,183 said citation metrics such as the h-index should exclude self-citations; 1,541 said researchers' self-citation rates should be reported; 968 said journals should set policies about appropriate levels of self-referencing; 565 said to do nothing and 318 chose 'other'.

rom journal

peatedly suggested

ighly cited researchers has b
rred as a reviewer for another,
ass citations to his own work.

l Biology (JTB) announced in a
unnamed editor from the bo

https://www.nature.com/articles/d41586-019-02479-7

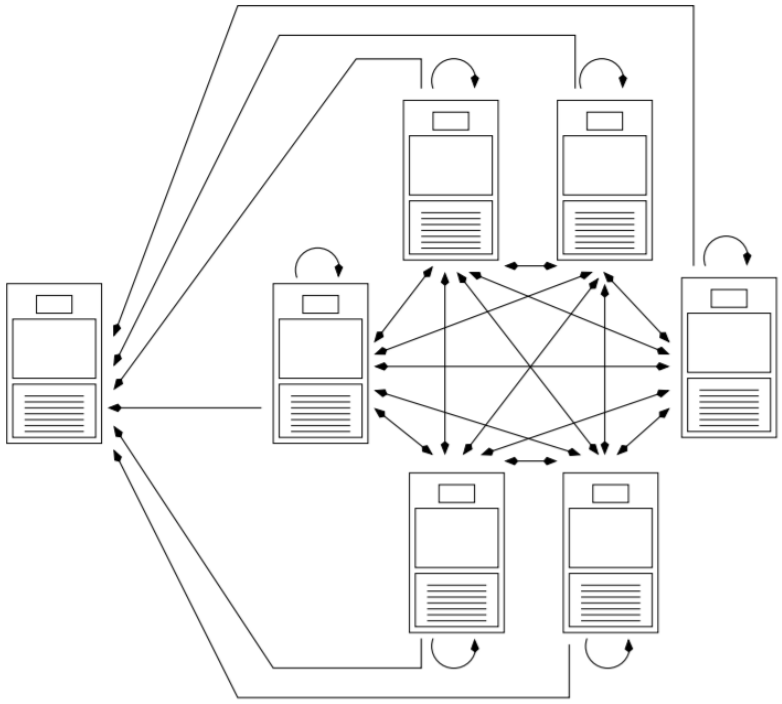
Introduction to self-citations

- **Excessive self-citation as a problem**
  - More and more researchers, papers and pressure
  - Scientific impact of papers as the most important factor
- **Quantitative measures** like h-index (Hirsch, 2005) for individuals
  - Oversimplified, „false precision“, one metric fits all
  - Often principles of good bibliometrics are ignored
  - Unintended effects: Gaming and abuse
- Momentum for: **Self-citation data should become more transparent, explainable, and accountable** (Flatt et al., 2017; Ioannidis et al., 2019)

What are self-citations?

- **A self-citation is any instance where a given author cites their own articles**
- When used appropriately, self-cites are equally important as cites from the surrounding community
  - Show how scholars build on their own work

Articles from **one** author: idealised example



Labbé (2010)

Related work

- Labbé (2010): **Gaming** of the h-index in Google Scholar. One invented researcher „*Ike Antkare*“ + 100 faked papers
- King et al. (2016): „Men cited their own papers 56 percent more than did women”
- Seeber et al. (2019): **Strategic increase** of self-citation (esp. Social science researchers) in the Italian Higher Education system
- Ioannidis et al. (2019): publicly available database of 100,000 top scientists in Scopus. “250 scientists have amassed more than 50% of their citations from themselves **or their co-authors**”
- Flatt et al. (2017): Introduction of a **simple metric (s-index)** to present the self-citation bahavior of a researcher in a transparent way

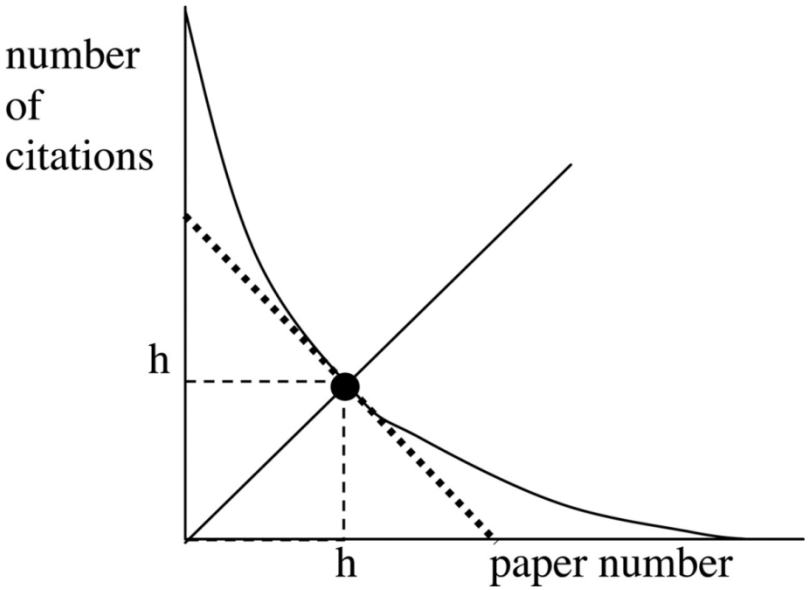
The self-citation index

- **Calculated similar to the *h*-index**

"A scientist has a self-citation index *s* equal to the total number of s papers that he or she has published that have at least the same amount of s self-citations." (Flatt et al., 2017)

- Goal: **elevate awareness** and subsequently direct more careful attention
- *s*-index brings **context**: Providing a simple **quantitative measure** of how much a given author has resorted to self-citation
- Each disciplinary community will adjust to an **acceptable use of self-citations**

H-index – S-index



Author 0001	paper nr	# of self-cites
	1	20
	2	10
	3	10
	4	10
	5	9
	6	8
	7	**7**
	8	5
	9	5
	10	5
	11	1
	12	0
	n	

S-Index = **7**

s-index

"Schematic curve of **number of citations** versus paper number, **with papers numbered in order of decreasing citations"**

Hirsch, 2005

Number of self-citations versus paper number, **with papers numbered in order of decreasing self-citations**

Self-citation Study

- 3-months BMBF project
- **Self-referencing behavior** among various academic disciplines and thousands of authors
- **Author-level tracking**: We consider here just author self-citations

Scientometrics (2020) 123:1157–1165
https://doi.org/10.1007/s11192-020-03413-9

Check for updates

Tracking self-citations in academic publishing

Ameni Kacem[1] · Justin W. Flatt[2,3] · Philipp Mayr[1]

Received: 26 February 2020 / Published online: 18 March 2020

Abstract
Citation metrics have value because they aim to make scientific assessment a level playing field, but urgent transparency-based adjustments are necessary to ensure that measurements yield the most accurate picture of impact and excellence. One problematic area is the handling of self-citations, which are either excluded or inappropriately accounted for when using bibliometric indicators for research evaluation. Here, in favor of openly tracking self-citations we report on self-referencing behavior among various academic disciplines as captured by the curated Clarivate Analytics Web of Science database. Specifically, we examined the behavior of 385,616 authors grouped into 15 subject areas like Biology, Chemistry, Science and Technology, Engineering, and Physics. These authors have published 3,240,973 papers that have accumulated 90,806,462 citations, roughly five percent of which are self-citations. Up until now, very little is known about the buildup of self-citations at the author-level and in field-specific contexts. Our view is that hiding self-citation data is indefensible and needlessly confuses any attempts to understand the bibliometric impact of one's work. Instead we urge academics to embrace visibility of citation data in a community of peers, which relies on nuance and openness rather than curated scorekeeping.

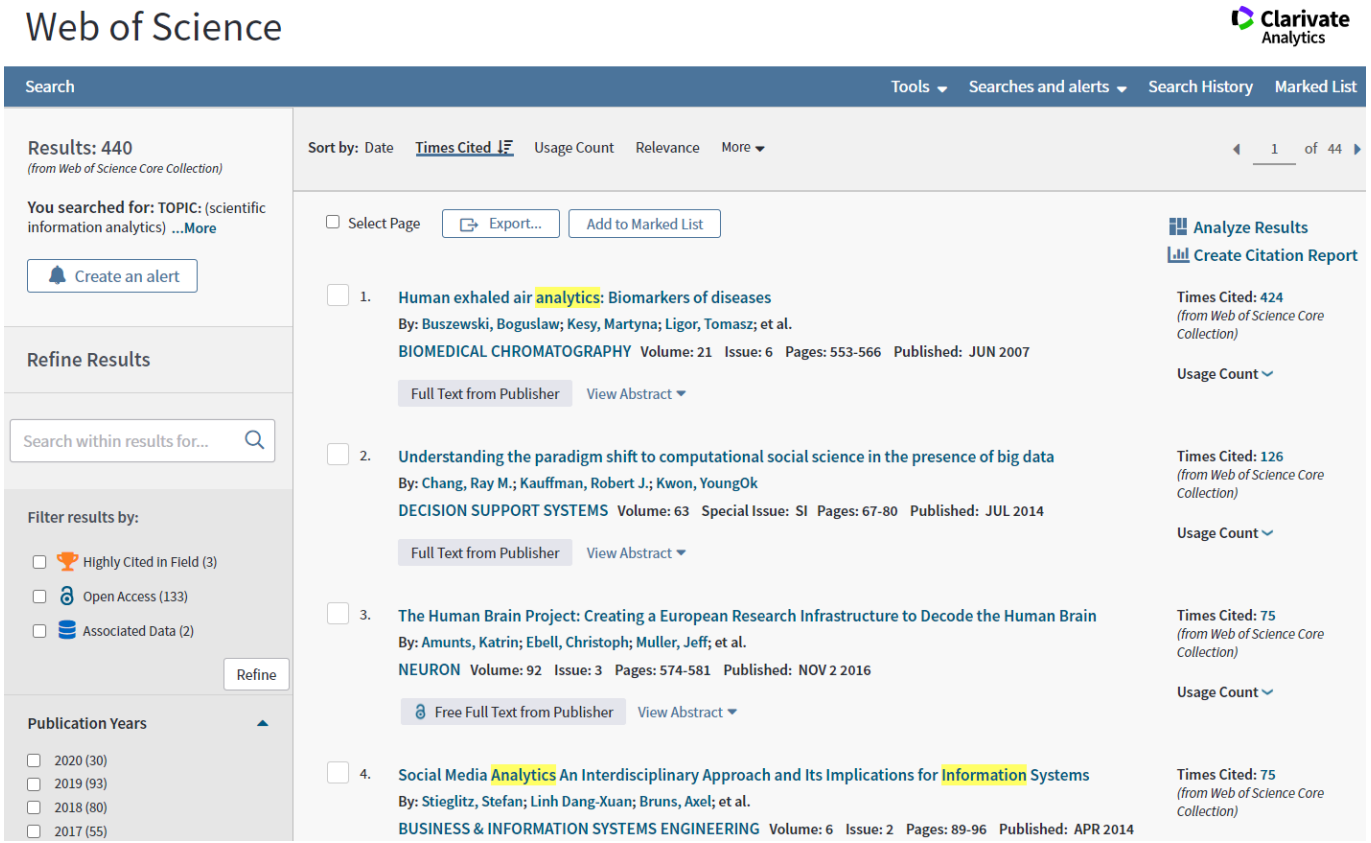
https://doi.org/10.1007/s11192-020-03413-9

Goal of the study

- Demonstrate an **easy way to track self-cites without distorting other metrics**, namely the h-index
  - not meant to criminalize self-referencing
  - we do not intend to suggest a certain threshold of acceptable behavior
- Provide a **tool** to clarify how researchers build on their own ideas; tool to find excessive behaviour

Database: Web of Science

- Web of Science (WoS) database (2016) by Clarivate Analytics
- 50,040,717 records (journal articles) for a period of publishing from 1965 to 2016

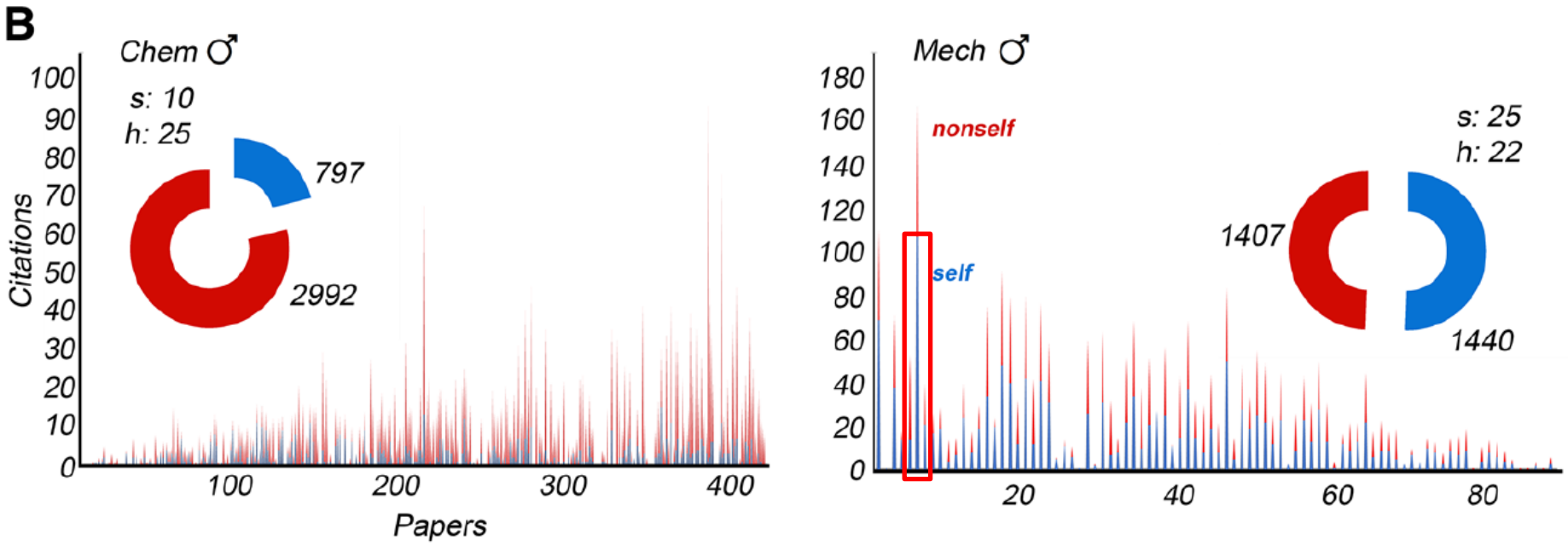


Approach

- Only authors with a **unique identifier**
  - ORCID  or ResearcherID
- **Curated version** of WoS with all author-self-citations detected
- Authors were categorized into **major 15 subject areas** (WoS subject categorization scheme) e.g., Physics, Chemistry, Biology, …
- We analyzed categories if they contained at **least 1000 unique authors**
- Citation queries were run using the Oracle SQL language
- We have computed ***s*-index scores for all authors**

Two examples

- Examples of two author profiles where citations have been clarified to separate self-(**blue**) and nonself-citations (**red**)
- Papers are sorted from **oldest to most recent**. Gender (manually verified), category, *s*-index, and *h*-index (excluding self-citations)



Results: Overview

A

Categories	Authors With self-citations	Authors Without self-citations	Papers	Self-citations	Nonself-citations	Total citations	% Self	% Nonself
Science &Technology	112379	52179	60200	933513	1152425	29199011	30351436	3.80	96.20
Physics	41640	25320	16320	629561	1037709	14815314	15853023	6.55	93.45
Chemistry	25725	20926	4799	552626	1072511	14653868	15726379	6.82	93.18
Engineering	52034	14424	37610	215179	198288	2799801	2998089	6.62	93.38
Biology	45717	17139	28578	243633	390591	10156418	10547009	3.70	96.30
Materials Science	32656	10148	22508	181476	231101	3487300	3718401	6.22	93.78
Neurosciences	7148	3879	3269	57217	85335	2240384	2325719	3.67	96.33
Computer Science	12327	5149	7178	85781	54281	831563	885844	6.14	93.86
Agriculture	11116	3520	7596	45766	46408	794825	841233	5.52	94.48
Geology	2856	1734	1122	26604	35269	695198	730467	4.83	95.17
Mathematics	10357	3935	6422	56351	66814	736364	803178	8.33	91.67
Ecology	16860	7977	8883	127162	181936	3567613	3749549	4.85	95.15
Psychology	3491	2450	1041	46472	62599	1349709	1412308	4.43	95.57
Economics	5643	1895	3748	20246	14251	504973	519224	2.75	97.25
Mechanics	5667	1693	3974	19386	23682	320921	344603	6.88	93.12
Total	385616	172368	213248	3240973	4653200	86153262	90806462	5.12	94.88

- 98% (377,533) of authors achieve an ***s*-index of 5 or less**
- 1.9% (7561) achieve a score between 6 and 10,
- **Only 0.1% (522) exceed 10**
- 213,248 authors have no self-citation at all

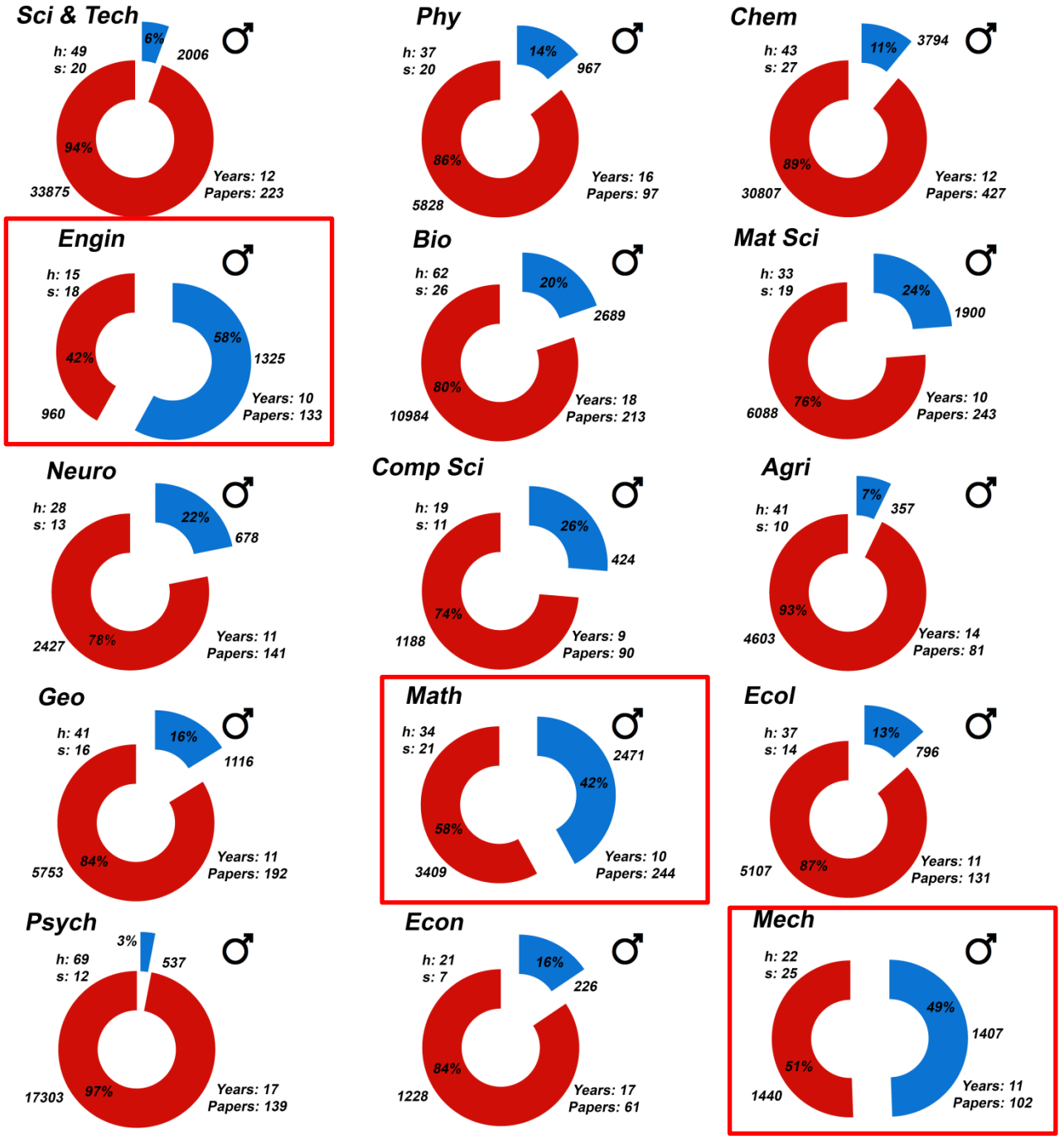
Results: top scorer identification

Self-citation Index	Chem	Bio	Mech	Math	Sci & Tech	Phy	Mat Sci	Engin	Geo	Ecol	Neuro	Psych	Comp Sci	Agri	Econ
0	4799	28578	3974	6422	60200	16320	22508	37610	1122	8883	3269	1041	7178	7596	3748
1	6047	8559	1012	1939	26771	8522	5049	8732	722	3601	1730	1027	3233	2054	1298
2	5204	4151	370	1142	12725	6477	2585	3444	479	2078	1005	624	1311	813	399
3	3581	2044	179	487	6249	4265	1236	1347	284	1129	545	371	389	377	143
4	2316	1075	83	212	3000	2676	616	564	138	565	266	220	130	156	34
5	1434	556	25	87	1615	1557	302	234	67	316	160	99	58	73	14
6	872	320	6	30	825	849	173	95	21	123	83	52	15	23	5
7	540	184	10	19	460	444	80	40	13	80	48	28	5	17	2
8	346	92	4	10	240	255	45	22	4	52	22	15	4	4	0
9	195	62	1	4	127	120	26	16	4	13	13	6	1	2	0
10	155	41	1	1	70	82	17	10	0	8	3	4	1	1	0
11	73	20	0	1	39	32	6	8	0	5	2	1	2	0	0
12	56	14	0	0	23	15	5	0	0	4	0	3	0	0	0
13	39	7	0	0	16	10	3	1	1	2	2	0	0	0	0
14	20	3	0	0	8	4	2	1	0	1	0	0	0	0	0
15	15	4	0	1	6	4	0	0	0	0	0	0	0	0	0
16	8	2	1	0	1	4	1	0	1	0	0	0	0	0	0
17	6	2	0	0	0	0	1	1	0	0	0	0	0	0	0
18	7	0	0	1	2	0	0	1	0	0	0	0	0	0	0
19	2	1	0	0	0	3	1	0	0	0	0	0	0	0	0
20	4	0	0	0	2	1	0	0	0	0	0	0	0	0	0
21	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0
22	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
25	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
26	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
27	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

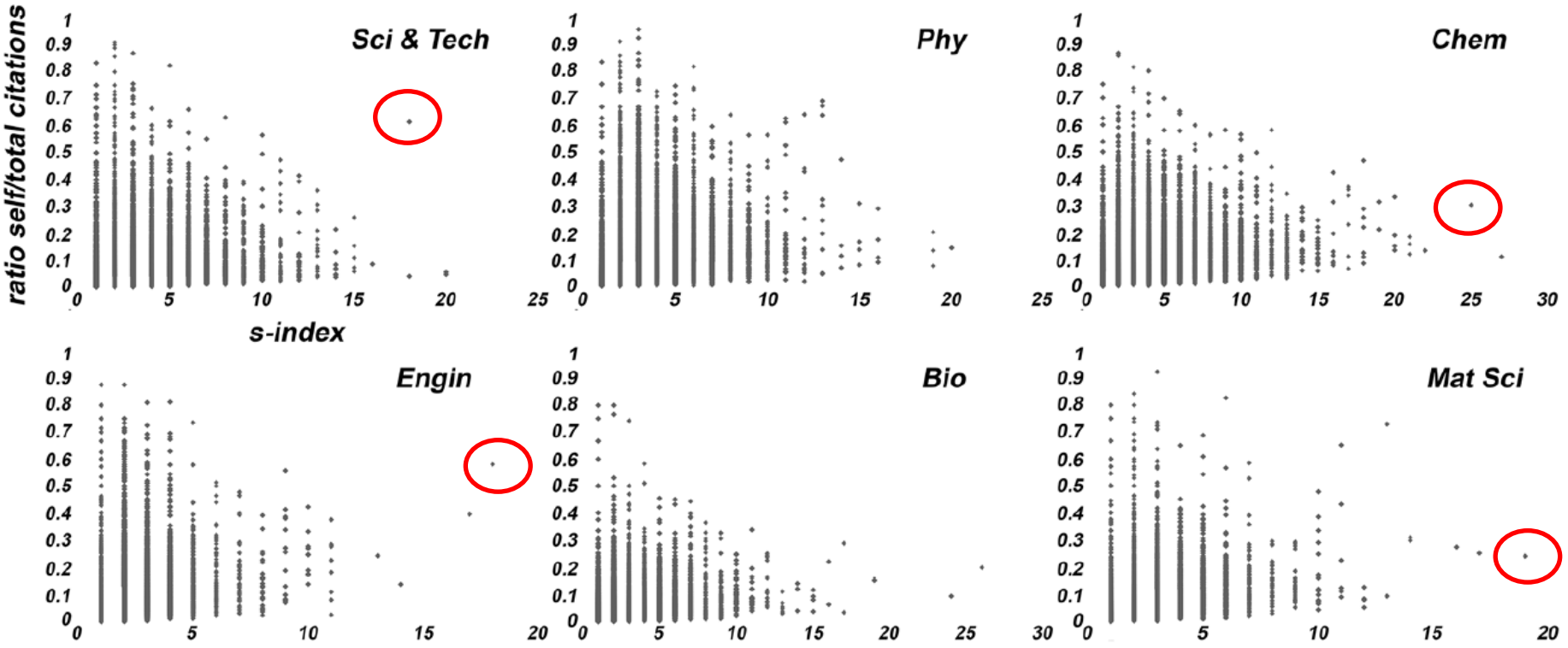
- Highest scores are highlighted in blue (Chem: 27, Comp Sci: 11)
- **High scorers**: the research categories most represented were Chemistry (236), Science and Technology (97), Physics (73), and Biology (55)

High scorer per discipline

- All **male**
- **beyond the "early"** phase of their careers
- **Productive** in terms of paper output and citations



Results: self/total citation and s-index



The ratio of self/total citation as a function of increasing s-index

Attention: **when an author has a high ratio of self/total citations along with a high s-index**

Limitations

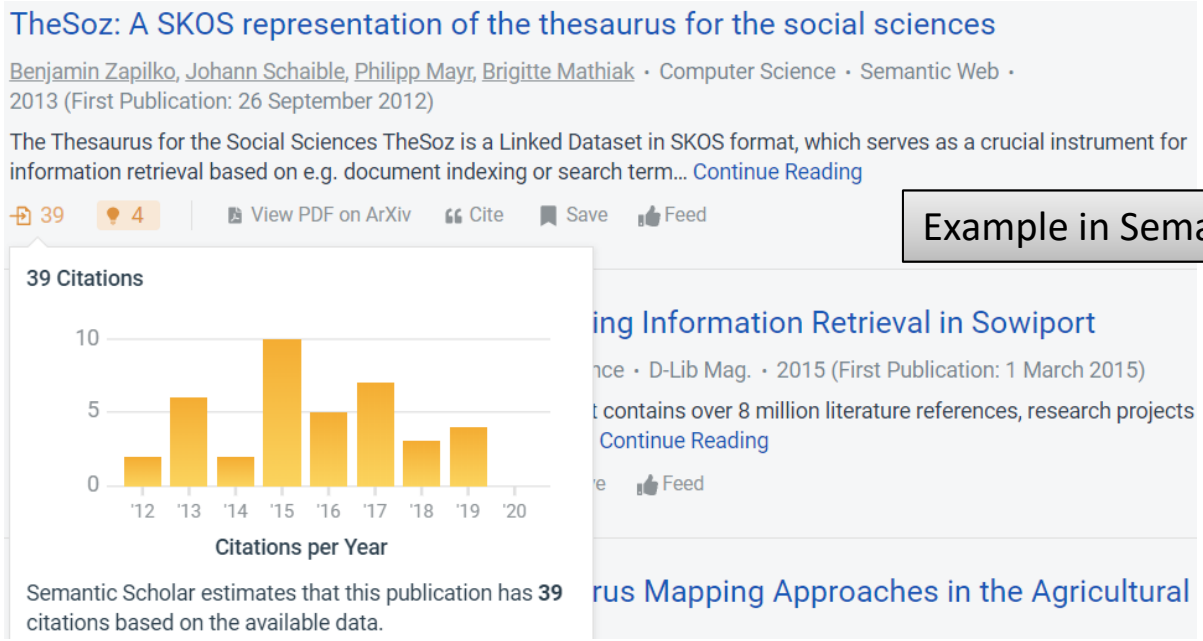
- The study bases on **authors with identifiers** (ORCID & ResearcherID) **only**
  - 9 million ORCIDs published, but still not standard
  - Often author identifiers are not assigned to papers
- **Just journal literature** in the highly-selective WoS database,
  - no proceedings
  - no other publications
  - not all disciplines (e.g. computer science) are fully coverered
- **Closed** dataset

Concluding the study

- Excessive self-citation cannot be eliminated, **but showing citation buildup for everyone will help to illuminate** (bibliometric footprint)
  - We have shown here how to account for self-citation **without introducing distortions**
- **Single global measures** like e.g., total citation counts, *h*-index or *s*-index are **not enough**
- Current obstacle is that self-citation information is currently **not freely accessible**, making large-scale studies problematic
  - Initiative for Open Citations (I4OC)
- Good news: Excessive self-citation is done **just by very few**, at least in more controlled settings, like WoS

Future work

"Future efforts centered on clarifying citations will **better inform** policymakers, funding agencies, hiring/promotion/award committees, and the general public about the value of published research."



Example in Semantic Scholar

Why is this interesting?

- Relevant **technical problems**:
  - Author name disambiguation: develop scalable approaches to disambiguate author names
  - Identify citation farms (e.g. like Ioannidis et al., 2019)
  - Improve field delineation to facilitate reporting beyond whole fields e.g. Physics or Computer Science
  - Include full text of the papers (e.g. like Semantic Scholar)
- **More elaborate and combined measures are needed**, include
  - Field-normalized data (see e.g. Leiden Manifesto)
  - Temporal aspects (e.g. to reduce gaming)
- **Produce and work on Open Datasets**
  - Open Citations/Crossref
  - Microsoft Academic
  - Arxiv, PubMedCentral
  - …

This talk bases on: Kacem, A., Flatt, J. W., & Mayr, P. (2020). Tracking self-citations in academic publishing. Scientometrics, 123(2), 1157–1165. https://doi.org/10.1007/s11192-020-03413-9

Thank you for your attention!

Contact:

https://philippmayr.github.io/

philipp.mayr@gesis.org

gesis Leibniz Institute for the Social Sciences