Open Universiteit

No papers, please!

Fighting scientific fraud by looking at people, not papers

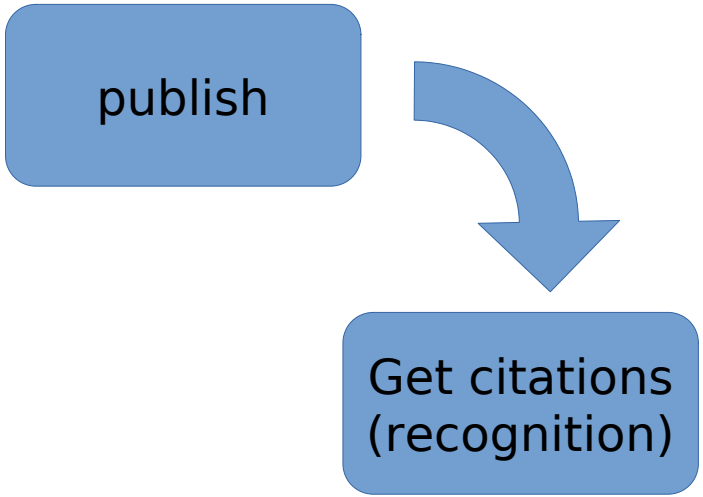
Hugo Jonker, Ewoud Westerbaan

How to be an academic rockstar

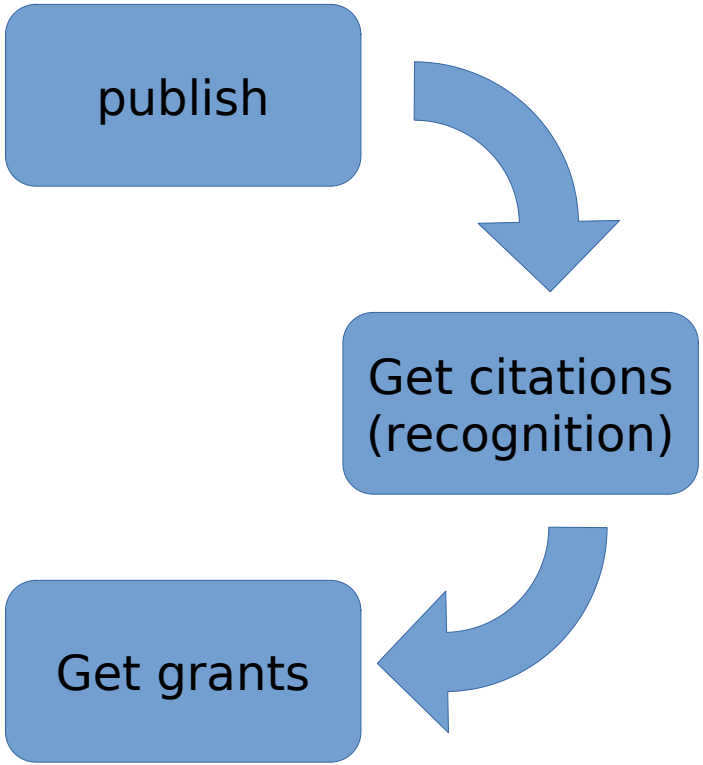
How to be an academic rockstar

How to be an academic rockstar

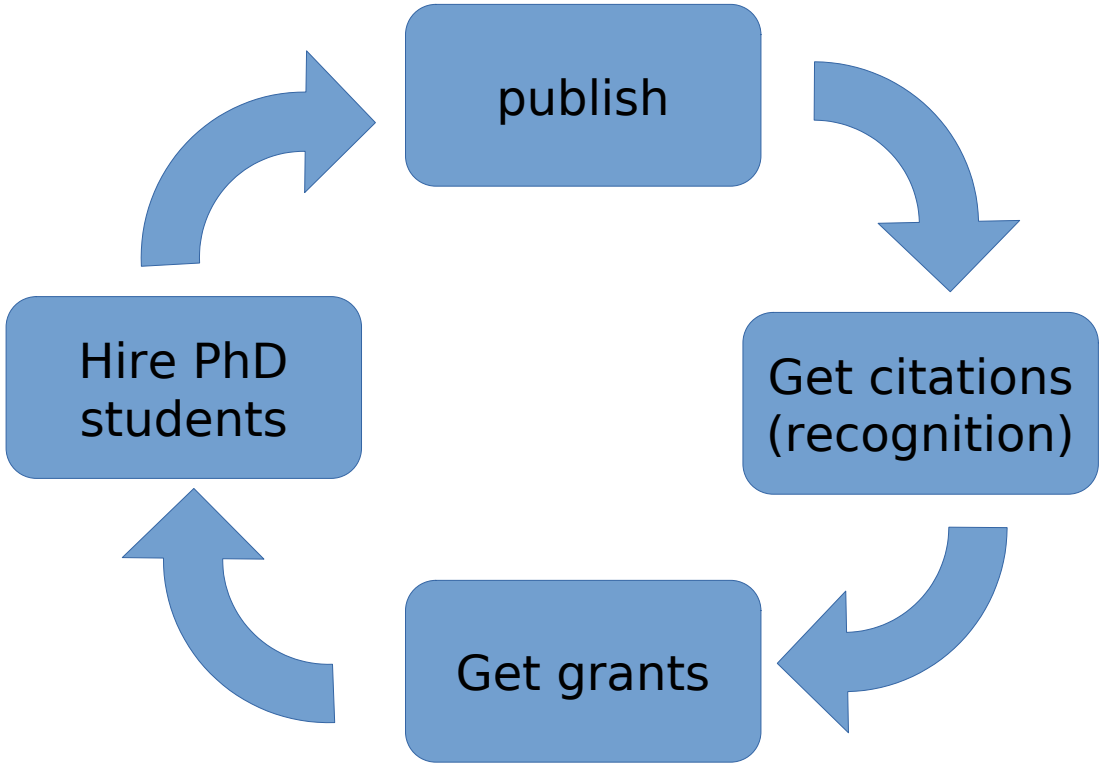
How to be an academic rockstar



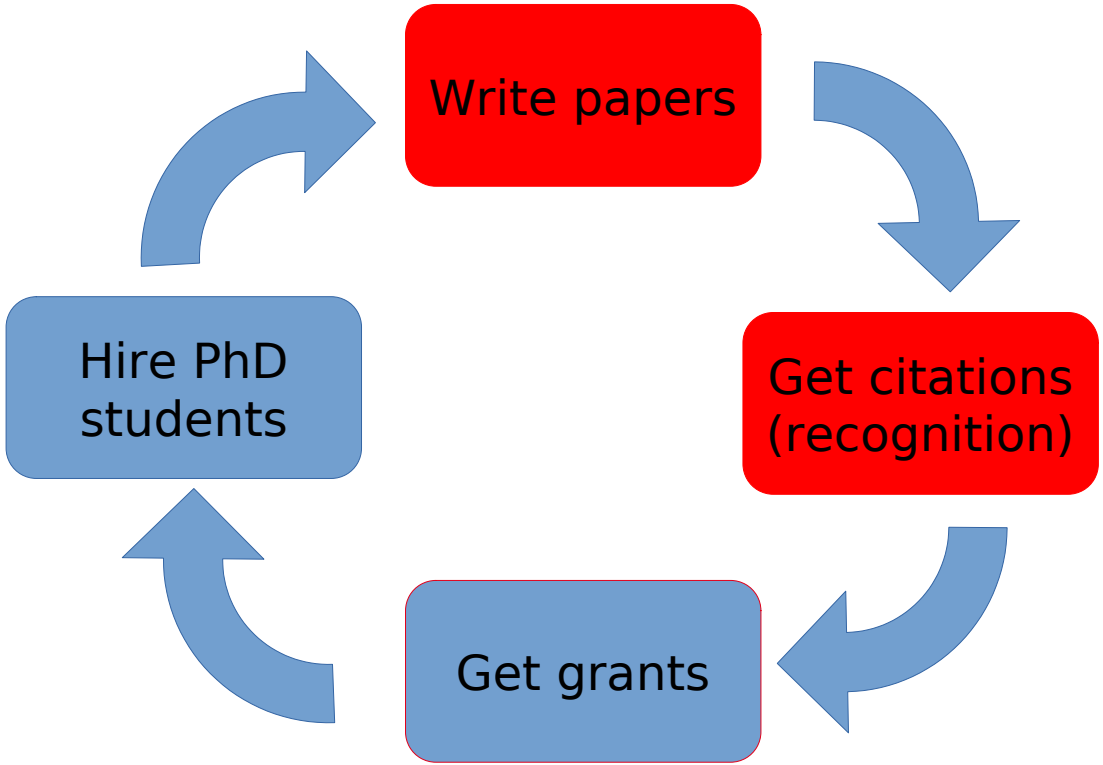
How to be an academic rockstar



How to be an academic rockstar



Cheats (see also [SPW17])



Existing defenses

Existing defenses

- Pre-registration
  - Prevents p-hacking

Existing defenses

- Pre-registration
  - Prevents p-hacking
- Plagiarism checkers
  - Detect copied texts

Existing defenses

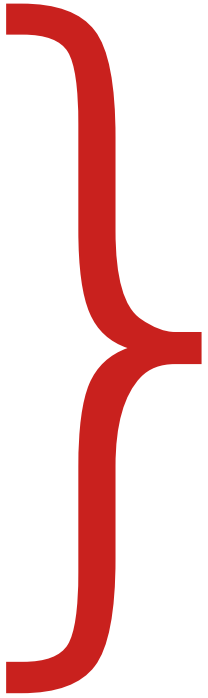
- Pre-registration
  - Prevents p-hacking
- Plagiarism checkers
  - Detect copied texts
- Image checkers
  - Detect manipulated images

Existing defenses

- Pre-registration
  - Prevents p-hacking
- Plagiarism checkers
  - Detect copied texts
- Image checkers
  - Detect manipulated images
- Statistical evaluation tools
  - Detect faked data

Existing defenses

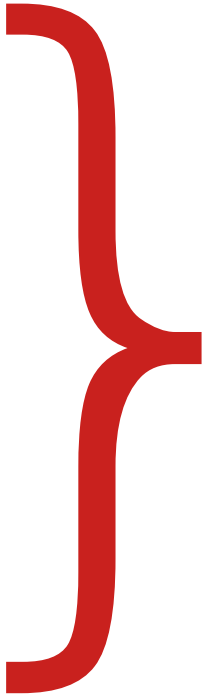
- Pre-registration
  - Prevents p-hacking
- Plagiarism checkers
  - Detect copied texts
- Image checkers
  - Detect manipulated images
- Statistical evaluation tools
  - Detect faked data



Focused on article...

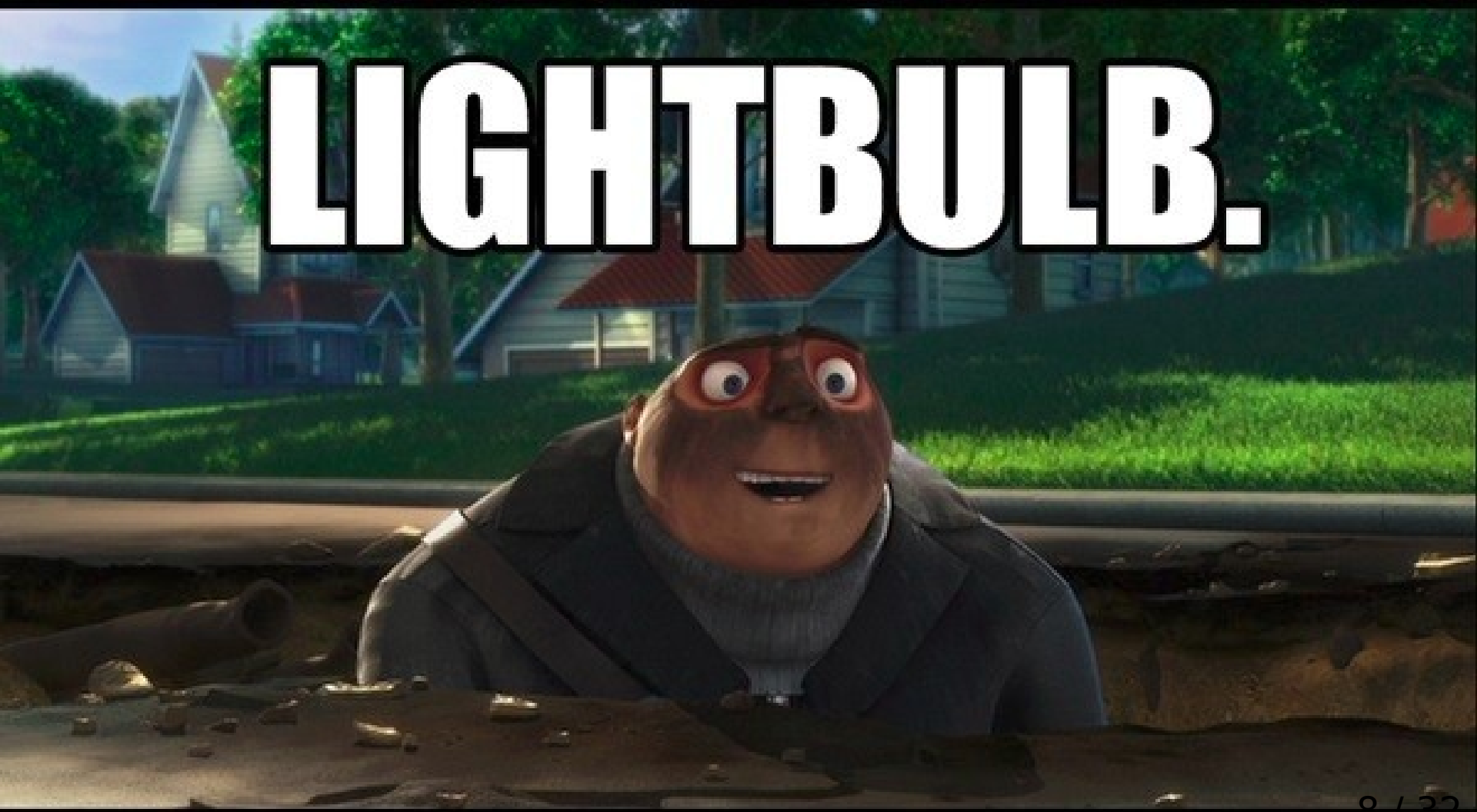
Existing defenses

- Pre-registration
  - Prevents p-hacking
- Plagiarism checkers
  - Detect copied texts
- Image checkers
  - Detect manipulated images
- Statistical evaluation tools
  - Detect faked data



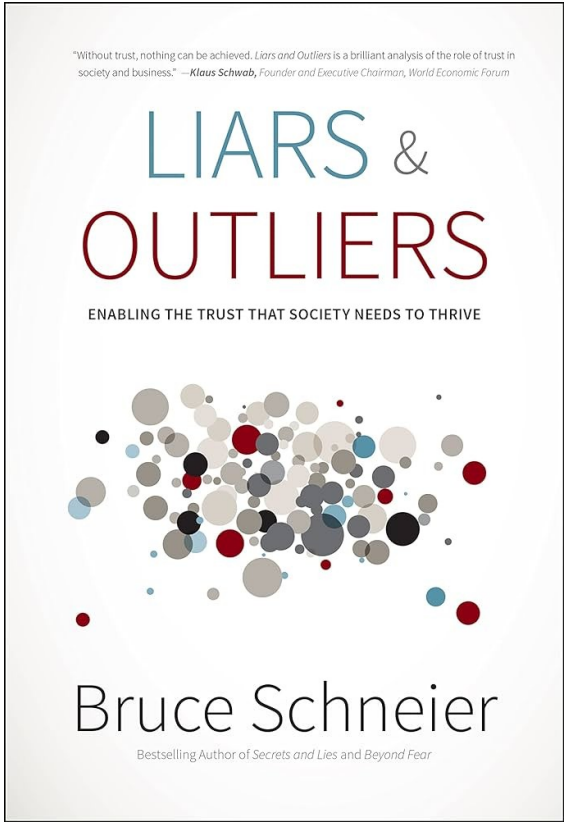
Focused on article...

but articles don’t commit fraud

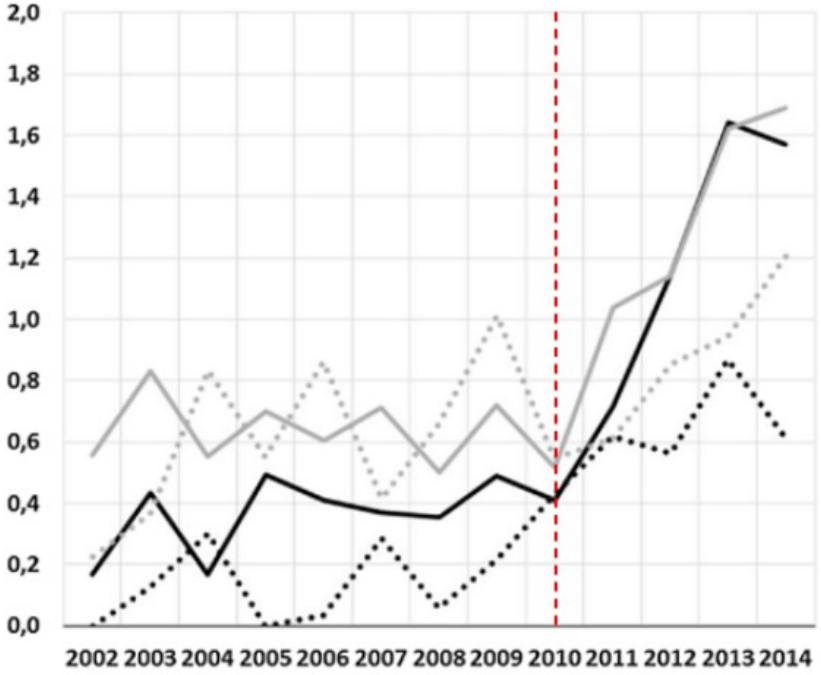

LIGHTBULB.

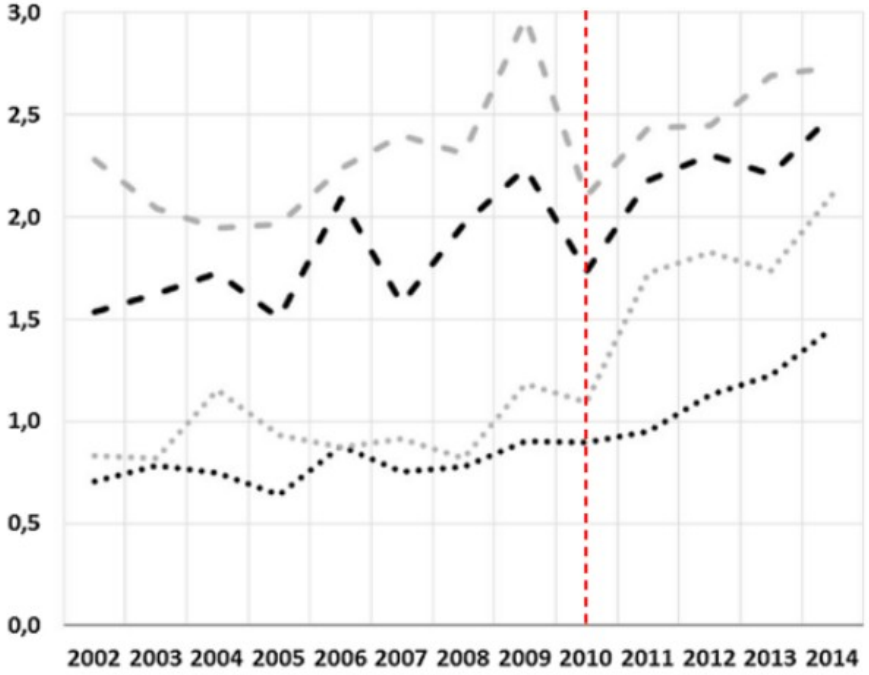
Detecting cheatERS instead of cheatS

- Idea: finding outliers in metrics
- Challenges
  - Correlation does not equal causation
    That is:
    - ¬outlier ↛ ¬fraudster
    - outlier ↛ fraudster
- Goal: focus manual investigation capacity



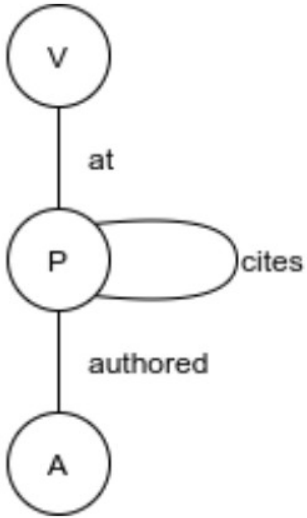
Example: Italy, 2010





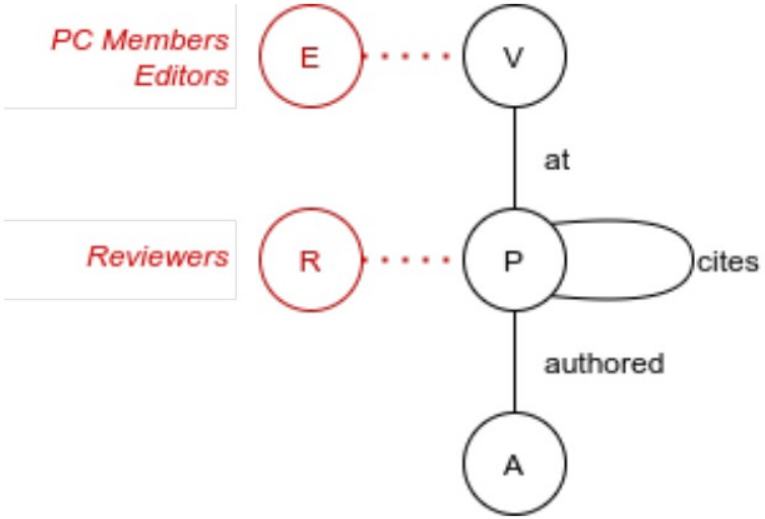
Finding outliers in publication metrics

- From [SPW17]:



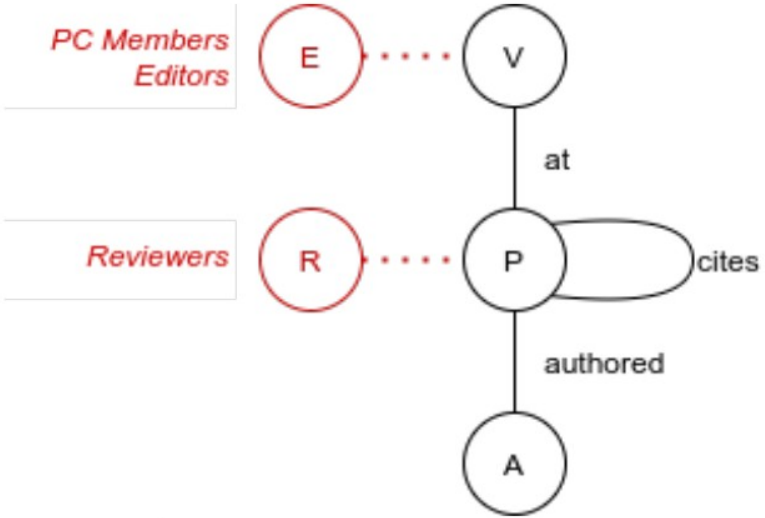
Case #1: reviewers with high citations

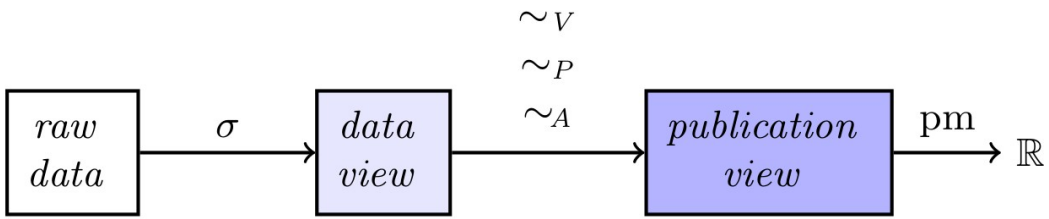
- From [SPW17]:
- Extending with editors, reviewers:



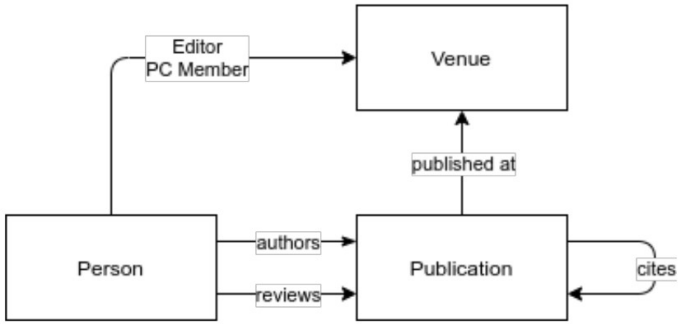
Case #1: reviewers with high citations

- From [SPW17]:
- Extending with editors, reviewers:
- Fill with data:



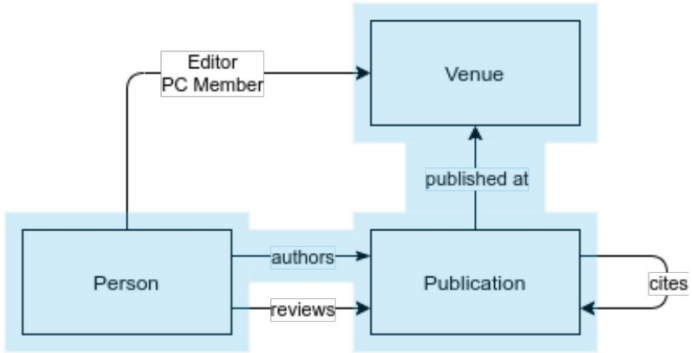


Getting data



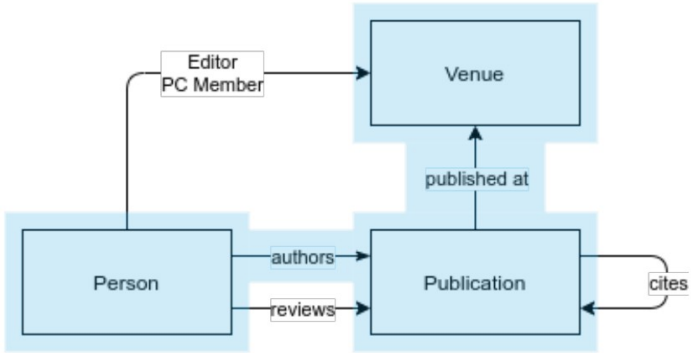
Desired data

Getting data

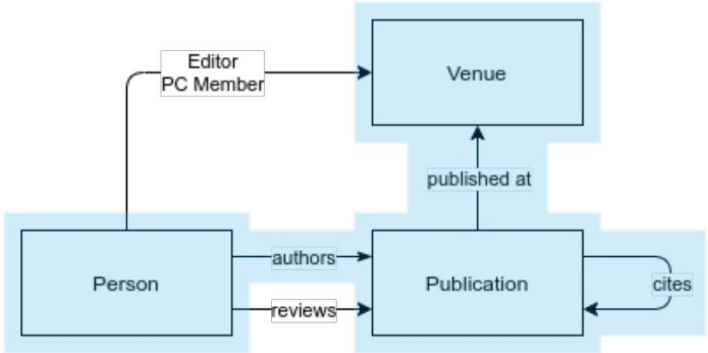


DBLP

Getting data

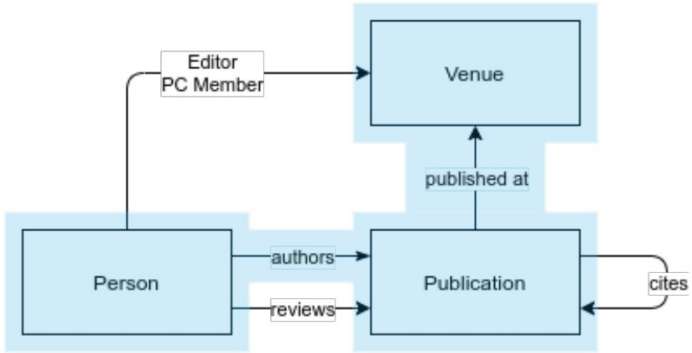


DBLP

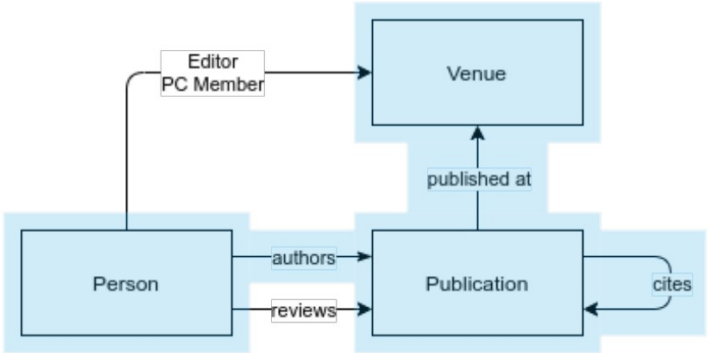


Aminer

Getting data



DBLP



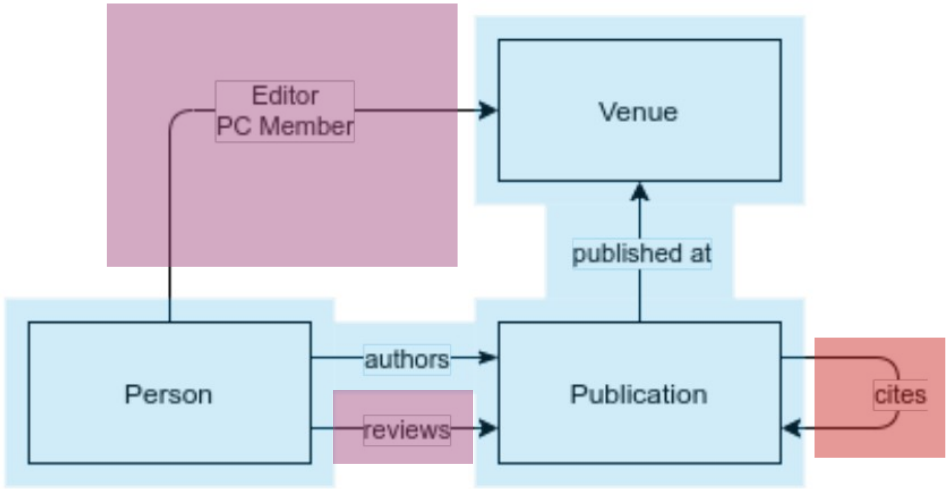
Aminer

Low quality :(

Augmenting DBLP data

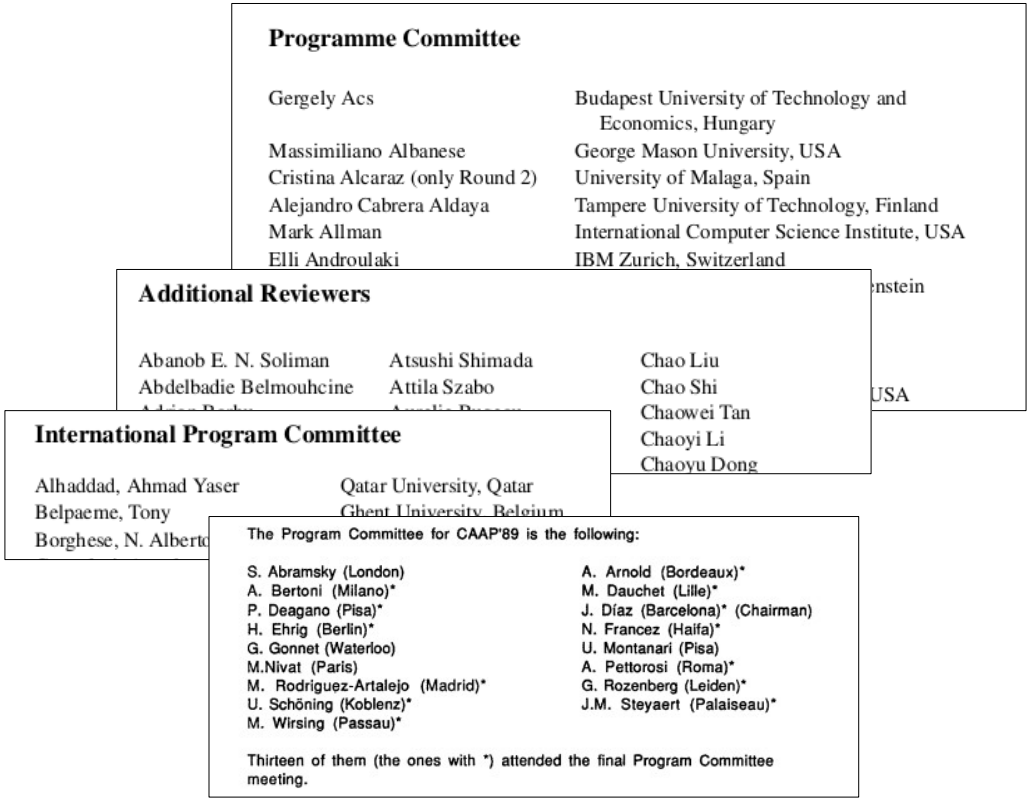
- OpenCitation database

- Scraping the web
- Scraping PDFs



Extracting info from PDF

- Layout varies
- Location may be important
- Affiliation is important



Programme Committee

Gergely Acs	Budapest University of Technology and Economics, Hungary
Massimiliano Albanese	George Mason University, USA
Cristina Alcaraz (only Round 2)	University of Malaga, Spain
Alejandro Cabrera Aldaya	Tampere University of Technology, Finland
Mark Allman	International Computer Science Institute, USA
Elli Androulaki	IBM Zurich, Switzerland

Additional Reviewers

Abanob E. N. Soliman	Atsushi Shimada	Chao Liu
Abdelbadie Belmouhcine	Attila Szabo	Chao Shi
		Chaowei Tan
		Chaoyi Li
		Chaoyu Dong

International Program Committee

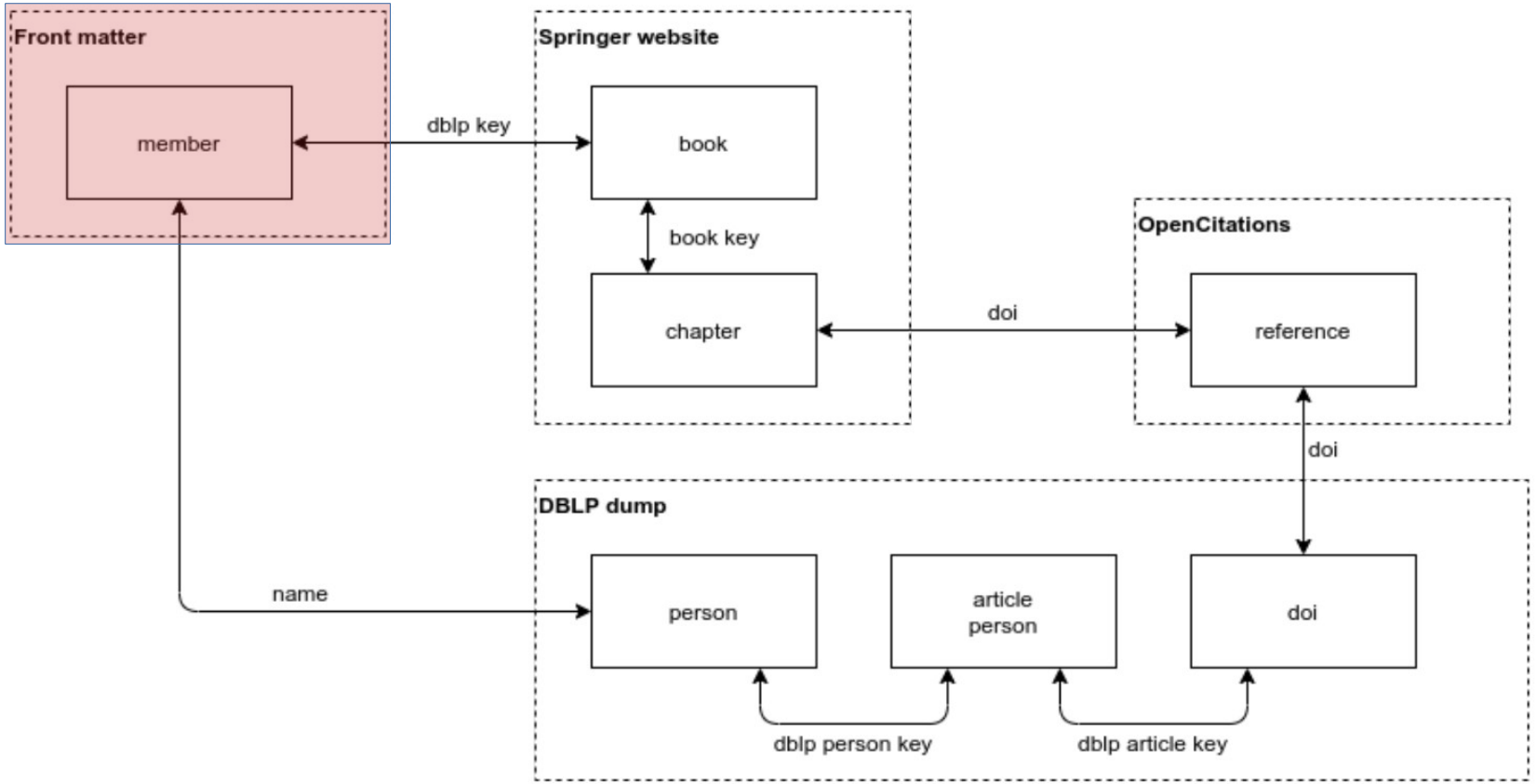
Alhaddad, Ahmad Yaser	Qatar University, Qatar
Belpaeme, Tony	Ghent University, Belgium
Borghese, N. Alberto	

The Program Committee for CAAP'89 is the following:

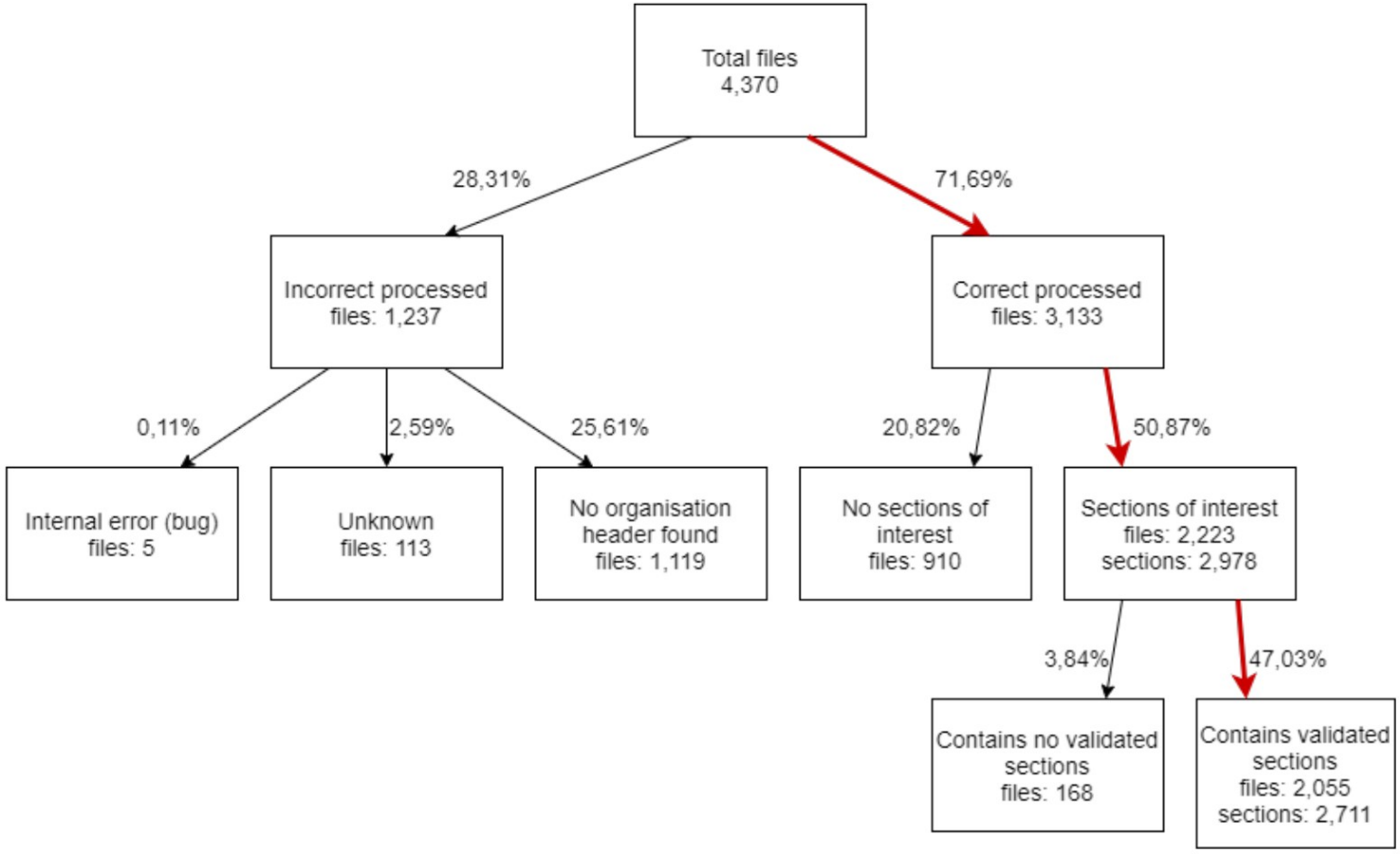
S. Abramsky (London)	A. Arnold (Bordeaux)*
A. Bertoni (Milano)*	M. Dauchet (Lille)*
P. Deagano (Pisa)*	J. Díaz (Barcelona)* (Chairman)
H. Ehrig (Berlin)*	N. Francez (Haifa)*
G. Gonnet (Waterloo)	U. Montanari (Pisa)
M.Nivat (Paris)	A. Pettorosi (Roma)*
M. Rodriguez-Artalejo (Madrid)*	G. Rozenberg (Leiden)*
U. Schöning (Koblenz)*	J.M. Steyaert (Palaiseau)*
M. Wirsing (Passau)*	

Thirteen of them (the ones with *) attended the final Program Committee meeting.

Data acquisition + integration

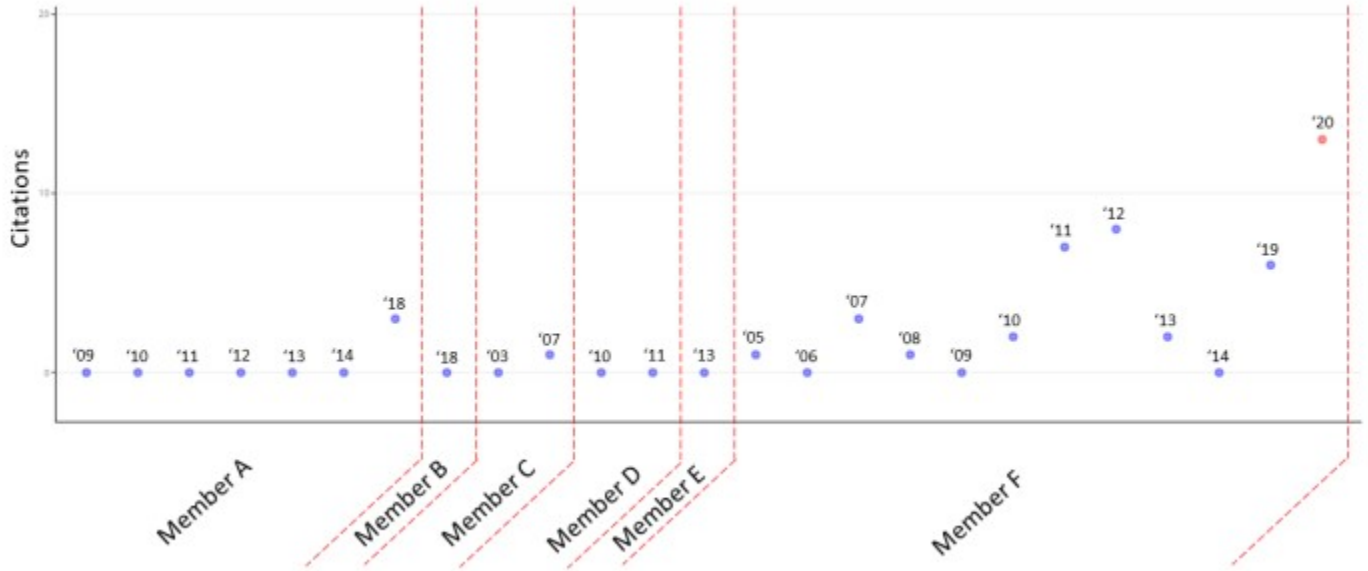


Extraction results



Does it work?

We have data! For example:



SPIRE conference, 27 years

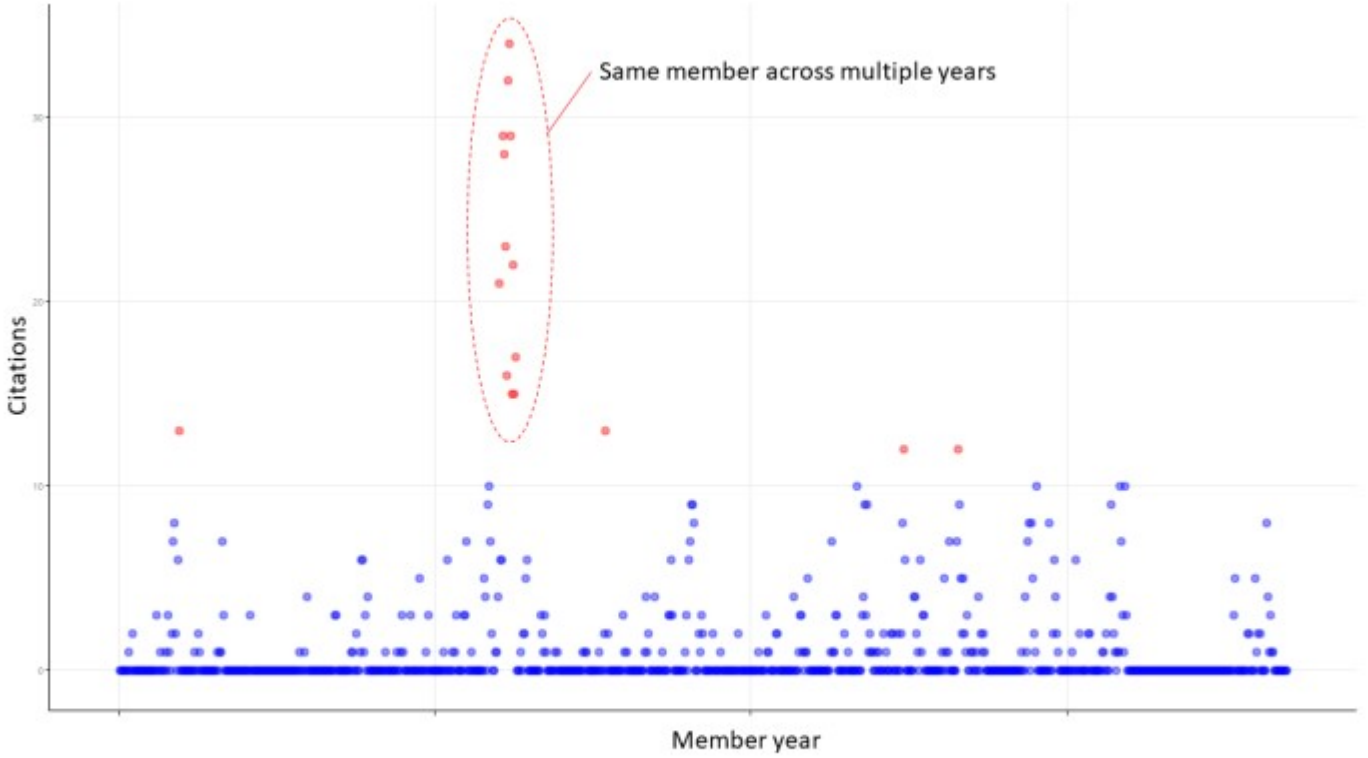
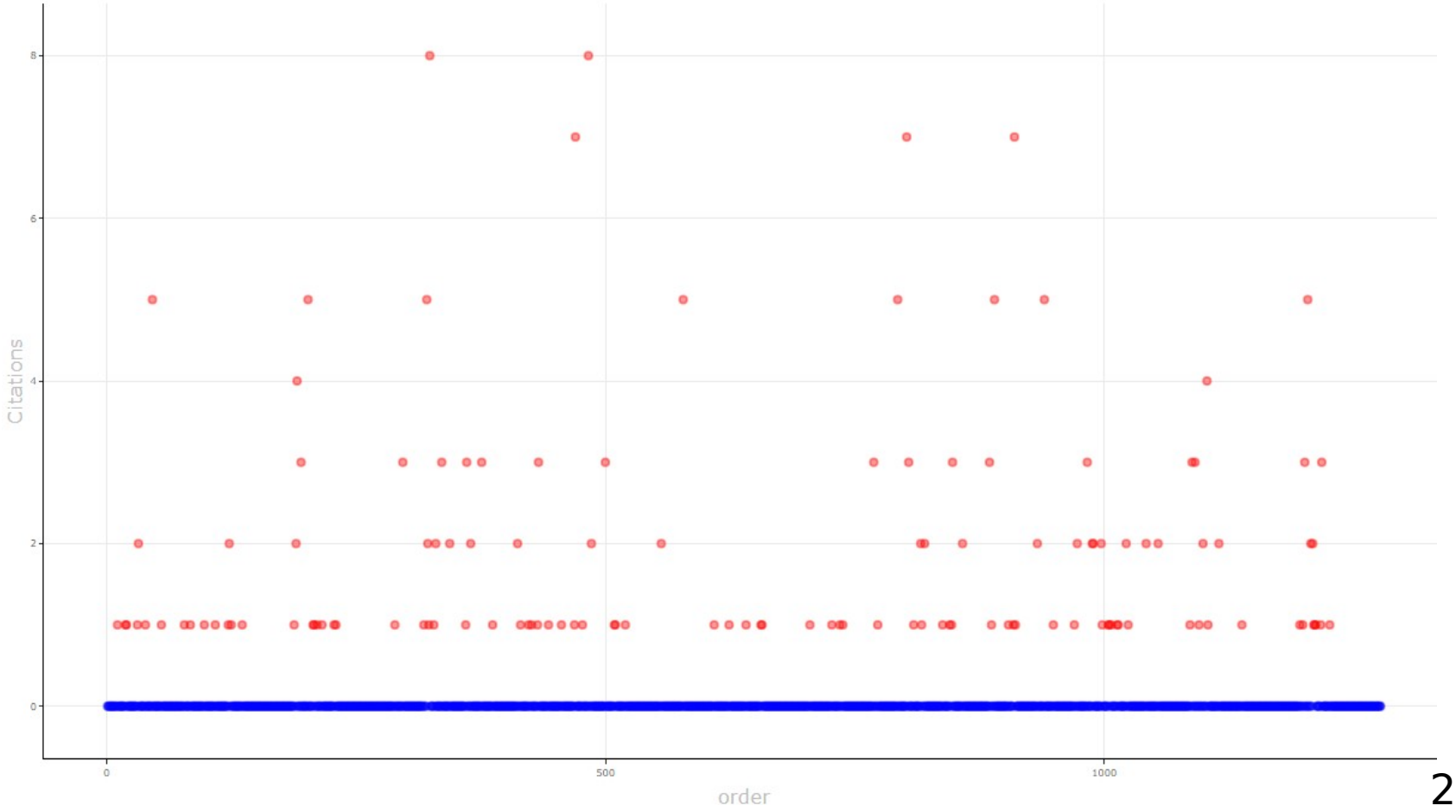


Figure 7.19: How often each SPIRE PC member (1993-2020) was cited by each SPIRE where they were PC members.

ESORICS



Case #2: coercive editors

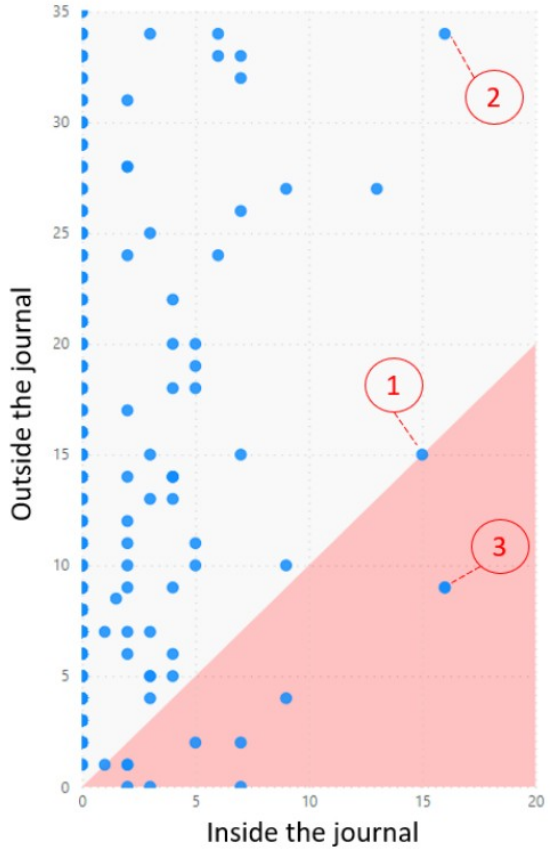
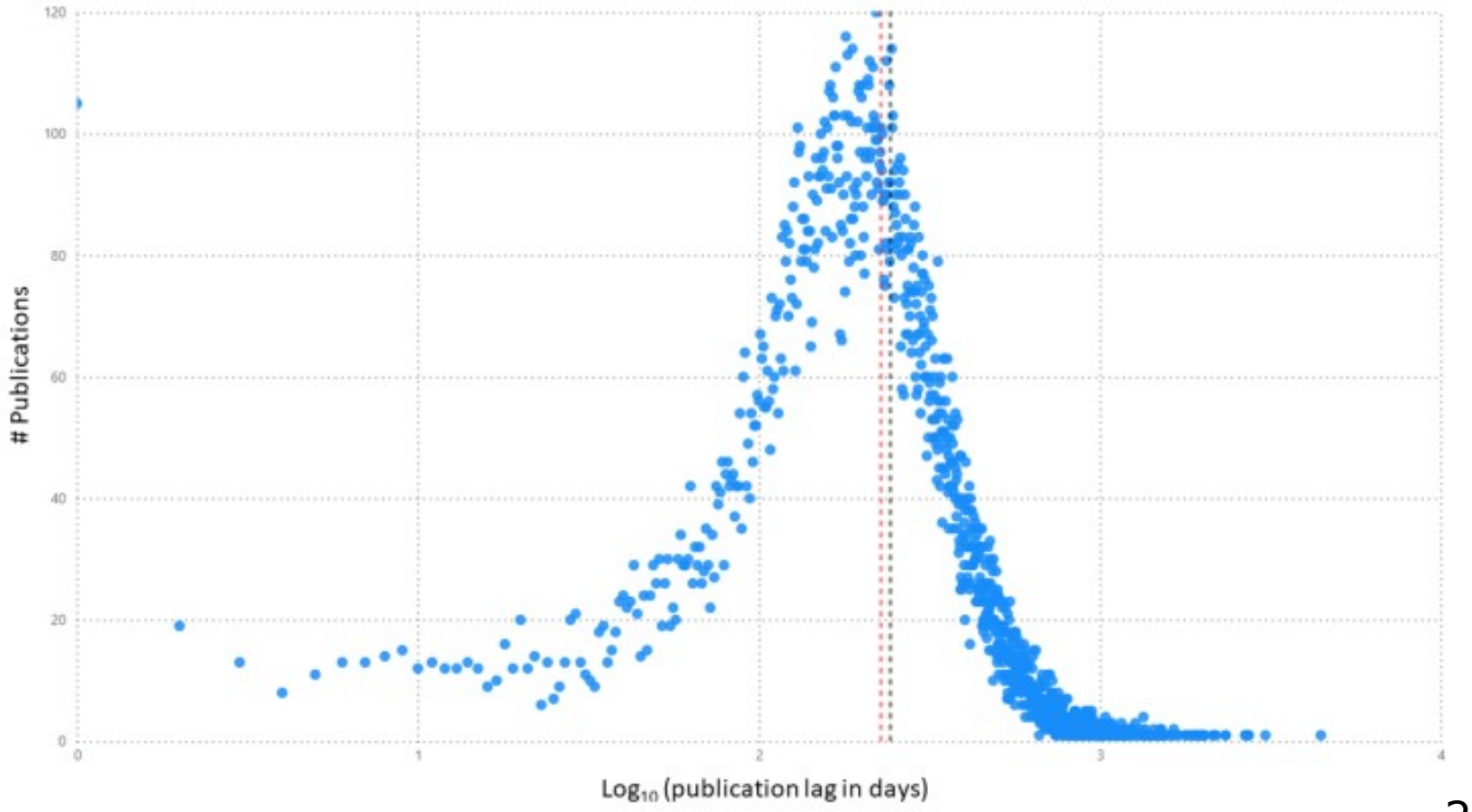


Figure 8.7: Number of unique coauthors inside vs outside the journal per editor

Case #3: *publication-lag*

publication-lag (36 journals)



publication-lag (1 journal)

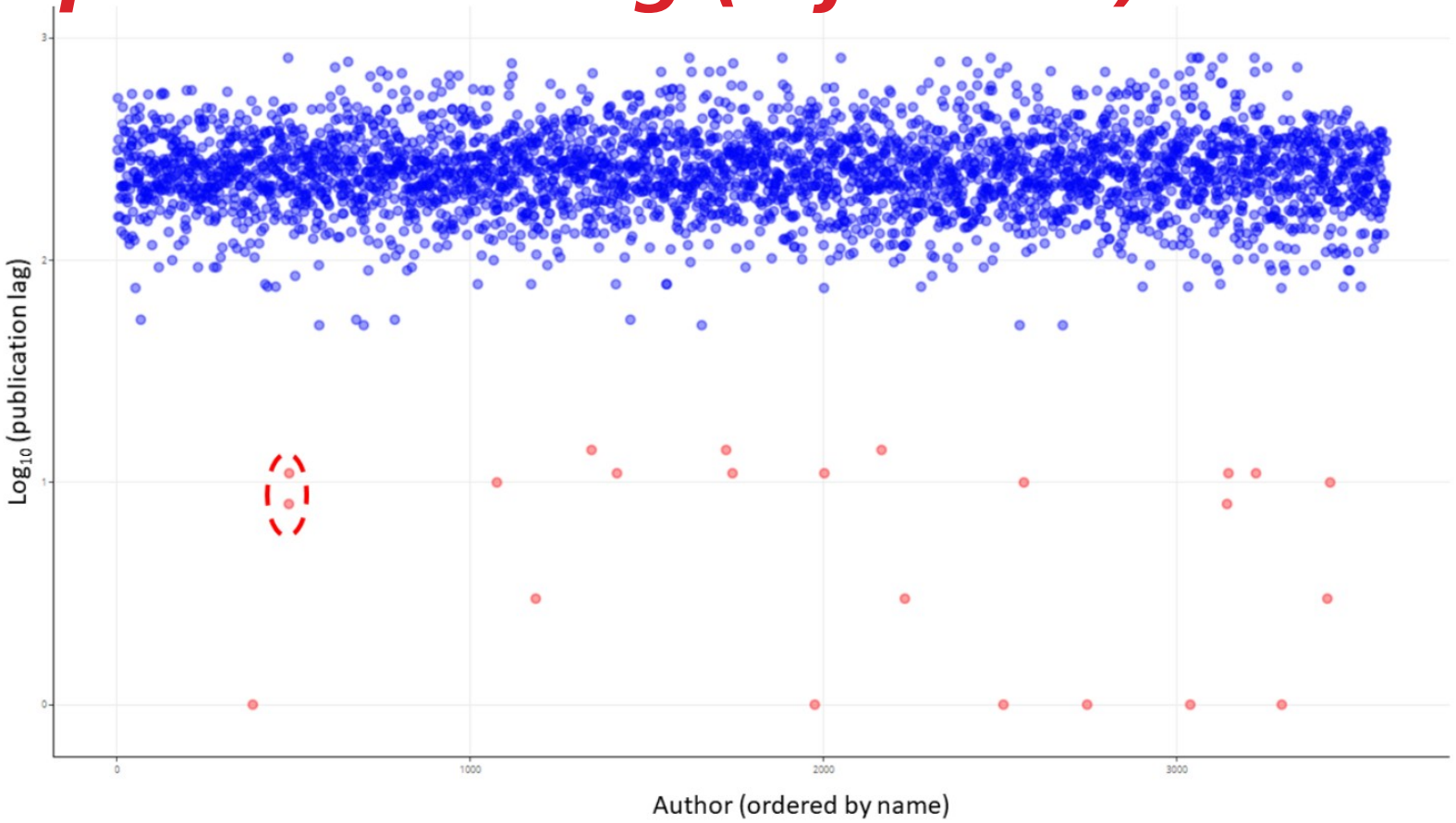


Figure 9.9: Publication lag of authors of Information and Software Technology.

Conclusions

Conclusions

- Data acquisition and integration: challenging
- Nevertheless: useable results
  - Paradigm shift: no hard evidence
- Downside: need to hardcode attack
  - Future work: graph-based approach

Thanks for your attention!