Analysing Author Self-Citations in Computer Science Publications

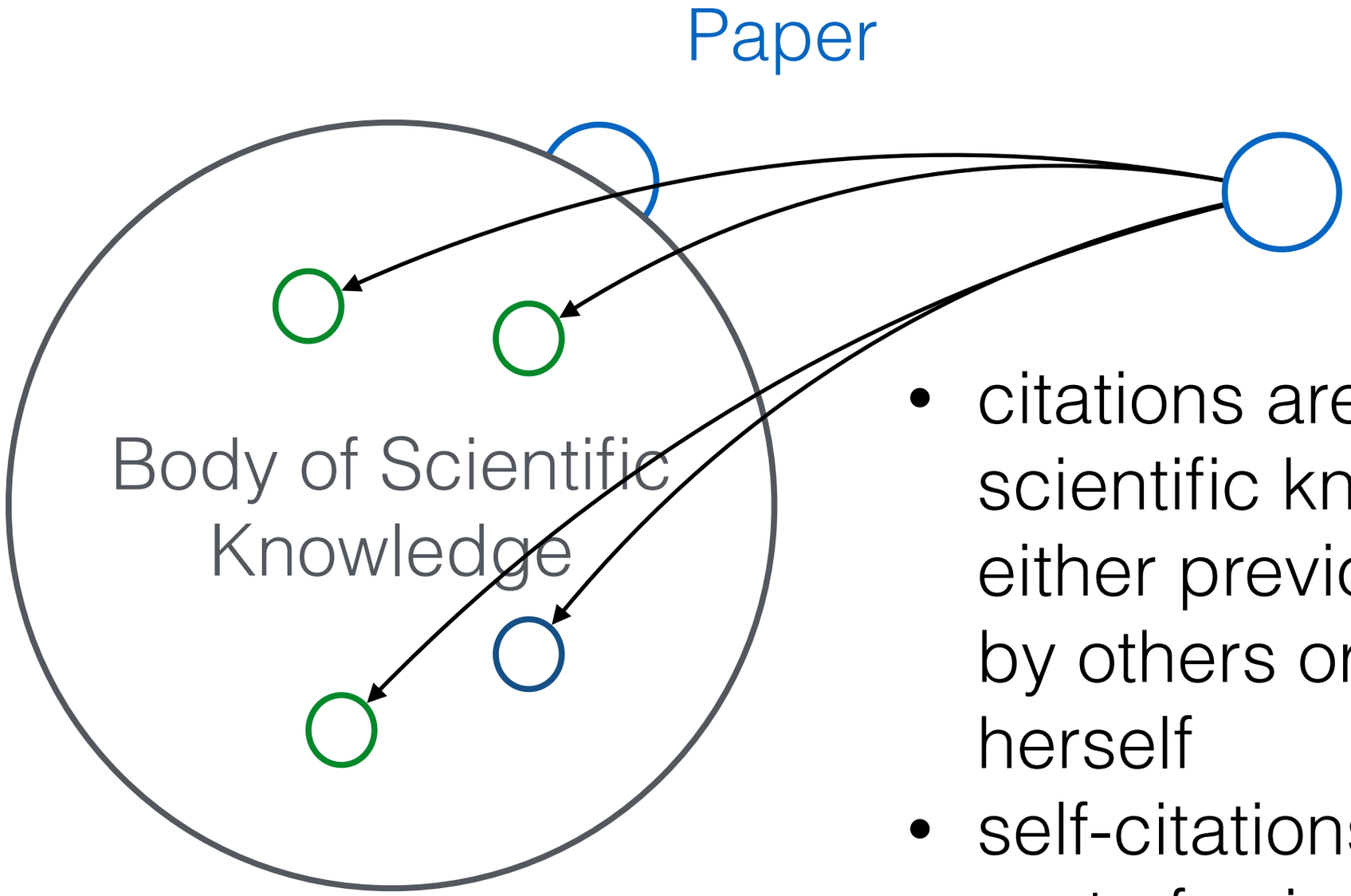
TIR@DEXA, 2018-09-04

Tobias Milz & Christin Seifert

tobias.milz@uni-passau.de

c.seifert@utwente.nl

Science



- citations are references to scientific knowledge — either previously derived by others or by the author herself
- self-citations are essential part of science

https://www.youtube.com/watch?v=We760YM5-iM

The bad side

The *h* index: playing the numbers game

Andy Purvis

Division of Biology, Imperial College L

Scientometrics (2011) 87:85–98
DOI 10.1007/s11192-010-0306-5

Detecting *h*-index manipulation through self-citation analysis

Christoph Bartneck · Servaas Kokkelmans

Article

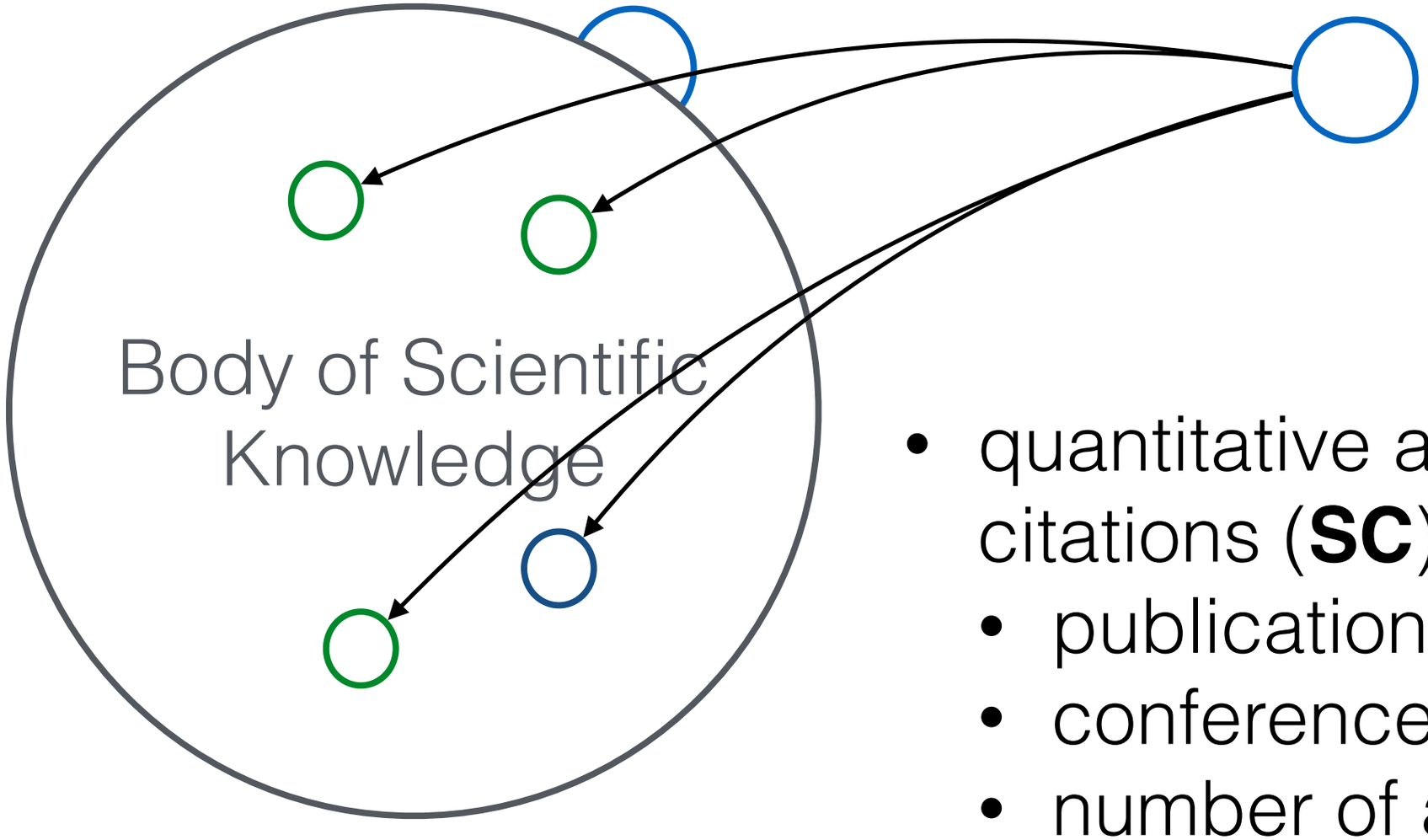
Scientific impact evaluation and the effect of self-citations: mitigating the bias by discounting h-index

Emilio Ferrara
Center for Complex Networks and Systems Research, School of Informatics and Computing, Indiana University Bloomington, USA

Alfonso E. Romero
Department of Computer Science, Centre for Systems and Synthetic Biology, Royal Holloway, University of London, UK

interesting to identify self-citations

Goal



- quantitative analysis of self-citations (**SC**) w.r.t.
  - publication year and age
  - conference rank
  - number of authors
  - gender?
- no judgment whether justified or not

Related work

- 1.8 million JSTOR papers across all disciplines, 1779 to 2011, SC-rate 24% [King et. al, 2017]
- 46,849 paper in computer science, Norway (or Norwegian co-authors), 1981-2000, SC-rate 24% [Asknes, 2013]
- varies across scientific disciplines, over time, across gender, across countries

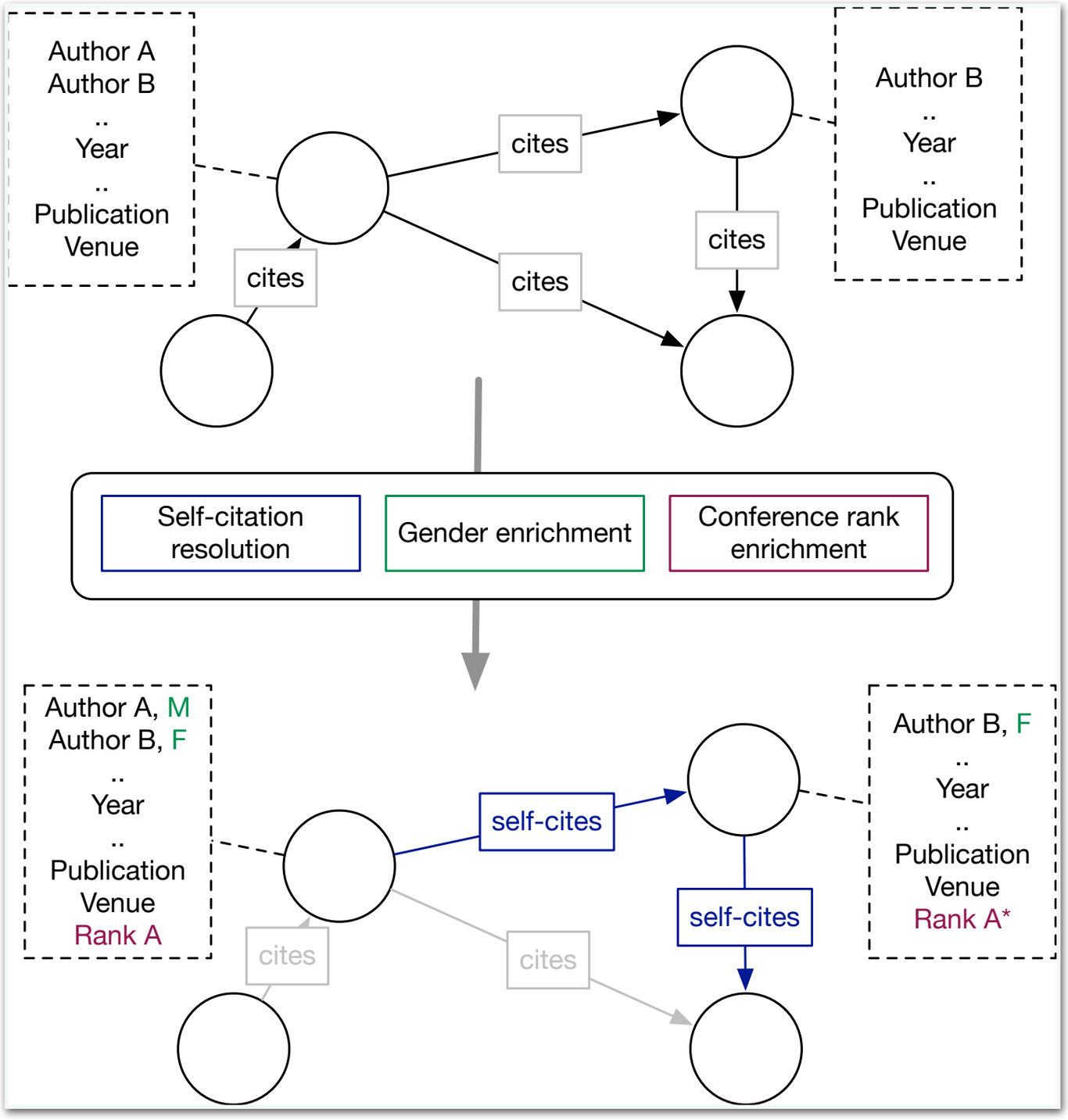
- Computer science in general today?

King, M.M., Bergstrom, C.T., Correll, S.J., Jacquet, J., West, J.D.: Men set their own cites high: Gender and self-citation across fields and over time. Socius 3 (2017)

Aksnes, D.W.: A macro study of self-citation. Scientometrics 56(2), 235–246 (Feb 2003)

Approach

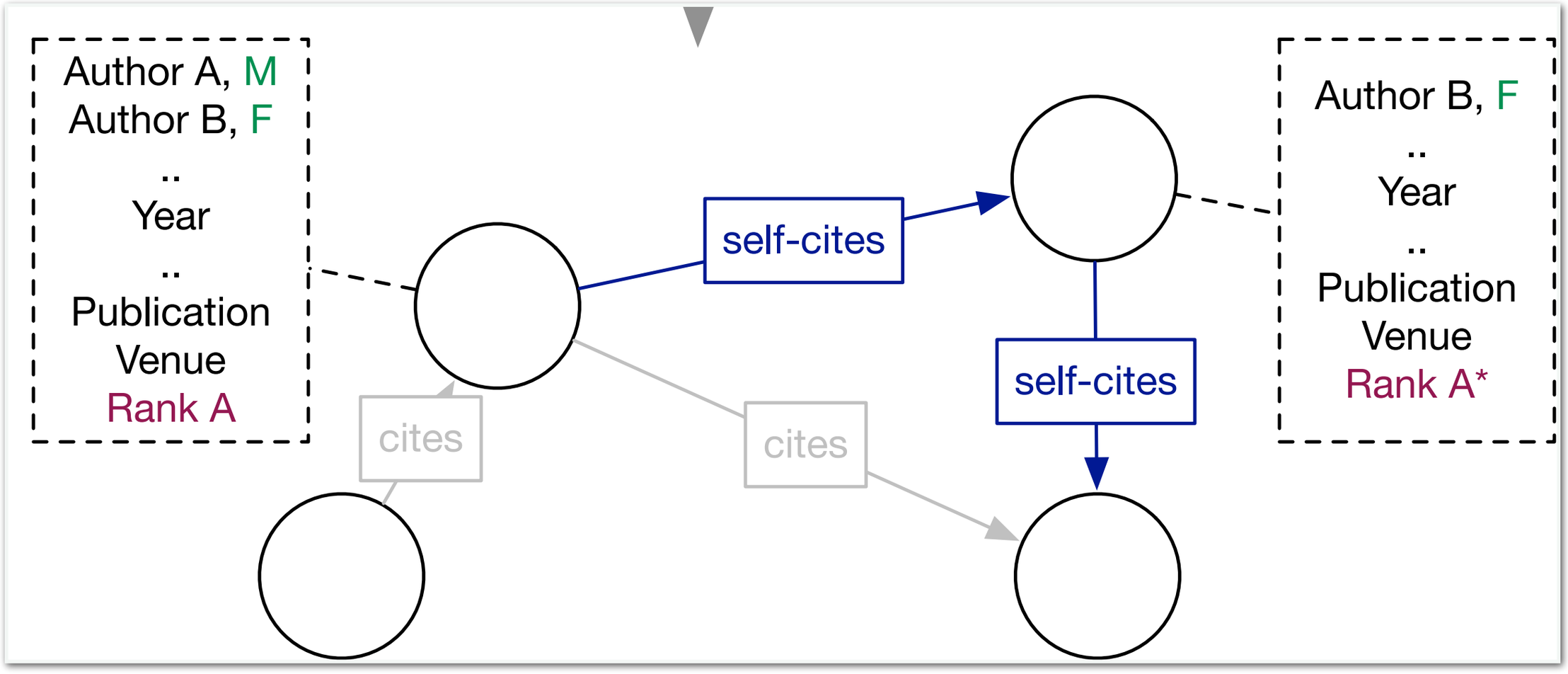
- corpus: DPLP
- rank enrichment with CORE database
- gender enrichment with name lists (US census, Wikipedia, baby names data bases)



Data Set

# P	# A	# P w Rank	# A w G(M/F)	Time Period
3,079,007 (100%)	1,766,540 (100%)	440,356 (14.30%)	954,705 (54.04%)	1936 – 2018

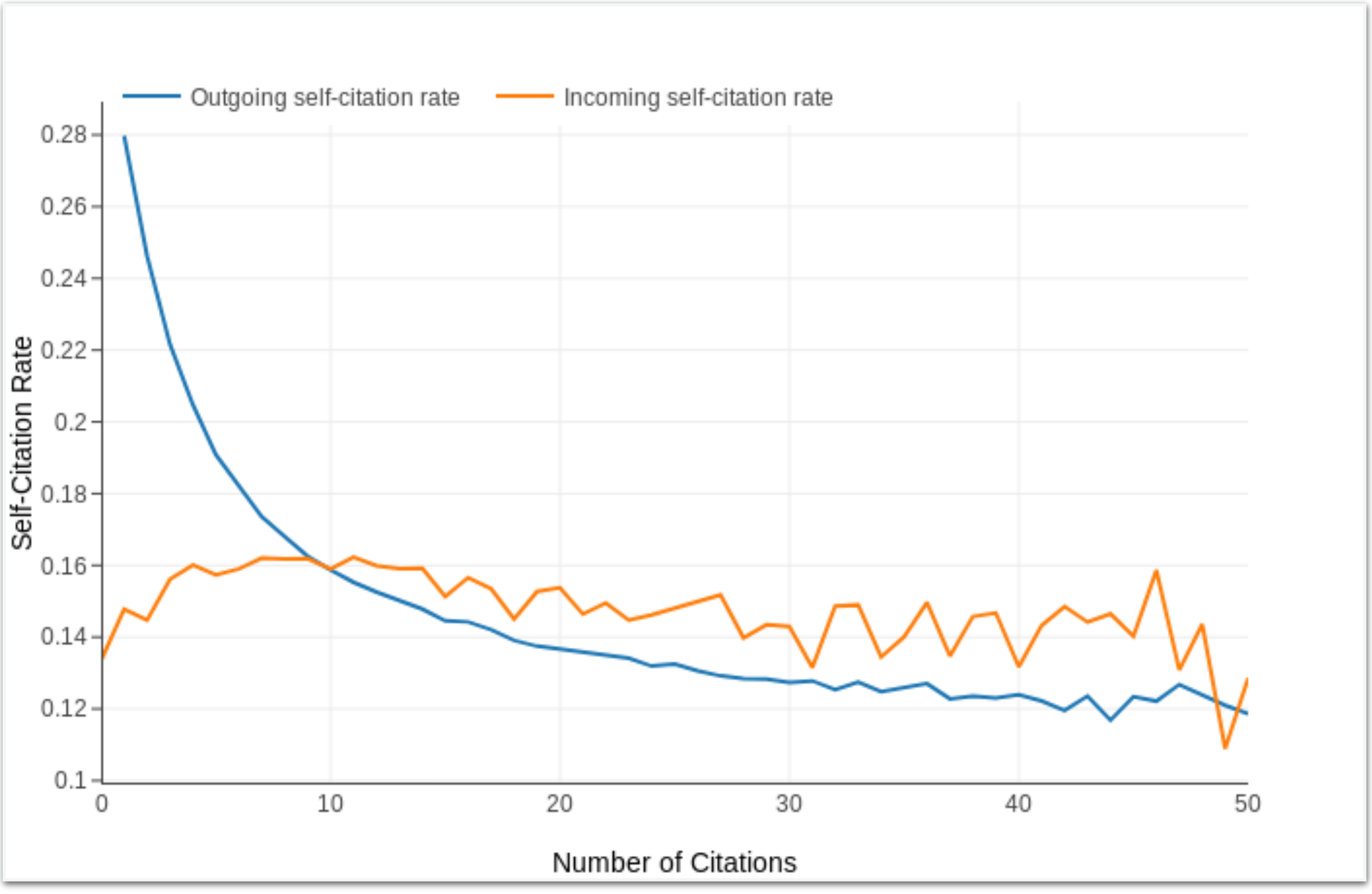
Counts



- author-based, paper-based, incoming, outgoing
- the next numbers are all paper (self-)citations
- counts with neo4j data base + queries

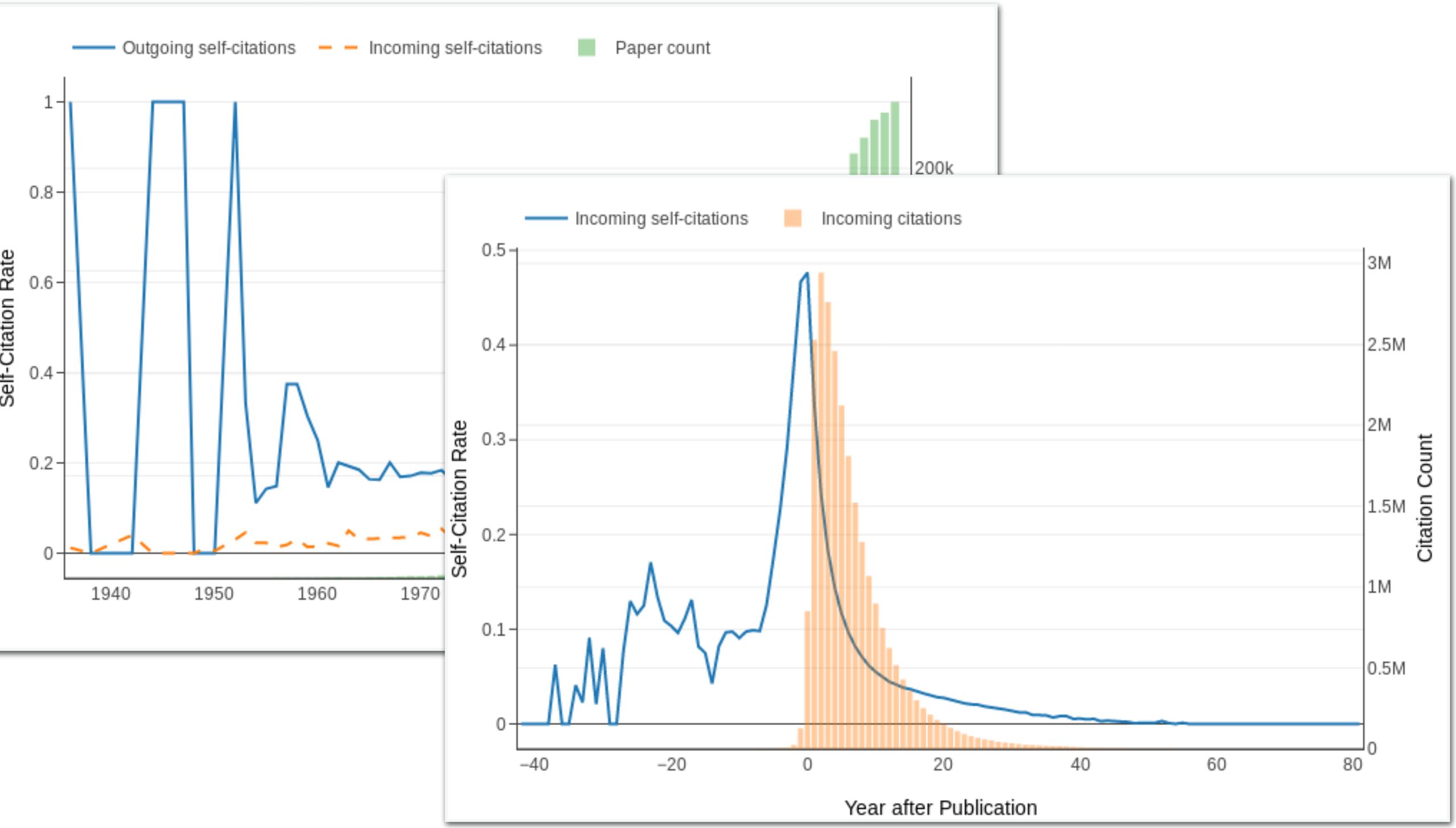
Results

SC Rate (number of self-citations / number of citations)



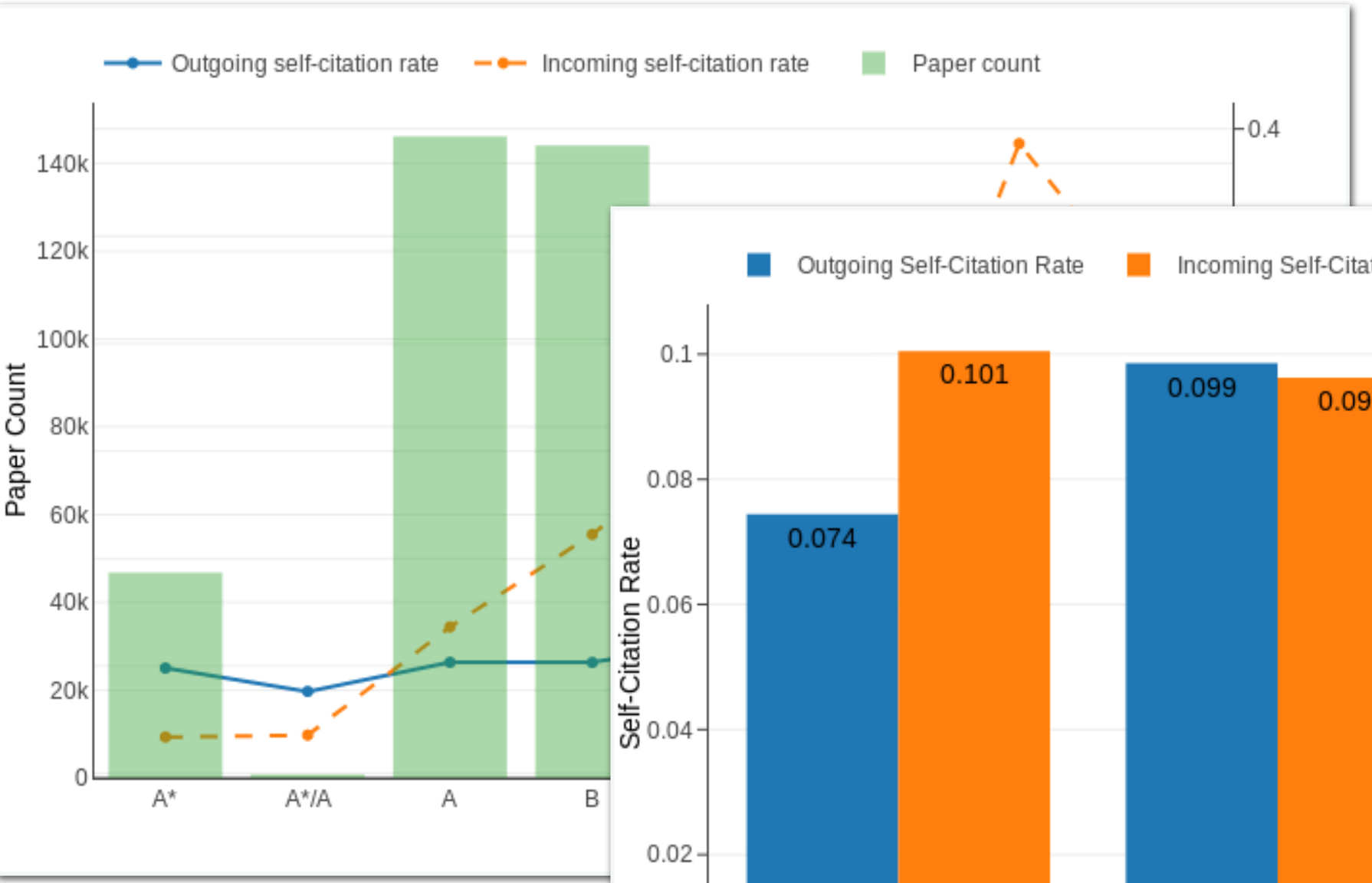
Results

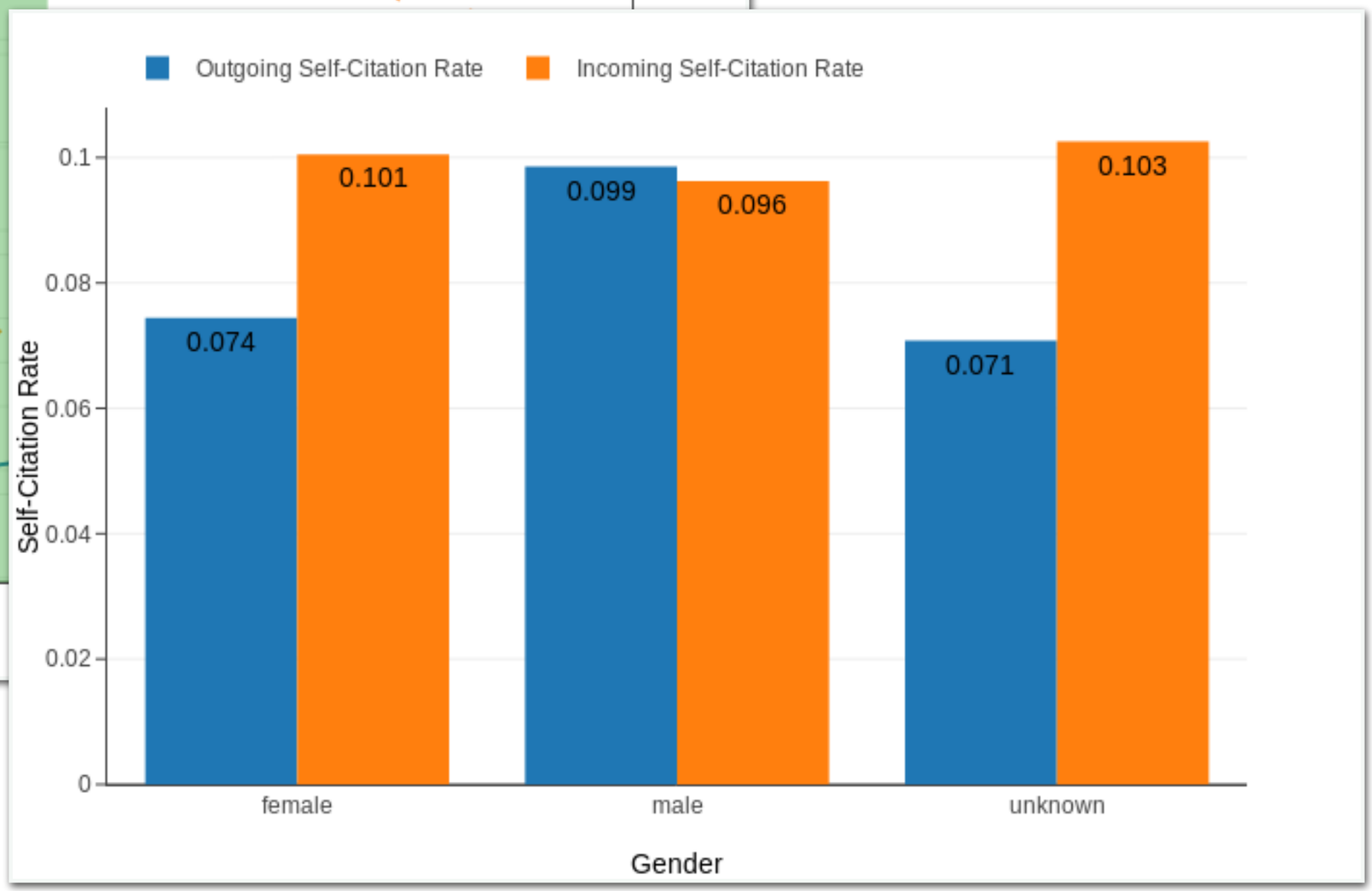
SC Rate by Year



Results

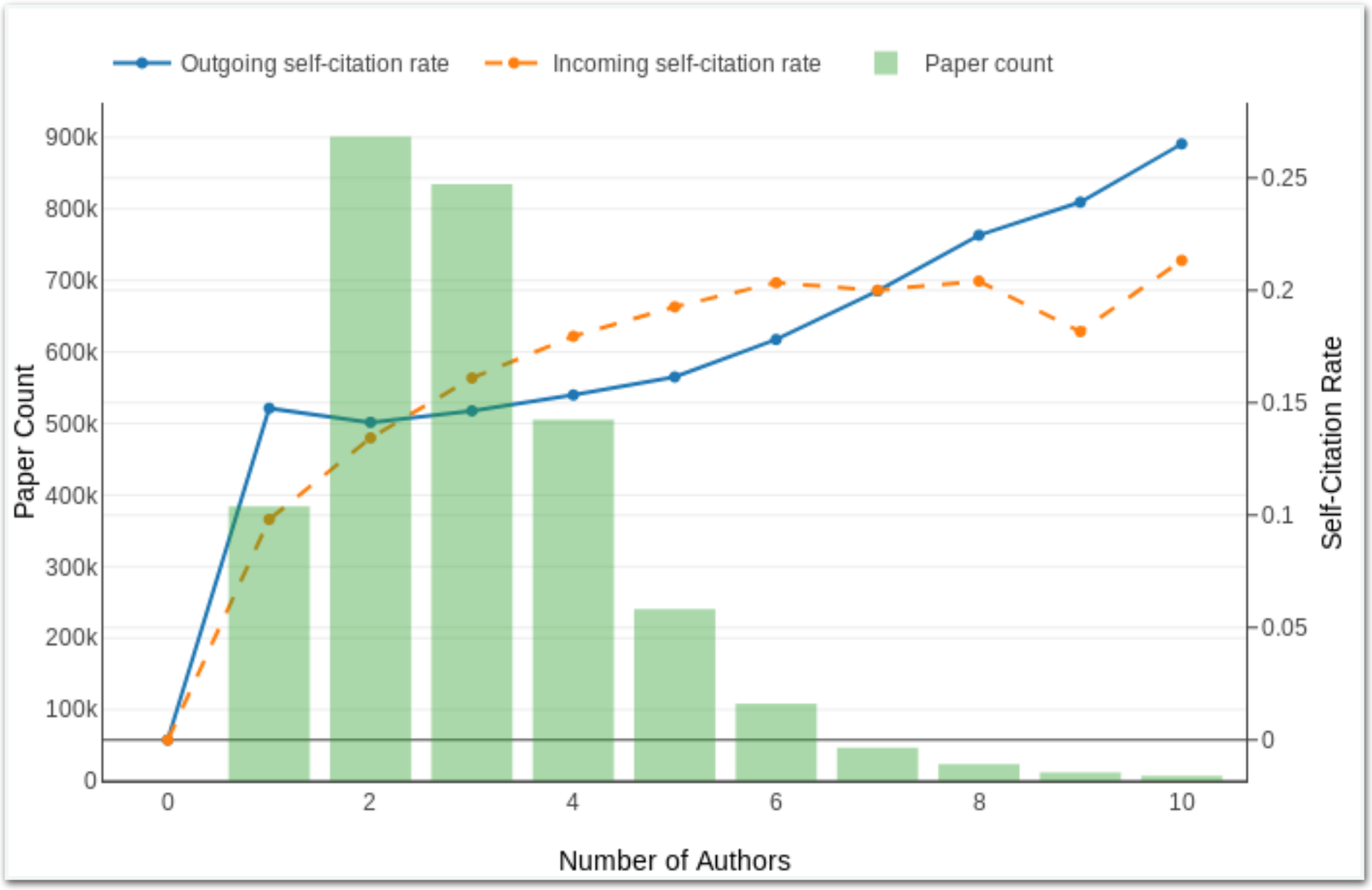
SC Rate by Rank and Gender





Results

SC Rate Author



Summary & Future Work

- in DBLP incoming and outgoing sc-rate approx. 14%
- highest rates of sc in first years after publication
- men cite their own work more, women papers get more self-citations (from papers they co-author)
- in sc-rate varies over conference ranks (out not)
- multi-author papers get more self-citations

Limitations and Future Work

- sources of error
  - rank resolution
  - gender name resolution
  - author name disambiguation

- future work:
  - classifying justified self-citations
  - source - target
    - Tobias Milz and Christin Seifert Who cites what in Computer Science? - Analysing Citation Patterns across Conference Rank and Gender In: Proc. International Conference on Theory and Practice of Digital Libraries (TPDL). 2018.

Thanks!

Contact:

C.Seifert@utwente.nl