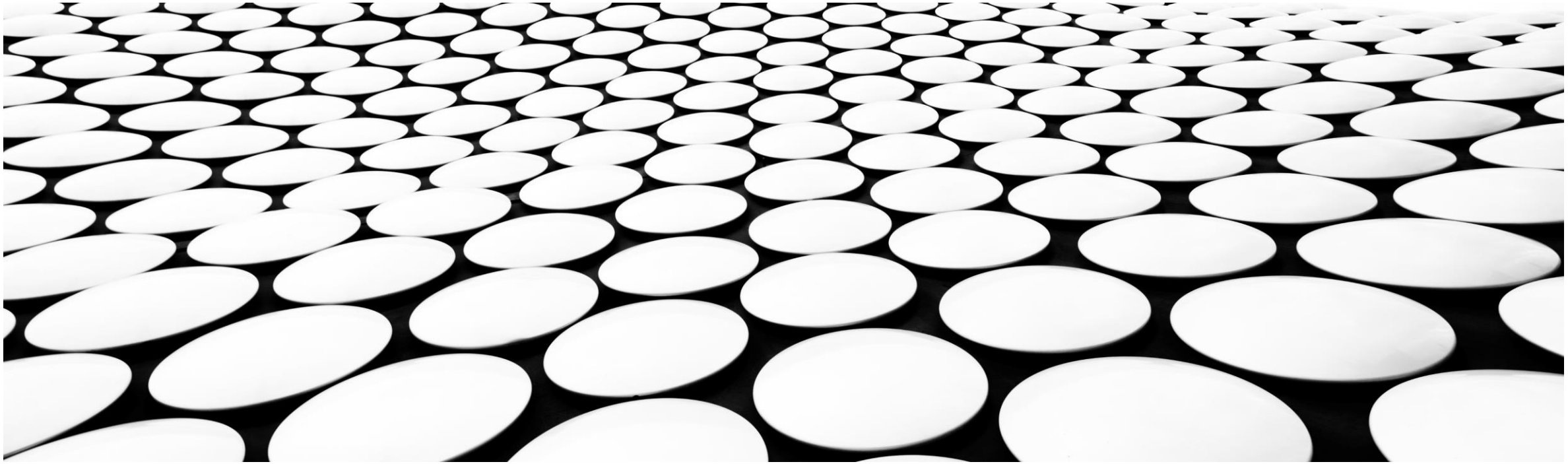
MITIGATING SCHOLARLY CORPUS BIASES WITH CITATIONS: A CASE STUDY ON CORD-19

MICROSOFT RESEARCH, REDMOND WA & ALLEN INSTITUTE FOR ARTIFICIAL INTELLIGENCE, SEATTLE WA



Mitigating Biases in CORD-19 for Analyzing COVID-19 Literature

Provisionally accepted The final, formatted version of the article will be published soon. Notify me

Anshul Kanakia[1], Kuansan Wang[1*], Yuxiao Dong[1], Boya Xie[1], Kyle Lo[2], Zhihong Shen[1], Lucy Wang[2], Chiyuan Huang[1], Darrin Eide[1], Sebastian Kohlmeier[2] and Chieh-Han Wu[1]

[1]Microsoft Research (United States), United States

[2]Allen Institute for Artificial Intelligence, United States

FRONTIERS IN RESEARCH METRIC AND ANALYTICS

Special Topic on

Coronavirus Research Landscape: Resources, Utilities, and Analytic Studies

DOI: 10.3389/frma.2020.596624

ETHICAL AI: AN URGENT TOPIC TO OUR SOCIETY

BIAS

FAIRNESS

SAFETY

TRANSPARENCY

ACCOUNTABILITY

THE WALL STREET JOURNAL.

By Michael Totty

Nov. 3, 2020 10:00 am ET

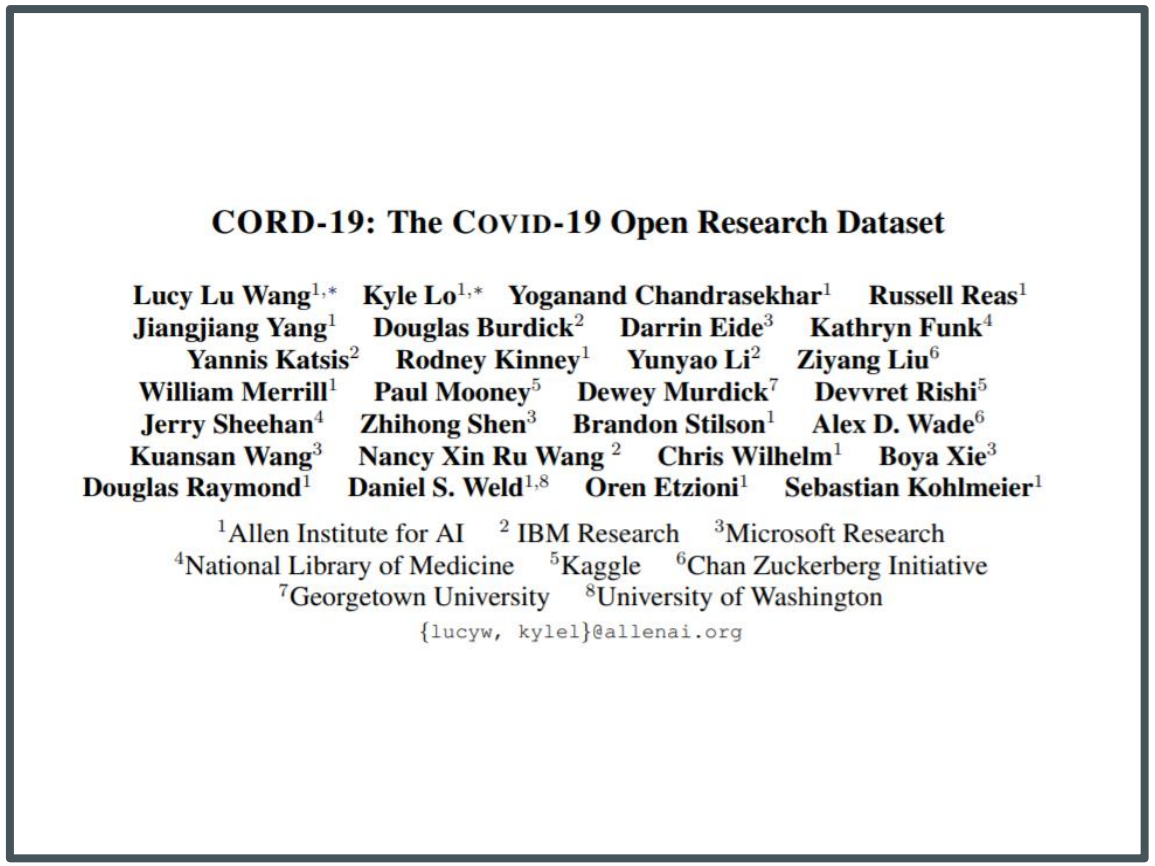
JOURNAL REPORTS: TECHNOLOGY

How to Make Artificial Intelligence Less Biased

AI systems can unfairly penalize certain segments of the population—especially women and minorities. Researchers and tech companies are figuring out how to address that.

The AI world is making a strong push to root out bias in AI systems, but it faces some significant obstacles. KEITH A. WEBB AND IMAGES FROM ISTOCK

TACKLING INFORMATION OVERFLOW WITH AI



CORD-19: The COVID-19 Open Research Dataset

Lucy Lu Wang[1,*] Kyle Lo[1,*] Yoganand Chandrasekhar[1] Russell Reas[1]
Jiangjiang Yang[1] Douglas Burdick[2] Darrin Eide[3] Kathryn Funk[4]
Yannis Katsis[2] Rodney Kinney[1] Yunyao Li[2] Ziyang Liu[6]
William Merrill[1] Paul Mooney[5] Dewey Murdick[7] Devvret Rishi[5]
Jerry Sheehan[4] Zhihong Shen[3] Brandon Stilson[1] Alex D. Wade[6]
Kuansan Wang[3] Nancy Xin Ru Wang[2] Chris Wilhelm[1] Boya Xie[3]
Douglas Raymond[1] Daniel S. Weld[1,8] Oren Etzioni[1] Sebastian Kohlmeier[1]

[1]Allen Institute for AI [2]IBM Research [3]Microsoft Research
[4]National Library of Medicine [5]Kaggle [6]Chan Zuckerberg Initiative
[7]Georgetown University [8]University of Washington
{lucyw, kylel}@allenai.org

- 3500+ articles/week by Mid March
  - 372,698 articles as of November 15,2020
- Full text corpus: first released on March 16, 2020
  - Publisher contributions + archival services
- Activities
  - Open QA Challenge on Kaggle
  - TREC tracks at NIST
- Methodology: keyword query into various databases

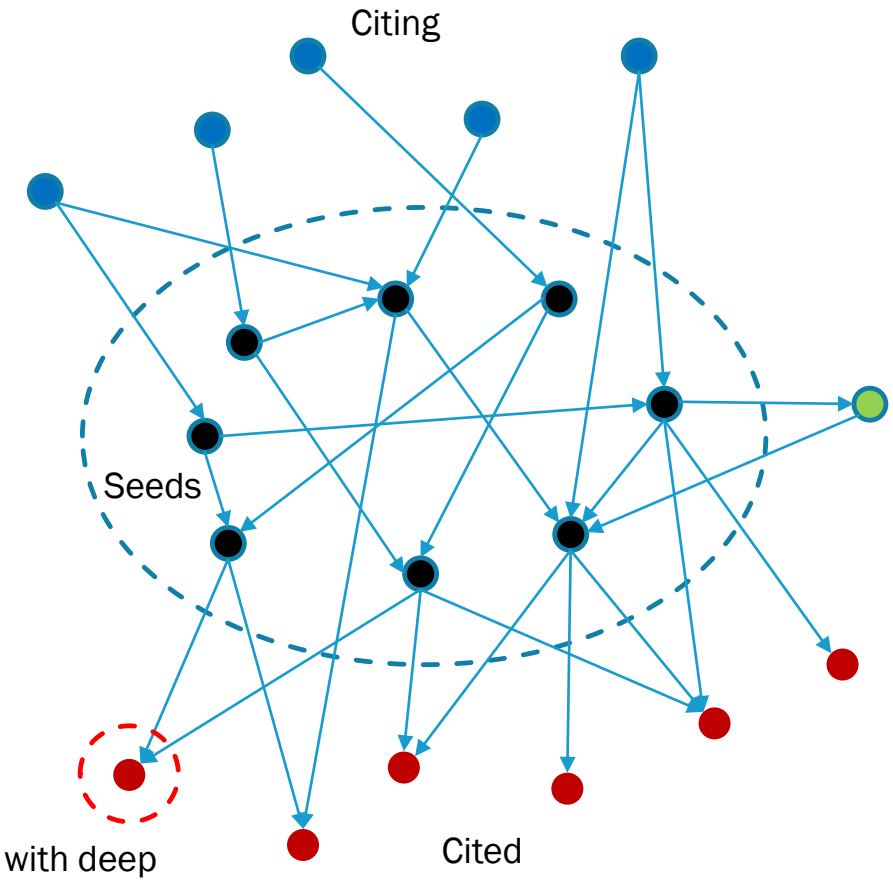
  "**COVID**" OR "**COVID-19**" OR "**CORONAVIRUS**" OR "**2019-nCov**" OR "**SARS-COV**" OR "**MERS-COV**" OR "**Severe Acute Respiratory Syndrome**" OR "**Middle East Respiratory Syndrome**"

ASSESSING BIASES BY CORPUS EXPANSION WITH CITATIONS

- Price, D., "Networks of Scientific Papers", *Science* 1965
- Notable network science models:
  - Preferential attachment, Albert and Barabasi, *Science* 1999
  - Individual node fitness, Caldarelli et al., *Physical Review Letters*, 2002
  - Latent space model, Papadopoulos et al., *Nature* 2012
  - Discrete choice, Overgoor et al., *WWW-2019*
- Methods evaluated:
  - Enclosure graph (Sinatra et al., "A century of physics", *Nature Physics*, 2015)
  - Closure graph: updated data available at Github: https://aka.ms/magcord19mapping

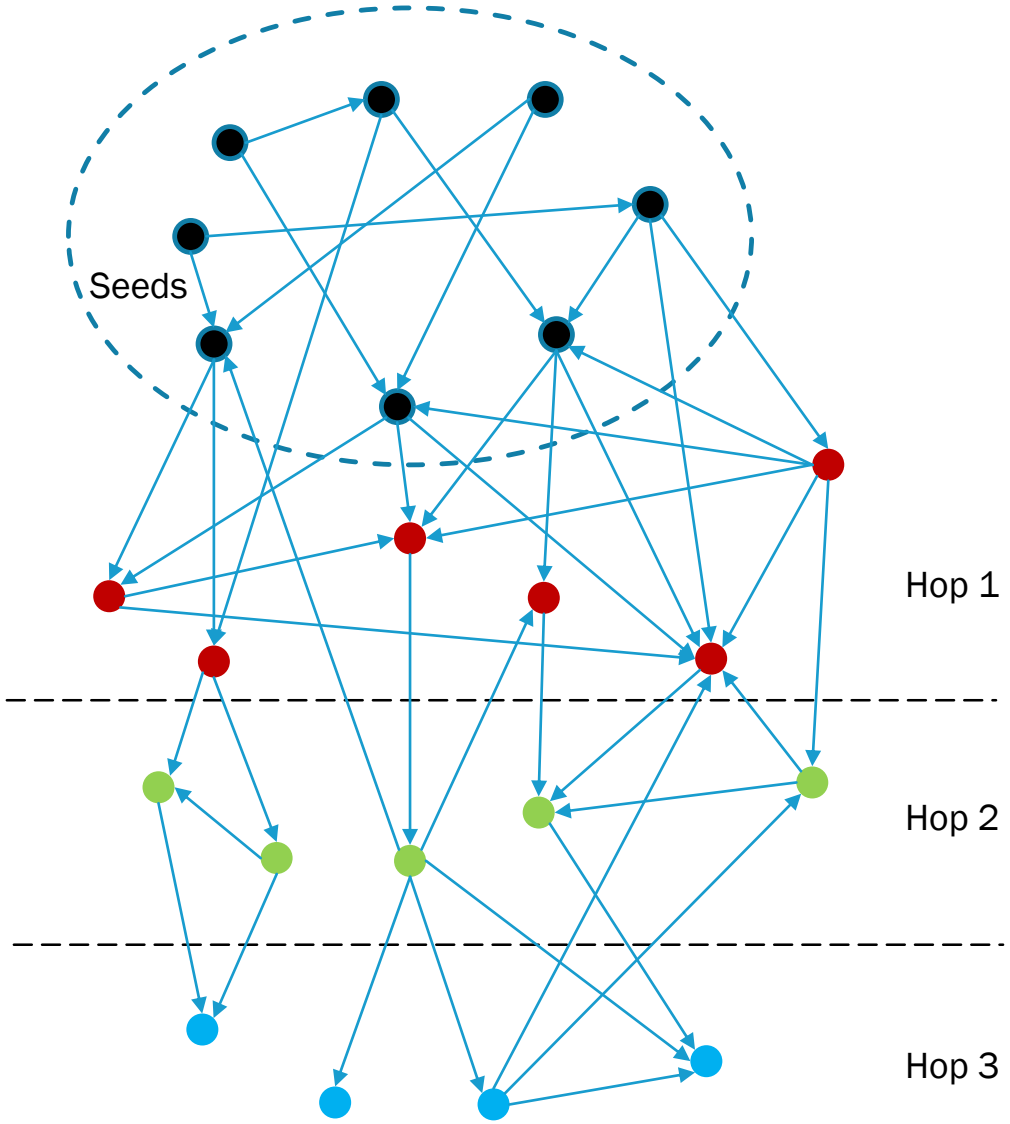
ENCLOSURE GRAPH

- From a seed collection of articles:
  - Expand to citing and cited (or both) articles
  - CORD-19 => CORD-19E
- Observations:
  - Bidirectional, single step traversal on citation networks
  - Cannot be made into iterative algorithm without topic overrun
    - Ex: citing a “tool” paper
    - Sensitive to seed quality

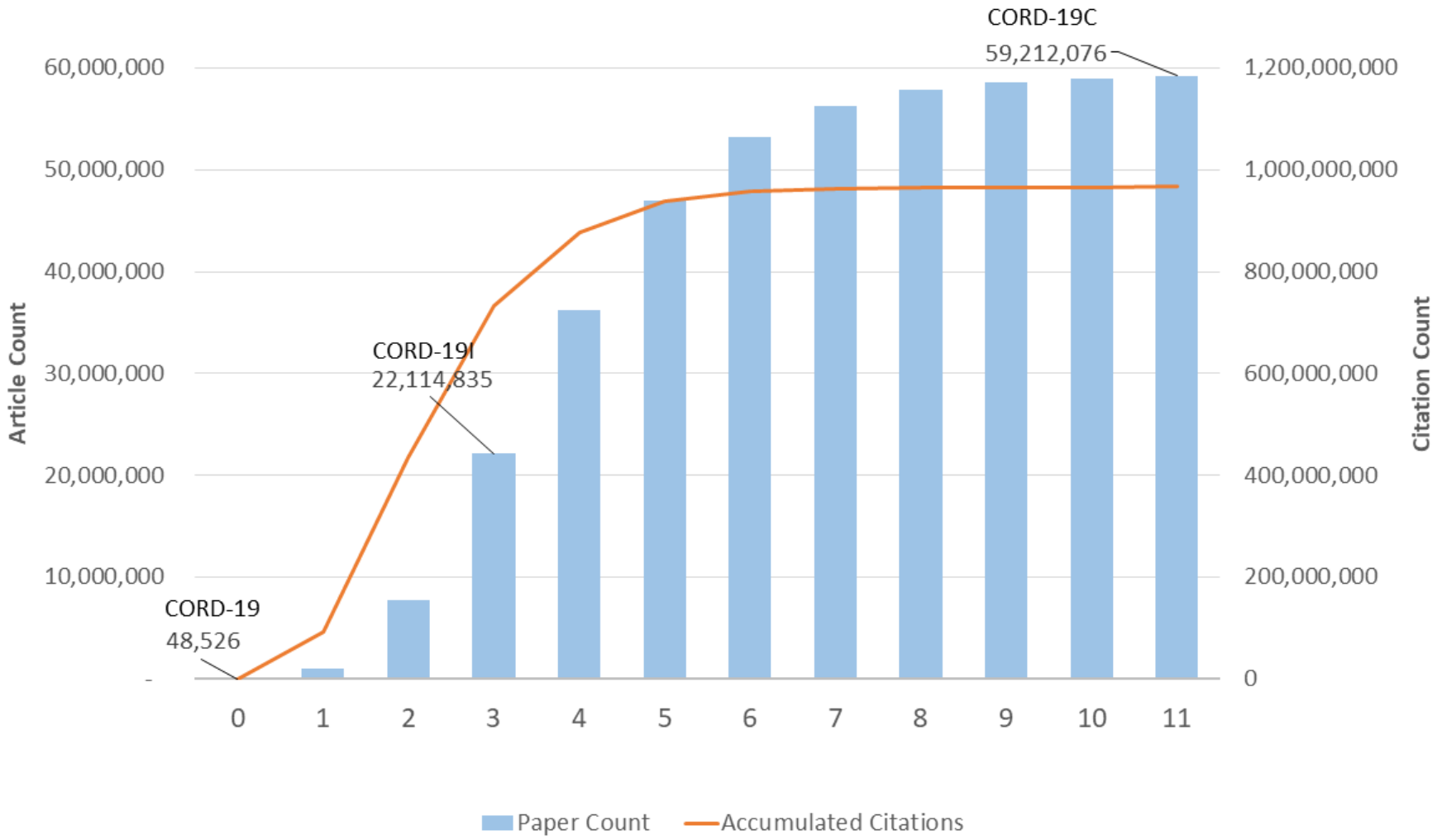


CLOSURE GRAPH

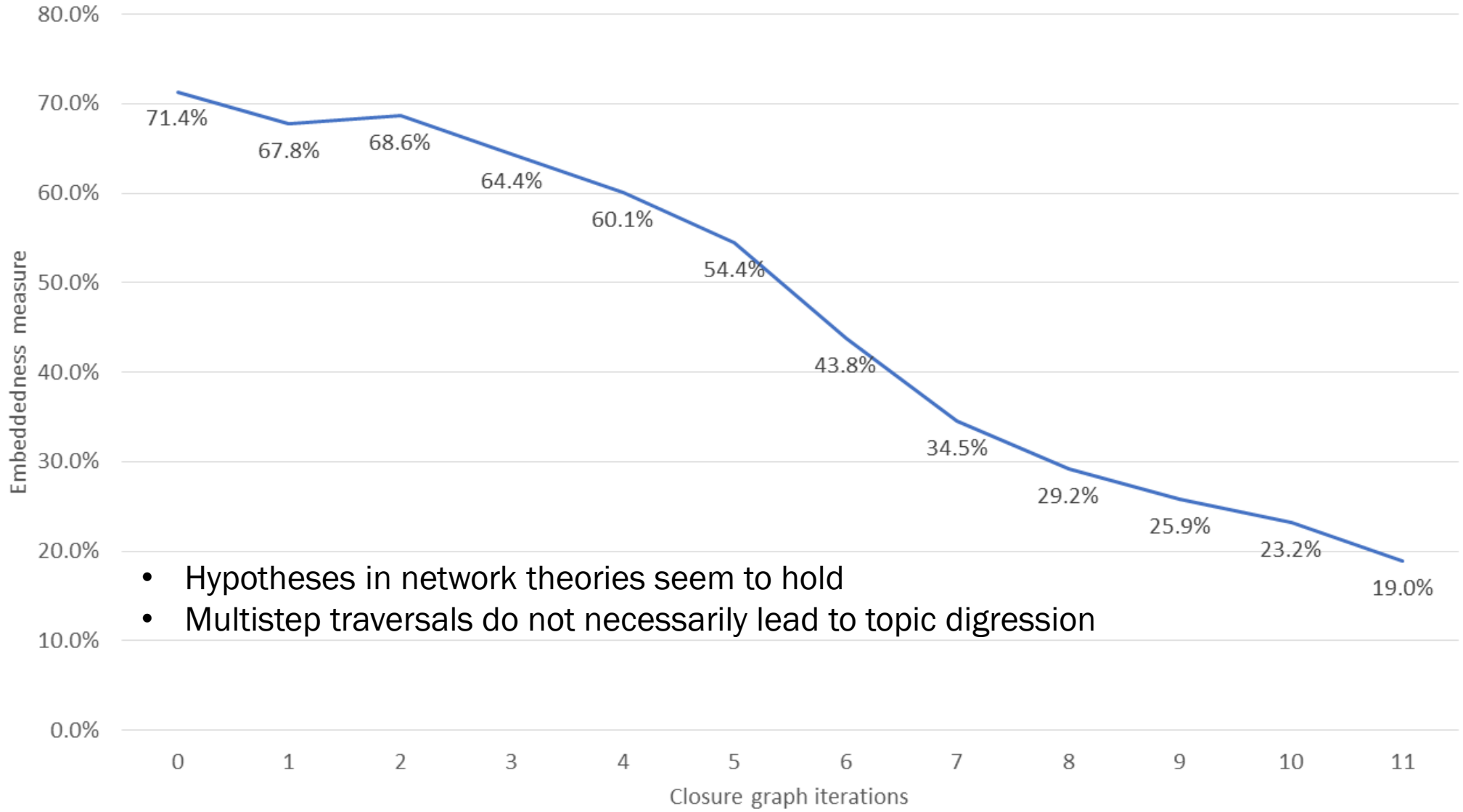
- Uni-direction, multiple step traversals
  - Continue until all references are in the expanded set
  - Mostly directed acyclic graph
- Motivations
  - Cover more background and lineage of knowledge
  - Discrete (discreet) Choice => Topic overrun less severe
- Surprise finding
  - Don’t need full closure!
  - “Inflection” point seems to capture themes
- CORD-19C: full closure; CORD-19I: inflection closure



CLOSURE GRAPH GROWTH VS HOPS

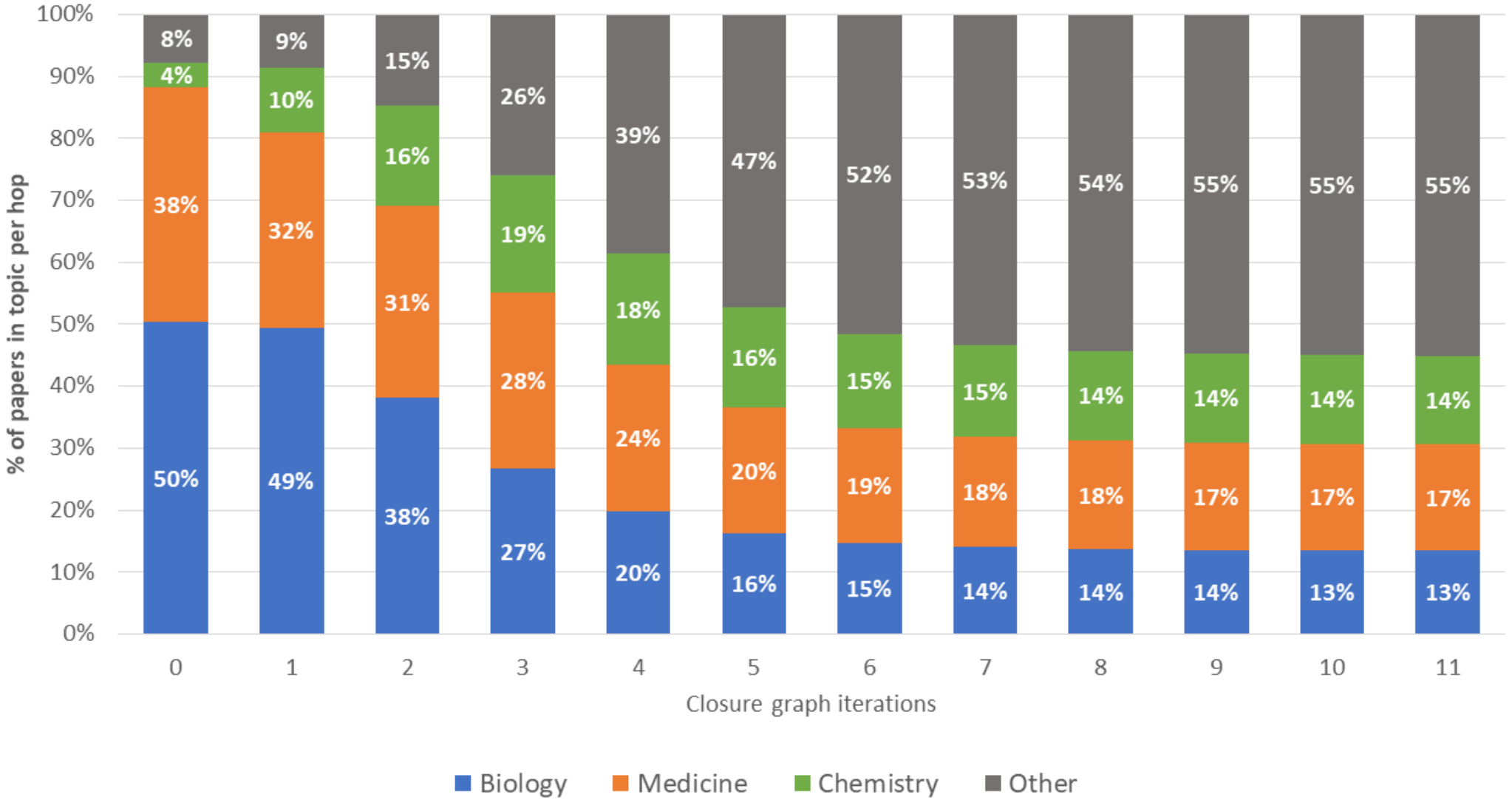


EMBEDDEDNESS: CITATIONS RECEIVED FROM WITHIN COLLECTION

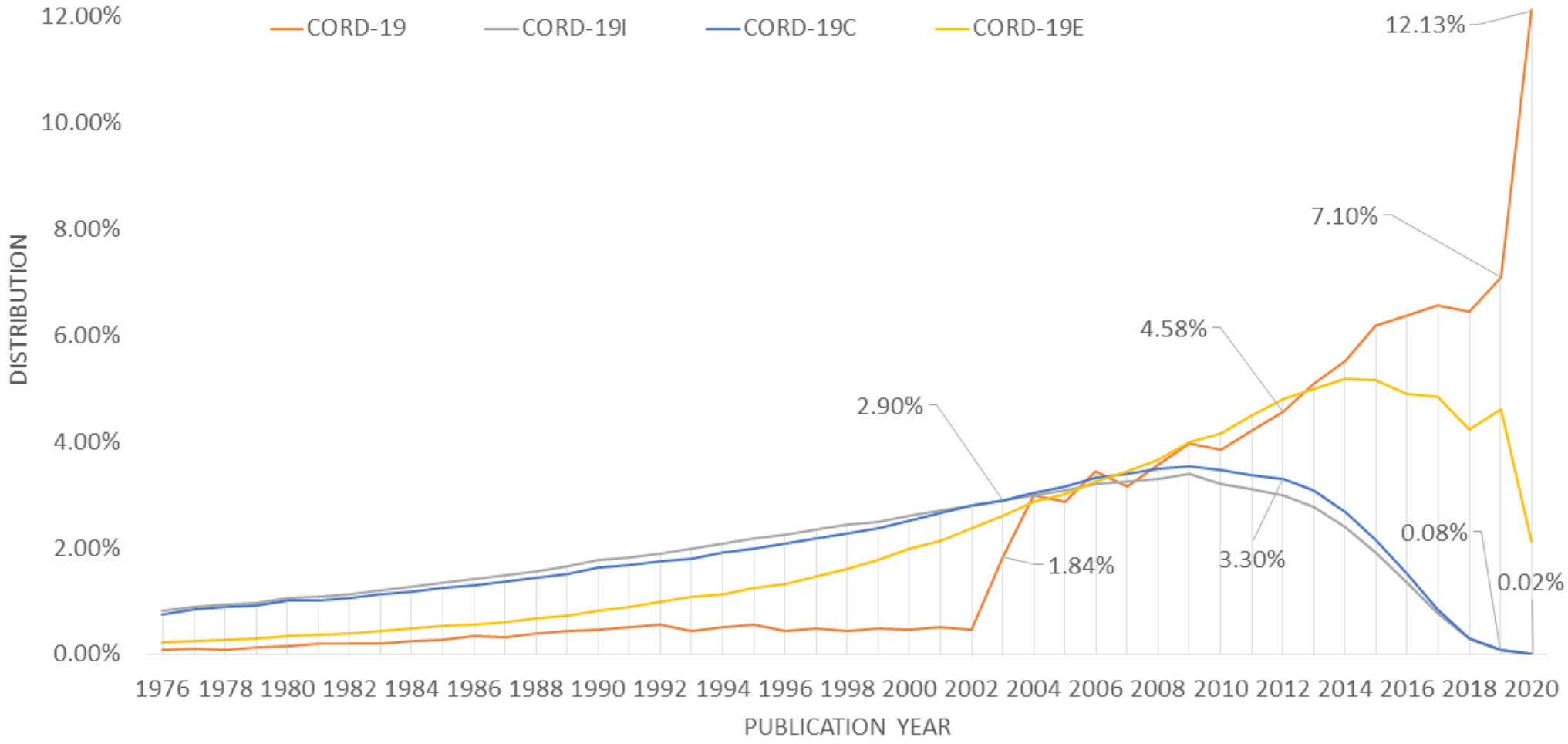


- Hypotheses in network theories seem to hold
- Multistep traversals do not necessarily lead to topic digression

TOPIC COVERAGE

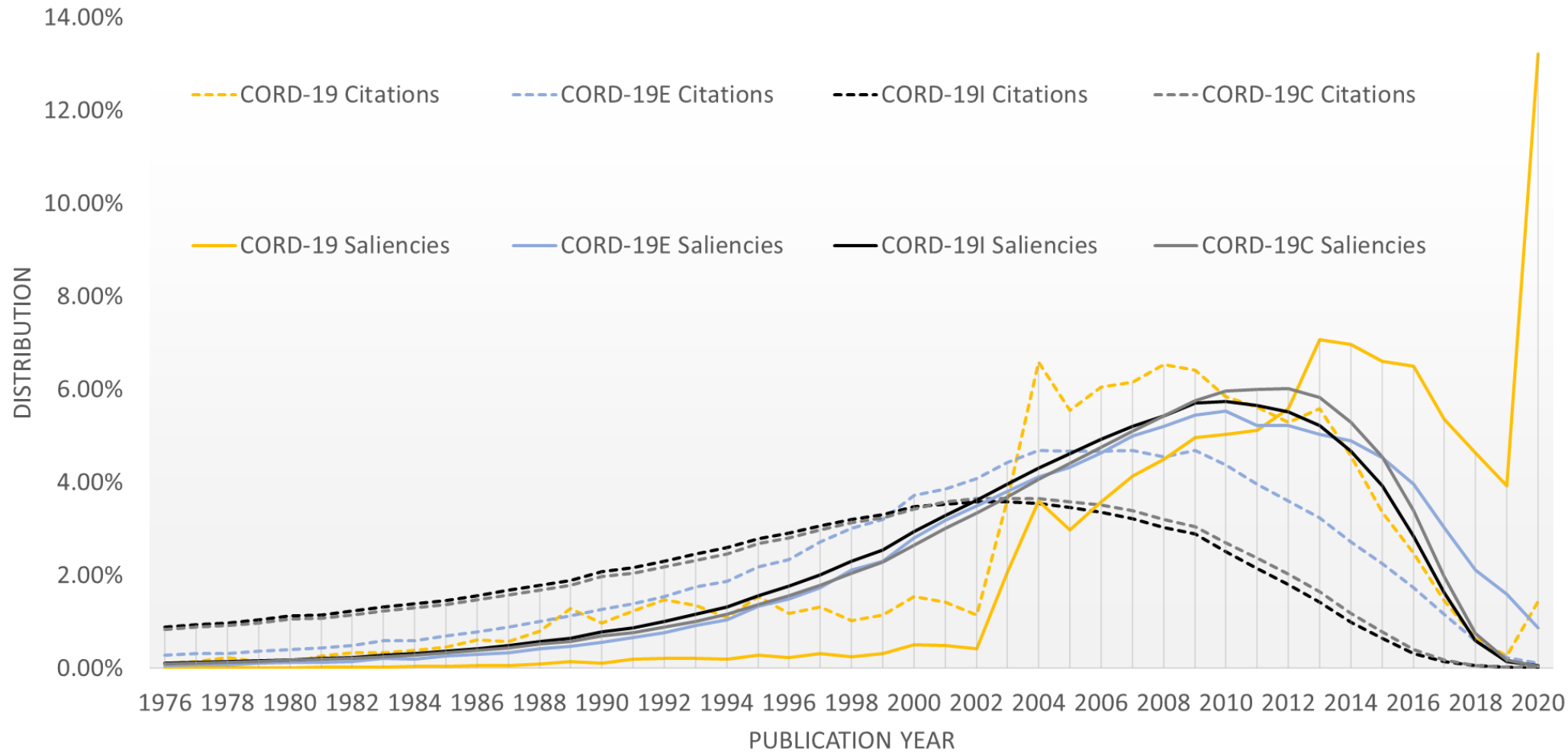


ARTICLE AGE DISTRIBUTION

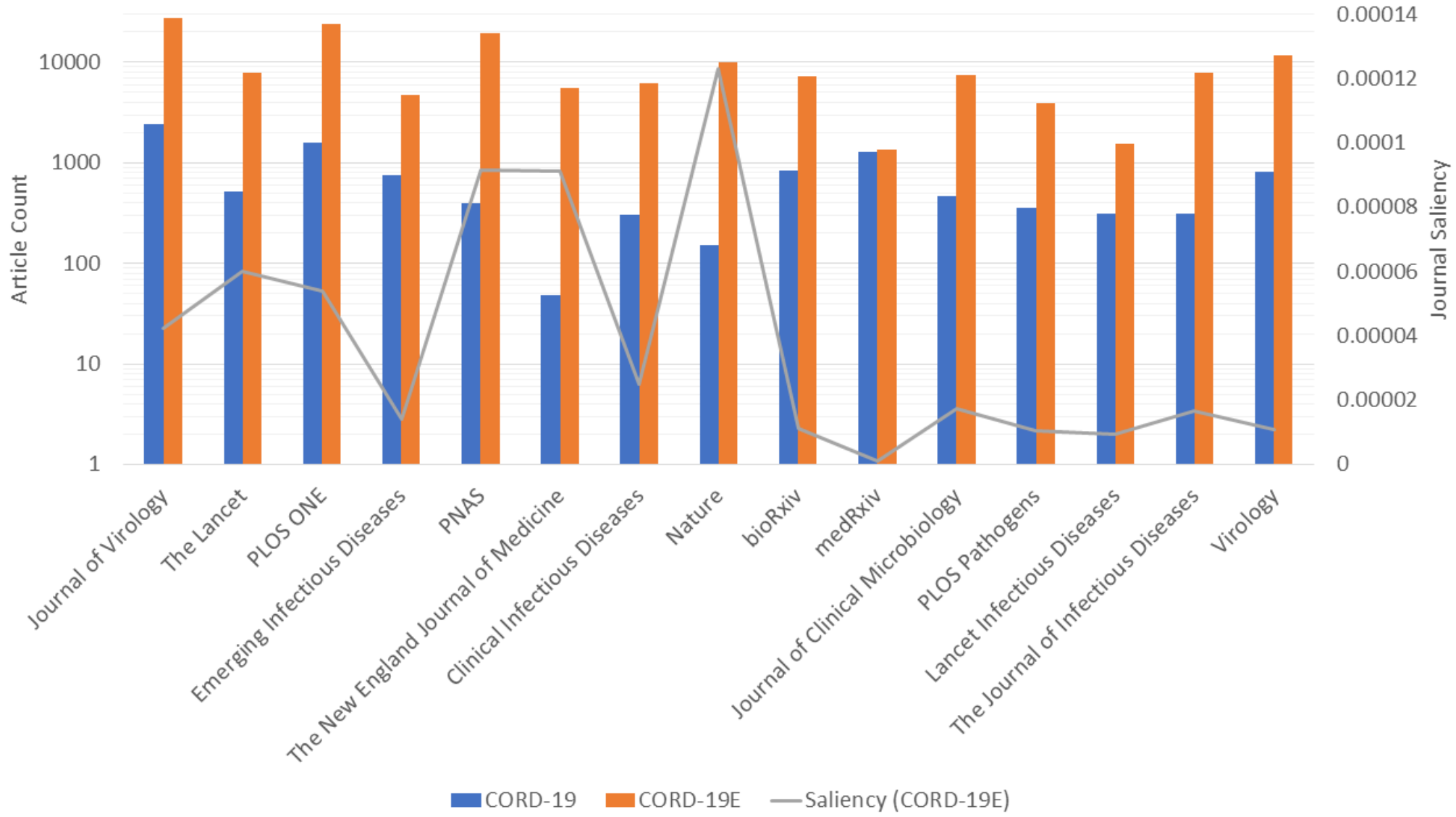


DISTRIBUTION OF ARTICLE IMPORTANCE

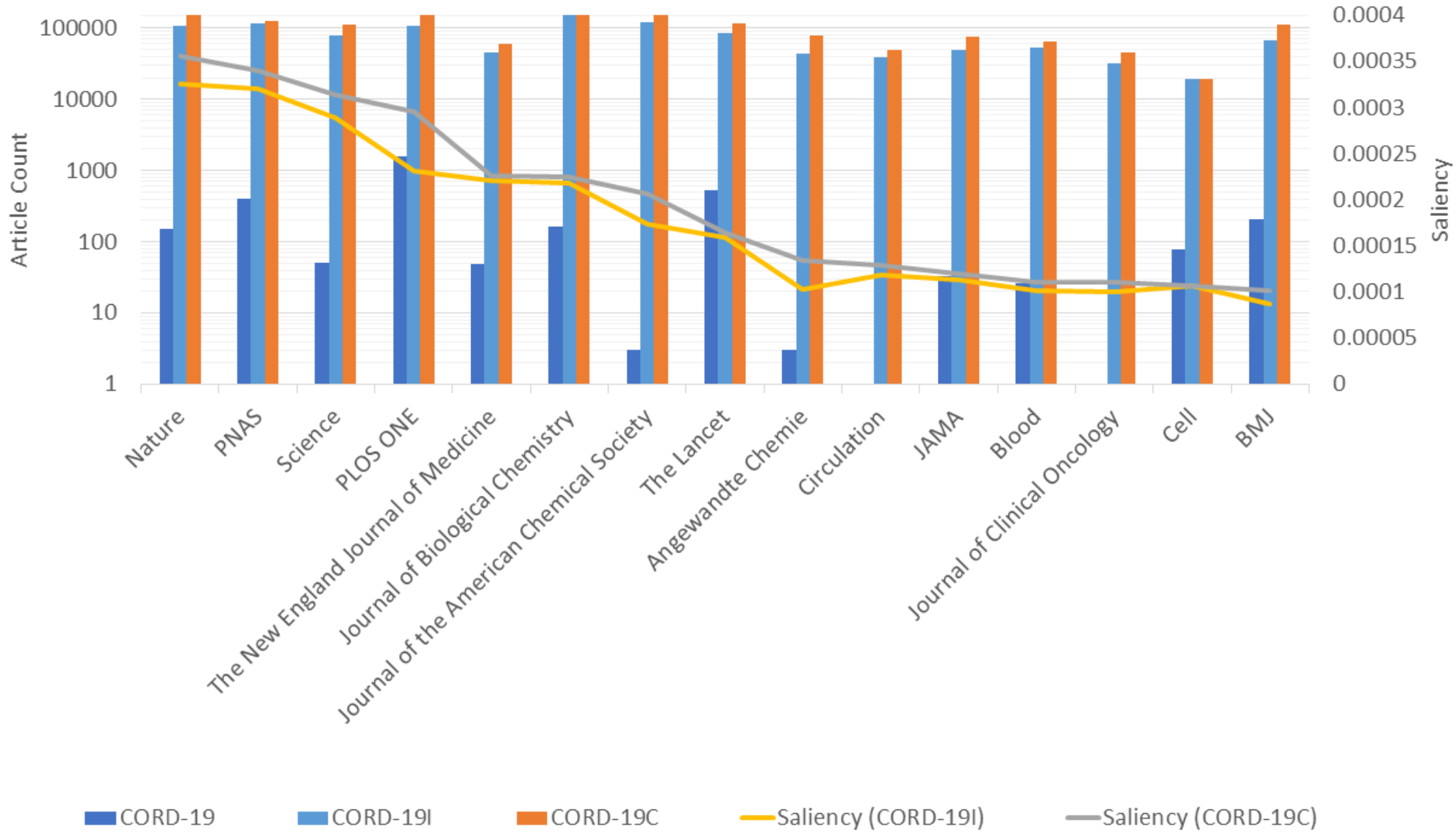
*Details on saliency can be found here



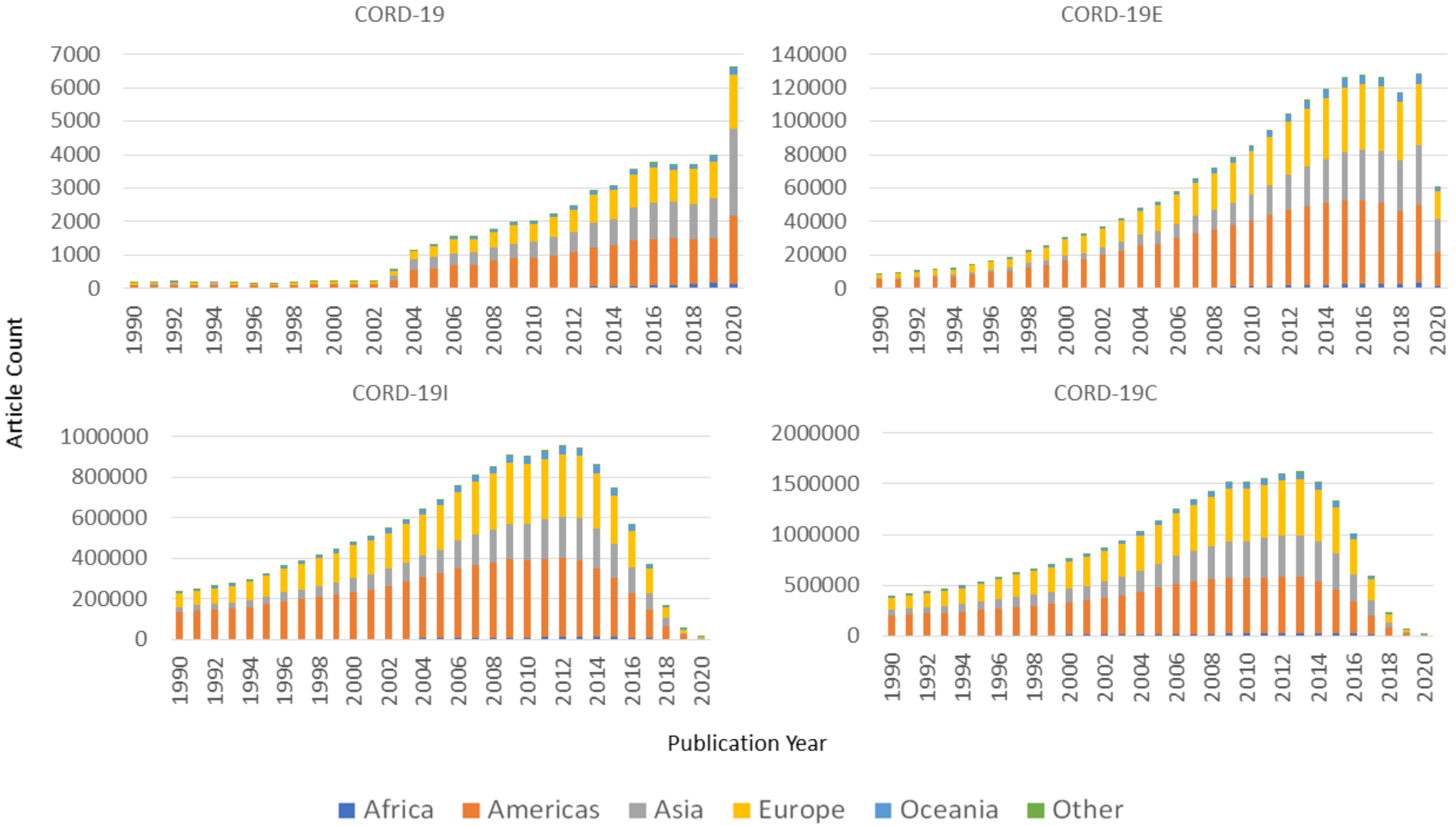
JOURNAL COVERAGE/RANKINGS: CORD-19 VS CORD-19E



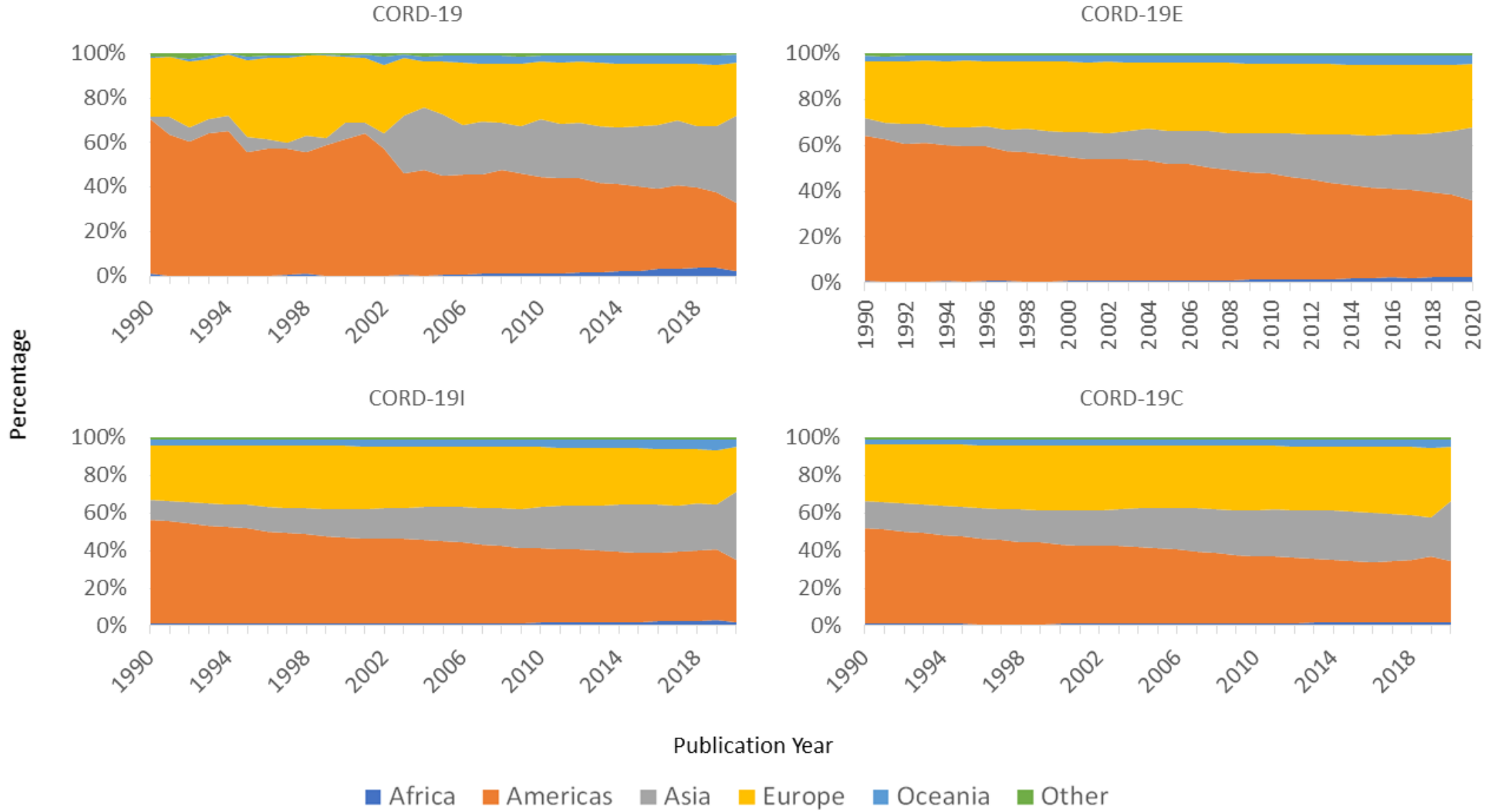
JOURNAL COVERAGE/RANKING: CORD-19 VS CORD-19I/C



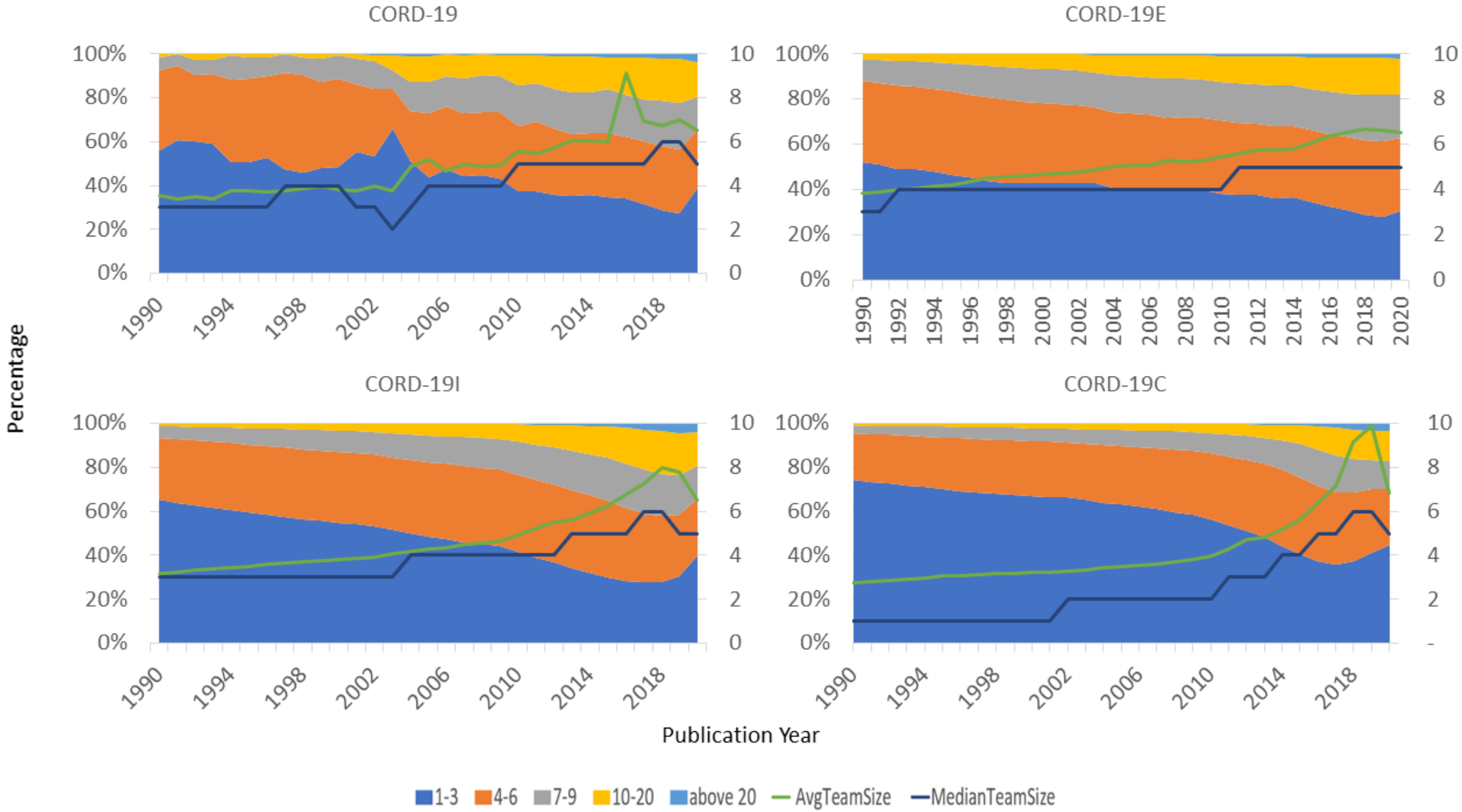
WHERE IS RESEARCH CONDUCTED



WHERE IS RESEARCH CONDUCTED



TEAM SIZE


CORD-19
CORD-19E
CORD-19I
CORD-19C
Percentage
Publication Year
1-3
4-6
7-9
10-20
above 20
AvgTeamSize
MedianTeamSize

CONCLUSIONS

1. Citations complement retrieval techniques
2. Analytics between CORD-19 and CORD-19E/I/C
   - Different: Topics, Age, Article/Journal impacts
   - Similar: Collaboration trends
3. Corpus expansion offers statistically smoother analytics
4. Closure graph findings support network science theories
   - Preferential attachments with fitness, discrete choice
5. Partial “inflection” closure is almost as good as full closure
6. Check out and build on our GitHub share https://aka.ms/magcord19mapping
7. Future studies
   - Not all citations are equal: *need citation classification*?
   - Probabilistic network traversal