video demo

End-User Web Scraping: Google Scholar Edition

Sarah Chasins

data scraping tool

input

demonstration of how to collect the first row of a relational dataset

From highly structured webpages

output

a script that collects the rest of the dataset

case study: Google Scholar data

vapnik	Statistical Learning Theory	1998	54228	VN Vapnik	Wiley-Interscience
vapnik	The Nature of Statistical Learning Theory	1995	53976	V Vapnik	Data mining and knowledge discovery
vapnik	Support-vector networks	1995	15513	C Cortes, V Vapnik	Machine learning 20 (3), 273-297
vapnik	A training algorithm for optimal margin classifiers	1992	6095	BE Boser, IM Guyon, VN Vapnik	Proceedings of the fifth annual workshop on Computational learning theory ...
vapnik	An introduction to variable and feature selection	2003	6059	I Guyon, A Elisseeff	The Journal of Machine Learning Research 3, 1157-1182
vapnik	Gene selection for cancer classification using support vector machines	2002	4058	I Guyon, J Weston, S Barnhill, V Vapnik	Machine learning 46 (1-3), 389-422
...	...	...	...	...	...

case study: Google Scholar data

current author	title	year	citations	authors	venue
vapnik	Statistical Learning Theory	1998	54228	VN Vapnik	Wiley-Interscience
vapnik	The Nature of Statistical Learning Theory	1995	53976	V Vapnik	Data mining and knowledge discovery
vapnik	Support-vector networks	1995	15513	C Cortes, V Vapnik	Machine learning 20 (3), 273-297
vapnik	A training algorithm for optimal margin classifiers	1992	6095	BE Boser, IM Guyon, VN Vapnik	Proceedings of the fifth annual workshop on Computational learning theory ...
vapnik	An introduction to variable and feature selection	2003	6059	I Guyon, A Elisseeff	The Journal of Machine Learning Research 3, 1157-1182
vapnik	Gene selection for cancer classification using support vector machines	2002	4058	I Guyon, J Weston, S Barnhill, V Vapnik	Machine learning 46 (1-3), 389-422
...	...	...	...	...	...

scale

authors limit

2000

papers per author limit

500

two central questions

did the tool generate a good script?

at what age do researchers peak?

did the tool generate a good script?

should we trust this data at all?

vapnik	Statistical Learning Theory	1998	54228	VN Vapnik	Wiley-Interscience
vapnik	The Nature of Statistical Learning Theory	1995	53976	V Vapnik	Data mining and knowledge discovery
vapnik	Support-vector networks	1995	15513	C Cortes, V Vapnik	Machine learning 20 (3), 273-297
vapnik	A training algorithm for optimal margin classifiers	1992	6095	BE Boser, IM Guyon, VN Vapnik	Proceedings of the fifth annual workshop on Computational learning theory ...
vapnik	An introduction to variable and feature selection	2003	6059	I Guyon, A Elisseeff	The Journal of Machine Learning Research 3, 1157-1182
vapnik	Gene selection for cancer classification using support vector machines	2002			

Title 1–20	Cited by	Year
Statistical Learning Theory VN Vapnik Wiley-Interscience	54369 *	1998
The Nature of Statistical Learning Theory V Vapnik Data mining and knowledge discovery	~~54116~~ *	1995
Support-vector networks C Cortes, V Vapnik Machine learning 20 (3), 273-297	15609	1995
A training algorithm for optimal margin classifiers BE Boser, IM Guyon, VN Vapnik Proceedings of the fifth annual workshop on Computational learning theory ...	6124	1992
An introduction to variable and feature selection I Guyon, A Elisseeff The Journal of Machine Learning Research 3, 1157-1182	6096	2003
Gene selection for cancer classification using support vector machines I Guyon, J Weston, S Barnhill, V Vapnik Machine learning 46 (1-3), 389-422	4076	2002
Estimation of dependences based on empirical data		

So checking up on the data afterwards is hard...

what do we expect?

2000 authors

up to 500 papers per author

what did we actually get?

rows: 157,159

what did we actually get?

rows: 157,159

unique authors: 1993

what did we actually get?

rows: 157,159

unique authors: 1993

what did we actually get?

rows: 157,159

unique authors: 1993

possible explanations:

1. tool doesn't work as well as I thought :((my problem)
2. data updates during scraping (problem inherent in long scraping tasks)
3. Scholar lists some authors twice (Scholar problem)
4. some authors share names (not a problem!)

what did we actually get?

rows: 157,159

unique authors: 1993

more thorough author analysis:

author names that appear separated by other author names:

Yves Deville : listed as author 183 and 191
Giovanni Pau : listed as author 355 and 1736
Henry Lin : listed as author 1024 and 1403
Fabrizio Messina : listed as author 1391 and 1396

authors whose citation counts jump in the middle of their runs:

Marco Ronchetti : listed as author 225 and 226
Joefon Jann : listed as author 810 and 811
Marcin Kubica : listed as author 1069 and 1070

remember papers were listed in order of decreasing citation count

Marco Ronchetti	Defects in Amorphous Solids: a Possible Approach	1984	ì	M Ronchetti	Computer Simulation in Physical Metallurgy, 129-143
Marco Ronchetti	Dynamical Properties of Classical Liquids and Liquid Mixtures	1984	ì	G Jacucci, M Ronchetti, W Schirmacher	Condensed Matter Research Using Neutrons, 139-161
Marco Ronchetti	Didattica per competenze: che supporto dalla tecnologia?		ì	S Giaffredo, M Ronchetti, A Valerio	
Marco Ronchetti	Insegnare l'informatica a non-informatici: emergenza annunciata		ì	S Giaffredo, L Mich, M Ronchetti	
Marco Ronchetti	Some considerations from ontological standpoint of modeling processes in the social domain		ì	A Ghosh, M Ronchetti, R Ferrario	
Marco Ronchetti	LEZIONI SUL TELEFONINO: PORTING IN AMBIENTE SYMBIAN		ì	M Ronchetti, J Stevovic	
Marco Ronchetti	Costruzione di un'interfaccia-utente per Lavagne Interattive Multimediali nel caso di simulazioni bidimensionali di fisica		ì	M Ronchetti, N Dorigatti	
Marco Ronchetti	A Service-Oriented Architecture for the NEEDLE (Next gEneration sEarch engine for Digital LibrariEs) Multimodal Search Engine		ì	M Ronchetti, MJN Krishnan, M Jarke	
Marco Ronchetti	Predizione contestuale di termini per fornire supporto a studenti con varie forme di disabilitì .		ì	A Zanella, M Ronchetti	
Marco Ronchetti	Spacetime: A Two Dimensions Search and Visualisation Engine Based on Linked Data		ì	M RONCHETTI, F VALSECCHI	
Marco Ronchetti	Dipartimento di Informatica e Telecomunicazioni Universitì degli Studi di Trento, 38050 Povo (Trento) Italy		ì	M Ronchetti	
Marco Ronchetti	Dipartirnento di Infolmatica e Studi Aziendali Universitli di Trento via F. Zeni 8, 1-38068 Rovereto (TN) ITALY		ì	G Kovacs, G Succi, F Baruchelli, M Ronchetti	
Marco Ronchetti	L'uso di video su Internet nella didattica universitaria.		ì	M Ronchetti	
Marco Ronchetti	Bond-orientational order in liquids and glasses	1983	1608	PJ Steinhardt, DR Nelson, M Ronchetti	Physical Review B 28 (2), 784
Marco Ronchetti	Icosahedral bond orientational order in supercooled liquids	1981	261	PJ Steinhardt, DR Nelson, M Ronchetti	Physical Review Letters 47 (18), 1297

aut
Yve
Giov
Hen
Fabi
aut
Mar
Joef
Mar

mber
s were
n order
easing
count

what did we actually get?

rows: 157,159

unique authors: 1,993

unique author runs: 2,000

splitting into runs based on new author or jump in citation count

what did we actually get?

what if the runs weren’t the first 2,000?

Scholar page at end of run confirms they really were the first 2,000

what did we actually get?

what if the runs weren't the first 2,000?

Scholar page at end of run confirms they really were the first 2,000

1. ~~tool doesn't work as well as I thought :((my problem)~~
2. data updates during scraping (problem inherent in long scraping tasks)
3. Scholar lists some authors twice (Scholar problem)
4. some authors share names (not a problem!)

what did we actually get?

can we eliminate explanation 2 also?

1. ~~tool doesn't work as well as I thought :((my problem)~~
2. data updates during scraping (problem inherent in long scraping tasks)
3. Scholar lists some authors twice (Scholar problem)
4. some authors share names (not a problem!)

what did we actually get?

label:computer_science marco ronchetti

Marco Ronchetti
Università di Trento
Verified email at unitn.it
Cited by 3075
e-learning computer science physics

Marco Ronchetti
Università di Trento
Verified email at unitn.it
Cited by 3075
e-learning computer science physics

Dates and citation counts are estimated and are determined automatically by a computer program.

what did we actually get?

label:computer_science yves deville

Yves Deville

Professor of Computer Science, Université catholique de Louvain, ICTEAM, EPL
Verified email at uclouvain.be
Cited by 3624

Computer Science Artificial Intelligence Constraints Optimization

Yves Deville

Professor of Computer Science, University of Louvain
Cited by 3536

Computer Science Artificial Intelligence Constraints Optimization

Dates and citation counts are estimated and are determined automatically by a computer program.

what did we actually get?

can we eliminate explanation 2 also?

1. ~~tool doesn't work as well as I thought :((my problem)~~
2. ~~data updates during scraping (problem inherent in long scraping tasks)~~
3. Scholar lists some authors twice (Scholar problem)
4. some authors share names (not a problem!)

I suspect 3 is true cause for all seven, but can't be positive.

what did we actually get?

David S. Johnson

Follow

Visiting Professor, Columbia University Computer Science Department
Algorithms, computer science, optimization, traveling salesman problem, bin packing
Verified email at research.att.com - Homepage

Title 1–20	Cited by	Year
Computers and intractability MR Garey, DS Johnson wh freeman	51116	2002
The traveling salesman problem: a guided tour of combinatorial optimization EL Lawler, JK Lenstra, AHGR Kan, DB Shmoys Wiley	3264	1985
Approximation algorithms for combinatorial problems DS Johnson Proceedings of the fifth annual ACM symposium on Theory of computing, 38-49	2193	1973
Some simplified< i> NP</i>-complete graph problems MR Garey, DS Johnson, L Stockmeyer Theoretical computer science 1 (3), 237-267	1914	1976
The complexity of flowshop and jobshop scheduling MR Garey, DS Johnson, R Sethi Mathematics of operations research 1 (2), 117-129	1820	1976
Optimization by simulated annealing: An experimental evaluation; part I, graph partitioning DS Johnson, CR Aragon, LA McGeoch, C Schevon Operations research 37 (6), 865-892	1401	1989
Unit disk graphs BN Clark, CJ Colbourn, DS Johnson Annals of Discrete Mathematics 48, 165-177	1165	1991
The traveling salesman problem: A case study in local optimization DS Johnson, LA McGeoch Local search in combinatorial optimization 1, 215-310	1034	1997
Approximation algorithms for bin packing: A survey EG Coffman Jr, MR Garey, DS Johnson Approximation algorithms for NP-hard problems, 46-93	952	1996

David S. Johnson

Follow

Visiting Professor, Columbia University Computer Science Department
Algorithms, computer science, optimization, traveling salesman problem, bin packing
Verified email at research.att.com - Homepage

Title 141–160	Cited by	Year
Red/Infrared Observations of WOLF: 424AB-are the Components Substellar TJ Henry, DS Johnson, DW McCarthy Jr, JD Kirkpatrick Astronomy and Astrophysics 254, 116	13	1992
Computers and Intractability, a Guide to the Theory of NP-Completeness," Freeman, San Francisco MR Garey, D JOHNSON to appear	13 *	1979
Wedding dress across cultures Berg	12	2003
Data Structures, Near Neighbor Searches, and Methodology: Fifth and Sixth DIMACS Implementation Challenges: Papers Related to the DIMACS Challenge on Dictionaries and Priority Queues (1995-1996) and the DIMACS Challenge on Near Neighbor Searches (1998-1999) American Mathematical Soc.	12	2002
Neural network implementation using a single MOST per synapse DE Johnson, JS Marsland, W Eccleston Neural Networks, IEEE Transactions on 6 (4), 1008-1011	12	1995
Hand and foot control system for an off-highway implement JM Moffitt, ML Morris, DE Johnson US Patent 5,197,347	12 *	1993
Nudist Society: An Authoritative, Complete Study of Nudism in America WE Hartman, M Fithian, D Johnson Crown Publishers	12	1970
Disjoint-Path Facility Location: Theory and Practice. L Breslau, I Diakonikolas, NG Duffield, Y Gu, MT Hajiaghayi, DS Johnson, ... ALENEX, 60-74	11	2011
Dress sense: emotional and sensory experiences of the body and clothes DC Johnson, HB Foster Berg Publishers	11	2007

papers per author

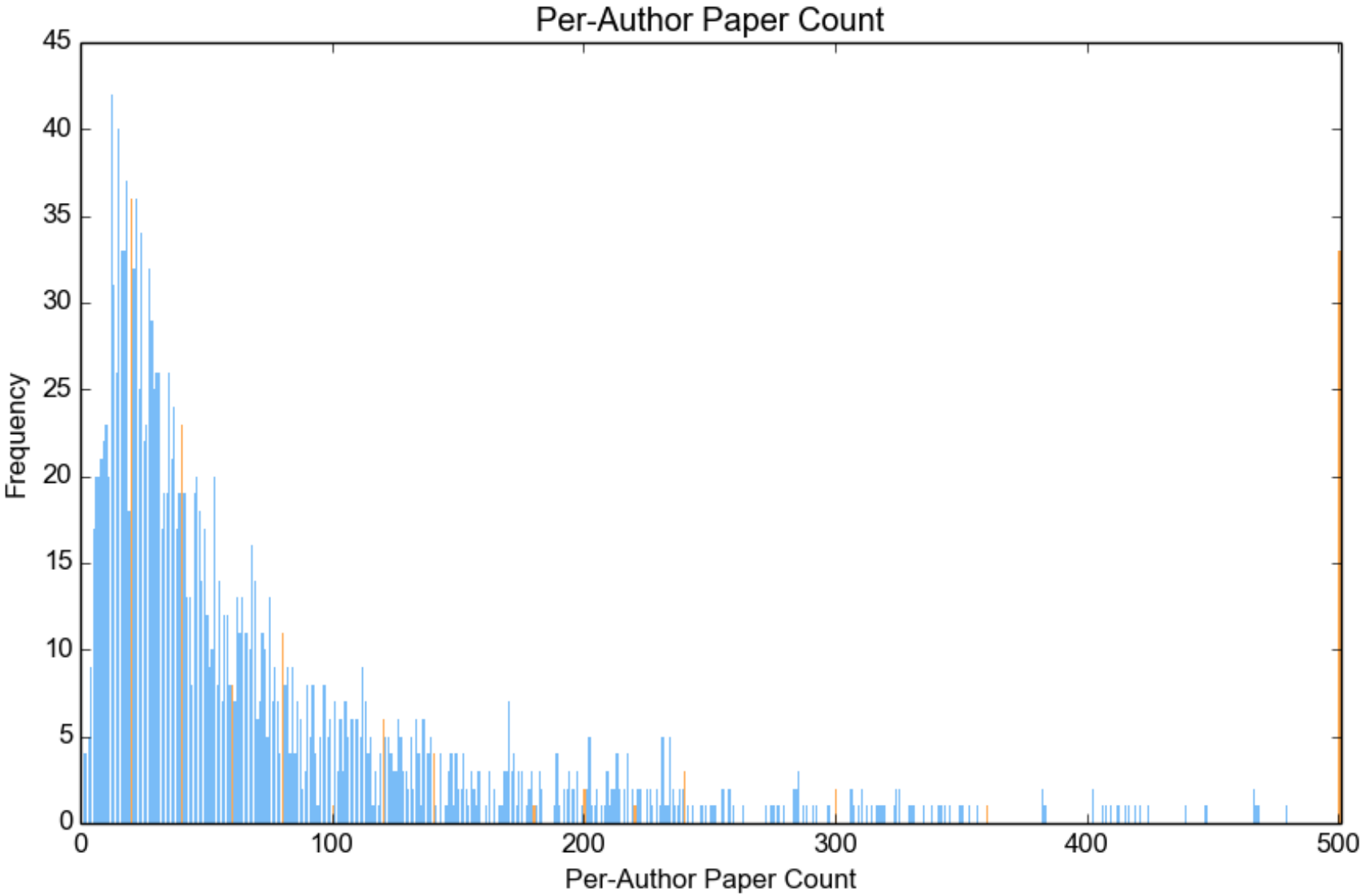
what we expect to see

many authors with few papers
a few authors with many papers
spike around 500, from truncation

what we don’t want to see

spikes around multiples of 20

papers per author



papers per author

one paper authors?

turns out, yes

Anuphan Sutthimarn

suansunandha rajabhat university

computer science

Verified email at ssru.ac.th - Homepage

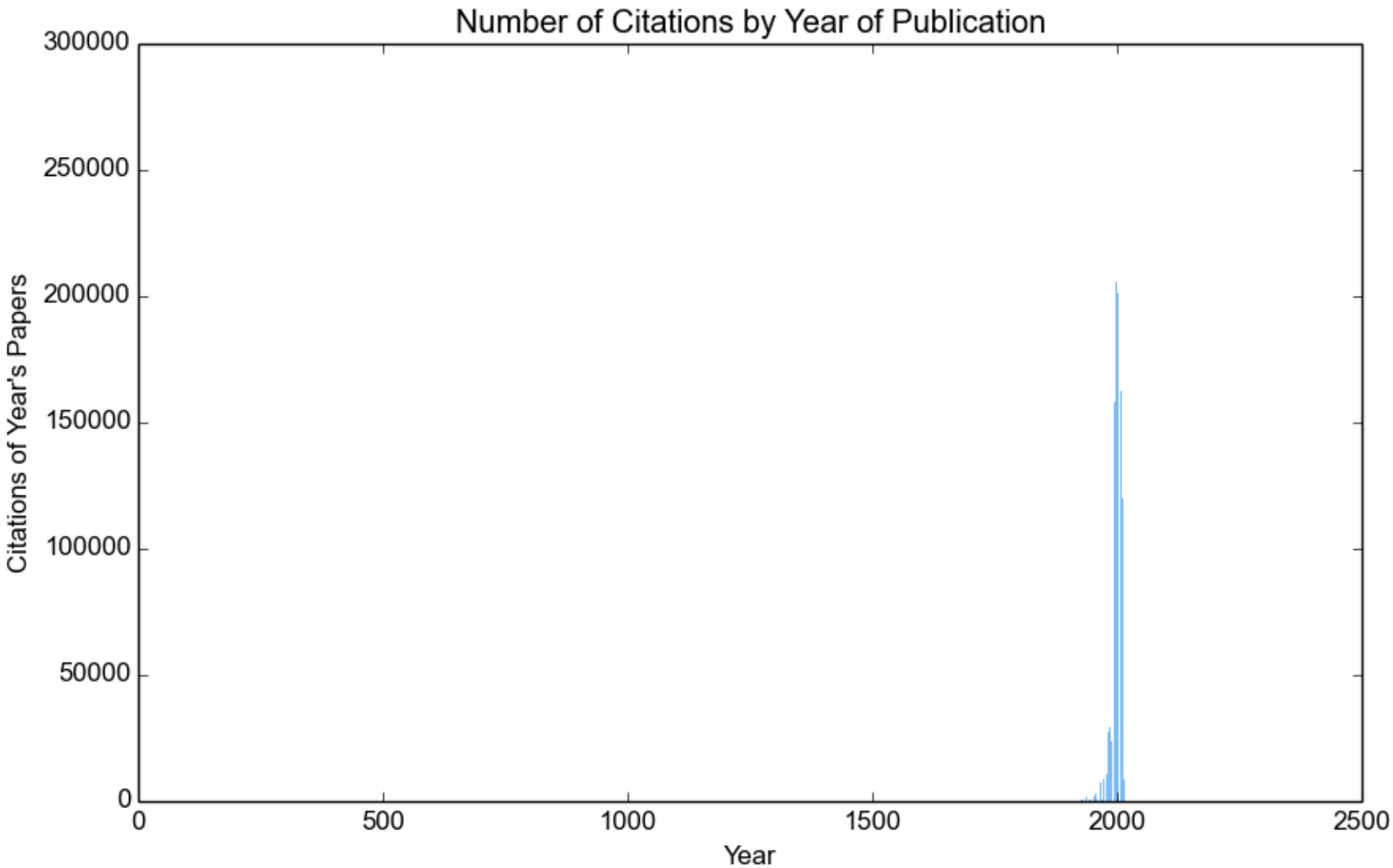
Follow

Title 1–1	Cited by	Year
Novel dark-bright optical solitons conversion system and power amplification K Sarapat, N Pornsuwancharoen, N Sangwara, K Srinuanjan, ... Optical Engineering 48 (4), 045004-045004-7	65	2009

Show more

at what age do researchers peak?

citations by year



citations by year

Celine Fabry (Bursztein)

Follow

No affiliation

Computer Science, Structural biology

Verified email at celine.im - Homepage

Title 1–14	Cited by	Year
A quasi-atomic model of human adenovirus type 5 capsid C Fabry, M Rosa-Calatrava, JF Conway, C Zubieta, S Cusack, ... The EMBO journal 24 (9), 1645-1654	120	2005
How good are humans at solving CAPTCHAs? a large scale evaluation E Bursztein, S Bethard, C Fabry, JC Mitchell, D Jurafsky Security and Privacy (SP), 2010 IEEE Symposium on, 399-413	104	2010
The failure of noise-based non-continuous audio captchas E Bursztein, R Beauxis, H Paskov, D Perito, C Fabry, J Mitchell Security and Privacy (SP), 2011 IEEE Symposium on, 19-31	38	2011
Structure of the dodecahedral penton particle from human adenovirus type 3 P Fuschiotti, G Schoehn, P Fender, CMS Fabry, EA Hewat, J Chroboczek, ... Journal of molecular biology 356 (2), 510-520	35	2006
Three-dimensional structure of canine adenovirus serotype 2 capsid G Schoehn, M El Bakkouri, CMS Fabry, O Billet, LF Estrozi, L Le, ... Journal of virology 82 (7), 3192-3203	30	2008
An Archaeal Peptidase Assembles into Two Different Quaternary Structures A TETRAHEDRON AND A GIANT OCTAHEDRON G Schoehn, FMD Vellieux, MA Dura, V Receveur-Bréchot, CMS Fabry, ... Journal of Biological Chemistry 281 (47), 36327-36337	25	2006
The C-terminal domains of adenovirus serotype 5 protein IX assemble into an antiparallel structure on the facets of the capsid CMS Fabry, M Rosa-Calatrava, C Moriscot, RWH Ruigrok, P Boulanger, ... Journal of virology 83 (2), 1135-1139	19	2009
and G. Schoehn. 2005. A quasi-atomic model of human adenovirus type 5 capsid CM Fabry, M Rosa-Calatrava, JF Conway, C Zubieta, S Cusack, ... EMBO J 8 (24), 1645-1654	10	7

no future dates, though...

citations by year

papers removed for having no year information

14,115 (9.0%)

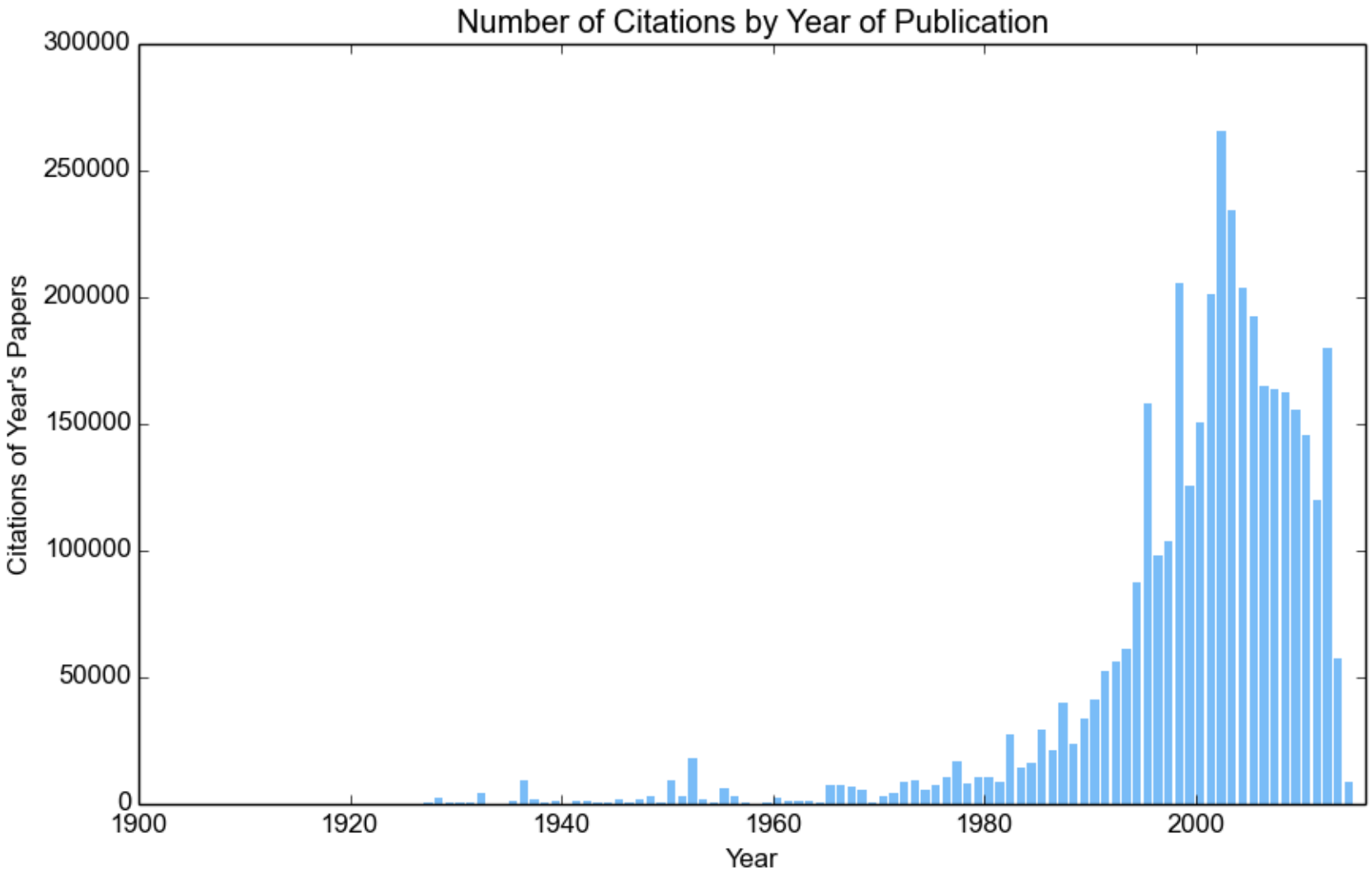
papers removed for being more than 50 years from author mean

169 (0.1%)

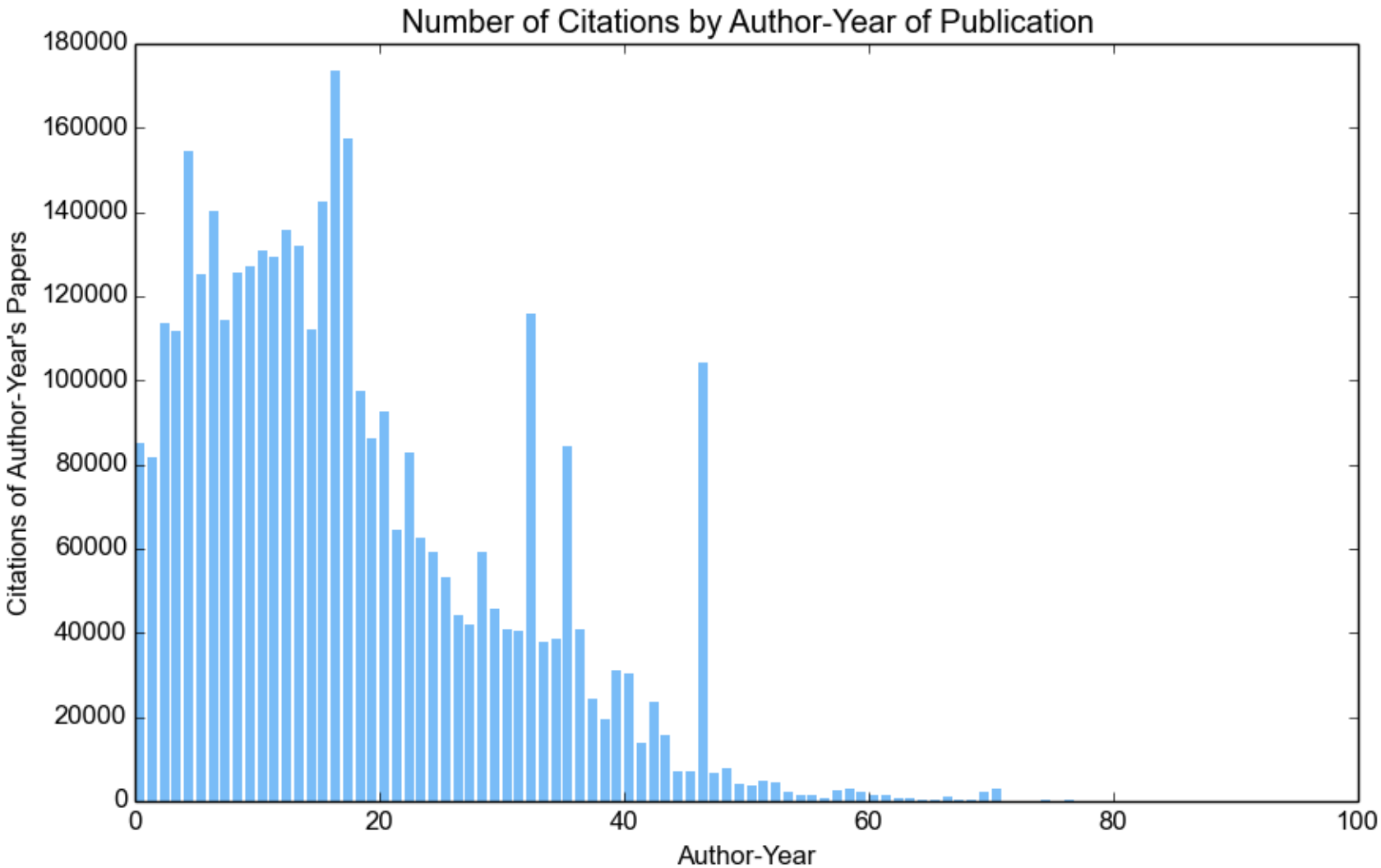
papers remaining

142,875 (90.9%)

citations by year



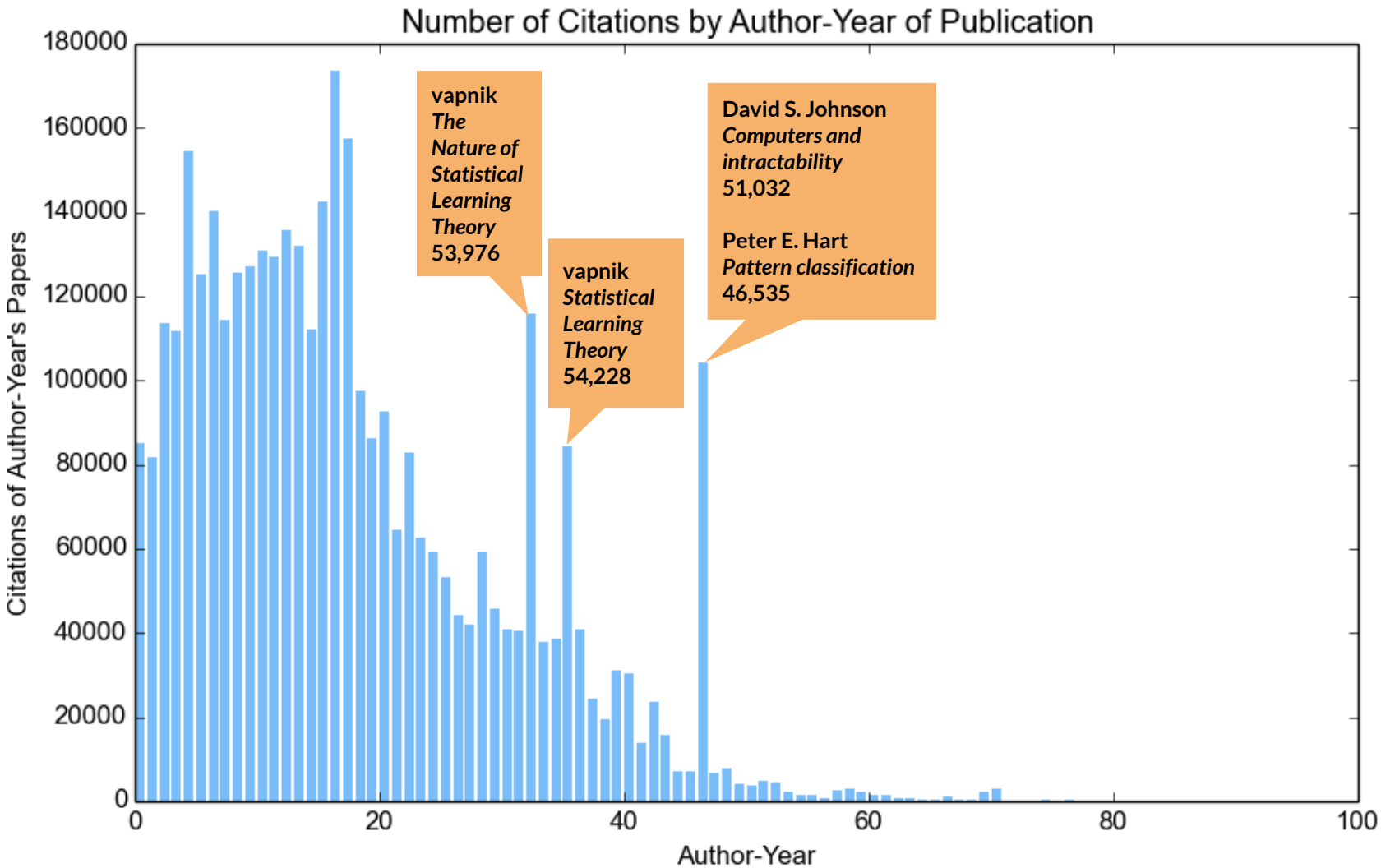
citations by author-year



citations by author-year

but this allows a few authors with high citation counts to skew results

citations by author-year



citations by author-year

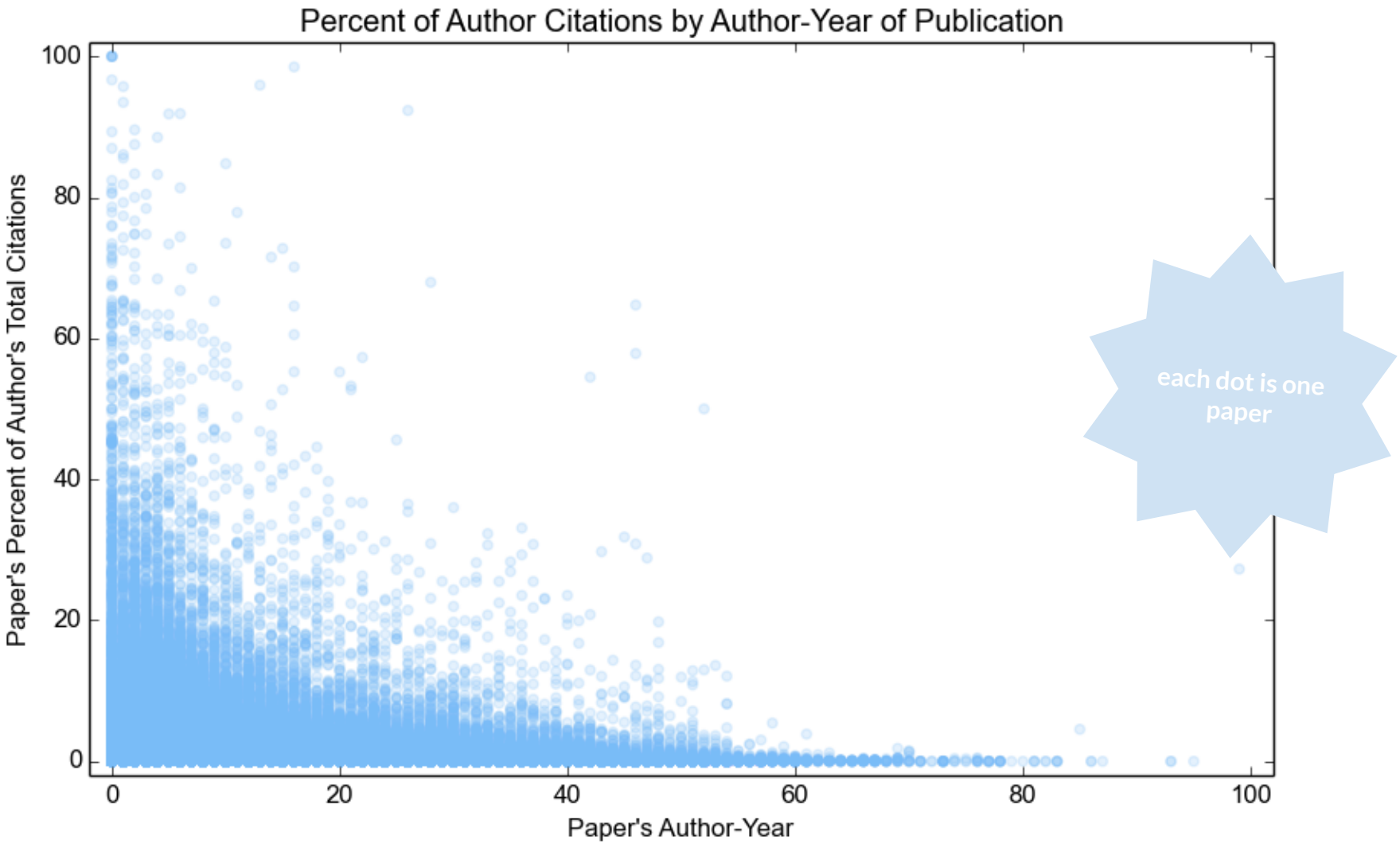
but this allows a few authors with high citation counts to skew results

alternatives

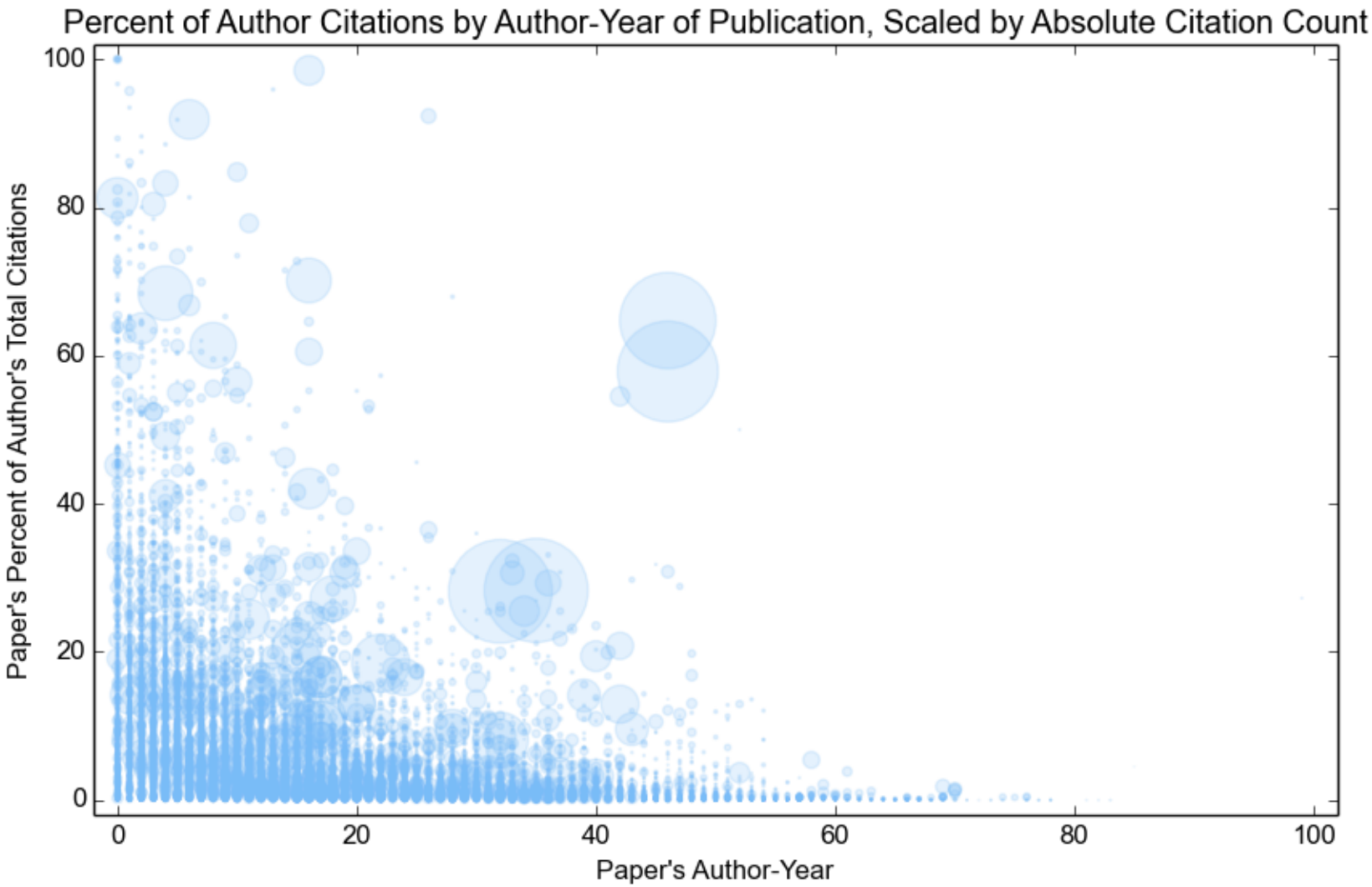
authors' percent citations by year

authors' highest cited paper years

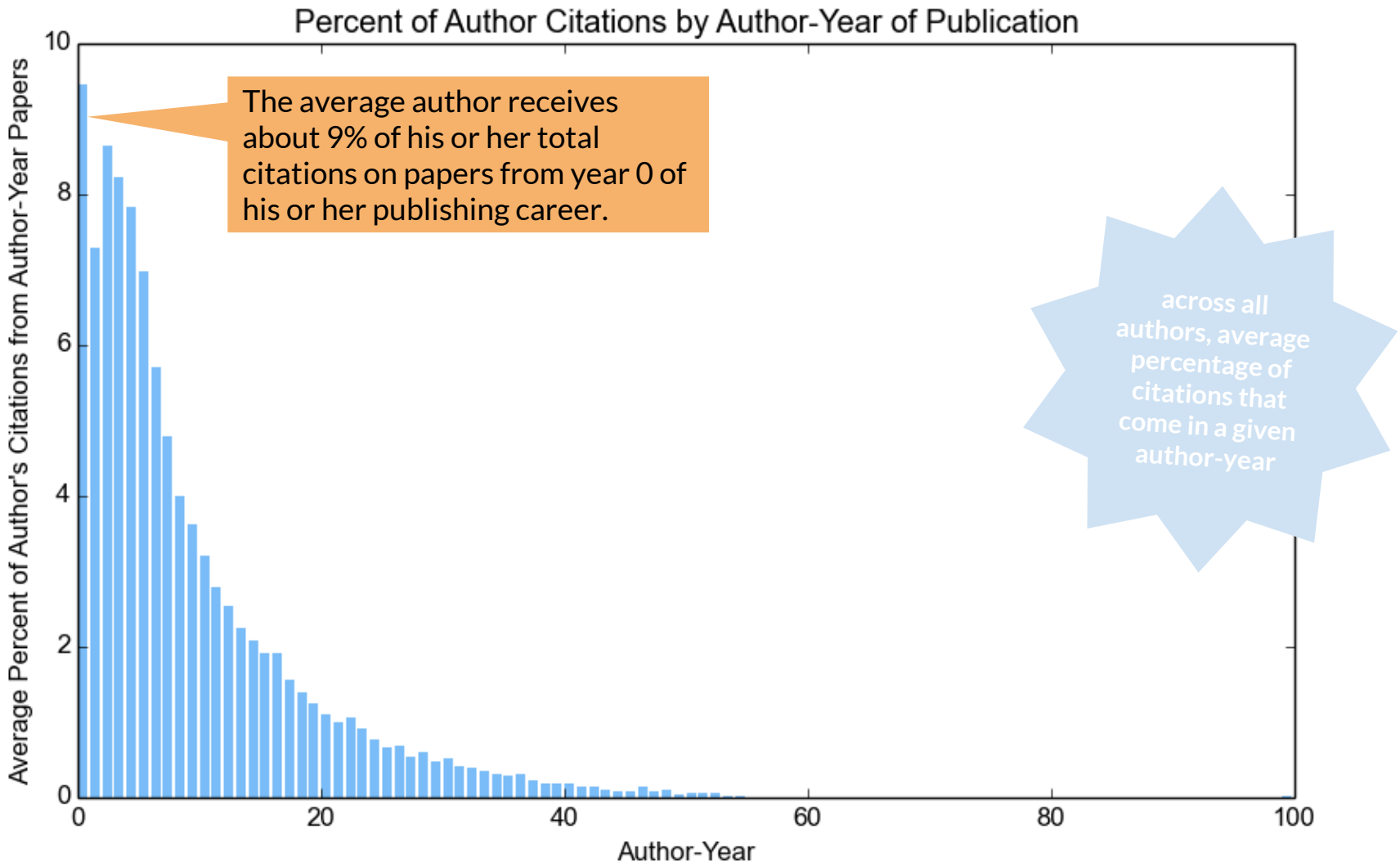
citations by author-year



citations by author-year



citations by author-year



citations by author-year

but this puts extra weight on early papers because some authors have short careers

for authors with 1 paper, 100% of citations in year 0...

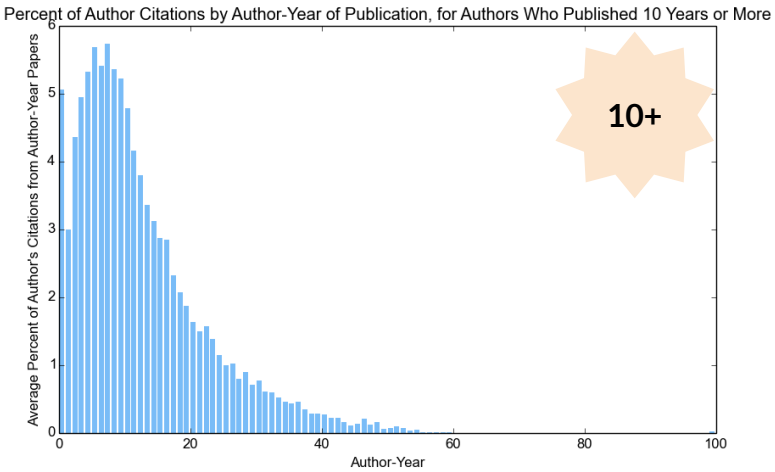
citations by author-year

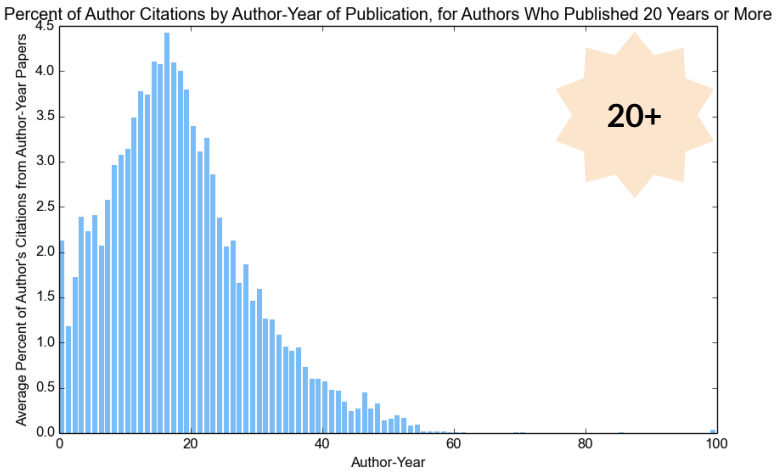
citations by author-year

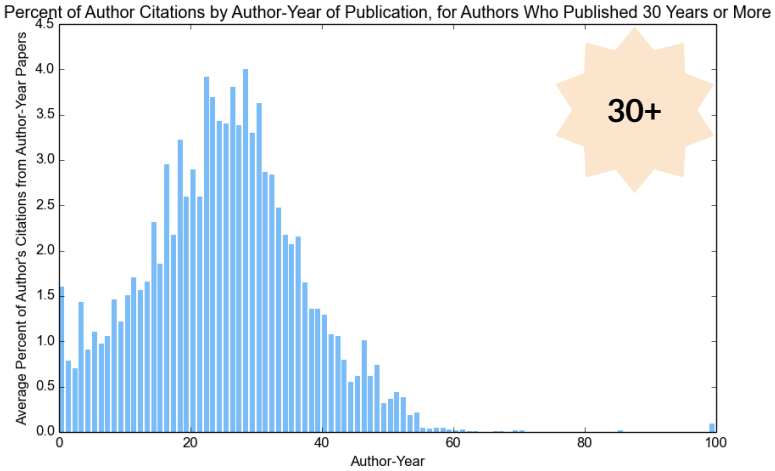
citations by author-year

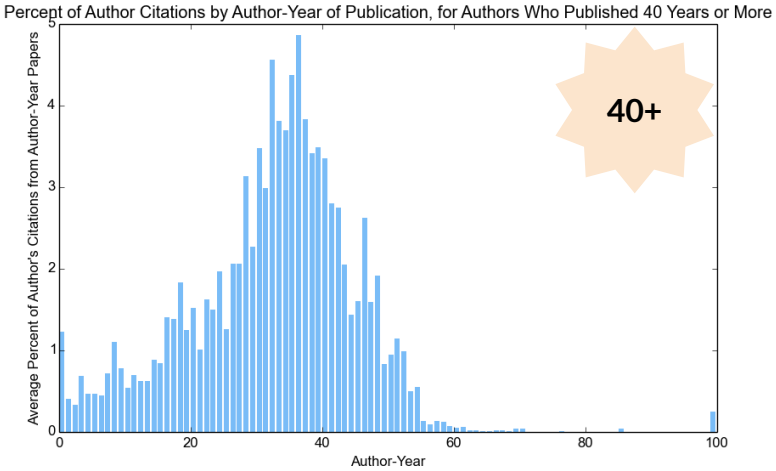
citations by author-year

citations by author-year

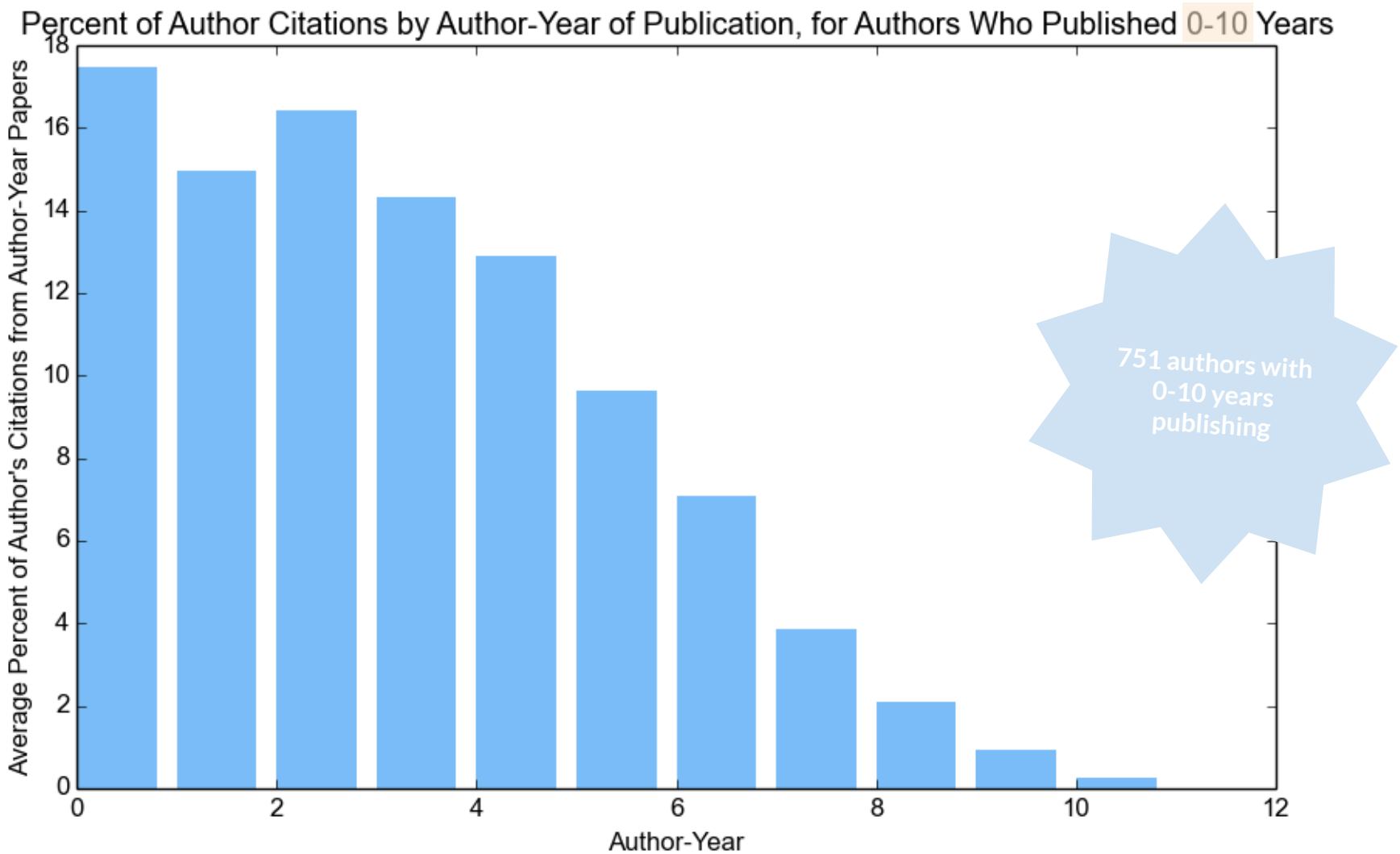




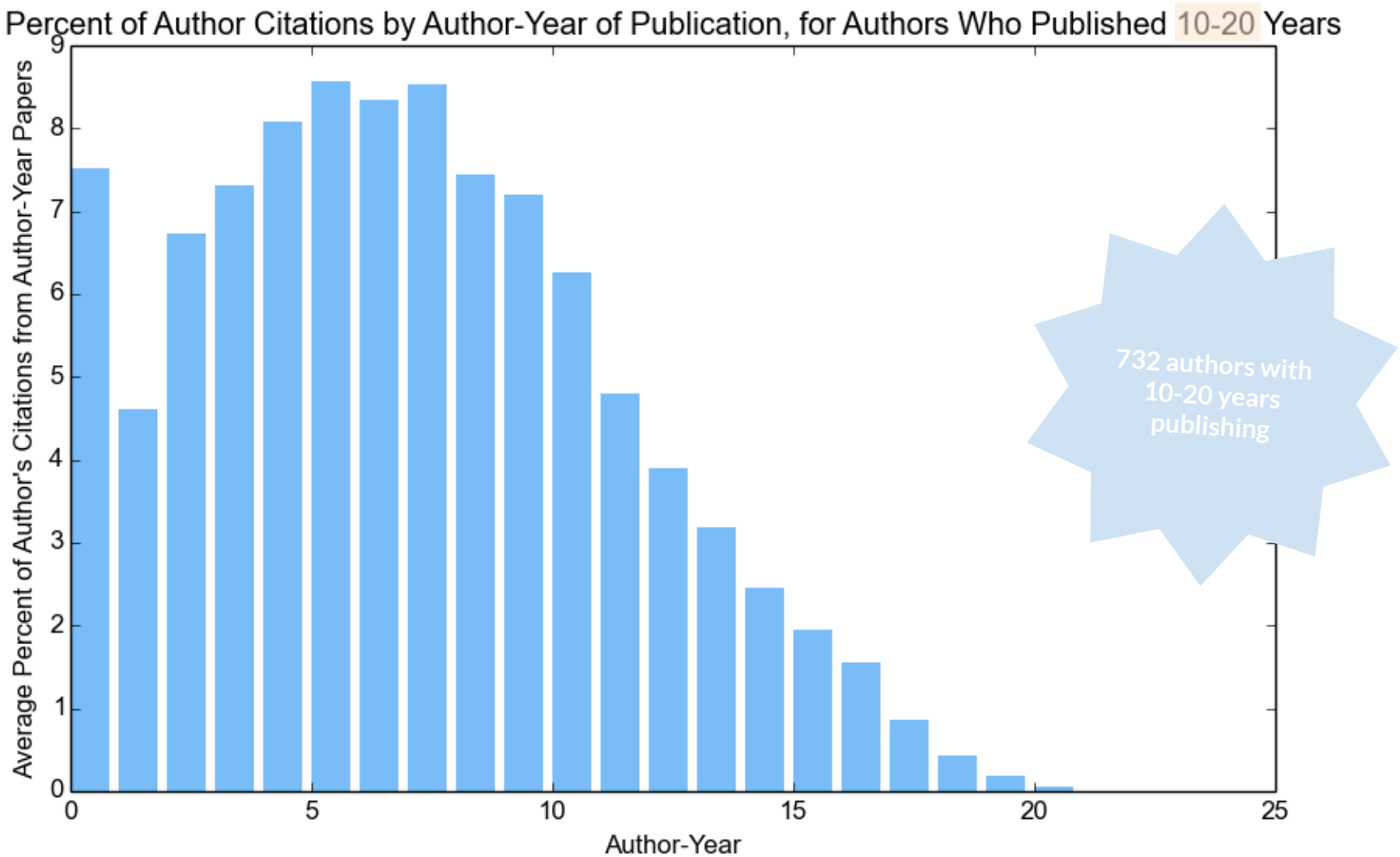




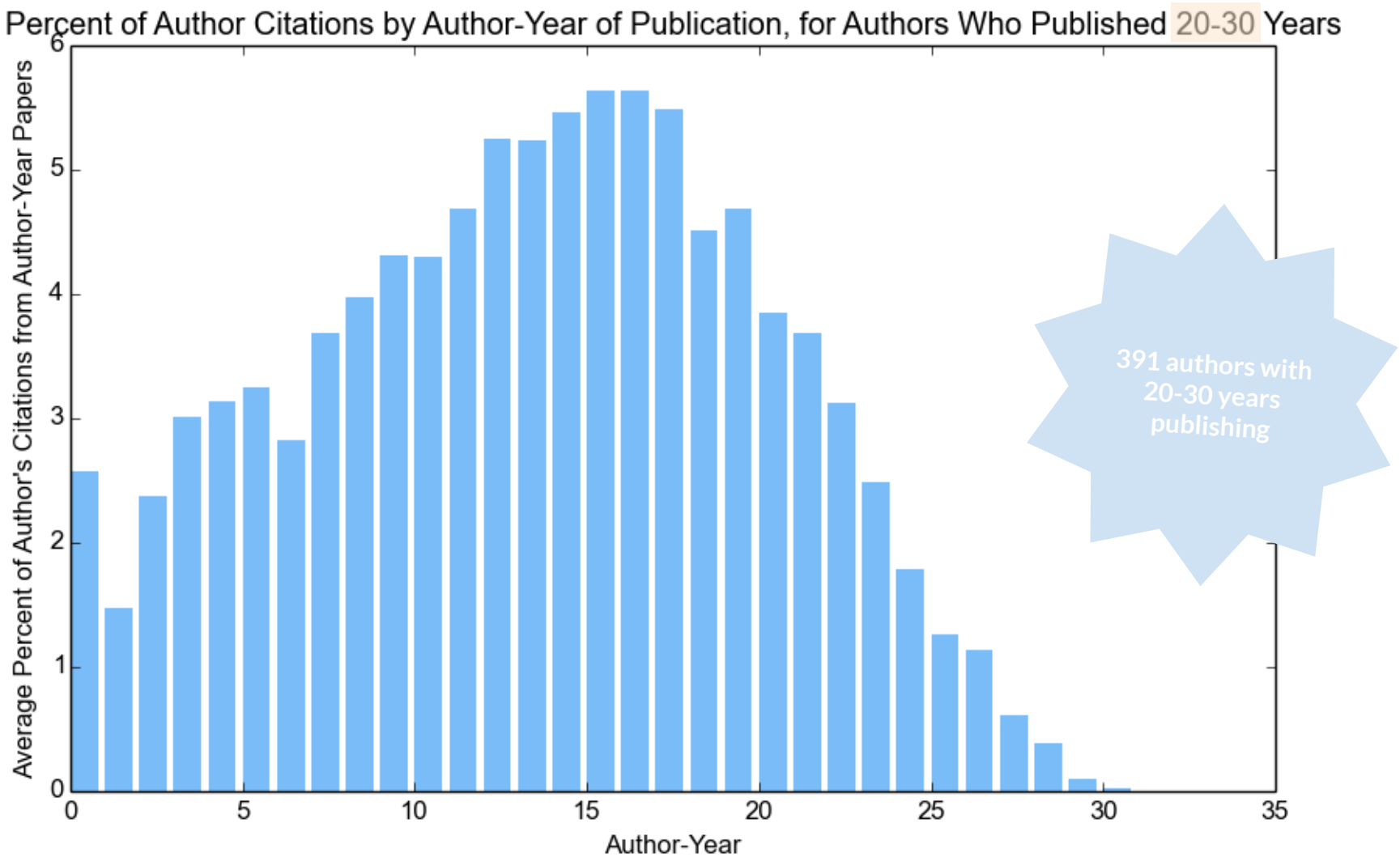
citations by author-year



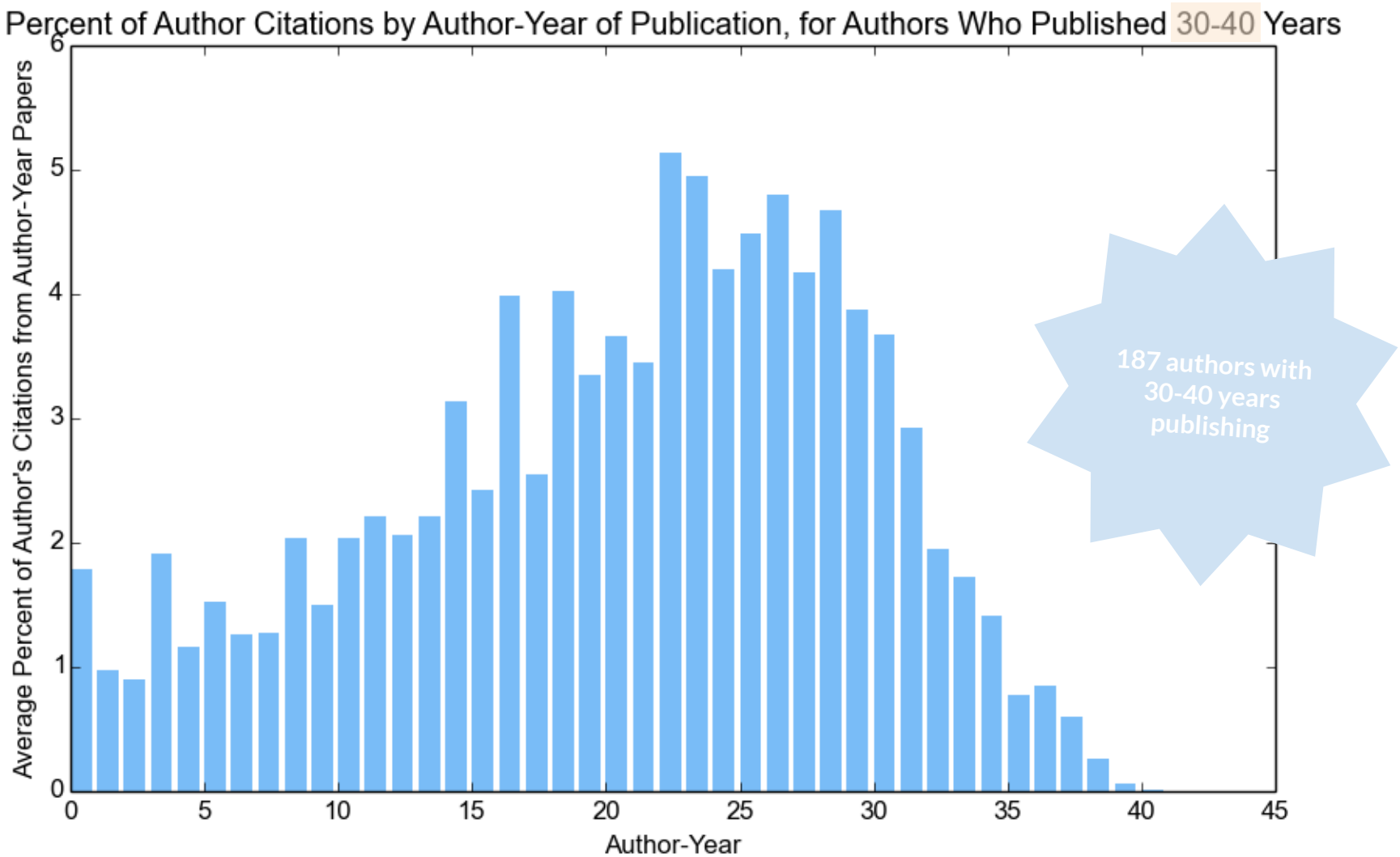
citations by author-year



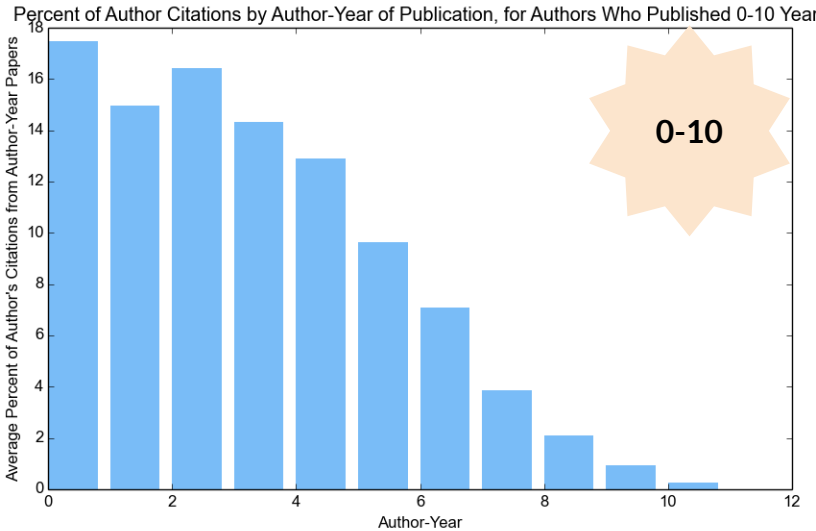
citations by author-year

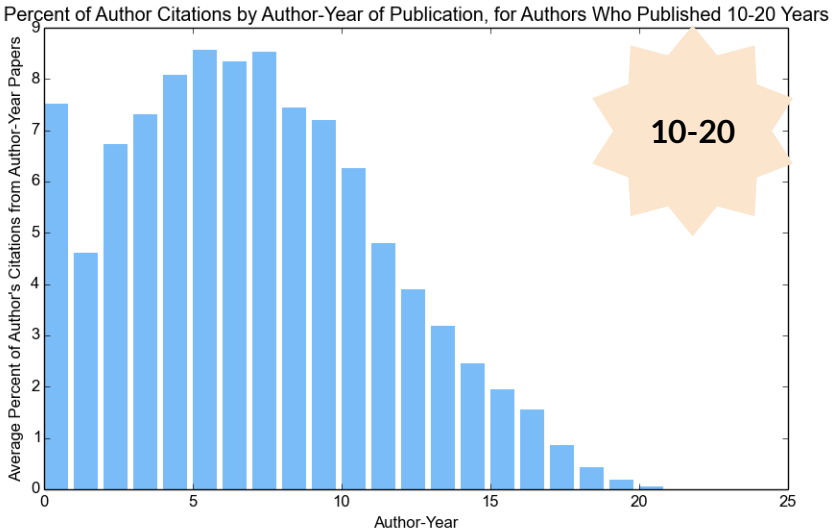


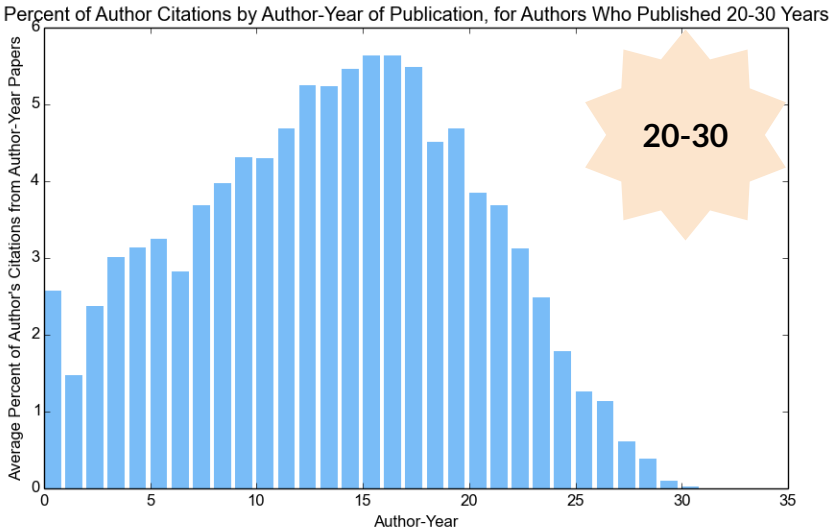
citations by author-year

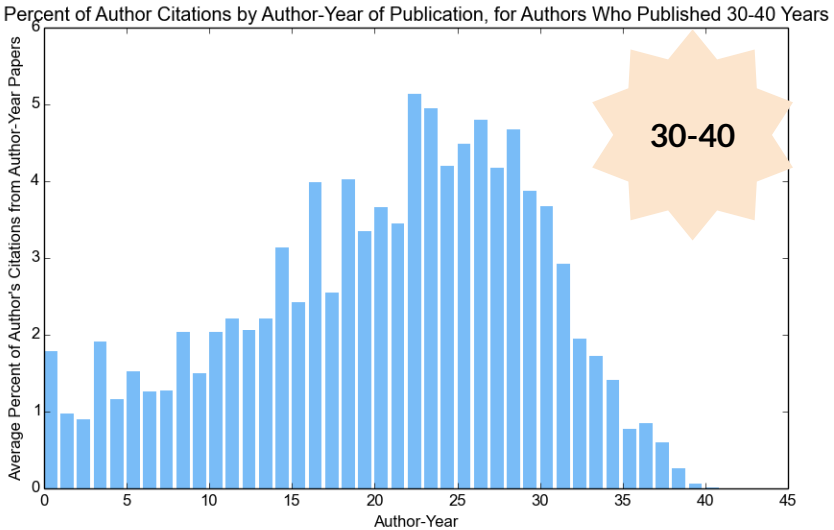


citations by author-year

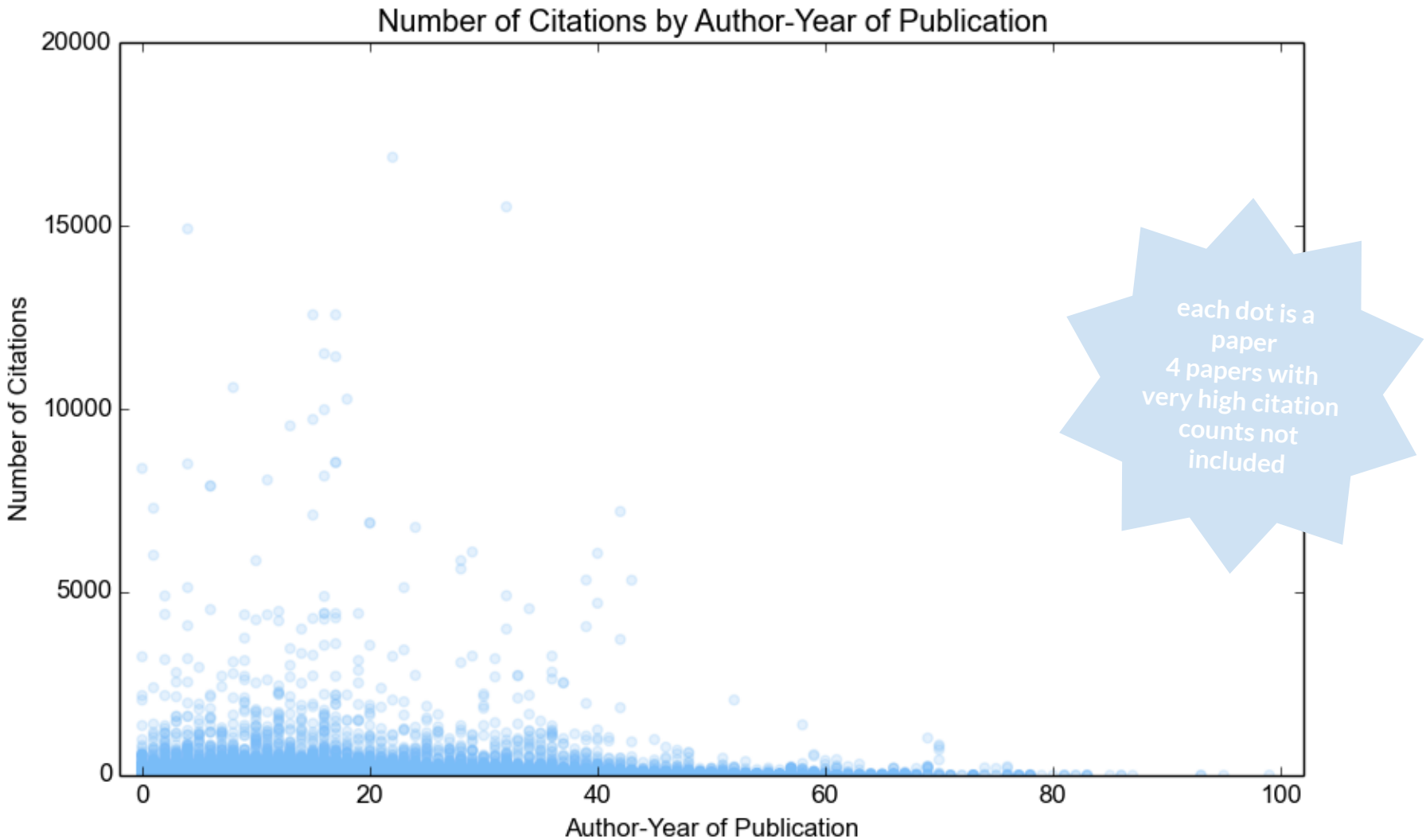




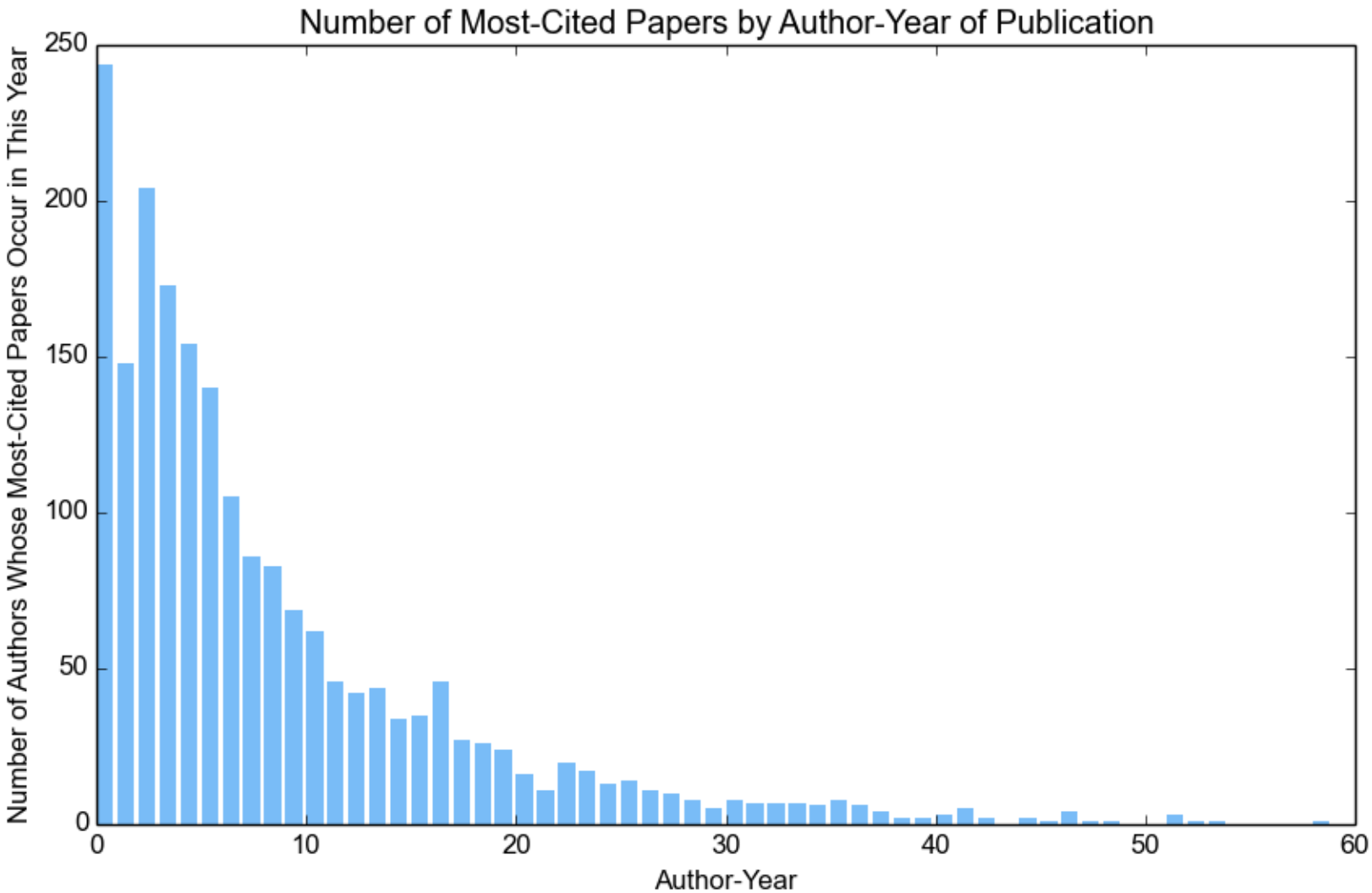




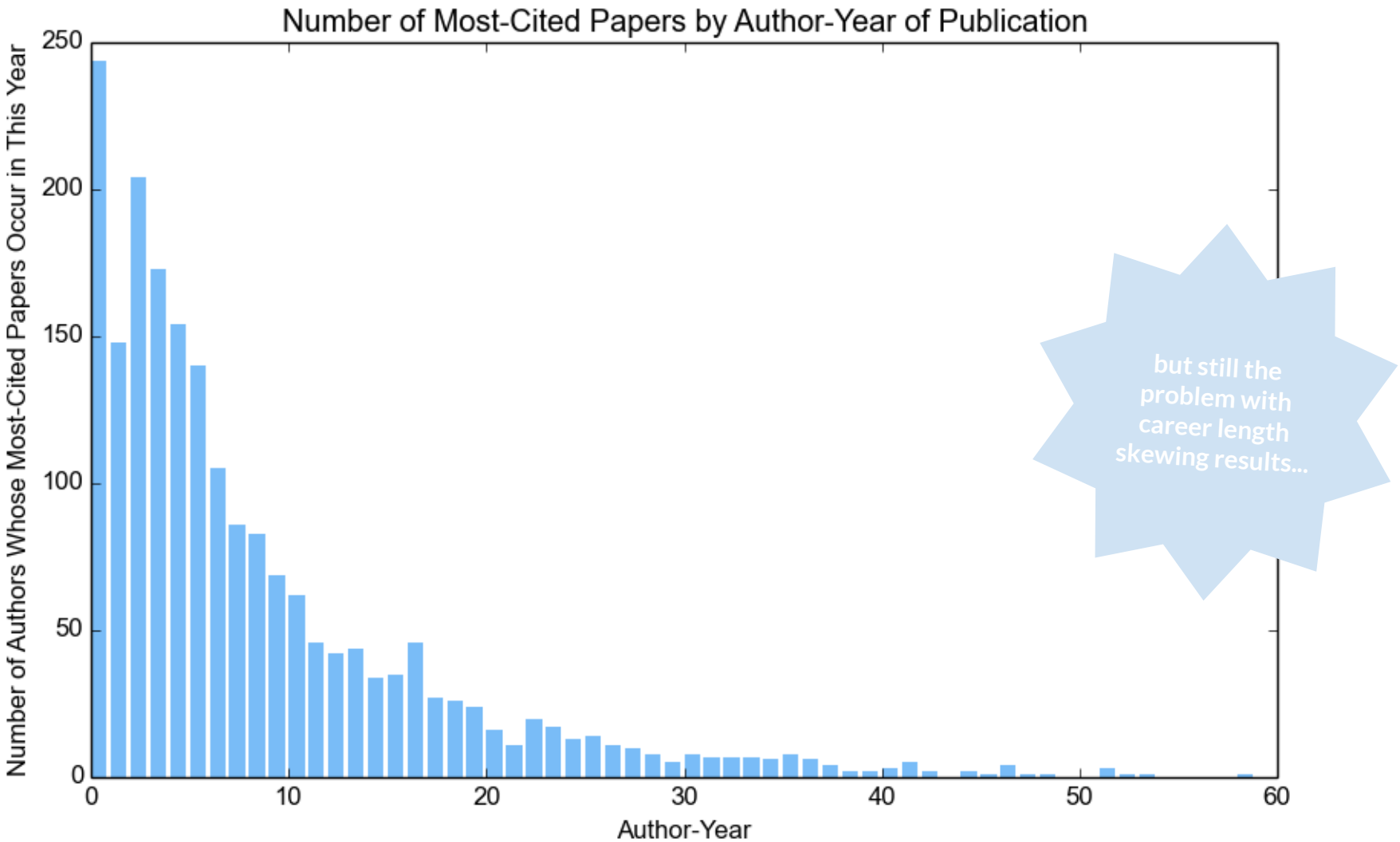
citations by author-year



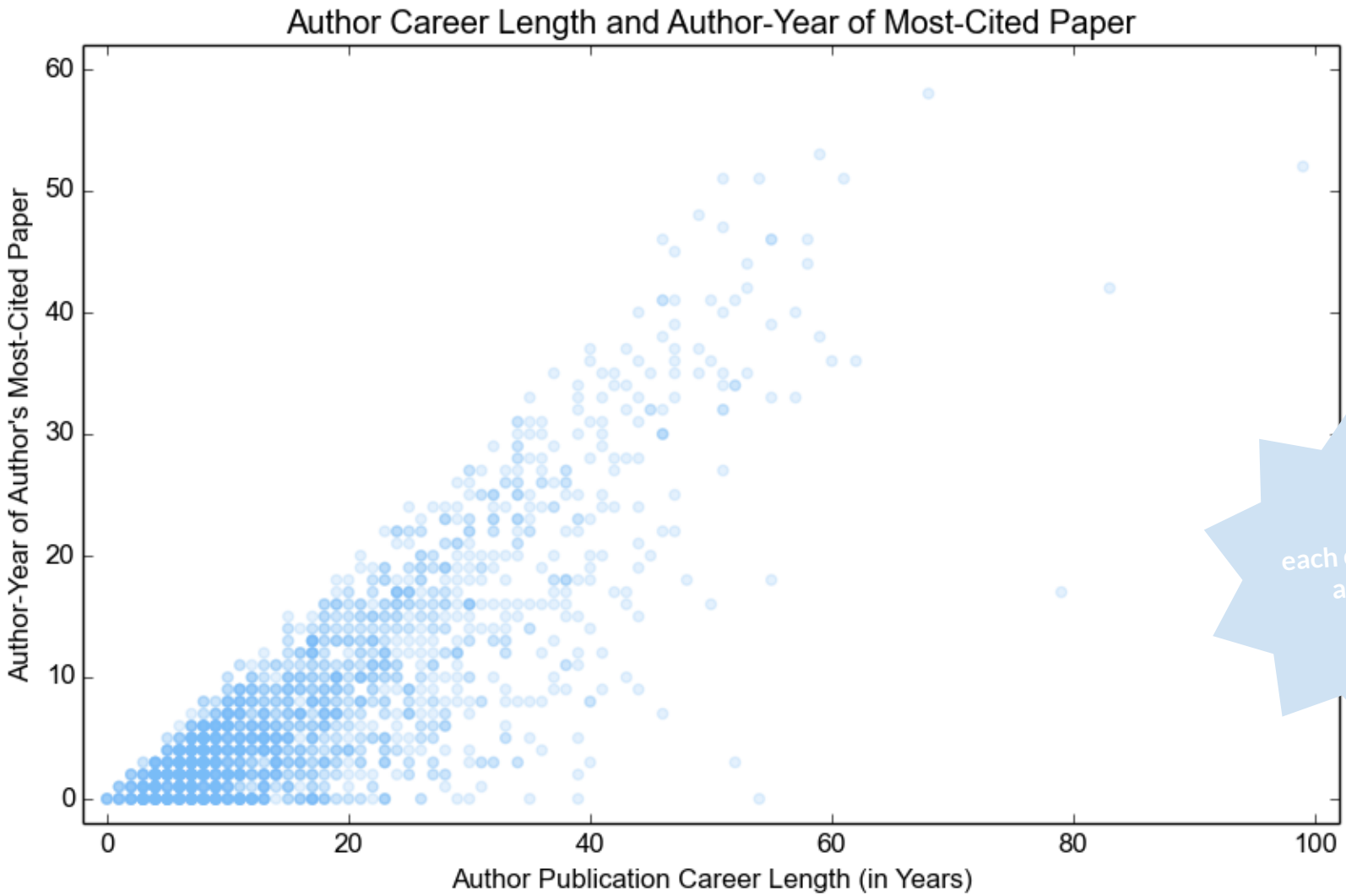
most-cited papers



most-cited papers

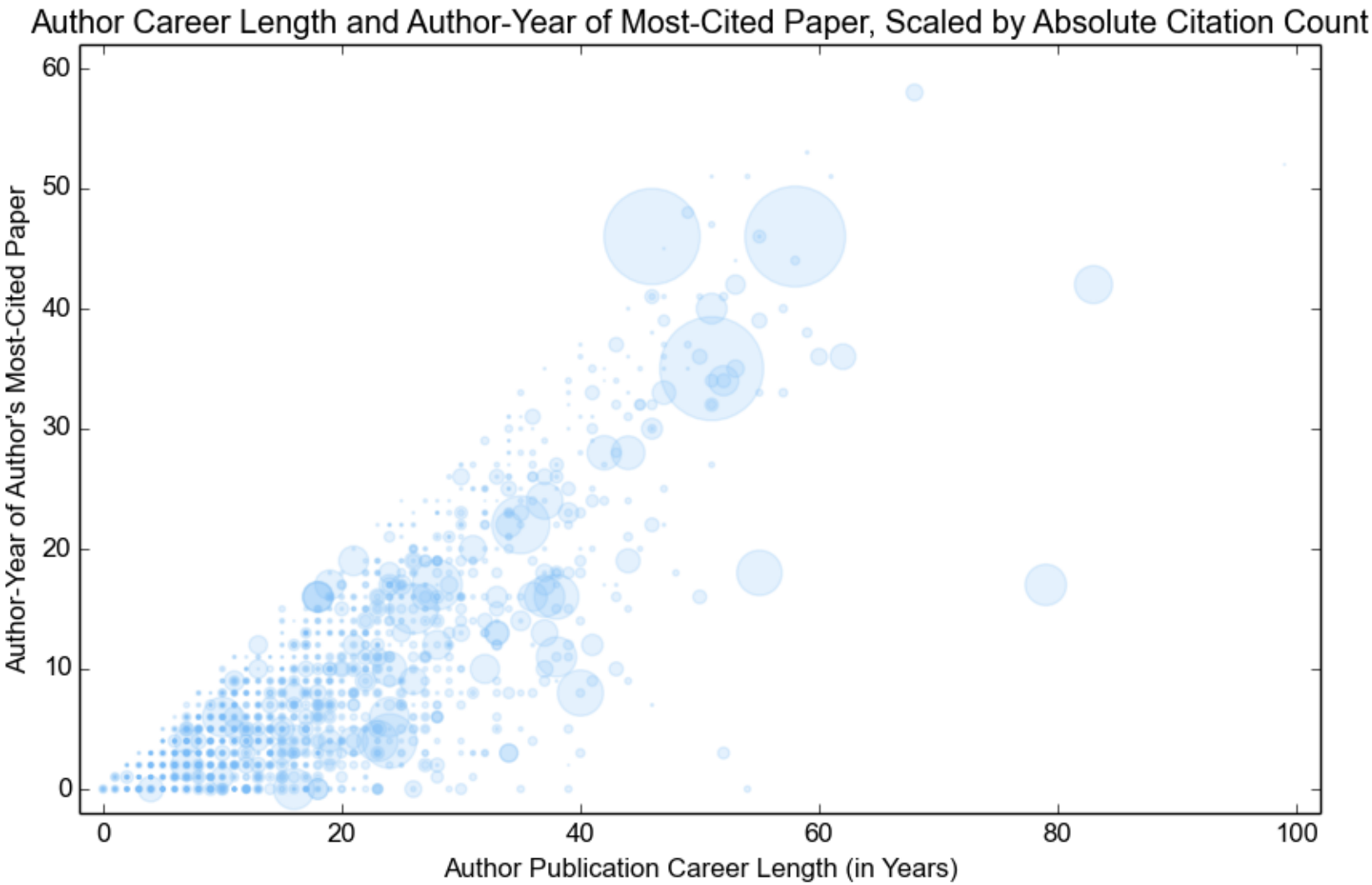


most-cited papers

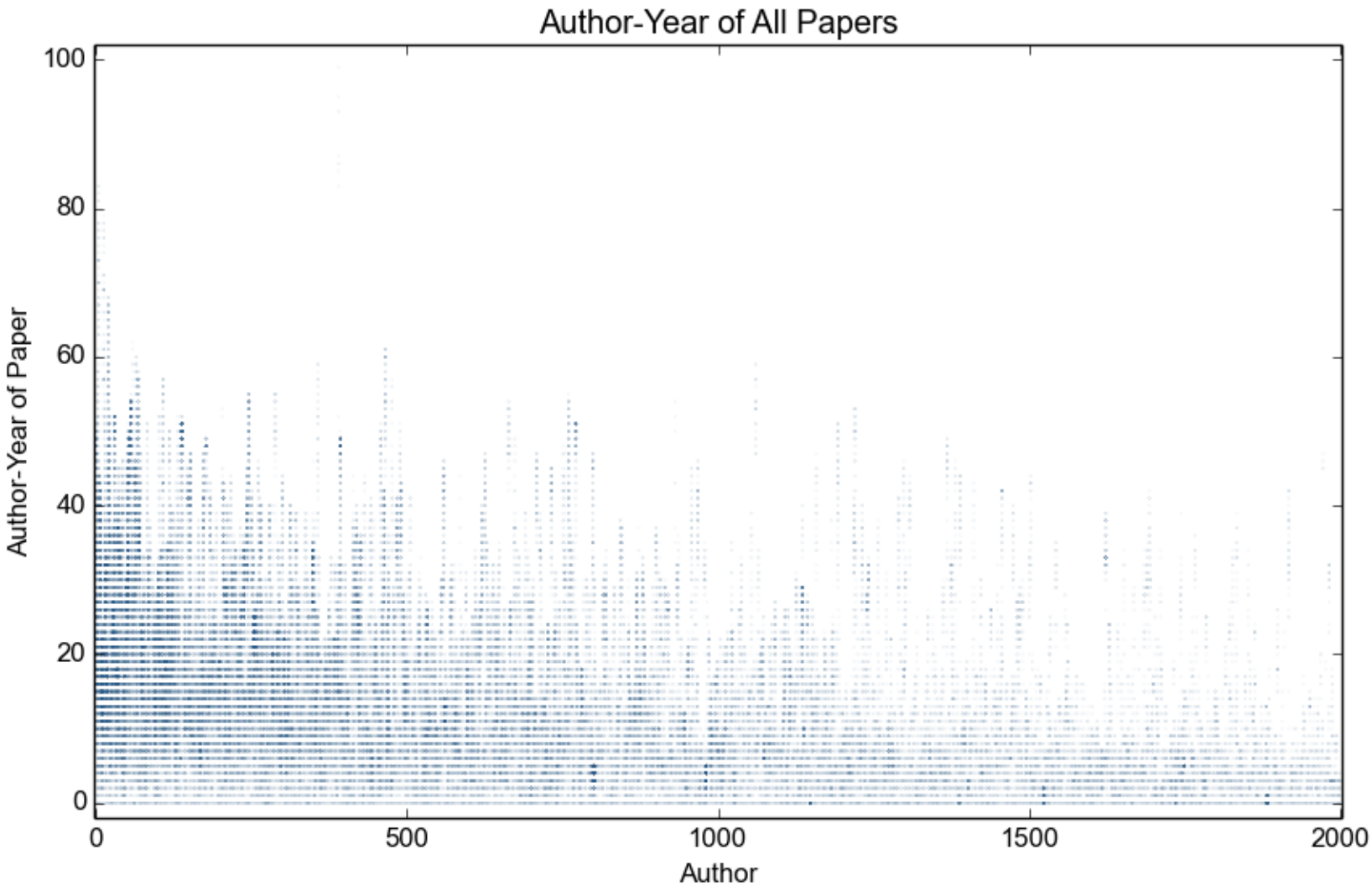


each dot is one author

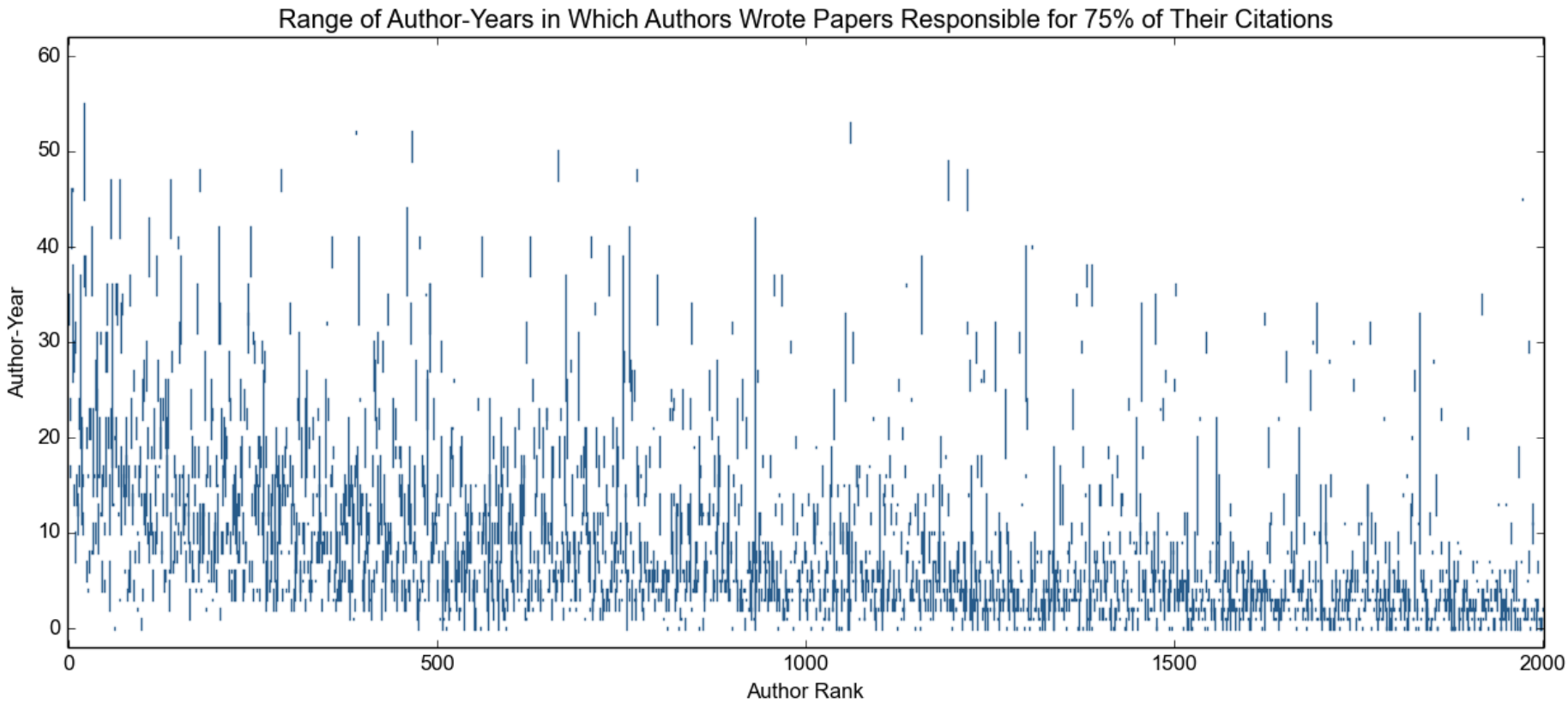
most-cited papers



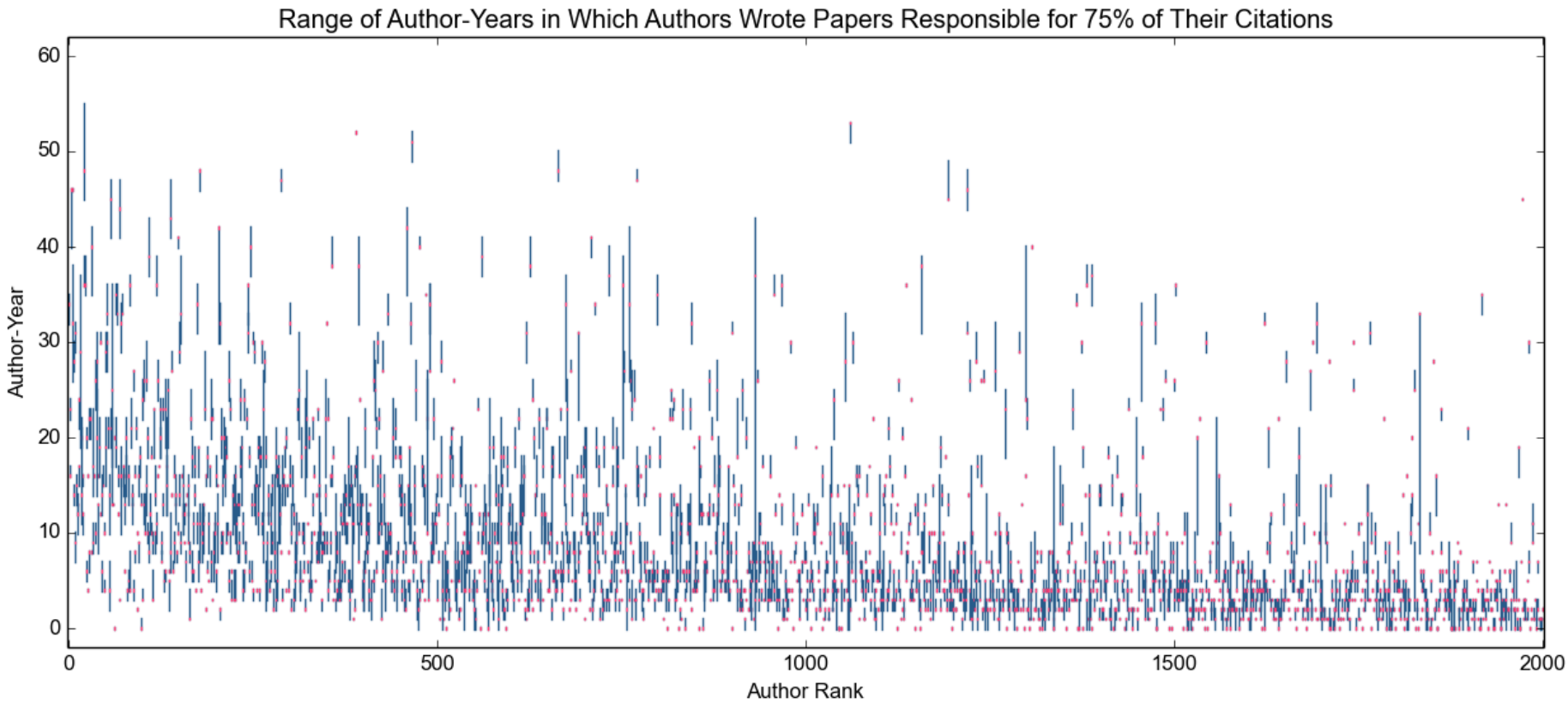
all papers


Author-Year of All Papers
Author-Year of Paper
Author
0
20
40
60
80
100
0
500
1000
1500
2000

all papers


Range of Author-Years in Which Authors Wrote Papers Responsible for 75% of Their Citations
Author-Year
Author Rank
0
10
20
30
40
50
60
0
500
1000
1500
2000

all papers


Range of Author-Years in Which Authors Wrote Papers Responsible for 75% of Their Citations
Author-Year
Author Rank
0
10
20
30
40
50
60
0
500
1000
1500
2000

truncation

recent papers may not have had time to accumulate citations

authors still working may not have reached true peak yet

truncation

big concern, but removing authors who've written in last 5 years leaves only 68

recent papers may not have had time to accumulate citations

authors still working may not have reached true peak yet

controlling for career length helps here

future work

remove the papers per author limit

good for analyzing my tool, not the author

peak question

future work

not all computer science authors tagged with “computer science” label

plans to search CS string and label, scrape common tags, then scrape larger set of authors

above approach -> larger data set

should allow better analysis of effects of truncation

future work

collect data on conference committees (DBLP)?

aligning data with citation count data may reveal correlation

other suggestions?