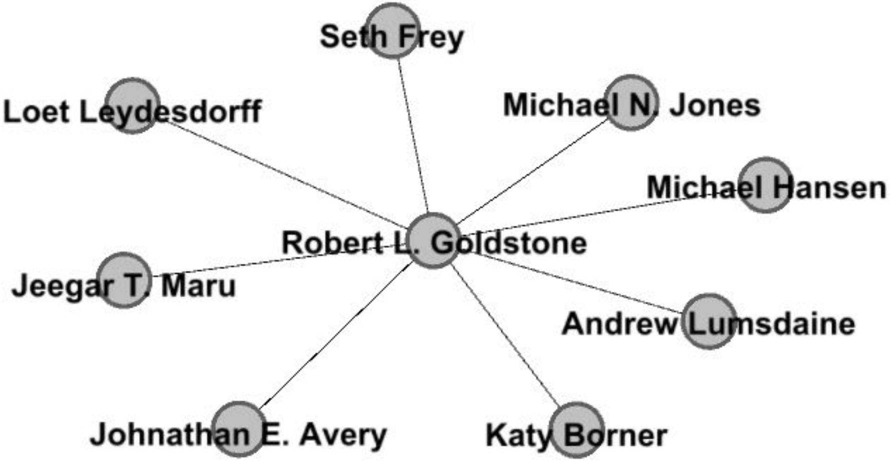
Author collaboration networks on arXiv

Morgan Klutzke

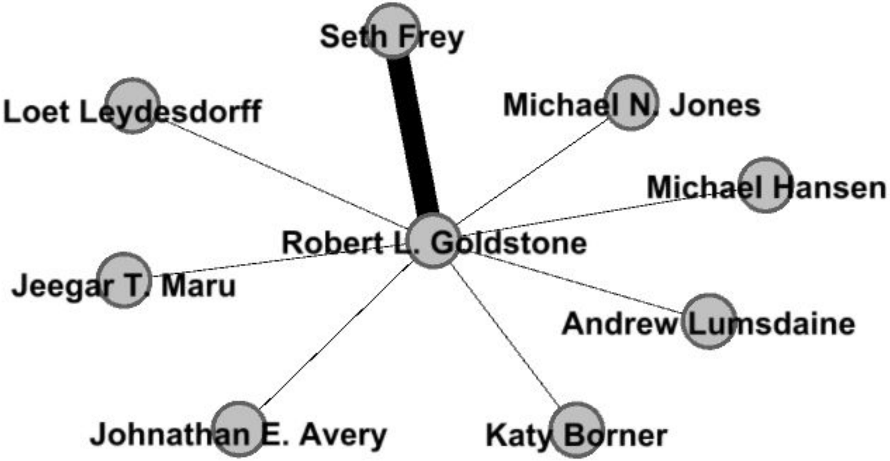
What’s an author collaboration network?

- Social network of co-authors for scholarly journal articles
  - Each node is an author
  - Each edge is a collaboration
  - Edges can be weighted by # of co-authored papers
- Collaboration networks tend to be scale-free
  - Growth
  - Preferential attachment
  - Power-law distribution of degrees



What’s an author collaboration network?

- Social network of co-authors for scholarly journal articles
  - Each node is an author
  - Each edge is a collaboration
  - Edges can be weighted by # of co-authored papers
- Collaboration networks tend to be scale-free
  - Growth
  - Preferential attachment
  - Power-law distribution of degrees



Data source: arXiv

- Open-access repository for preprints
- Download the dataset yourself at kaggle.com/Cornell-University/arxiv
- Includes metadata for 1.8 million scholarly articles

Cornell University Library
arXiv.org

- Fields represented:
  - Physics
  - Mathematics
  - Computer science
  - Quantitative biology
  - Quantitative finance
  - Statistics
  - Electrical engineering & systems science

Process for data manipulation

Original data format

authors	categories	...
“C. Balazs, E. L. Berger, P. M. Nadolsky, C.-P. Yuan”	“hep-ph”	...
“Ileana Streinu & Louis Theran”	“math.CO cs.CG”	...
“Hongjun Pan”	“physics.gen-ph”	...
...	...	...

… +1.7 million more rows

Desired edgelist format

author1	author2	weight	categories
“C. Balazs”	“E. L. Berger”	1	“hep-ph”
“C. Balazs”	“P. M. Nadolsky”	1	“hep-ph”
“C. Balazs”	“C.-P. Yuan”	1	“hep-ph”
“E. L. Berger”	“P. M. Nadolsky”	1	“hep-ph”
“E. L. Berger”	“C.-P. Yuan”	1	“hep-ph”
“P. M. Nadolsky”	“C.-P. Yuan”	1	“hep-ph”
“Ileana Streinu”	“Louis Theran”	1	“math.CO cs.CG”
...	...	...	...

Process for data manipulation

Original data format

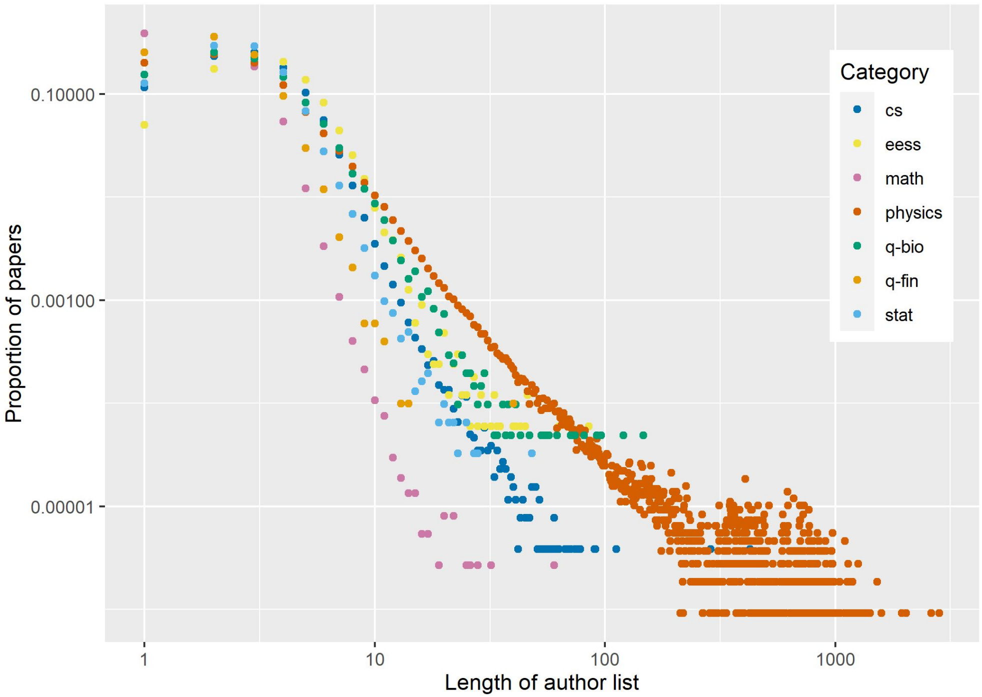
authors	categories	...
“C. Balazs, E. L. Berger, P. M. Nadolsky, C.-P. Yuan”	“hep-ph”	...
“Ileana Streinu & Louis Theran”	“math.CO cs.CG”	...
“Hongjun Pan”	“physics.gen-ph”	...
...	...	...

… +1.7 million more rows

Desired edgelist format

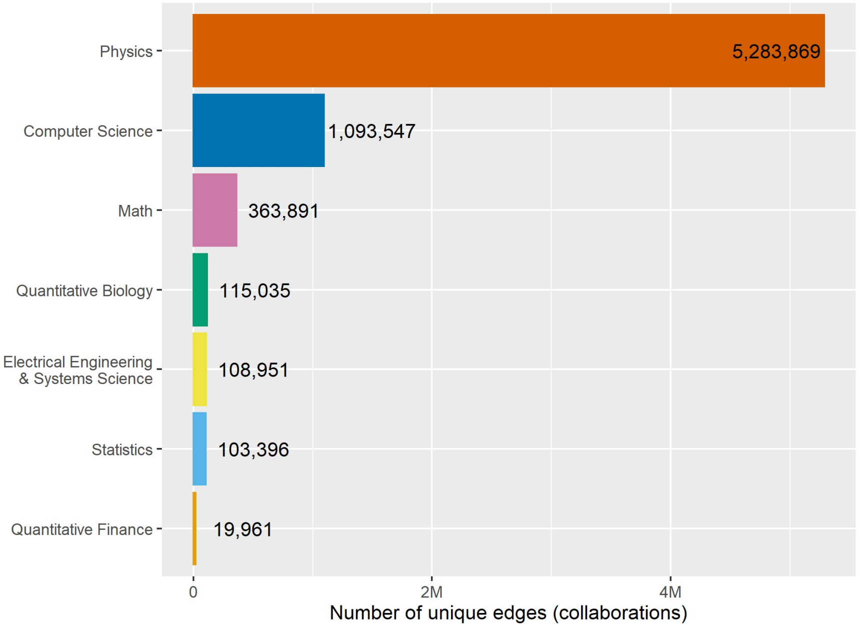
author1	author2	weight	categories
“C. Balazs”	“E. L. Berger”	1	“hep-ph”
“C. Balazs”	“P. M. Nadolsky”	1	“hep-ph”
“C. Balazs”	“C.-P. Yuan”	1	“hep-ph”
“E. L. Berger”	“P. M. Nadolsky”	1	“hep-ph”
“E. L. Berger”	“C.-P. Yuan”	1	“hep-ph”
“P. M. Nadolsky”	“C.-P. Yuan”	1	“hep-ph”
“Ileana Streinu”	“Louis Theran”	1	“math.CO cs.CG”
...	...	...	...

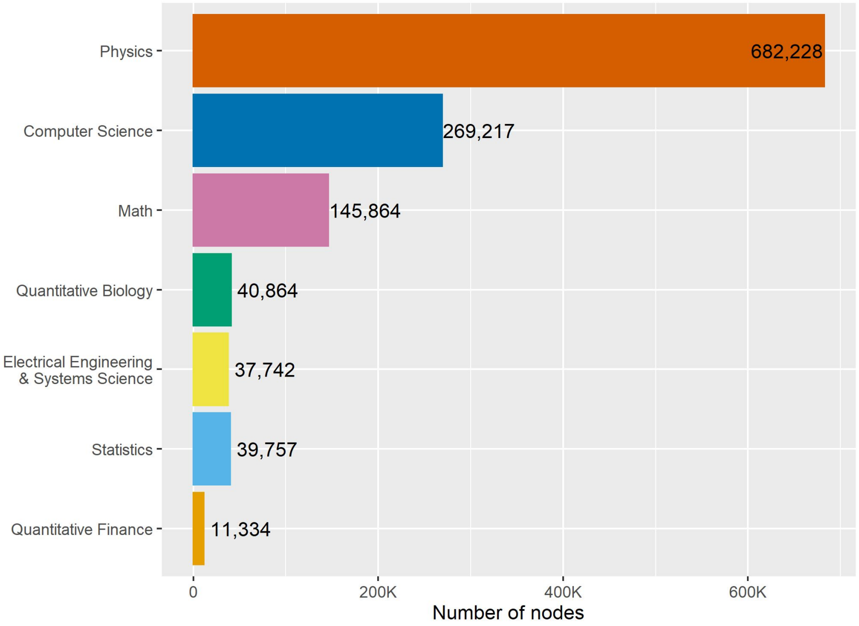
Distribution of author list lengths



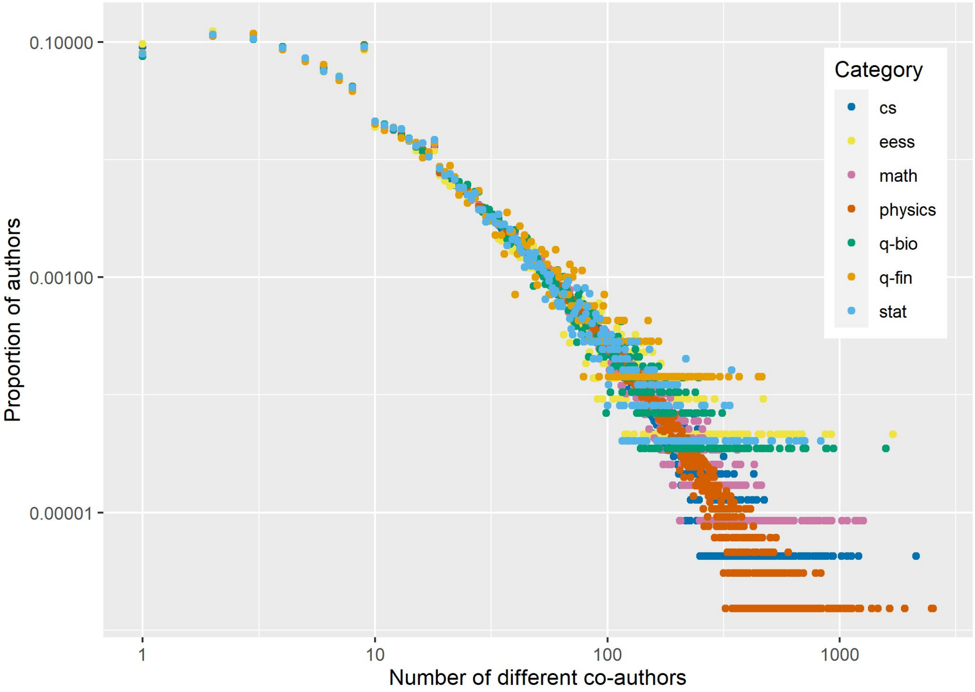
- Average is 4 authors per paper
- Maximum is 2,829 authors
- Decided to truncate after 10 authors
  - Less than 4% of papers have more than 10 authors
  - Still not optimal, especially for physics papers

Network properties by category



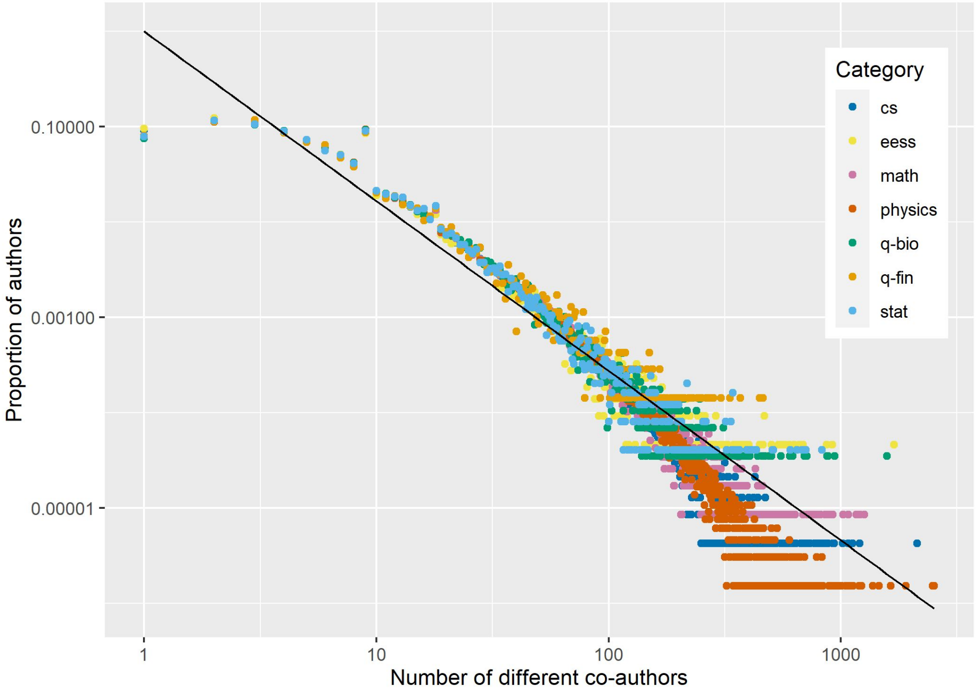


Degree distribution



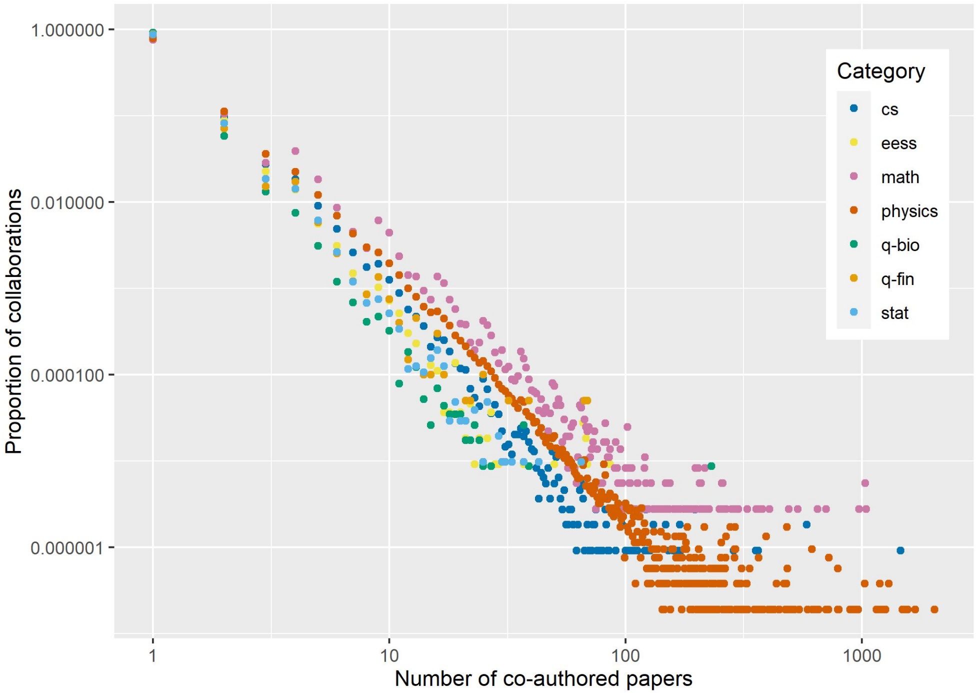
- Average degree is 13
- Connectivity (sort of) follows power-law distribution
- $P(x) \sim x^{-\alpha}$

Degree distribution



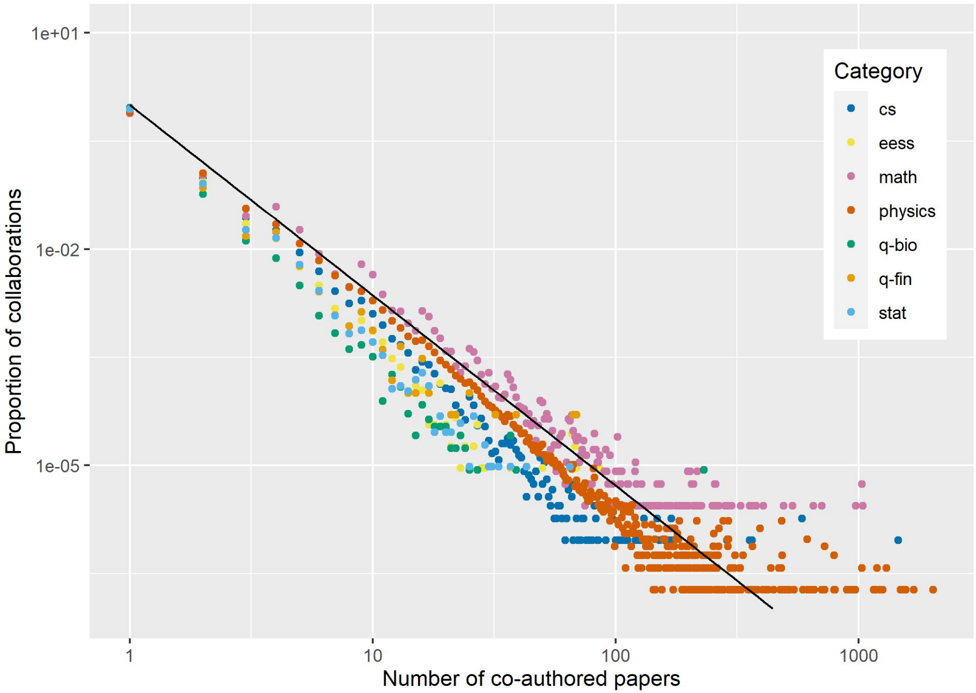
- Average degree is 13
- Connectivity (sort of) follows power-law distribution
- $P(x) \sim x^{-\alpha}$
- Fitted α = 1.78

Collaboration strength distribution



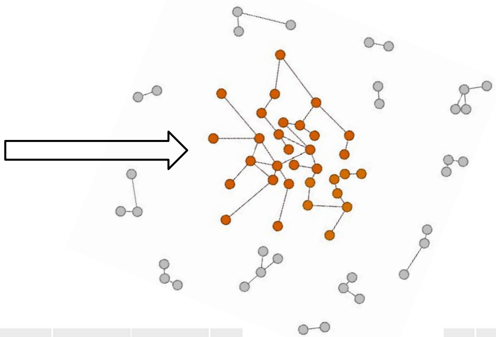
- Average edge weight is 1.58
- Number of co-authored papers per collaboration follows the power-law distribution better

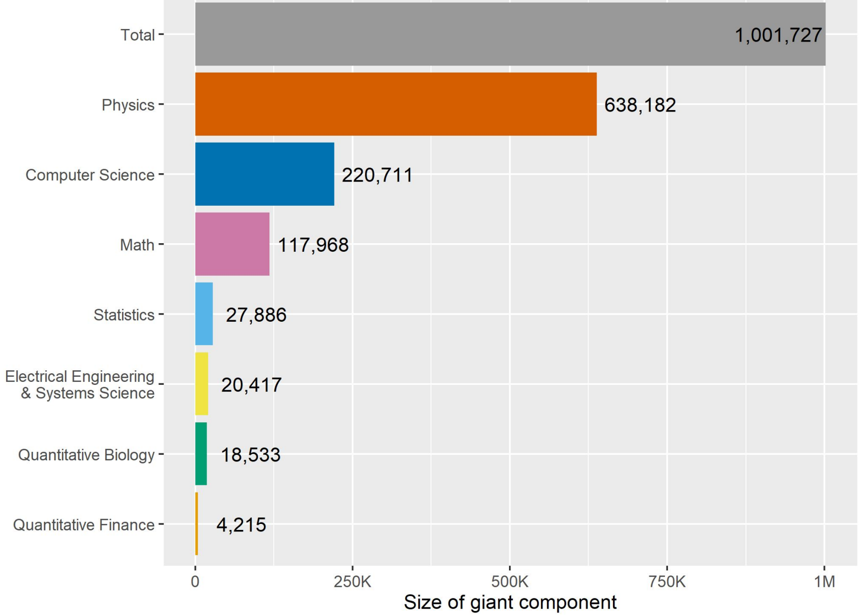
Collaboration strength distribution

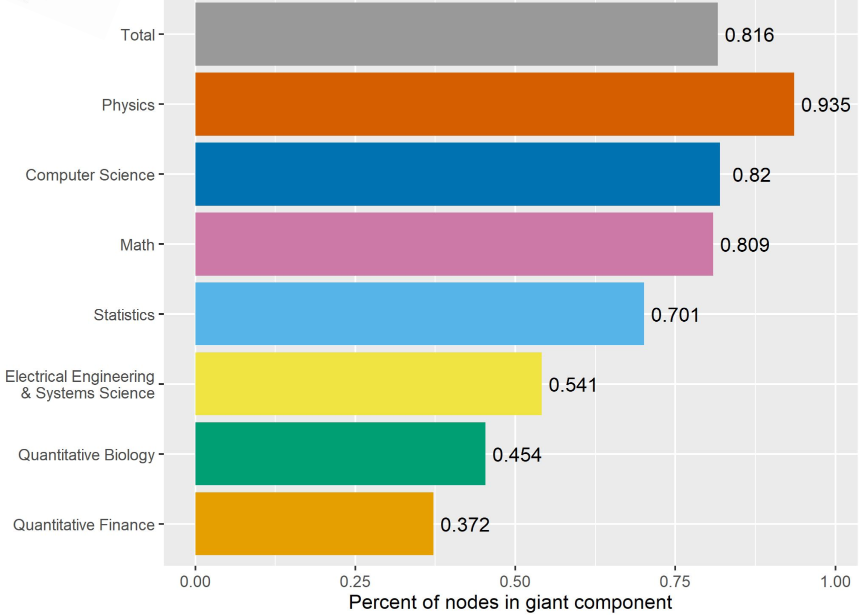


- Average edge weight is 1.58
- Number of co-authored papers per collaboration follows the power-law distribution better
- Fitted α = 2.64

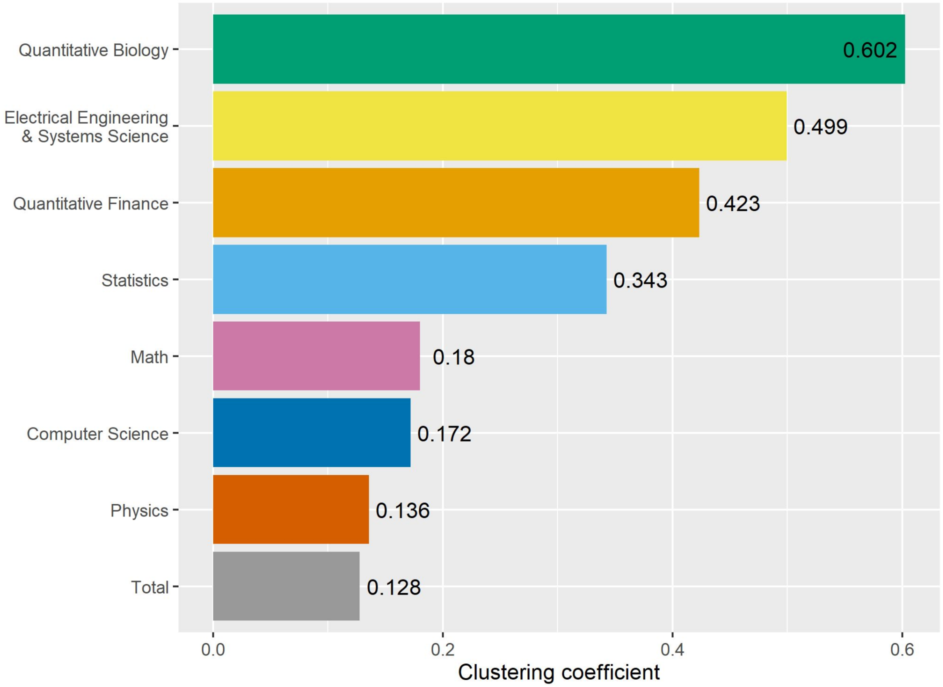
Giant components



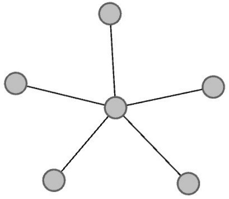




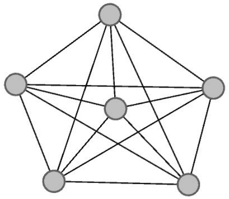
Clustering coefficients



- Measures likelihood that two nodes connected to the same node are connected themselves



CC = 0



CC = 1

In an ideal world I would...

- Look at evolution of the network over time
  - Confirm/refute preferential attachment
- Link more datasets to get more author information
  - Could use Scopus API
- Use a supercomputer
  - Wouldn’t need the arbitrary 10 author cutoff
  - Would be able to calculate shortest path lengths, diameter of giant component
  - Could maybe make visuals without the software crashing

Field	Total edges	Mean authors per paper	Mean collaborators per author (degree)	Mean co-authored papers per collaboration (edge weight)	Size of giant component	Giant component as percentage of authors	Clustering coefficient
Total	11,184,997	4.167	13.02	1.578	1,001,727	81.6%	0.128
Physics	8,523,195	5.161	13.16	1.613	638,182	93.5%	0.136
Computer Science	1,533,660	3.422	12.55	1.402	220,711	82.0%	0.172
Mathematics	703,100	1.970	13.21	1.932	117,968	80.9%	0.180
Quantitative Biology	132,670	3.561	13.39	1.153	18,533	45.4%	0.602
Electrical Engineering & Systems Science	138,726	4.112	12.64	1.273	20,417	54.1%	0.499
Statistics	128,605	2.991	12.85	1.244	27,886	70.1%	0.343
Quantitative Finance	25,041	2.366	12.88	1.254	4,215	37.2%	0.423