SNA: research design

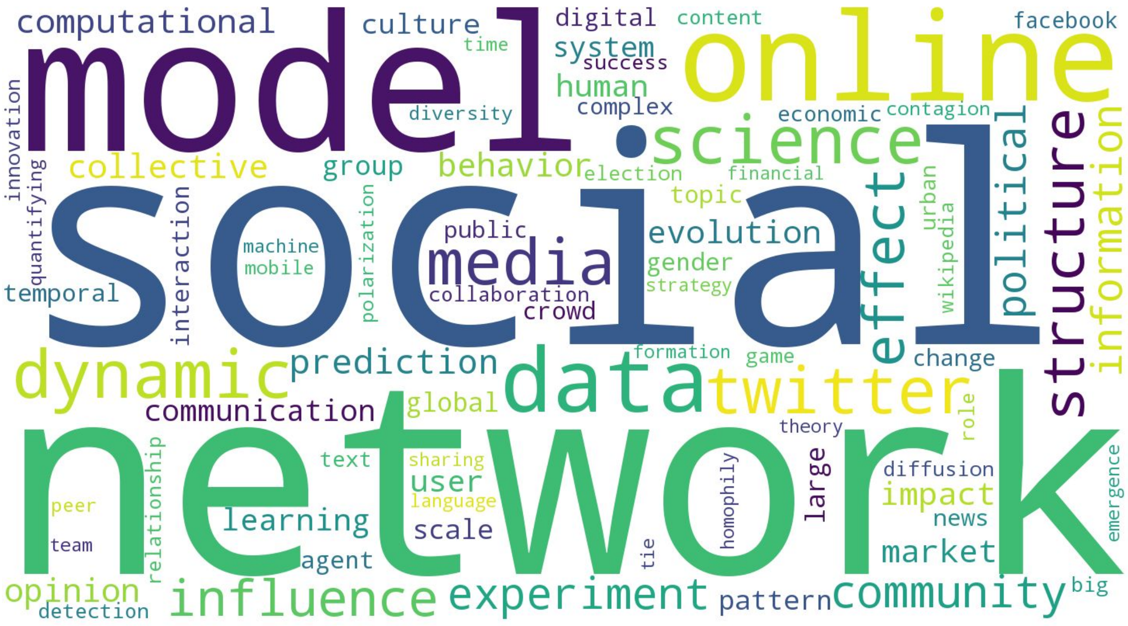
Sofia Dokuka

SICSS-HSE, 16 June 2021, Moscow

Outline

- ➢ Why networks are important part of CSS
- ➢ Data structures: Network data VS Raw data
- ➢ Surveys
- ➢ Observation
- ➢ Big data
- ➢ Recommended literature

Why networks are important part of CSS



991 submissions and 1907 authors from IC2S2 2015-2018

Why networks are important part of CSS

Science. Author manuscript; available in PMC 2009 Sep 16.
PMCID: PMC2745217

Published in final edited form as:
NIHMSID: NIHMS98137

Science. 2009 Feb 6; 323(5915): 721–723.
PMID: 19197046

doi: 10.1126/science.1167742

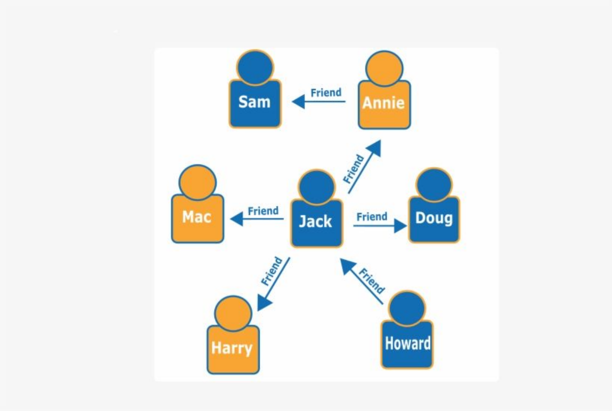
Life in the network: the coming age of computational social science

David Lazer, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne

▸ Author information ▸ Copyright and License information   Disclaimer

Network: definition

A social network consists of a finite set or sets of ***actors*** and the relation or ***relations*** defined on them [Wasserman and Faust 1994].



Data structures: Raw data

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	relatives	not_alone	Deck	Title	Age_Class	Fare_Per_Pers
0	0	3	0	2	1	0	0	0	1	0	8	1	6	0
1	1	1	1	5	1	0	3	1	1	0	3	3	5	1
2	1	3	1	3	0	0	0	0	0	1	8	2	9	0
3	1	1	1	5	1	0	3	0	1	0	3	3	5	1
4	0	3	0	5	0	0	1	0	0	1	8	1	15	1
5	0	3	0	4	0	0	1	2	0	1	8	1	12	1
6	0	1	0	6	0	0	3	0	0	1	5	1	6	3
7	0	3	0	0	3	1	2	0	4	0	8	4	0	0
8	1	3	1	3	0	2	1	0	2	0	8	3	9	0
9	1	2	1	1	1	0	2	1	1	0	8	3	2	1
10	1	3	1	0	1	1	2	0	2	0	7	2	0	0

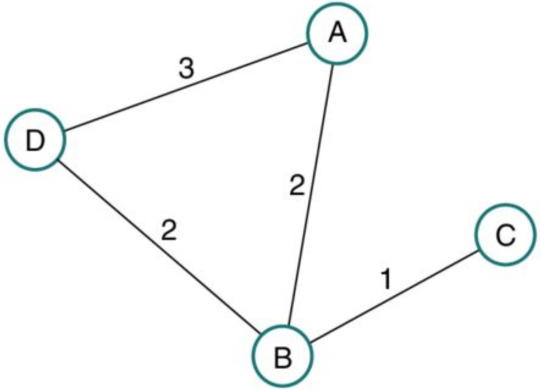
Data structures: Network data VS raw data

Rows	A	B	C	D
A	0	2	0	3
B	2	0	1	2
C	0	1	0	0
D	3	2	0	0

a) Square adjacency matrix as a grid

```
Rows,A,B,C,D
A,0,2,0,3
B,2,0,1,2
C,0,1,0,0
D,3,2,0,0
```

b) Matrix written to text file



c) Imported to Cytoscape Network

Data structures: Network data VS raw data

- Network data are RELATIONAL;

- Network require special mechanisms of data collection and processing;

- Network-related information might include multiple datasets.

Focus of the network studies

- The association between the position of an actor in social network and her/his outcomes;
- The group structure and dynamics;
- Network processes (formation, dissolution, diffusion).

Surveys

- Started a long before CSS era (1970 popular surveys started);
- Survey within a given community (nodes for a single bounded population such as classroom or company);
- GSS with a short set of network-related questions.

Network evolution (1)

The relationship between the network structure and academic achievements of students.

- 226 first year undergraduates from ETH Zurich;
- 5 longitudinal surveys on social networks and individuals characteristics;
- network questions on friendship, advice, collaboration, etc.

Stadtfeld, C., Vörös, A., Elmer, T., Boda, Z., & Raabe, I. J. (2019). Integration in emerging social networks explains academic failure and success. *Proceedings of the National Academy of Sciences*, *116*(3), 792-797.

Network evolution (2)

SAOM results (coevolution of two social networks)

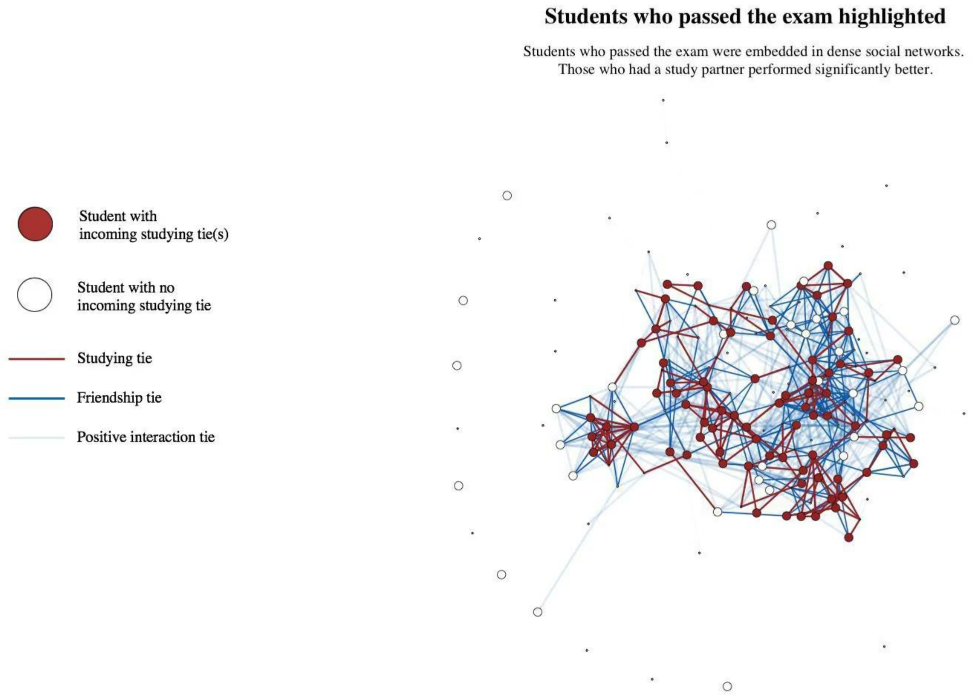
- Students are most likely to start studying with someone they already consider a friend.
- Weak evidence for the reverse process that study partners are more likely to become friends.
- Multiplex interpersonal relations are more stable than one-dimensional relationships (i.e., just being friends or study partners).

OLS:

- Successful students have a higher number of incoming ties in each network and are more likely to be named as a study partner by someone else.

Stadtfeld, C., Vörös, A., Elmer, T., Boda, Z., & Raabe, I. J. (2019). Integration in emerging social networks explains academic failure and success. *Proceedings of the National Academy of Sciences*, *116*(3), 792-797.

Network evolution (3)



ETH zürich

Network evolution (4)

Further authors covered methodological details of their study:

- data collection steps;
- organizational issues;
- participants recruitment and financial compensation;
- data quality, rr;
- survey structure, software;
- digital trace data collection;
- other important notes

Vörös, A., Boda, Z., Elmer, T., Hoffman, M., Mepham, K., Raabe, I. J., & Stadtfeld, C. (2021). The Swiss StudentLife Study: Investigating the emergence of an undergraduate community through dynamic, multidimensional social network data. *Social Networks*, *65*, 71-84.

Core discussion networks of Americans (1)

- ‘With whom did you discuss important matters over the last six months’;
- GSS;
- ‘Name interpreter’ focus on the first five persons;
- Describe the relationships between first five nominated individuals;
- Information about the whole ‘CDN’: age, gender, education, kinship, etc.

Marsden, P. V. (1987). Core discussion networks of Americans. *American sociological review*, 122-131.

Core discussion networks of Americans (2)

- Nearly quarter of Americans have 0 or 1 CDN size;
- The majority of the sample support 2-5 CD ties;
- The mode is 3;
- 1.5 mean for kin contacts and 1.4 for non-kin contacts;
- The CDNs are densely linked (d = 0.61);
- In general, differences are greatest for subgroups defined by age and education;
- Network size as well as diversity drops with age, density rises with age;
- More educated people support larger networks, they also support more diverse networks.

Marsden, P. V. (1987). Core discussion networks of Americans. *American sociological review*, 122-131.

Observation (1)

Knowledge network building framework of curiosity to capture styles of curious information seeking. 149 participants explore Wikipedia for over 5 hours spanning 21 days.

Busybody - preference for sampling diverse concepts, characterized by “distraction” and “never-dwelling anywhere”. The busybody will “frisk about, and rove about, at random, wherever they please".

Hunter - preference for sampling closely connected concepts. The hunter “wishes [they] had a few hundred helpers and good, well-trained hounds that [they] could drive into the history of the human soul to round up [their] game".

Lydon-Staley, D. M., Zhou, D., Blevins, A. S., Zurn, P., & Bassett, D. S. (2021). Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nature human behaviour*, *5*(3), 327-336.

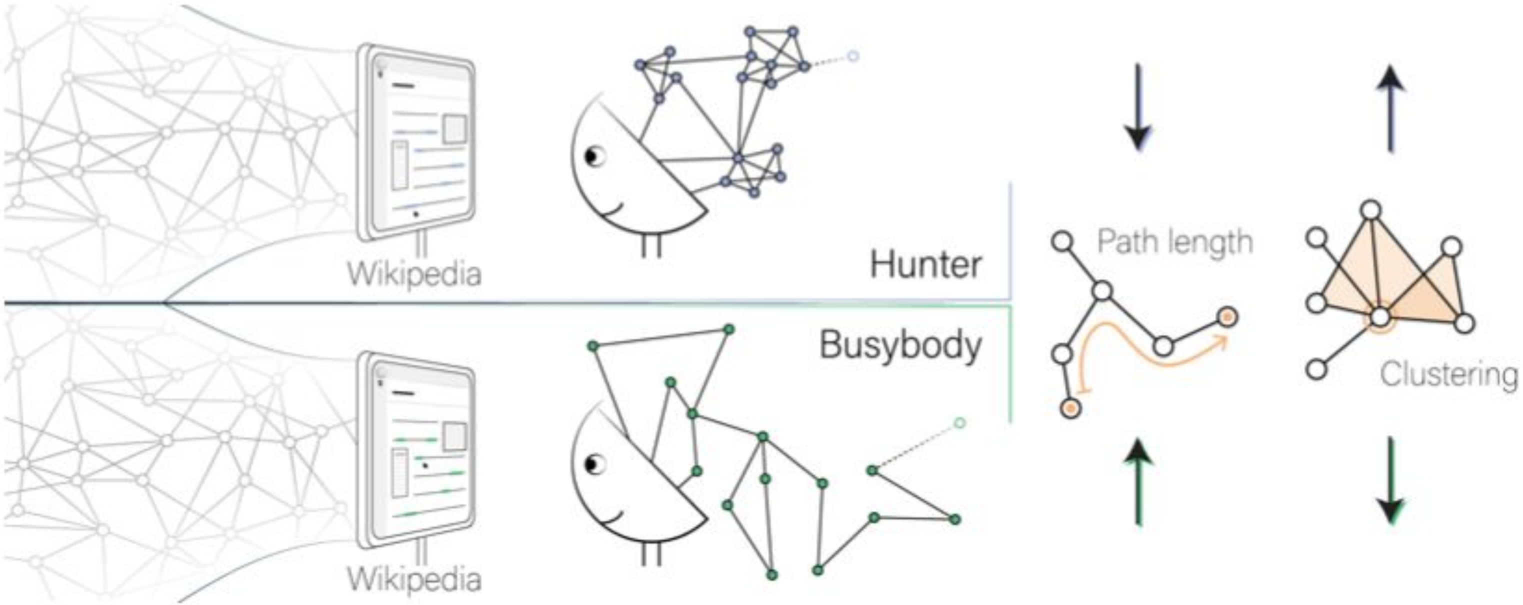
Observation (2)

Network analysis used to quantify the curiosity behavior. Network nodes represent distinct concepts and network edges represent the manner in which the concepts are related.

- Busybody will be marked by the creation of loose knowledge networks of sparsely connected, seemingly unrelated concepts. In network terms, their networks will have small edge weights, low clustering, and high characteristic path length.
- Hunter, in contrast, will create the tight knowledge networks consisting of closely-connected concepts and their networks will have large edge weights, high clustering, and low characteristic path length.

Lydon-Staley, D. M., Zhou, D., Blevins, A. S., Zurn, P., & Bassett, D. S. (2021). Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nature human behaviour*, *5*(3), 327-336.

Observation (3)



Lydon-Staley, D. M., Zhou, D., Blevins, A. S., Zurn, P., & Bassett, D. S. (2021). Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nature human behaviour*, *5*(3), 327-336.

Observation (4)

- Each Wikipedia page was treated as a distinct concept or node in a knowledge network. NLP was used to determine the weight of network edges reflecting similarities in semantic content between any two pages. On the resulted network the metrics were computed.
- Network metrics were correlated with self-reported curiosity (five-dimensional curiosity scale).

Lydon-Staley, D. M., Zhou, D., Blevins, A. S., Zurn, P., & Bassett, D. S. (2021). Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nature human behaviour*, *5*(3), 327-336.

Observation (5)

- Deprivation sensitivity is positively associated with the average clustering coefficient (R^2 = 20%), suggesting that participants high in deprivation sensitivity examine closely related concepts during information seeking to a greater extent than participants low in deprivation sensitivity.
- Social curiosity is negatively associated with the average clustering coefficient.
- Deprivation sensitivity is negatively associated with the characteristic path length such that participants high in deprivation sensitivity, while exploiting local information, also have networks that are easily traversable from one end to the next.

Lydon-Staley, D. M., Zhou, D., Blevins, A. S., Zurn, P., & Bassett, D. S. (2021). Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nature human behaviour*, *5*(3), 327-336.

Big data

‘The science of science (SciSci) offers a quantitative understanding of the interactions among scientific agents across diverse geographic and temporal scales: It provides insights into the conditions underlying creativity and the genesis of scientific discovery, with the ultimate goal of developing tools and policies that have the potential to accelerate science.’

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... & Barabási, A. L. (2018). Science of science. *Science*, *359*(6379).

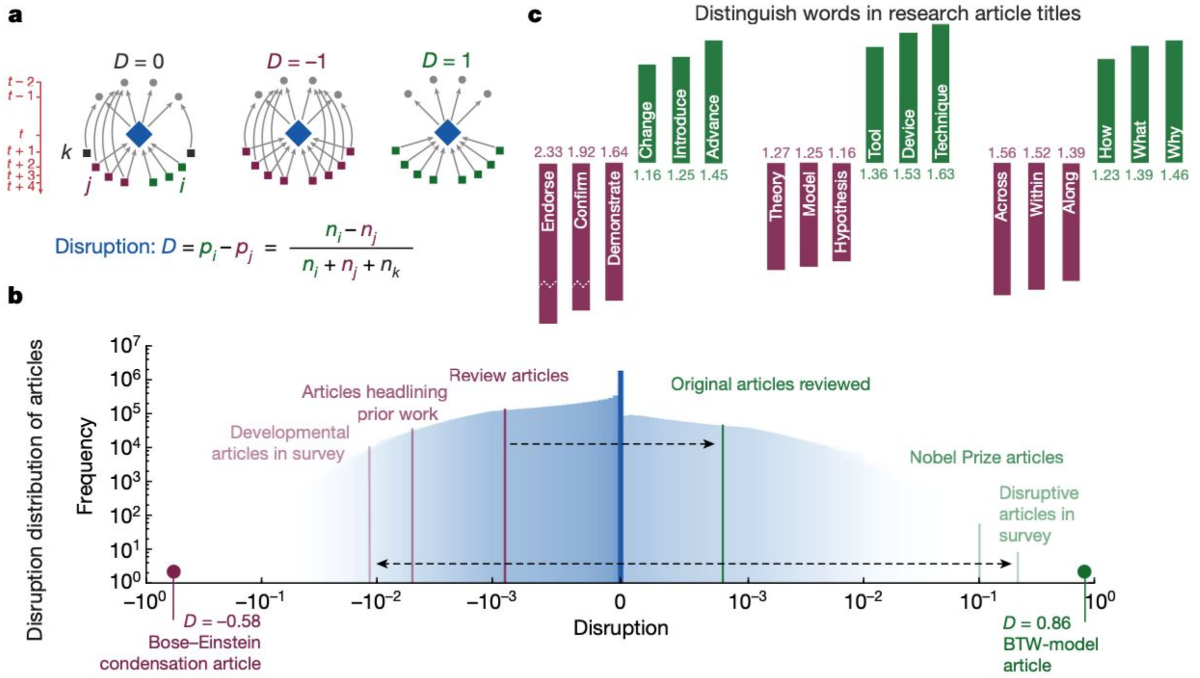
Big data (1)

Data sources:

- the Web of Science (WOS) with more than 42 million articles published between 1954 and 2014, and 611 million citations among them;
- 5 million patents granted by the US Patent and Trademark Office from 1976 to 2014, and 65 million citations added by patent applicants;
- 16 million software projects and 9 million forks to them on GitHub (2011–2014).

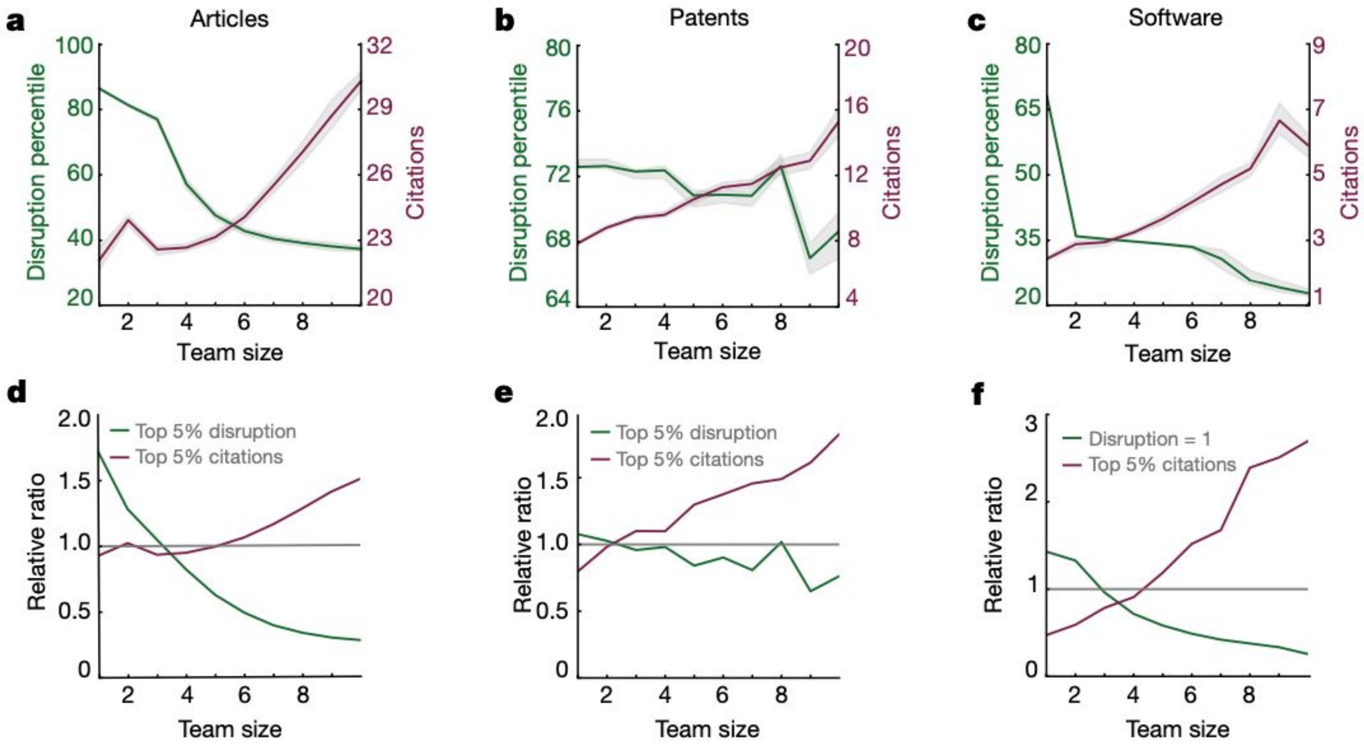
Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378-382.

Big data (2)



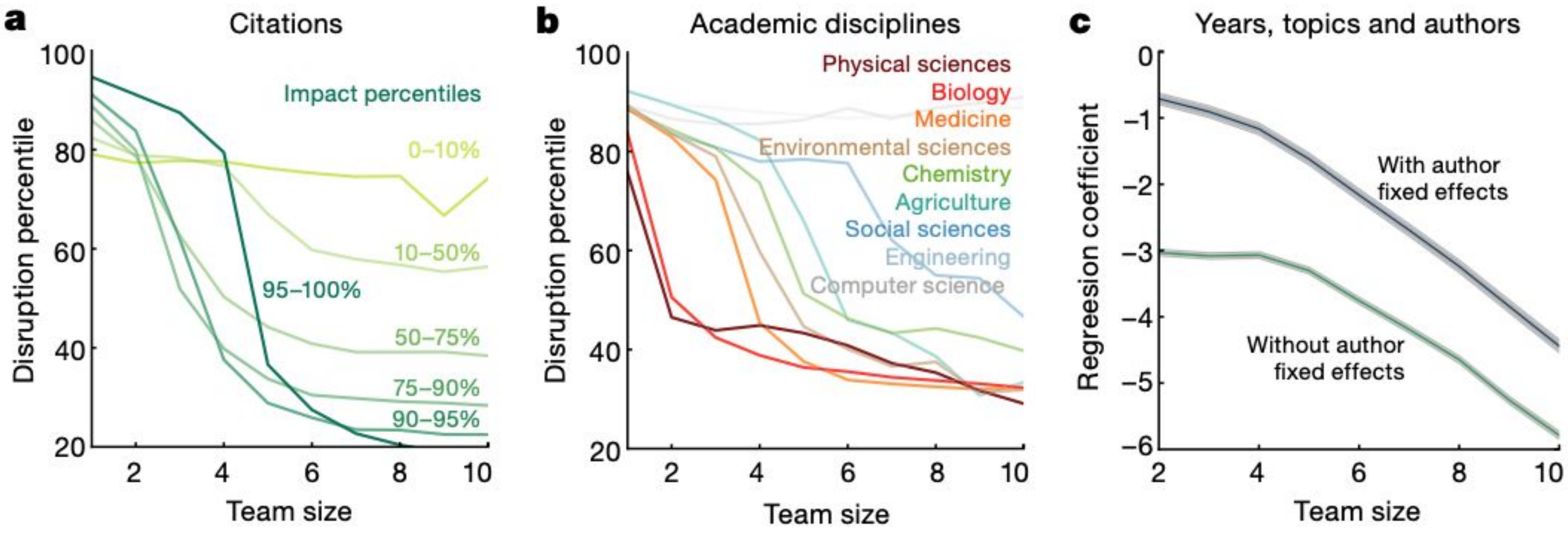
Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378-382.

Big data (3)



a
Articles
Disruption percentile
Citations
100
80
60
40
20
32
29
26
23
20
2
4
6
8
Team size
b
Patents
Disruption percentile
Citations
80
76
72
68
64
20
16
12
8
4
2
4
6
8
Team size
c
Software
Disruption percentile
Citations
80
65
50
35
20
9
7
5
3
1
2
4
6
8
Team size
d
Top 5% disruption
Top 5% citations
Relative ratio
2.0
1.5
1.0
0.5
0
2
4
6
8
Team size
e
Top 5% disruption
Top 5% citations
Relative ratio
2.0
1.5
1.0
0.5
0
2
4
6
8
Team size
f
Disruption = 1
Top 5% citations
Relative ratio
3
2
1
0
2
4
6
8
Team size

Big data (3)



a
Citations
Impact percentiles
0–10%
10–50%
95–100%
50–75%
75–90%
90–95%
Disruption percentile
Team size
b
Academic disciplines
Physical sciences
Biology
Medicine
Environmental sciences
Chemistry
Agriculture
Social sciences
Engineering
Computer science
Disruption percentile
Team size
c
Years, topics and authors
With author fixed effects
Without author fixed effects
Regreesion coefficient
Team size

Big data (4)

- Systematic difference the contributions of small and large teams in the creation of scientific papers, technology patents and software products.
- Small teams disrupt science and technology by exploring and amplifying promising ideas from older and less-popular work.
- Large teams develop recent successes, by solving acknowledged problems and refining common designs.

Recommended literature and educational activities

To read

http://faculty.ucr.edu/~hanneman/nettext/
https://www.sciencedirect.com/science/article/abs/pii/S0022249613000126
https://uk.sagepub.com/en-gb/eur/social-network-analysis/book231856
https://www.cambridge.org/ru/academic/subjects/sociology/sociology-general-interest/social-network-analysis-methods-and-applications?format=PB&isbn=9780521387071
http://networksciencebook.com
https://www.amazon.com/Networks-Introduction-Mark-Newman/dp/0199206651
https://arxiv.org/pdf/cond-mat/0303516v1.pdf
https://www.cs.cornell.edu/home/kleinber/networks-book/networks-book.pdf

To watch

https://www.coursera.org/learn/python-social-network-analysis
https://www.coursera.org/learn/social-economic-networks
http://leonidzhukov.net/hse/2014/socialnetworks/

Recommended software

Python
https://networkx.org
http://snap.stanford.edu
https://igraph.org
https://graph-tool.skewed.de
R
https://cran.r-project.org/web/packages/sna/sna.pdf
https://cran.r-project.org/web/packages/network/network.pdf
https://cran.r-project.org/web/packages/statnet/statnet.pdf
https://cran.r-project.org/web/packages/ergm/ergm.pdf

Other software
https://sites.google.com/site/ucinetsoftware/home
http://www.casos.cs.cmu.edu/projects/ora/download.php
http://vlado.fmf.uni-lj.si/pub%20/networks/pajek/default.htm
https://gephi.org

Recommended mailing lists

https://www.insna.org/socnet

https://lists.nongnu.org/mailman/listinfo/igraph-help

https://mailman13.u.washington.edu/mailman/listinfo/statnet_help

Recommended resources (great links)

https://kateto.net

https://github.com/briatte/awesome-network-analysis