Lecture 4: I) Producing Scientific Knowledge: papers; power laws

Economists model production functions as a relation between inputs and outputs: Y= AF(K,L), where K= capital; L=labor; and A measures shift in productivity due to increased knowledge or something. The most common form is the Cobb-Douglas, which is ln-linear with constant returns to scale: ln Y = (1-Θ) ln K + Θ ln L, where 1- Θ is often taken to be labor's share of output. Another common form is fixed coefficient function used in input-output analysis. L/Y and K/Y are constant. If we measure output as papers, and researchers writes 3 papers/year, the model would be Papers = 3 Researcher. In fact the distribution of papers is a power law, not C-D or fixed coefficient.

The output from a scientific investigation is a scientific paper. Published papers in peer reviewed journals meet one market test – some expert reviewer thought paper worth appearing in print – sort of like a start-up firm that raised some capital and produced something that we can count. But a paper is a **unique** amalgam of information – new ideas, old ideas, new facts, old facts – that is a distinct product, and may have different values over time and for different people. Some papers will be path-breaking; some ordinary; and some will be valueless → need measure of quality of paper – a price in the market.

Since everyone who reads a paper pays a time price, one measure of value could be number of downloads x average time spent reading x wage of people who read it. This would be an "expenditure measure" of the value, just as number of people who buy soup at restaurant x price of soup measures value of soup in national income accounts (but if it takes longer/shorter to do restaurant soup than home soup, must do time account also)

Most widely used measure of quality of paper is **citations**. Paper with 10 citations is presumably more valuable than paper with one citation. Citers only read the paper (in principle) but also found something useful in it so citations have a price-type indicator (citations measure quantity of use and the price is set at numeraire 1.).

BUT AT ANY GIVEN TIME – SAY 2 YEARS COVERED USED IN MEASURING IMPACT FACTOR OF JOURNAL , LOTS OF PAPERS GET NO CITES

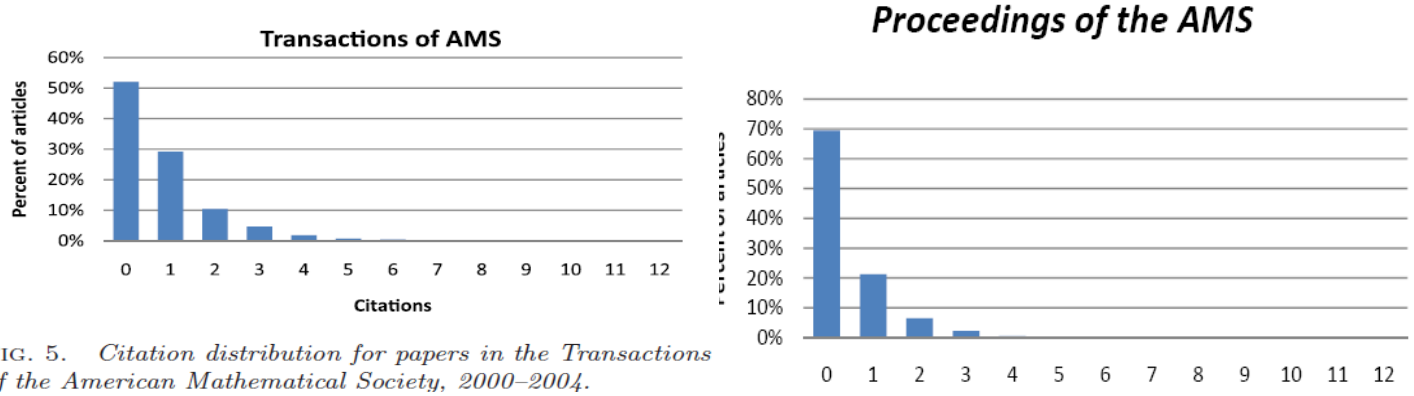
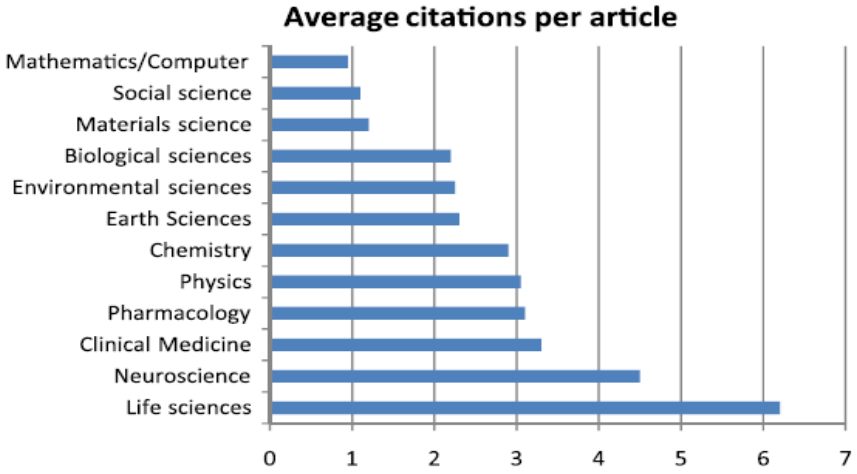


FIG. 5. *Citation distribution for papers in the Transactions of the American Mathematical Society, 2000–2004.*

Math has huge number of 0-cited articles in 2 year window. Proceedings, which publishes shorter articles has impact factor of 0.434, while Transactions has impact factor of 0.846. But wide dispersion of cites in both cases.
Why so many 0-cited papers? Math papers give fewer citations/norm so greater chance of none. There may be deeper reasons, knowledge changes more slowly so no need to cite new things?



For math citations see R Adler, J Ewing and P Taylor 2009 Citation Statistics A Report from International Mathematical Union (IMU) in Cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS) Statistical Science 2009, Vol. 24, No. 1, 1–14

But while publication is current event; citations are a future measure that follows a "life cycle" time pattern. Papers get few citations 1-2 years out, more around 5-7 years out, and then fewer citations except for "runaways" – paper that got few citations but suddenly got lots of attention.

Alternative way to see life cycle of citations is through REFERENCES. References in articles published in 2003 have effectively 0 for 2003 increasing number over next serveral years and tail off late. Math has smaller number of references and flatter curve.

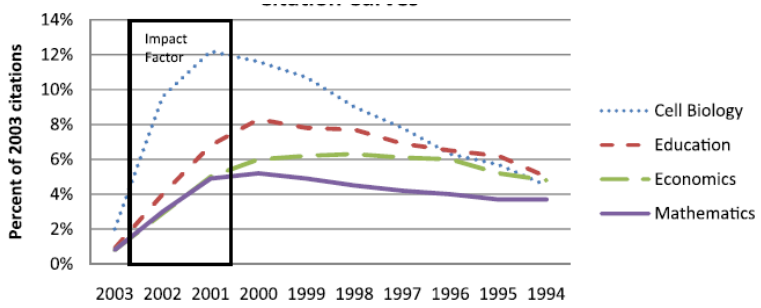


FIG. 1. *The age of citations from articles published in 2003 covering four different fields. Citations to articles published in 2001–2002 are those contributing to the impact factor; all other citations are irrelevant to the impact factor. Data from*

To assess "true impact paper need citations over some extended period of time. But not very helpful for decisions today. Researchers ignored your great 2000 paper on the coming collapse of Wall Street so university fired you as useless … until suddenly Wall Street collapsed and the paper got widely cited, but then it was too late.

Is there a more immediate indicator of the value? Take the journal which publishes paper. Journal metric is IMPACT FACTOR – which Thomson-Reuters measures as the average number of citations received per paper published in journal during the two preceding years.

A = # times that articles published in 2006 and 2007 were cited by indexed journals during 2008.
B = # "citable items" published by that journal in 2006 and 2007.
2008 impact factor = A/B.

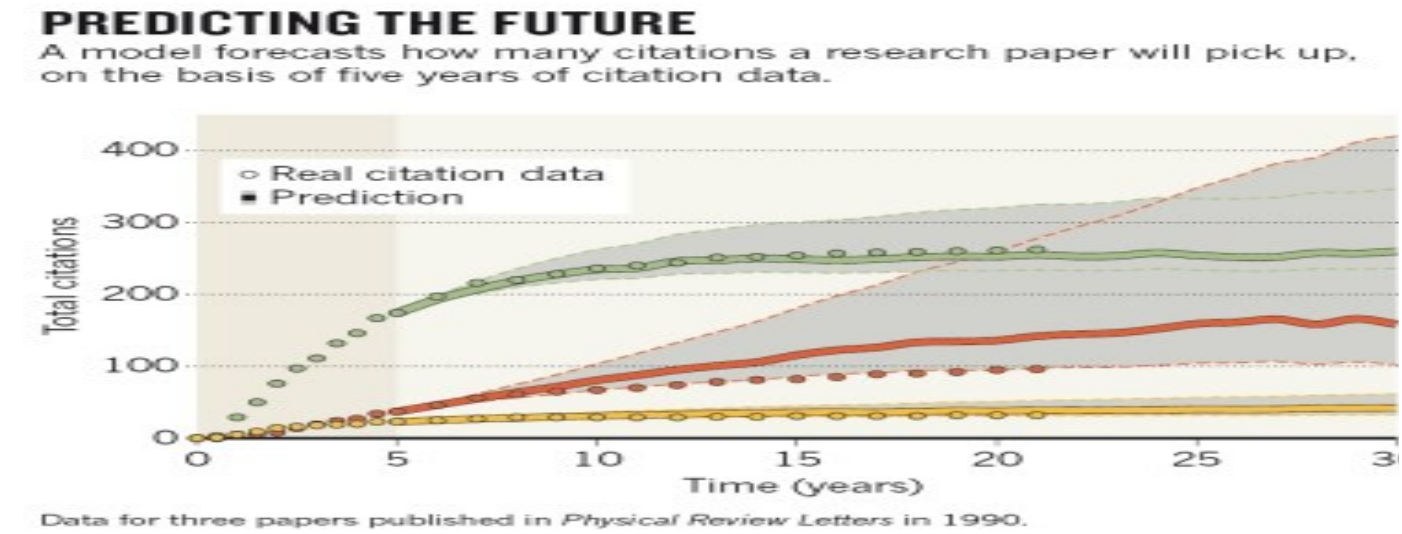
No particular reason for 2 year metric. Initially Eugene Garfield (founder of Institute for Scientific Information (ISI predecessor to T-R) used one year and five years as example. Could determine earliest number of years of citations that correlates highly with some measure of "lifetime citations". If impact factors highly correlated by year, measure picks out journals which are highly cited regularly and thus which may have higher quality papers. May be better to take acceptance ratio = accepted articles/submitted articles as measure of journal standards?

Would assume impact factor would be negatively correlated with acceptance ratio: greater chance of being accepted implies fewer submissions and less quality competition. Anyone for a model that would predict submission behavior in response to impact factor leading to some equilibrium sorting?.

Another alternative to impact factor would be to see if early citations – say from first year or two – give a good prediction of future/lifetime citations. This says estimate $\sum C = f(C_1)$. Could compare to $\sum C = f(\text{IMPACT FACTOR})$. Is it better to have 5 early cites in a journal with impact factor 2 than 1 cite in a journal with impact factor 3? Power laws are scale free so that if you know something about one part of the distribution, shape generalizes. There is some disagreement about predictive power of early cites on later career (Allison, Long, Krauze, 1982, American Sociology Review pp 615-625), but NO studies that contrast early cites vs Impact factors.

Allison and Long (Departmental effects on scientific productivity American sociological review,1990 469-47) ask whether top departments hire people who are gaining lots of citations or whether being hired at top department produces lots of citations. They conclude it is more the latter than the former. Good paper would be to use larger data set//statistical model to see if their conclusion is correct.

In Oct 2013, Wang, Song, Barabasi ("Quantifying Long-Term Scientific Impact" Science 4 Oct, pp 127-132 claimed that they had a powerful predictor for future citations from early citations that would go a long way to resolving the problem of assessing research based on early citations, a longterm decay factor, and a fitness measure and claim great success from 5 years of citation for future citations based on the mean and standard deviation of the first five years citation data.



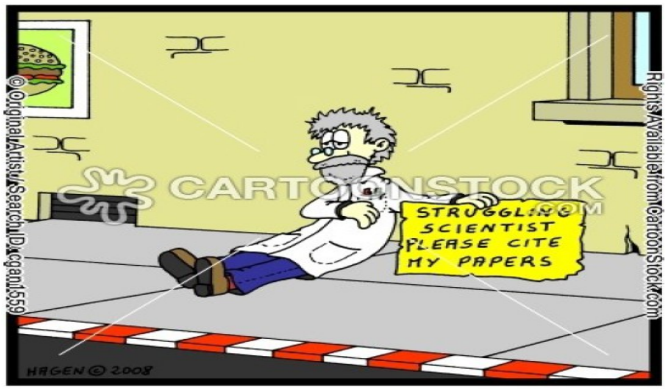
But Wang, Mei and Hicks ("Comment on …, Science vol 345 11 July 2014 report that the prediction power is horrid, "even worse than simply using short-term citations to approximate long-term citations"

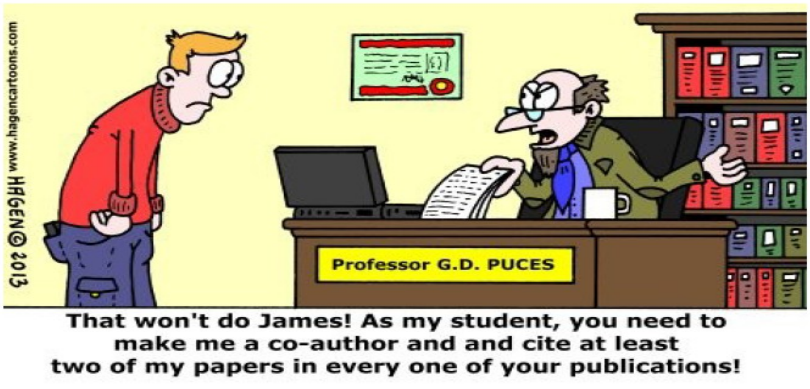
Table 1. Prediction power evaluation.

	WSB	Naïve	WSB-with-prior ‡	WSB	Naïve
$T_{Train} = 5$		*All papers (N = 1973)*		*μ* < 5 (N = 1682)*	
Mean absolute percentage error	6.71×10^{303}†	0.56	0.75	1.98×10^{57}	0.54
Spearman correlation	0.51	0.74	0.64	0.58	0.76
Percentage of correctly identified top 10% papers	23.74	58.29	59.60	31.95	57.99
$T_{Train} = 10$		*All papers (N = 1973)*		*μ* < 5 (N = 1942)*	
Mean absolute percentage error	1.91×10^{7}	0.34	0.27	0.38	0.34
Spearman correlation	0.90	0.91	0.91	0.90	0.91
Percentage of correctly identified top 10% papers	67.68	76.41	74.75	71.28	75.90

†Conservative estimations. The largest number our software can handle is 1.797693×10^{308}, so numbers larger than this threshold will be treated as $+\infty$. In our calculations, we truncate these larger numbers and only record them as 1.797693×10^{308}, so the actual prediction error is even larger than reported here. ‡With $T_{Train} = 5$, WSB-with-prior failed to find finite values of α and β for optimal solutions, so we adopt the four sets of α and β values reported by Shen *et al*. (*4*) and report their best evaluation statistics. These four sets of (α, β) are (4.237, 4.061), (4.759, 4.440), (6.130, 4.924), and (10.706, 5.379). Among them, the smallest MAPE (0.75) is yielded by (4.759, 4.440), the highest Spearman correlation (0.64) is yielded by (4.237, 4.061), and the highest percentage of correctly identified top 10% papers (59.60%) is yielded by (6.130, 4.924).

Moreover, model cannot deal with late blooming papers, such as superconductivity papers after the discovery of high-temperature superconductivity in the 1980s, or delayed impact, like the explosion of citations to Erdős and Rényi's work 4 decades after their publication, following the emergence of network science .





Another refinement on citations: two people cite my paper – me and Albert E. Self-cites often eliminated as not as meaningful as cites by others. If Albert E cites my paper, wow! But we can generalize from self-cites: how about my co-author cites my paper; my students; people in my network, etc. Sifan Zhou finds that men cite men more and women cite women more, so there is a gender bias in cites. Saving grace to all this is that a very highly cited paper has to break out of all the network connections.

"Sociological network issues" involved with journal publications. I bet that QJE (Harvard edited ec journal) publishes many papers with Harvard-MIT-Cambridge connections while JPE (Chicago edited ec journal) published lots of papers with Chicago connections? Does Science do more US based and Nature more UK/EU based papers?

How about weighting citations by the citations of paper citing us? Per Google pagerank algorithm.

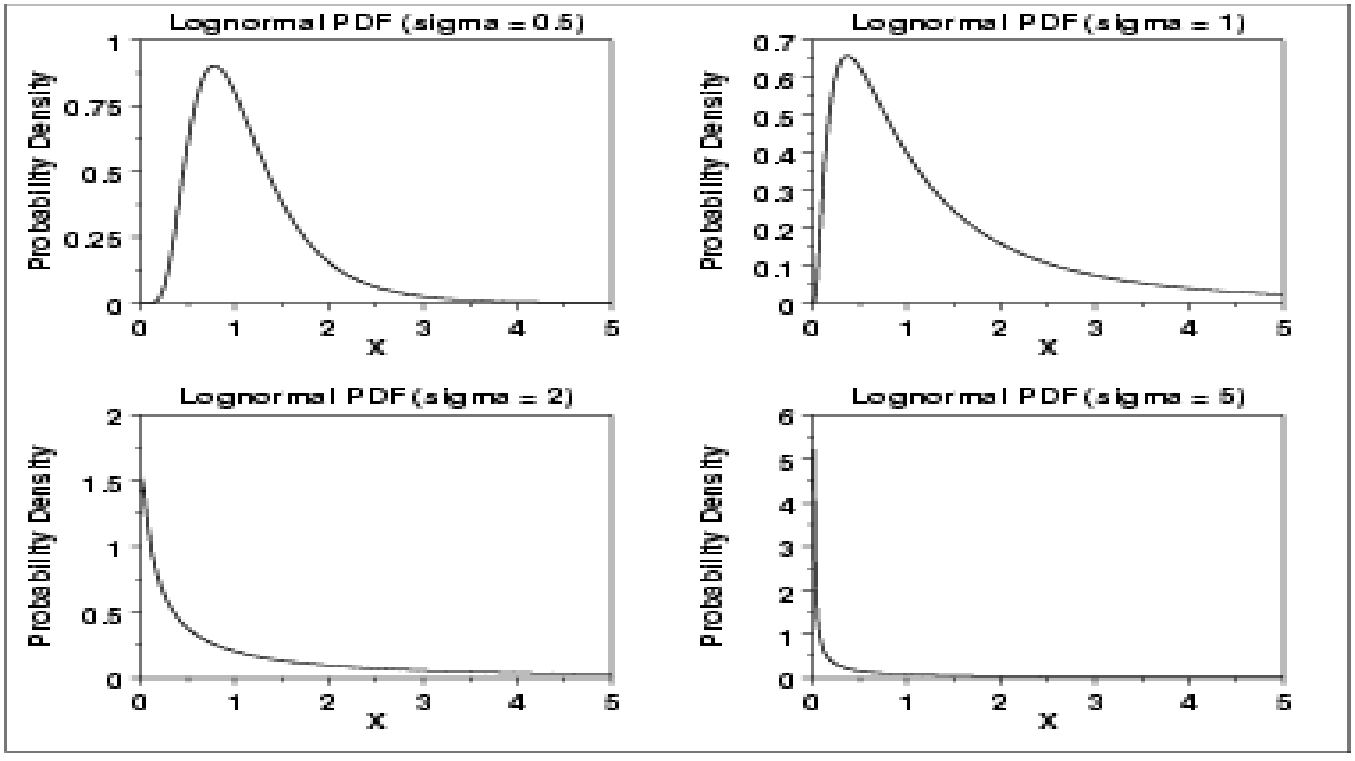
2.KEY FINDING IS POWER LAW best represents data. This is an empirical relation between variable Y and variable S in which $Y= S^{a}$, where the relation is determined by fixed power a or -a, aka as the scaling parameter. Power law is log linear per Cobb-Douglas but with coefficient that makes it different from Cobb-Douglas relation.

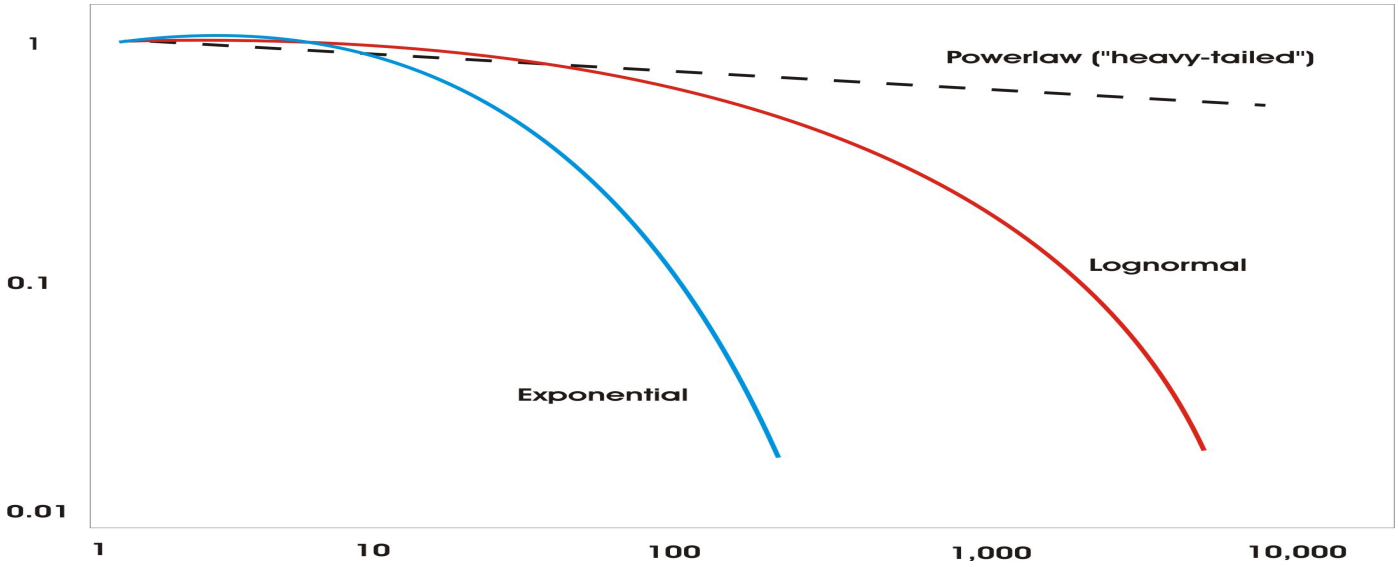
The term power reflects the dependence of Y on S by powers: a could be 2 in which case we have a quadratic, 3, 4, any number. If Y is the frequency of an event and S measures the size of the event, the coefficient linking them is usually a negative number -a. Big events rare; small events frequent.

The **power law $Y= S^{-a}$** = $1/S^{a}$ gives the inverse relation between the frequency and the size of events. It says that the frequency of an event, say ten times as large as S, 10S, is $1/(10S)^{a}$, which makes that event $1/10^{a\ th}$ as likely. With a =1 the large event 10S is $1/10^{th}$ as likely as the smaller event S. With a =2 the large event is $1/100^{th}$ as likely.

Power law distribution differs from normal distribution where middling events are most common. Taking ln of both sides gives line in lns of variables: ln Y = ln B - a lnS so that dlnY/dlnS = - a (constant elasticity relation). This is scale free since the same pattern applies regardless of whether we have large or small units or changes in units, but it often just fits the upper tail of a distribution. Economists who use constant elasticity functions for demand, supply, and in the Cobb-Douglas production function but constant elasticities are more for convenience or first-order Taylor Series expansion around a more complicated relation.

Bibliometric analysis of the number of papers scientists write, citations to people and to papers and MANY other quantitative measures of scientific relations such as number of collaborators follow a power law AT THE UPPER TAIL. But there are other functional forms that can fit some of the data – log normal also has a "fat" tail that depends on its standard deviation. The stretched exponential function exp – t^{B} adds the B parameter to stretch the tail of exponential, where where 0< B <l. B=1 is exponential.





Statistical problem is that power law depends critically on upper tail, but upper tail has few observations, so danger of getting imprecise estimates. Since lots of distributions have long tails there must be some mechanism generating this shape just as there is a random shock mechanism generating normal or lognormal distribution.

Mitzenmacher credits economists for discovering power laws as Pareto distribution. Another famous power law is Zipf Law, that George Zipf , professor of German at Harvard, used to relate frequency to **rank**. Lada Adamic (Power-law, Pareto - a ranking tutorial –www.hpl.hp.com/research/idl/papers/ranking/ranking.html) shows that Pareto and Zipf are alternative cumulative distribution representations of the same power law with independent and dependent variables reversed. She related the three main tail distributions to power law.

The Pareto distribution: Let $P(S > s) = s^{-k}$ – the probability that people have incomes above s. The **cumulative distribution** $1- s^{-k}$ is the proportion of people below s in the ranking of income (ie cumulative distribution is position/rank in a distribution). The frequency/density distribution is $P(S = s)$ $\mathbf{ks^{-(k+1)}}$ **.**

Power law linking frequency to size of objects is $Y= BS^{-a}$, so this is just Pareto with a = k+ 1.

Zipf relates size to rank R: $S = BR^{-b}$. But since rank is position in a distribution, when object of size s has rank R, there are R objects with size >s. Rewrite Zipf as $R = B^{1/b} s^{-1/b}$. Divide by # of objects T so $R/T = (B^{1/b}/T)$ $\mathbf{s^{-1/b}}$. For instance R/T is the proportion of cities with size > s. The Zipf coefficient b is thus 1/ Pareto coefficient. All three forms represent the same **power law coefficient for density a = k +1= 1+ 1/b**

A Summary Table

Distribution	"dependent" Measure	Right hand side measure	Coefficient for density
Power law	Density	Size S^{-a}	-a
Pareto	Upper tail cumulative	s^{-k}	
	Density	-(k+1)	-(k+1)
Zipf	Size	Rank/upper tail R^{-b}	
	Rank/upper tail	Size $s^{-1/b}$	
	Density	$s^{-1/b -1}$	-1/b -1

Power Law for Numbers of Papers – In 1926 Alfred Lotka,, statistician, applied mathematical scientist, creator of the Lotka-Volterra predator-prey model in ecology, one of the first analysts of human capital, founder of mathematical demography, developed Lotka's inverse square law of scientific productivity (http://www.jehps.net/juin2008/Veron.pdf is a fascinating short intellectual bio of Lotka).

N(S) = # of scientists who write S papers, $N = A S^{-2}$ so that ln N = ln A -2 ln S. Let A =100. Then this says $N = 100/ S^2$ so that 1 scientist will write 10 papers while 100 scientists will write 1 paper. The top scientist will be ten times as productive as one of the other scientists.

Example	# sci	total
10 papers	1 10	
9	1	9
8	2	16
7	2	14
6	3	18
5	4	20
4	6	24
3	11	33
2	25	50
1	10 0	100
total:	155	303

Does this distribution favor the view that the few are critical to science or the collective view that the many are important?

Top 10% scientists = 15.5 write ~97 papers or about 1/3rd of papers
But low producers (those who write 3 or less papers) produce 60% of papers

What if low producers rely on ideas of top producers?

Lotka analyzed chemists and found a power law coefficient of 1.88. Ensuing analysis gives estimates for many fields: entomology (1.9) and psychology (2.8) in Africa (Gupta, 1987, 1989), geophysics (2.1) (Gupta 1992), journal of oil seeds research (2.07) – Kalyane and Sen (1995); library and information science (Sen, Taib, Hassan, 1996), 3.23; 6 risk and insurance journals, 2.22-2.44 ,Chung and Peulz (1992), economics, 1.84; Cox and Chung (1991); 1990, finance (–) ; Worthen (1978), medicine (–); Schorr (1974), library science (---); Newby, Greenberg, Jones (2003) … and you can find many more in recent years.

Without some theory of what differences in coefficients tell us about a field or or what might generate the differences, thee is no ordering of the facts beyond that data fit a power law. Perhaps power law coefficient is larger in fields with longer papers? Perhaps it depends on the number of authors, which has risen? Possible paper.

Do Power Laws Mean Science is Super-Star?

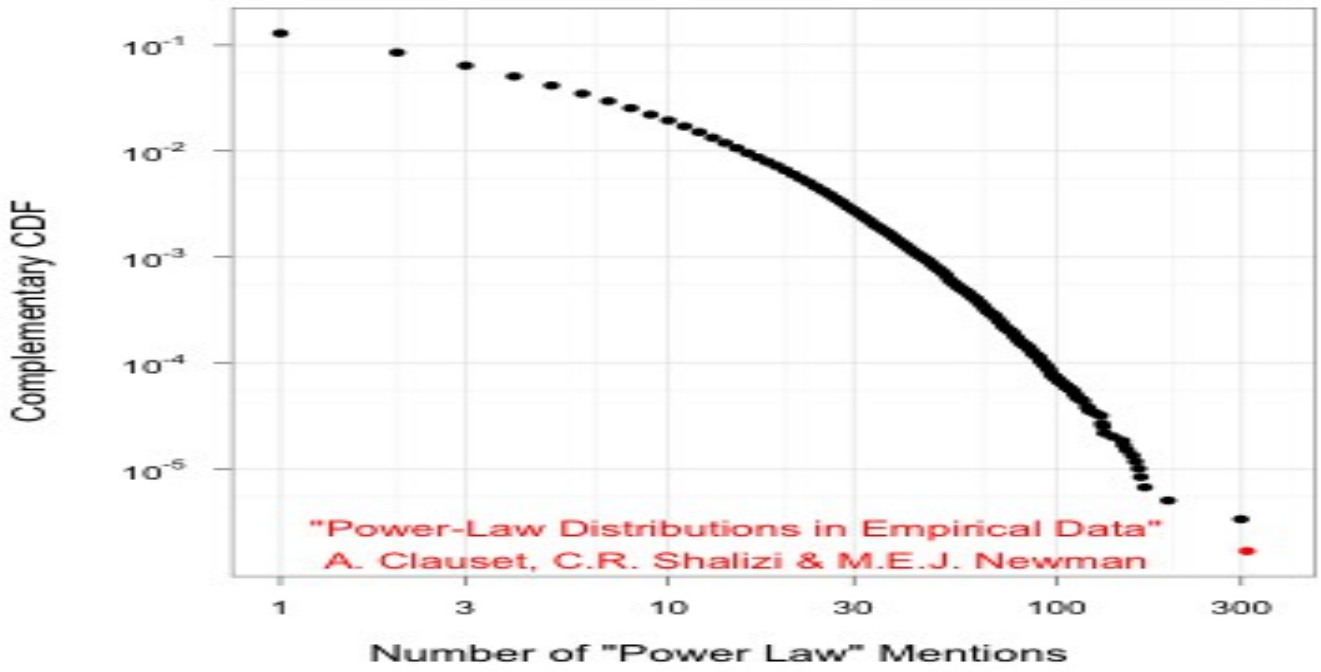
Power law production in science fuels debate over the role of individuals vs collective in production (of scientific knowledge) just as like Pareto income distribution raises issues of the wealth makers vs the 99.9%.

Superstar view --Bibliometric data shows that most papers/citations are from small number of people, which suggests that top performers – superstars – are all that matters. The implication is that to encourage scientific innovation, we must attract the best and brightest and reward them accordingly. (It's Watson and Crick, not Maurice Wilkins and Rosalind Franklin nor Pauling etc or others that discovered the double helix.) History of science is history of great persons … but also of dual discoveries and tournaments between comparably able folk.

Collective Enterprise view -Ideas are social, generated by combining/mutating previous knowledge often dependent on networks of connections. The problems worth study are set by scientific community. Individuals respond to the incentives. Merton's "Matthew effect" that the most renowned person gets more credit for a solution than others explains part of the concentration of attention on the few. Multiple discoveries reflects competition among similarly able teams, any one of which could get the answer. If we want to encourage science, must build good network structure and teams and distribute rewards to all.

The "knock-out" test of marginal productivity: What happens if "superstar" scientist goes extinct? Azoulay et al 2010 finds that it reduces the output of co-workers but does this affect the power law or does it leave "space" for someone else to move into that slot? His team's most recent work suggests the latter.

Power laws everywhere, even in number of times mention power law: Aaron Clauset,Cosma Shalizi, and Mark Newman have 310 mentions of "power law" in their paper, "Power-Law Distributions in Empirical Data."[2])



Source: http://messymatters.com/powerlaws/

3.Power Laws and citations

Analysis of citations as power law goes back to Derek de Solla Price's work. *Little Science, Big Science*, New York: Columbia University Press (1963), *Science since Babylon*, New Haven: Yale University Press (1961) and his "Networks of Scientific Papers" Science article (1965)

Citation statistics help determine careers.

> It is the Government's intention that the current method for determining the quality of university research—the UK Research Assessment Exercise (RAE)—should be replaced after the next cycle is completed in 2008. Metrics, rather than peer-review, will be the focus of the new system and it is expected that bibliometrics (using counts of journal articles and their citations) will be a central quality index in this system. [Evidence Report 2007, p. 3]

Twenty or so years ago, an ad hoc committee reviewing a proposal to tenure someone at Harvard received a letter from an outside scholar, who had not read the scientist's work but who produced an analysis of that persons' citations along with comparison to others, and based his comments on tenure on the statistical analysis of whether the life cycle of citation counts suggested the person would have as many/more/less than comparators! The candidate did not get the job because university president read one of his books and decided it was not up to snuff.*[1]

Citations vary by field. The lifespan of citations reflects the speed with which the field is changing. Anne Preston's book Leaving Science noted that since woman leave for child-bearing/rearing they should favor slow-moving fields where citations have longer "half-life" but in fact concentrate in biological sciences where evidence and techniques change rapidly.

---

1 'Up to snuff' originated in the early 19th century. In 1811, the English playwright John Poole wrote Hamlet Travestie, a parody of Shakespeare, in the style of Doctor Johnson and George Steevens, which included the expression. "He knows well enough The game we're after: Zooks, he's up to snuff." & "He is up to snuff, that is, he is the knowing one." A slightly later citation of the phrase, in Grose's Dictionary, 1823, lists it as 'up to snuff and a pinch above it', and defines the term as 'flash'. This clearly shows the derivation to be from 'snuff', the powdered tobacco that had become fashionable to inhale in the late 17th century. The phrase derives from the stimulating effect of taking snuff. The association of the phrase with sharpness of mind was enhanced by the fashionability and high cost of snuff and by the elaborate decorative boxes that it was kept in.

Figure 6 shows the distribution of citation ages from citing publications. This refers years in the past of each citation in the reference list of a given paper. Figure 7 shows the ages of citations to cited publications. For a paper published in 1980 that is cited once in 1982, twice in 1988 and three times in 1991, the citation age distribution has discrete peaks at 2, 8 and 11 years, with respectively weights 1/6, 1/3, and 1/2.
(Redner, "How popular is your paper (The European Physical Journal B - Condensed Matter and Complex Systems Volume 4, Number 2, 131-134) is a very popular paper in this area.)

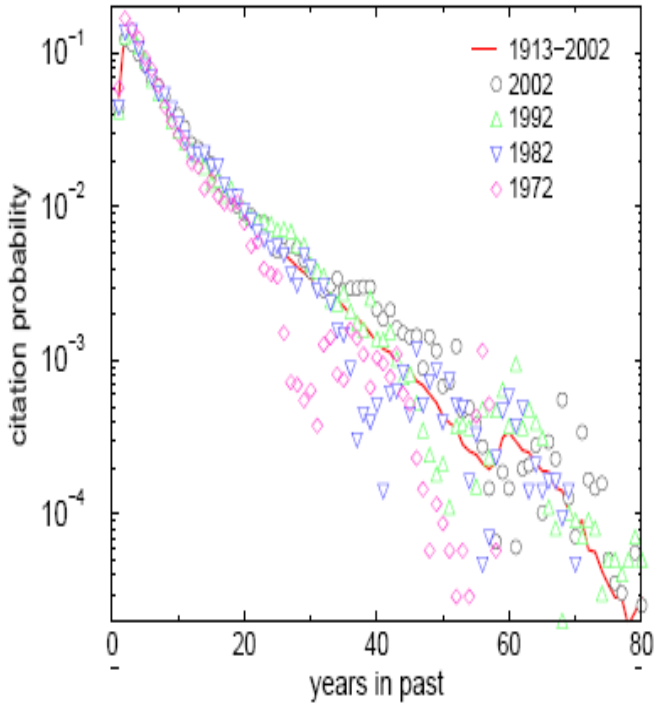


FIG. 6: The distribution of the ages of citations contained in the reference lists of publications that were published in selected years. Also shown is this same citing age distribution for the period 1913-2002.

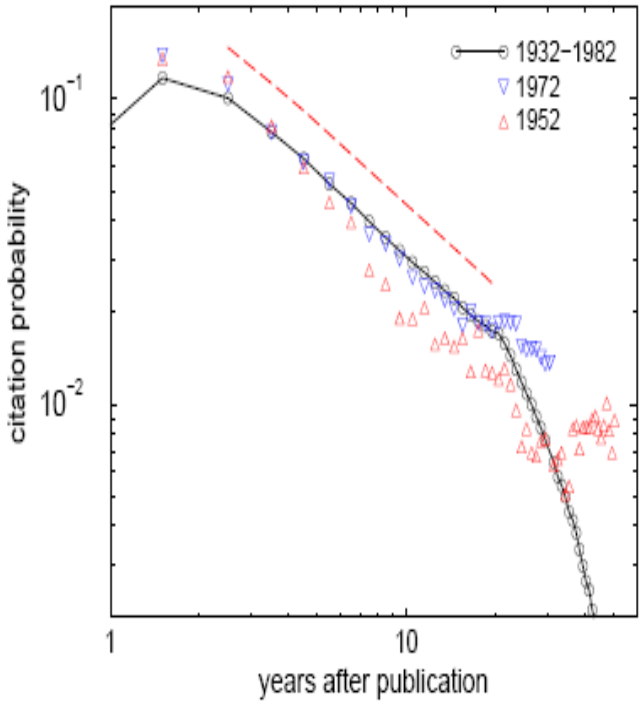


FIG. 7: Distribution of the ages of citations to cited papers in selected years, as well as the integrated data over the period 1932-1982. The dashed line is the best fit to the data in the range 2 – 20 years (displaced for visibility).

Numerical data for the distribution of citations are examined for: (i) papers published in 1981 in journals which are catalogued by the Institute for Scientic Information (783,339 papers) and (ii) 20 years of publications in Physical Review D, vols. 11-50 (24,296 papers). A Zipf plot of the number of citations to a given paper versus its citation rank appears to be consistent with a power-law dependence for leading rank papers, with exponent close to −1/2.

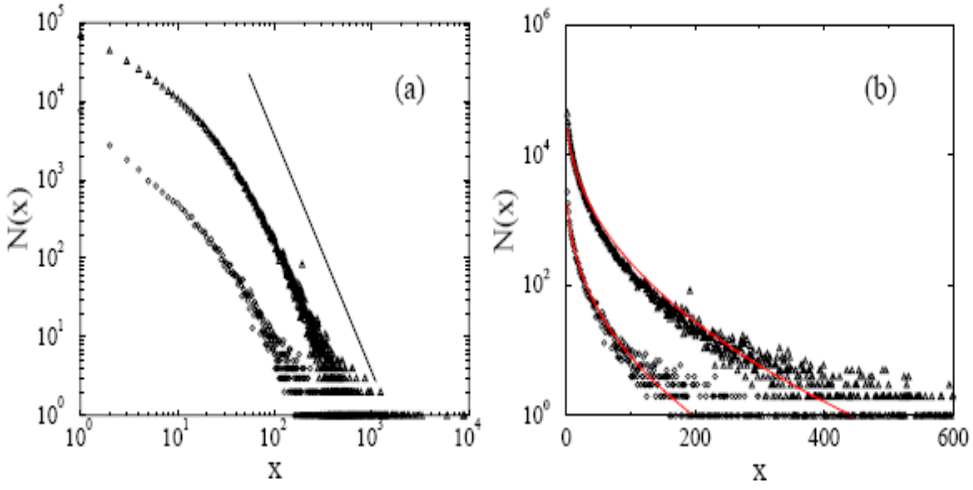


Fig. 1. (a) Citation distribution from the 783,339 papers in the ISI data set (△) and the 24,296 papers in the PRD data set (○) on a double logarithmic scale. For visual reference, a straight line of slope −3 is also shown. (b) Same as (a), except on a semi-logarithmic scale. The solid curves are the best fits to the data for $x \leq 200$ (PRD) and $x \leq 500$ (ISI).

This, in turn, suggests that the number of papers with x citations, N(x), has a large-x power law decay N(x)~ x−3.

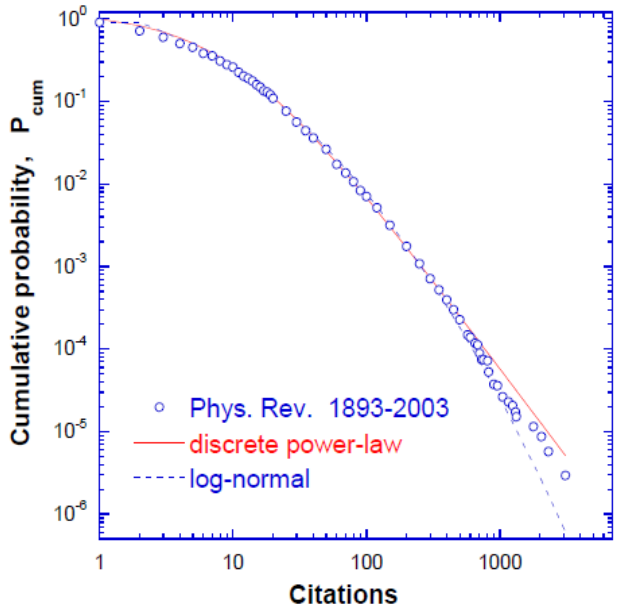


Fig. 1. Cumulative probability distribution (cdf) of citations to 353,268 papers published in Physical Review journals during 1893-2003 and cited by 2003. Only PR to PR citations were counted. The data were adapted from Ref. [13]. The continuous red line shows a fit with the discrete-power-law cdf (Eq.1) with $\gamma = 3.15, w = 10.2$. The dashed blue line shows a fit with the log-normal cdf (Eq.3) with $\mu = 1.15, \sigma = 1.42$.

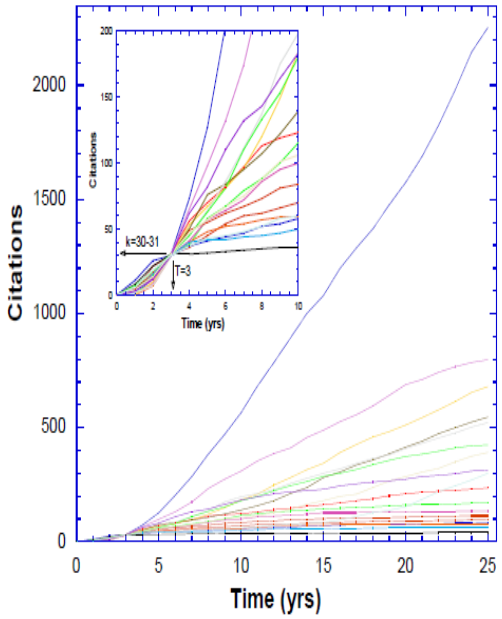


Fig. 4. Citation dynamics of 89 Physics papers published in 1984. We chose all those papers that by 1986 (three years after publication) had 30 or 31 citations. Although the initial citation dynamics of these papers is very similar, it quickly diverges in such a way that after 25 years (in 2008) the number of citations varies between 40 and 2254.

But even power laws have a problem at upper tail. Golosovsky and Solomon" Runaway events dominate the heavy tail of citation distributions" measured citation distribution for 418,438 Physics papers published in 1980-1989 and cited by 2008: "Discrete power law function with exponent of 3.15 beats log-normal fit and fits 99.955% of the data. However, the extreme tail of the distribution deviates upward even from the power-law fit and exhibits a dramatic "runaway" behavior. The onset of the runaway regime is revealed macroscopically as the paper garners 1000-1500 citations, however the microscopic measurements of autocorrelation in citation rates are able to predict this behavior in advance. Over time, the papers in the tail grow at a much faster rate than the rest of the distribution, indicating the runaway effect."

Many indicators built on the Thomson Scientific impact factors. For instance, there is a recursive impact factor that gives citations from journals with high impact greater weight than citations from low-impact journals (see http://eigenfactor.org/).

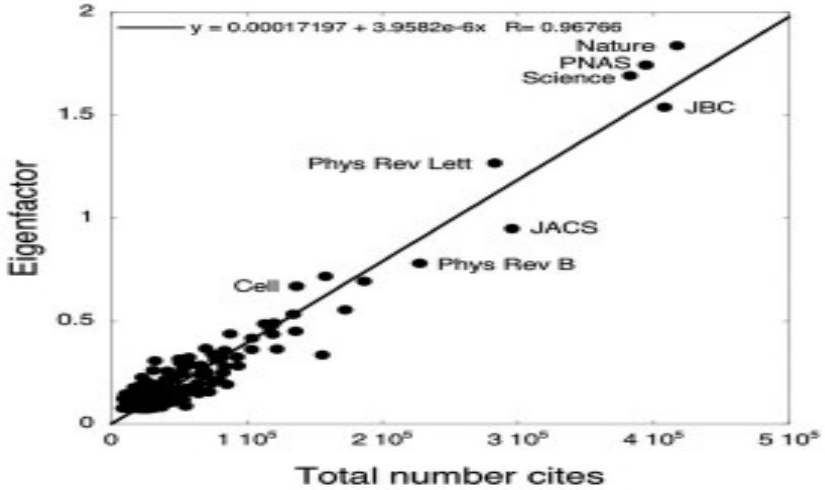


Fig. 1. Plot of the 2007 Eigenfactor rating against total number of citations listed in the *Journal Citation Reports*®.

Citations widely used to judge journals and departments

Here is ranking of Physics Departments by John Perdew and Frank Tipler of Tulane University and published in Physics Today, October 1996, p. 15

Top 20 U.S. physics departments by number of citations per scientific paper published (1981-94)

University	Papers	Citations	Impact	NRC ranking
1. Princeton University	4,252	88,150	20.7	2
2. Harvard University	3,541	72,372	20.4	1
3. Tulane University	265	5,338	20.1	115.5
4. UC Santa Barbara	4,306	83,256	19.3	10
5. University of Chicago	2,439	45,729	18.8	7
6. Brandeis University	559	10,339	18.5	42.5
7. UC Santa Cruz	709	13,068	18.4	47.5
8. Calif. Institute of Tech.	4,027	72,393	18.0	5
9. Univ. of Pennsylvania	3,047	53,854	17.7	17
10. Rockefeller University	523	8,597	16.4	30
11. Stanford University	6,659	105,736	15.9	9
12. Yale University	1,971	31,109	15.8	13
13. S.U.N.Y. at Stony Brook	3,052	43,871	14.4	22.5
14. Mass. Inst. of Tech.	9,382	132,948	14.2	3.5
15. UC Berkeley	5,474	75,411	13.8	3.5
16. Cornell University	4,776	63,605	13.3	6
17. UC Riverside	661	8,497	12.9	68.5
18. Michigan State Univ.	1,995	25,585	12.8	32
19. Tufts University	623	7,953	12.8	77
20. Illinois (Urb.-Cham.)	6,627	84,229	12.7	8

Note the difference between NAS-NRC ranking and ranking by rank by citations per paper.

And note that MIT would rank better by total citations.

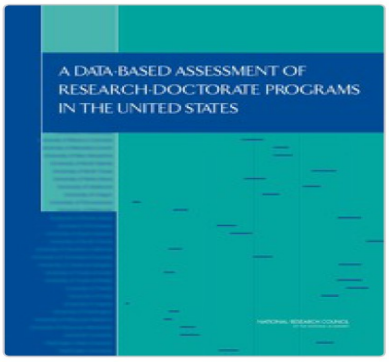
More papers-->more cites. Is cites per paper be the right metric? Should we expect a negative relation between # papers and cites per paper? And does more papers by department mean some of cites are self-department and should be downvalued?

How does this connect with the Allison and Long paper that said citations follow department rather than department chooses most cited? VERY NEAT AREA FOR PAPER
The 2014 NAS-NRC ratings show Tulane doing better

1--3	Harvard University Physics
☐ 1-6	Princeton University Physics
☐ 1-9	University of California-Berkeley Physics
☐ 2-12	Massachusetts Institute of Technology Physics
☐ 2-14	University of California-Santa Barbara Physics
☐ 3-18	Harvard University DEAS-Applied Physics
☐ 3-27	University of Hawaii at Manoa Physics
☐ 4-22	California Institute of Technology Physics

☐ 4-21	Pennsylvania State University-Main Campus Physics
☐ 4-22	University of Chicago Physics
☐ 4-23	University of Pennsylvania Physics and Astronomy
☐ 5-24	Columbia University in the City of New York Physics
☐ 6-30	Boston University Physics
☐ 8-35	Cornell University Physics
☐ 7-41	Yale University Physics
☐ 8-40	Stanford University Physics
☐ 8-40	University of California-Irvine Physics
☐ 10-41	California Institute of Technology Applied Physics
☐ 10-47	Carnegie Mellon University Physics
☐ 9-45	Tulane University of Louisiana Physics

For underlying data see



A Data-Based Assessment of Research-Doctorate Programs in the United States (with CD)

(https://www.nap.edu/catalog/12994/a-data-based-assessment-of-research-doctorate-programs-in-the-united-states-with-cd)

The Data Table in Excel includes data from more than 5,000 doctoral programs offered at 212 universities across the United States. This rich resource allows evaluation and comparison of programs in areas such as faculty research activity, student support and outcomes, and diversity of the academic environment. Three formats of the spreadsheet are available. The Windows and Excel 2004 and 2011 for Mac versions are optimized for users through the use of macros that enable customized filtering and click-through to background data

In 2005 Hirsch, J. E. published "An index to quantify an individual's scientific research output". 102 (46): 16569-16572, which creates an index that downweights having a single paper cited multiple times (because some may give mundane statistic - the newest digit on PI) and measures persons “productivity” in terms of a # of papers each of which has been cited at least h times, to reflect both the number of publications and the number of citations per publication. This paper was 9th most cited PNAS in Jan 2010. "In terms of a "usage' metric, Hirsch's h-index paper (3) is exceptional in its number of downloads (111,126 downloads versus 262 citations since it was published in November 2005). But they are all closely related AND as of 2018 citations to Hirsch are 7,384!!!

Four Spanish economists have estimated citation power laws for 23 aggregate fields and 250 sub-fields from the Thomson web of science data set – 8.5 million articles and 65 million citations from 1998-2007 – and found that 77% had power law distributions with most parameters > 3. But references given > citations received because data set is not complete to all possible references.

Table 3. The Distribution of Citations Received and References Made

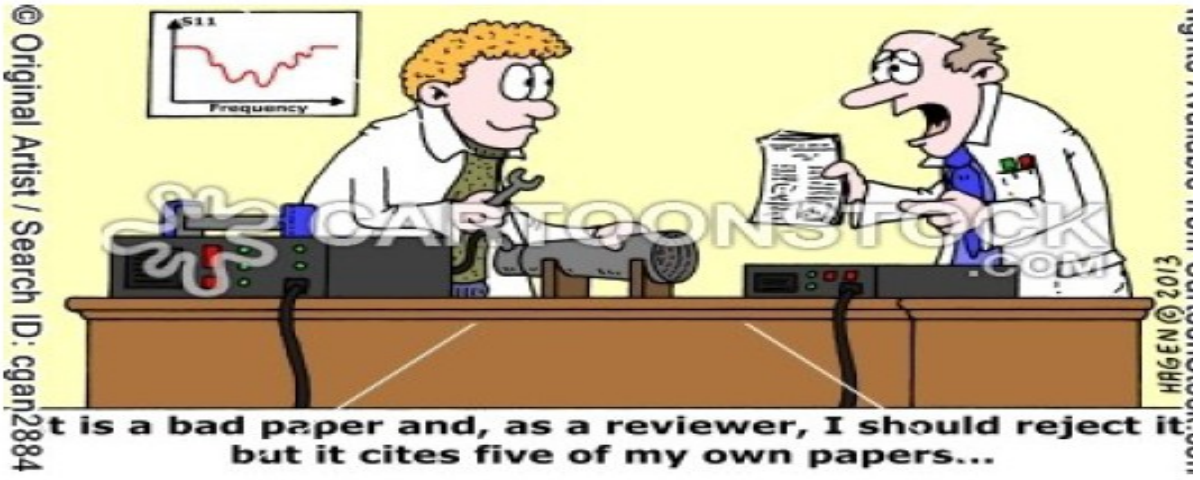
Panel A: The Entire Dataset

	Citations				References		Refs./Citat
	% of zeroes	Median	%5 Most cited 95-Percetile	% Over Total	Median	95-Percetile	
LIFE SCIENCES							
(1) Clinical Medicine	23.7	3	39	41.3	24	57	2.6
(2) Biol & Biochemistry	17.1	6	48	34.2	33	67	2.7
(3) Neurosci & Behav Sci	15.5	7	54	33.1	37	76	2.7
(4) Molec Biol & Genetics	14.9	8	79	38.6	38	73	2.0
(5) Psychiatry/Psychology	27.0	3	33	38.8	34	76	4.6
(6) Pharma & Toxicology	21.2	4	31	33.9	28	59	3.6
(7) Microbiology	16.5	6	43	30.9	32	65	2.9
(8) Immunology	12.8	8	60	32.8	35	66	2.2
PHYSICAL SCIENCES							
(9) Chemistry	25.6	3	31	35.6	23	60	3.4
(10) Physics	28.9	2	28	41.7	18	50	3.2
(11) Computer Science	55.7	0	11	55.4	16	44	7.2
(12) Mathematics	44.4	1	11	42.4	15	39	6.7
(13) Space Science	23.2	4	42	37.2	30	74	3.0
OTHER NATURAL							
(14) Engineering	45.2	1	14	42.6	15	43	5.5
(15) Plant & Animal Sci	30.1	2	22	36.3	28	64	5.4
(16) Material Science	38.9	1	19	43.2	16	43	4.0
(17) Geosciences	29.4	2	28	36.2	30	76	4.9
(18) Environment/Ecology	24.8	3	30	34.3	31	70	4.4
(19) Agricultural Sciences	33.3	2	21	36.7	24	53	5.0
(20) Multidisciplinary	45.0	1	20	50.9	14	56	4.4
SOCIAL SCIENCES							
(21) Social Sci, General	42.4	1	15	41.3	30	78	9.6
(22) Econ & Business	44.3	1	18	47.8	24	71	6.7
ARTS&HUMANITIES							
(23) Arts & Humanities	83.0	0	2	84.6	14	67	33.3

In the Lancet, " there was a strong association between increasing title length and citation rate, with the highest-scoring articles having more than twice as many words in the title than the lowest-cited articles." (Jacques, T. and N. Sebire, "The Impact of article titles on citation hits: an analysis of general and specialist medical journals" J.R. Soc Med, Special Reports, 2010, 1, 2

Sonnert finds publications are highly correlated with peer evalutions with the only other element that mattered being graduate school prestige! (What Makes a Good Scientist?: Determinants of Peer Evaluation among Biologists Social Studies of Science, Vol. 25, No.1 (Feb., 1995), pp. 35-55)

Citations as Choice Variable



Assume more citations helps your career. You want to decide whether to research/publish in hot growing area or in some more somnolent area. Will your paper be more/less cited in growth field or stable field?

New papers have two effects on citations of older papers. More new papers → more citations to older papers. I**n most markets when more competitors enter, this harms current producers by driving down prices and profits, but in science, the more people that enter the greater the likelihood someone will cite you.**

But new papers tend to cite new papers. Your analysis of XYZ has been replaced by Jones & Wang's analysis. They cited you but now people cite them. If your paper generated 10 "progeny" papers that do your 10 major results/ideas better than you did, your paper may disappear. Your career may be in trouble … unless we generate a new statistic that considers the references to the papers that referred to you.

This creates a differential equation model in which new papers both add to and reduce the citations to older papers. Let CITE (t,t+j) be the number of citations to an article published in year t j years after publication. Let ART (t+ j) be the number of articles published in year t+j. Then the reduced form of the impact of future articles published on your year t article is:

CITE (t, t+j) = a ART(t+j) for each year, with + a implying more articles published increases cites to you and negative meaning more articles published reduces cites to you. Expect more articles published has positive effect on cites to given article until the "newer version" comes along and cites to the t year paper falls. But of course cumulated cites can only go up.

5. Power Econometrics and Mechanisms

There are statistical problems in fitting power laws because the number of rare events - scientists who produce lots and lots of papers - is sparse. See A. Clauset, C.R. Shalizi, and M.E.J. Newman, "Power-law distributions in empirical data" SIAM Review 51(4), 661-703 (2009). (arXiv:0706.1062, doi:10.1137/070710111) if the estimation process interests you.

1) Fitting a log-log line by least squares not very informative. Sds etc not valid as depend on normality "fitting lines on log-log graphs is what Pareto did back in the day when he started this in *1890s"*

2)Use maximum likelihood to estimate the scaling exponent. Sampling distribution is an inverse gamma to get confidence intervals.

3)Use Kolmogorov -Smirnov goodness-of-fit statistic to estimate where the scaling region begins. goodness of fit of a *distribution*, use a statistic meant for *distributions*, not R-squared etc as you get more data

4)If you care compare with non-power law distributions

5)Doing things with cumulative often better

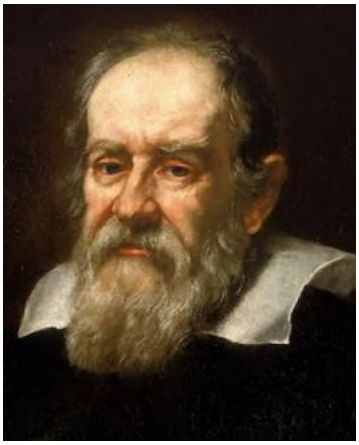
PART II Team science – network analysis

There are two "narratives" about research discoveries:

the standing on shoulders of giants collective enterprise view that the scientific community and market for ideas produces knowledge. According to this view, researchers work on problems set by the scientific community by combining/ mutating previous knowledge obtained through networks of connections. Your new results/paper is a predictable outcome from the 15 papers that you cite. Your experiment/innovation was on the drawing board on many other scientists as the next step in the research program. You just got there first."

The "Matthew effect" (in which the most renowned person gets more credit for a solution than others) explains part of the concentration on the few. Human desire to have "heroes and stories" or the efficiency of tournaments may help explain an overemphasis on individuals. Multiple discoveries proves that outcomes result from competition among similarly able teams, any one of which could get the answer. If we want to encourage science, must build good network structure and teams and distribute rewards to all.

the great scientist view that a few brilliant minds drive scientific progress. No one but Newton, Darwin, Einstein, etc could have conceived the ideas associated with them. Without Watson-Crick, the world might have waited for years to understand how DNA replicates. The division between the great scientists and the others is discontinuous following a very steep power law.



The Fermi Gamma-Ray Space Telescope Discovers the Pulsar in the Young Galactic Supernova Remnant CTA 1

A. A. Abdo,[1,2] M. Ackermann,[3] W. B. Atwood,[4] L. Baldini,[5] J. Ballet,[6] G. Barbiellini,[7,8] M. G. Baring,[9] D. Bastieri,[10,11] B. M. Baughman,[12] K. Bechtol,[3] R. Bellazzini,[5] B. Berenji,[3] R. D. Blandford,[3] E. D. Bloom,[3] G. Bogaert,[13] E. Bonamente,[14,15] A. W. Borgland,[3] J. Bregeon,[5] A. Brez,[5] M. Brigida,[16,17] P. Bruel,[13] T. H. Burnett,[18] G. A. Caliandro,[16,17] R. A. Cameron,[3] P. A. Caraveo,[19] P. Carlson,[20] J. M. Casandjian,[6] C. Cecchi,[14,15] E. Charles,[3] A. Chekhtman,[2,21] C. C. Cheung,[22] J. Chiang,[3] S. Ciprini,[14,15] R. Claus,[3] J. Cohen-Tanugi,[23] L. R. Cominsky,[24] J. Conrad,[20,25] S. Cutini,[26] D. S. Davis,[2,27] C. D. Dermer,[2] A. de Angelis,[28] F. de Palma,[16,17] S. W. Digel,[3] M. Dormody,[4] E. do Couto e Silva,[3] P. S. Drell,[3] R. Dubois,[3] D. Dumora,[29,30] Y. Edmonds,[3] C. Farnier,[23] W. B. Focke,[3] Y. Fukazawa,[31] S. Funk,[3] P. Fusco,[16,17] F. Gargano,[17] D. Gasparrini,[26] N. Gehrels,[22,32] S. Germani,[14,15] B. Giebels,[13] N. Giglietto,[16,17] F. Giordano,[16,17] T. Glanzman,[3] G. Godfrey,[3] I. A. Grenier,[6] M.-H. Grondin,[29,30] J. E. Grove,[2] L. Guillemot,[29,30] S. Guiriec,[23] A. K. Harding,[22] R. C. Hartman,[22] E. Hays,[22] R. E. Hughes,[12] G. Jóhannesson,[3] A. S. Johnson,[3] R. P. Johnson,[4] T. J. Johnson,[22,32] W. N. Johnson,[2] T. Kamae,[3] Y. Kanai,[33] G. Kanbach,[34,*] H. Katagiri,[31] N. Kawai,[33,35] M. Kerr,[18] T. Kishishita,[36] B. Kiziltan,[37] J. Knödlseder,[38] M. L. Kocian,[3] N. Komin,[6,23,†] F. Kuehn,[12] M. Kuss,[5] L. Latronico,[5] M. Lemoine-Goumard,[29,30] F. Longo,[7,8] V. Lonjou,[29,30] F. Loparco,[16,17] B. Lott,[29,30] M. N. Lovellette,[2] P. Lubrano,[14,15] A. Makeev,[21] M. Marelli,[19] M. N. Mazziotta,[17] J. E. McEnery,[22] S. McGlynn,[20] C. Meurer,[25] P. F. Michelson,[3] T. Mineo,[39] W. Mitthumsiri,[3] T. Mizuno,[31] A. A. Moiseev,[40] C. Monte,[16,17] M. E. Monzani,[3] A. Morselli,[41] I. V. Moskalenko,[3] S. Murgia,[3] T. Nakamori,[33] P. L. Nolan,[3] E. Nuss,[23] M. Ohno,[36] T. Ohsugi,[31] A. Okumura,[42] N. Omodei,[5] E. Orlando,[34] J. F. Ormes,[43] M. Ozaki,[36] D. Paneque,[3] J. H. Panetta,[3] D. Parent,[29,30] V. Pelassa,[23] M. Pepe,[14,15] M. Pesce-Rollins,[5] G. Piano,[41] L. Pieri,[10] F. Piron,[23] T. A. Porter,[4] S. Rainò,[16,17] R. Rando,[10,11] P. S. Ray,[2] M. Razzano,[5] A. Reimer,[3] O. Reimer,[3] T. Reposeur,[29,30] S. Ritz,[22,32] L. S. Rochester,[3] A. Y. Rodriguez,[44] R. W. Romani,[3] M. Roth,[18] F. Ryde,[20] H. F.-W. Sadrozinski,[4] D. Sanchez,[13] A. Sander,[12] P. M. Saz Parkinson,[4] T. L. Schalk,[4] A. Sellerholm,[25] C. Sgrò,[5] E. J. Siskind,[45] D. A. Smith,[29,30] P. D. Smith,[12] G. Spandre,[5] P. Spinelli,[16,17] J.-L. Starck,[6] M. S. Strickman,[2] D. J. Suson,[46] H. Tajima,[3] H. Takahashi,[31] T. Takahashi,[36] T. Tanaka,[3] J. B. Thayer,[3] J. G. Thayer,[3] D. J. Thompson,[22] S. E. Thorsett,[4] L. Tibaldo,[10,11] D. F. Torres,[44,47] G. Tosti,[14,15] A. Tramacere,[3,48] T. L. Usher,[3] A. Van Etten,[3] N. Vilchez,[38] V. Vitale,[41] P. Wang,[3] K. Watters,[3] B. L. Winer,[12] K. S. Wood,[2,*] H. Yasuda,[31] T. Ylinen,[20,49] M. Ziegler[4,*]

Team-based science Wuchty, Jones, and Uzzi analyzed 19.9 million papers in WOS from 1955 to 2000 and report rise in multi-authored papers in all fields x arts & humanities. Is history science or humanities? Check # of authors.

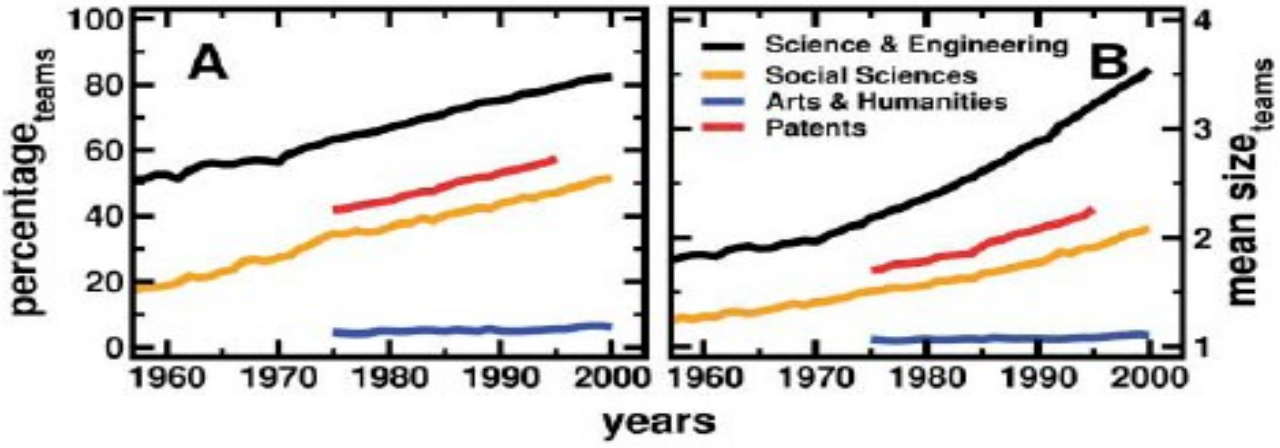


Fig. 1. The growth of teams. These plots present changes over time in the fraction of papers and patents written in teams (**A**) and in mean team size (**B**). Each line represents the arithmetic average taken over all subfields in each year.

Some facts about collaborations (Social Studies of Science, Oct 2005: (Lee and Bozeman "Impact of Scientific Collaboration on Productivity")

TABLE 2
Research time

Work setting	*N*	Mean percentage of research time	SD
Research time working alone	405	15.93	20.01
Research time working with researchers and graduate students in my immediate work group	405	51.10	23.85
Research time working with researchers in my university, but outside my immediate work group	405	11.44	12.66
Research time working with researchers who reside in nations other than the USA	405	5.11	7.73
Research time working with researchers in US universities other than my own	405	8.21	10.63
Research time working with researchers in US industry	405	5.23	7.94
Research time working with researchers in US government laboratories	405	2.98	6.53

Shapiro,et al JAMA feb 1994, notice # of tasks – last author as manager;would be great to relate hours to cites, etc

Table 3.—First Authors' Assessments of the Contributions of Authors to Specific Tasks

Task*	All Authors Together (n=1014), %	First Authors (n=184), %	Second Authors (n=175), %	Middle Authors (n=479), %	Last Authors (n=176), %
Initial conception	42	90	34	19	64
Design	47	97	41	24	61
Provision of resources	68	72	62	62	85
Data collection	54	89	62	45	34
Analysis and interpretation of data	52	98	56	30	61
Writing and revision	57	100	55	33	80
Total No. of tasks contributed to					
0 or 1	24	0	17	42	10
2 or 3	32	3	46	40	29
4, 5, or 6	43	97	37	18	61

*For each task, $P<.0001$ for differences among author positions.

Table 4.—First Authors' Estimates of the Number of Hours Authors Spent Contributing Directly to the Research

No. of Hours*	All Authors Together (n=966), %	First Authors† (n=176), %	Second Authors† (n=167), %	Middle Authors† (n=456), %	Last Authors† (n=167), %
0-10	20	0	13	31	20
11-50	26	4	27	31	35
51-500	36	43	42	31	37
>500	18	53	18	7	8

*Hours categories are collapsed from the questionnaire's categories of 0, less than 1, 1 to 5, 6 to 10, 11 to 25, 26 to 50, 51 to 100, 101 to 500, 501 to 1000, and more than 1000.
†$P<.0001$ for differences among these four author positions.

Four questions about teams

Question 1: Are teams more/ less productive than individuals? Could measure whether produce more papers in given period but that requires counter-factual of what would have produced separately. And requires some fractional count (Wuchty, et al Science, May 2007)

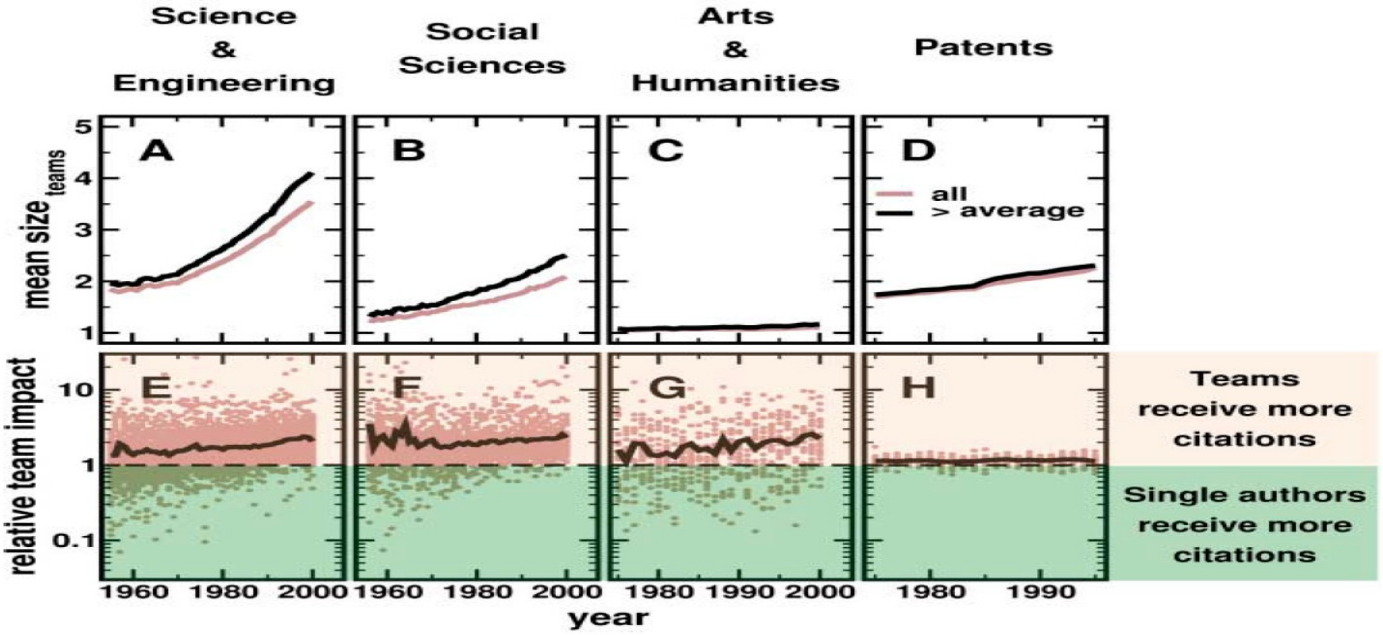


Fig. 2. The relative impact of teams. (**A** to **D**) Mean team size comparing all papers and patents with those that received more citations than average in the relevant subfield. (**E** to **H**) The RTI, which is the mean number of citations received by team-authored work divided by the mean number of citations received by solo-authored work. A ratio of 1 indicates that team- and solo-authored work have equivalent impact on average. Each point represents the RTI for a given subfield and year, whereas the black lines present the arithmetic average in a given year.

They calculate RTI – relative team impact = cites to "team authored" **paper**/cites to solo authored paper 1955-2001 : RTI 1.7 to 2.1 BUT authors per paper 1.9 to 3.5 so cites/**author** 0.9 to 0.6 fell while 2 authors' have relative rise from 1.3 to 1.74

Table 2 - Impact of author and foreign collaboration

Science	Authors	Countries	Papers (P)	Citations (C)	Ave. C/P
All	One and many	Any	376,226	2,411,789	6.4
	One	One[1]	74,481	285,536	3.8
	One	Many[2]	1,505	7,705	5.1
	Many	One	236,592	1,525,400	6.5
	Many	Many	63,648	593,148	9.3

Table 3 - Impact of author and domestic institution collaboration

Science	Authors - domestic Institutions	Papers (P)	Citations (C)	Ave. C/P
All	One - One	72,350	278,514	3.9
	One-Many	2,131	7,022	3.3
	Many - One	175,741	1,086,179	6.2
	Many - Many	60,851	439,221	7.2

Table 7 - Authors from the same institution

Authors	Papers (P)	Citations (C)	Ave. C/P
1	72,350	278,514	3.8
2	85,486	444,010	5.2
3	49,751	304,714	6.1
4	23,263	168,804	7.3
5	10,048	86,576	8.6
6	4,217	42,285	10.0
7	1,718	21,245	12.4
8	663	8,708	13.1
9	283	4,600	16.3
10	132	2,160	16.4

How much is a collaboration worth? J. S. Katz, Diana Hicks (Scientometrics,Nov 1997)

Is there an optimal team size? James D. Adams, Grant C. Black, J. Roger Clemmons, Paula E. Stephan Research Policy 34 (2005) 259–285 Scientific teams and institutional collaborations: Evidence from U.S. Universities, 1981–1999

Table 12
Determinants of research

Variable or statistic	Log (Citations over 5 years)		
	Eq. (12.4)	Eq. (12.5)	Eq. (12.6)
Time period	1981–1995	1981–1995	1981–1995
Fields included	All 12 main fields	All 12 main fields	All 12 main fields
Year dummies included	Yes, significant	Yes, significant	Yes, significant
Field dummies included	Yes, significant	Yes, significant	Yes, significant
Log (stock of federally funded R&D)	0.553 (69.6)**	0.546 (68.0)**	0.557 (69.6)**
Log (authors per paper)	0.312 (10.2)**		0.264 (8.4)**
Log (university-field authors per paper)		0.548 (10.8)**	
Top 110 U.S. university share per paper			1.276 (4.0)**
Foreign share per paper			1.237 (5.2)**
U.S. corporate share per paper			0.094 (0.2)
Root M.S.E.	0.688	0.687	0.686
Adjusted R^2	0.82	0.82	0.82
Number of observations	8,504	8,504	8,504

But the production function presumably differ between large and small papers. Calculations do not include capital equipment, need for experts with different skills, time spent producing papers, and "opportunity cost". ENDOGENEITY OF CO-AUTHORSHIP. Probably useful to examine same author, with others etc.

Question 2: What makes for productive team? NAS-NRC 2015 study

Whooley et al. "Evidence for a collective intelligence factor in the performance of human groups(Science, 30 Sept 2010)

What is collective intelligence? Ability to perform wide variety of tasks related to a measure of the performance of the group on other tasks that is independent from the IQ of its members.

How would you show this? IQ asks paper/pencil questions in different areas – math, reading, problem solving – and correlate results across test domains and with other tasks – producing single general measure that is related to many tasks. Use factor analysis based on correlation of answers to reduce dimensionality of data. Single factor explain 30-50% of variance – general intelligence, not math, reading, three-dimensional, etc.

To measure CQ, Wooley et al assign 120 people to 40 three person teams, give them tasks, and see if some groups do better. Tasks drawn from McGrath Task Circumplex – solving puzzles, dividing limited resources, making moral judgments. Second study with 152 people with groups of different size. To calculate group intelligence, used set of tests:

The tests were checkers game against computer, architectural design problem.

Brainstorming(Quadrant I). Groups spent 10 minutes brainstorming possible uses for a brick. Groups received one point for each non-redundant idea they generated, independent of quality of the ideas.

Group Matrix Reasoning (Quadrant II). Groups completed the even-numbered questions of RAPM questions as a group. Groups were scored on the number of items answered correctly.

Group Moral Reasoning (Quadrant II). Using the "Disciplinary Action Case" (*3*), groups decided on disciplinary actions in a fictitious case in which a college basketball player bribed an instructor to change his grade on an exam. The groups were given a list of five issues

Based on task Circumplex

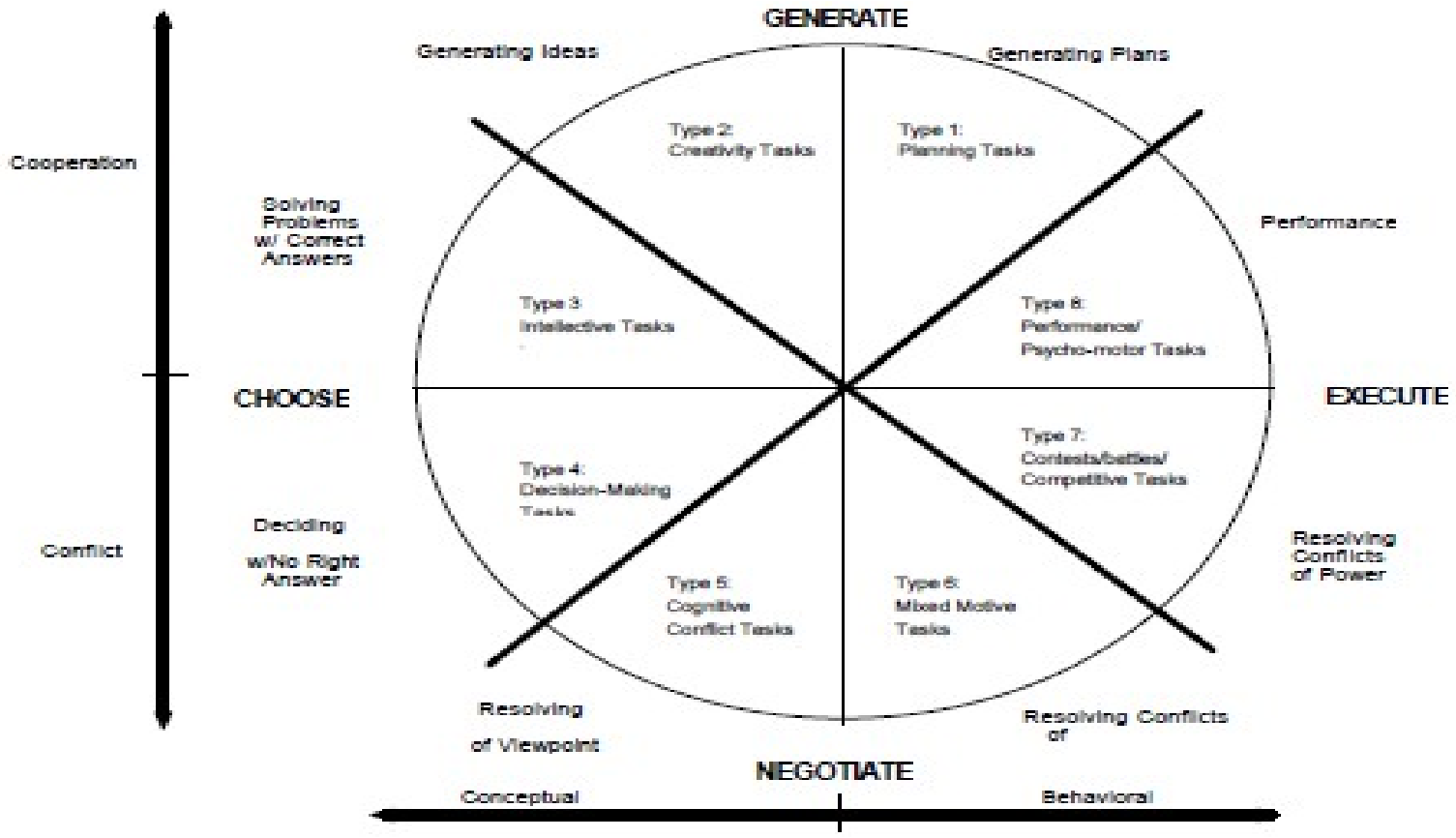


Figure 2
The Task Circumplex
(McGrath, 1984)

Table S2
Results of OLS Regression Analyses of Effects of Average Member Intelligence and Collective Intelligence on Criterion Tasks in Study 1 and Study 2

	Study 1: Video Game (n=40)			**Study 2: Architectural Design** (n=152)			
	Step 1	**Step 2**	**Step 3**	**Step 1**	**Step 2**	**Step 3**	**Step 4**
Number of members[a]				-0.04	-0.03	-0.20	-0.27*
Average Member Intelligence	0.18	0.08			0.18	0.05	
Maximum Member Intelligence			0.01				0.12
Collective Intelligence		0.51**	0.53**			0.36*	0.37*
F	1.21	7.14**	6.94**	0.20	2.62	6.59**	6.58**
R^2	0.03	0.28**	0.27**	0.04	0.16	0.34**	0.35**
change R^2		0.25**	0.24**		0.12	0.18*	0.19*

* significant at p<.05, two-tailed ** signficant at p<.001, two-tailed

[a] Number of members is constant for Study 1 and thus not a variable in the analysis

What makes the groups more effective?

Table S4

Results of OLS Regression Analyses of Effects of Percent Female, Average Member Social Sensitivity and Speaking Turn Variance on Collective Intelligence (n=46 groups)

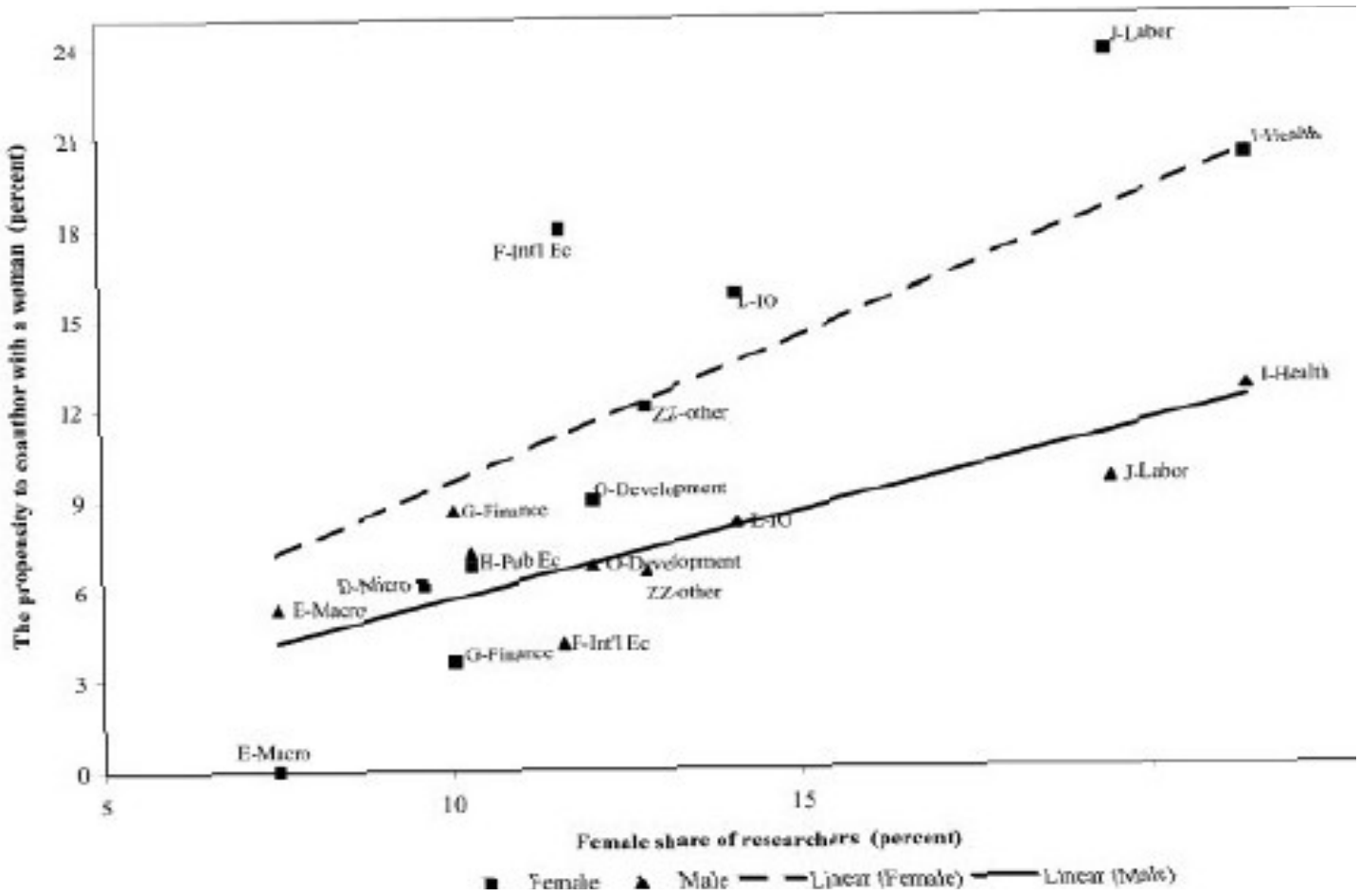
	Collective Intelligence *(c)*			
	Step 1	**Step 2**	**Step 3**	**Step 4**
Number of members	0.09	0.19	0.21	0.28
Percent Female		0.40*	0.26	0.25
Social Sensitivity			0.37*	0.33*
Speaking Turn Variance				–0.27
F	0.28	3.01	3.91*	3.87*
R^2	0.08	0.16	0.27	0.34
change R^2		0.08	0.11*	0.07

* Coefficient is significant at p<.05, two-tailed

Missing from analysis: Financial/other incentive for better performance/experimentation with different reward systems. And discussion of division of credit

Q3 Who writes with whom? Homophily Economics: women tend to write more papers with women; (Boschini &Sjogren, Is team formation neutral? Journal of Labor Economics, 2007, 25, 325-365

T



Ethnic groups write more with persons of same ethnicity.

Ethnicity	Authors' ethnicity distribution by position (%)				Probability of all authors same ethnicity (%)			Ratio (6)/(5)
	First	Second	Third	Fourth	Random	Realized	Difference (6) - (5)	
Panel A: Two-author paper								
CHN	16.63	9.15			1.52	4.16	2.64	2.73
ENG	49.80	60.21			29.99	33.56	3.57	1.12
EUR	12.76	14.65			1.87	2.27	0.40	1.22
HIN	7.71	6.53			0.50	1.61	1.10	3.19
HIS	4.57	3.76			0.17	0.43	0.26	2.50
JAP	2.24	1.31			0.03	0.27	0.24	9.23
KOR	2.39	1.02			0.02	0.14	0.11	5.58
RUS	3.55	3.15			0.11	0.40	0.29	3.55
VNM	0.35	0.23			0.00	0.01	0.01	11.13
Panel B: Three-author paper								
CHN	16.30	10.49	8.08		0.14	1.72	1.58	15.36
ENG	49.76	45.42	62.19		14.06	18.47	4.42	1.62
EUR	12.76	10.58	14.74		0.20	0.31	0.11	1.90
HIN	7.92	5.41	5.87		0.03	0.43	0.41	21.36
HIS	4.82	3.55	3.78		0.01	0.10	0.10	19.56
JAP	2.60	1.59	1.37		0.00	0.12	0.12	212.52
KOR	2.26	1.30	0.93		0.00	0.03	0.03	117.43
RUS	3.22	2.43	2.83		0.00	0.04	0.04	19.31
VNM	0.37	0.28	0.21		0.00	0.00	0.00	-

Q4What induces people to collaborate with others in a team?

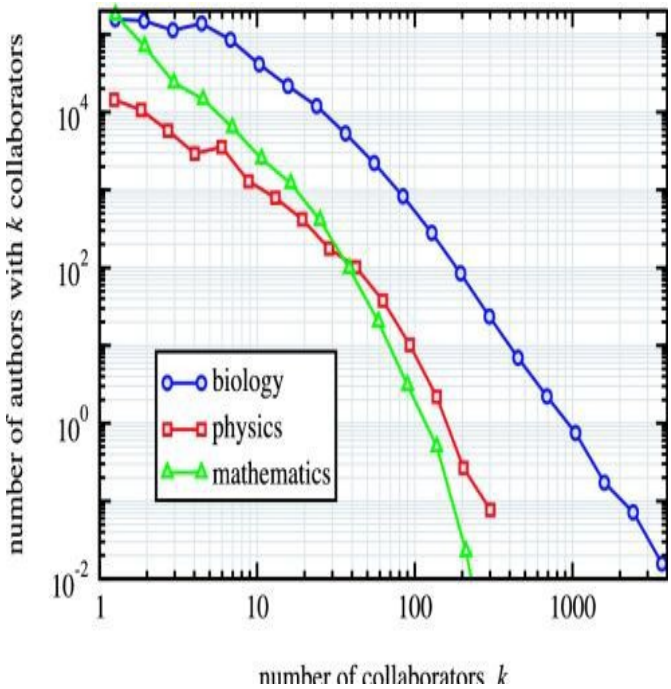
Recently Melin (2000) surveyed 195 university professors about their motives for collaboration and the chief benefits of collaboration. In their answers to open-ended questions, the respondents' most often-reported (41%) motive for collaboration is that the 'co-author has special competence'. Other common motives included 'co-author has special data or equipment (20%)', 'social reasons: old friends, past collaboration (16%)', 'supervisor–student relation (14%)', and 'development and testing of new methods (9%)'. With regard to the benefits of collaboration, the respondents pointed to 'increased knowledge (38%)', 'higher scientific quality (30%)', 'contact and connections for future work (25%)', and 'generation of new ideas (17%)'. Melin concluded that scientists collaborate for strong pragmatic reasons.

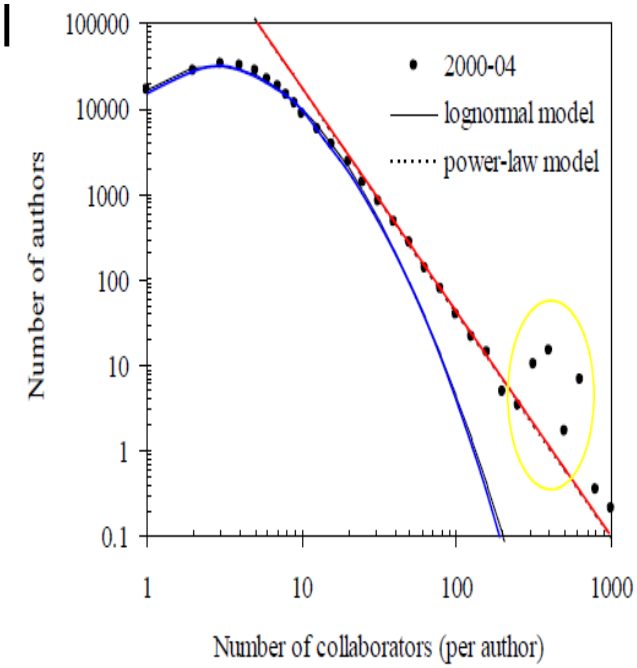
What are the pragmatics? Working alone, A and B can produce 1 paper each in a year , so the output is 2 papers. Working together, if they produce > 2 papers in a year of similar quality, the team is better.

Working alone, A and B's 1 paper per year of given quality generate 5 cites each, for 10 total cites for both.

Working together, they produce 1 better paper that generates 10 cites. But the fact that multi-authored papers have more cites is not sufficient to demonstrate the superiority of teams. Cites per author says that these two are equivalent. But if each gets credit for the 10 cites, it benefits us (but not science) to work together.

Power laws and more in collaborations (Newman, PNAS, Jan 22, 2004)Milojevik, http://arxiv.org/ftp/arxiv/papers/1004/1004.5176.pdf





Acknowledgments – another form of collaboration – Giles and Council, PNAS Dec 21,2004

Table 2. Number of citations to the most acknowledged individuals

Author	Acknowledgments	Citations
Olivier Danvy	268	847
Oded Goldreich	259	3,277
Luca Cardelli	247	3,847
Tom Mitchell	226	3,336
Martin Abadi	222	3,507
Phil Wadler	181	3,780
Moshe Vardi	180	3,786
Peter Lee	167	1,790
Avi Wigderson	160	2,566
Matthias Felleisen	154	1,622
Benjamin Pierce	152	1,484
Noga Alon	152	2,640
John Ousterhout	152	3,693
Frank Pfenning	148	1,639
Andrew Appel	144	2,064

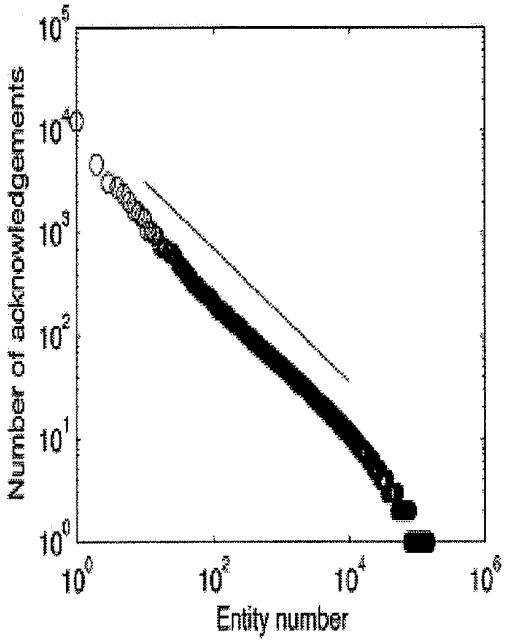


Fig. 3. The distribution of acknowledgments in the CiteSeer document collection follows a power law with the exponent −0.65. A line with −0.65 slope is drawn for reference.

Q5 How is credited allocated in team science? – In terms of jobs, pay, etc. There is division among authors. Could ask reviewers. Granting agencies likely give heavy weight to PI. "Support person, not projcct"

NETWORK ANALYSIS

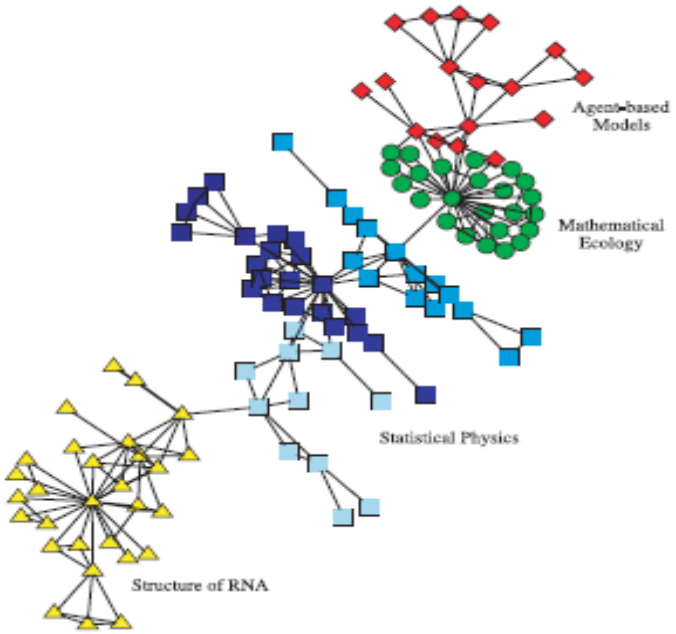


Fig. 1. An example of a small coauthorship network depicting collaborations among scientists at a private research institution. Nodes in the network represent scientists, and a line between two of them indicates they coauthored a paper during the period of study. This particular network appears to divide into a number of subcommunities, as indicated by the shapes of the nodes, and these subcommunities correspond roughly to topics of research, as discussed by Girvan and Newman (37).

Table 1. Summary statistics for the three coauthorship networks analyzed here

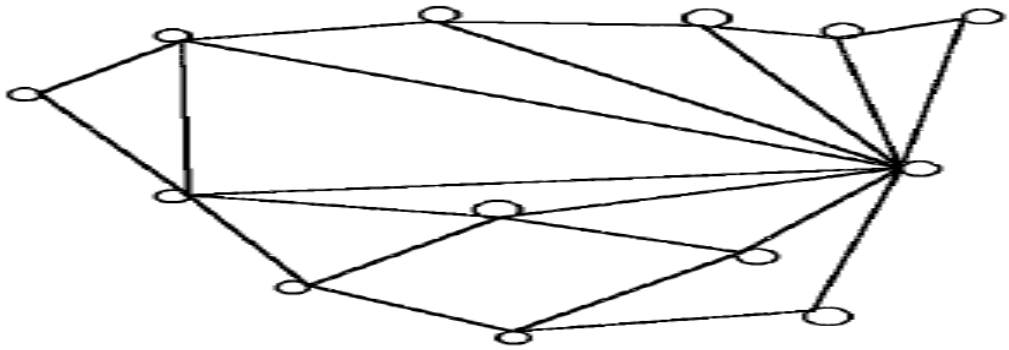
	Biology	Physics	Mathematics
Number of authors	1,520,251	52,909	253,339
Number of papers	2,163,923	98,502	—
Papers per author	6.4	5.1	6.9
Authors per paper	3.75	2.53	1.45
Average collaborators	18.1	9.7	3.9
Largest component	92%	85%	82%
Average distance	4.6	5.9	7.6
Largest distance	24	20	27
Clustering coefficient	0.066	0.43	0.15
Assortativity	0.13	0.36	0.12

The statistics are, from top to bottom, total number of authors appearing in the corresponding databases; total number of papers appearing; mean number of papers published by an author; mean number of coauthors on a paper; mean number of different individuals an author collaborated with; largest connected group of individuals in the network; mean vertex–vertex distance between connected individuals in the network; largest such distance; the clustering coefficient, which is the mean probability that two coauthors will also be coauthors of one another; and the degree assortativity coefficient, which is the Pearson correlation coefficient of the degrees (i.e., number of collaborators) of adjacent vertices in the network. The material shown here is after Newman (12) and Grossman (9).

What are the networks and what do they tell us about the way science produces knowledge?

Networks link scientists, papers. On the one side, they emphasize the collective nature of science and the position/location of people or documents in the system/network. The notion is that you need many bits of knowledge from diverse people to produce results and it gets communicated along networks. But they also pinpoint the tendency for a small number of people to have key positions in a network – for instance with a disproportionate numbers of collaborators, supporting the great person view of scientific progress.

The math is graph theory which has two elements: vertexes and edges that link vertexes.



Papers could be vertexes and the edges could be co-authors
Papers could be vertexes and the edges could be citations – with arrows to show directions

Authors could be vertexes and edges could be papers
Authors could be vertexes and edges could be citations

Could have authors and papers be edges but that is not commonly done.

Google graph theory and you can learn the basics easily. Hundreds of tutorials, including one on spectral graph theory https://www.youtube.com/watch?v=8XJes6XFjxM&list=PLW3Tw6vi-WwA_Zh8y4WPtclgtDnz-H701&index=4

If you can get from any vertex to any other you have a

CONNECTED GRAPH. Most scientists are connected through co-authors or citations. A group that is not connected could be a strange sect: creationist scientists might all cite papers and coauthor with their gang only. Marxists might cite different studies/have different co-authors than others. Since unlikely to have completely connected, often measure **the largest component** in terms of the % who are connected. You can devise other indicators to determine cliques or groups.

Graphs have natural distance metric d(i,j) is **minimum** number of edges to get from i to j. For a connected graph the average distance is the sum of all distances (counted once) divide by order of graph. **Characteristic path length** or average distance: Average of d(i,j) over all i,j — if you have distances of 5, 6, 7 you have characteristic path length of 6. If characteristic path length is large then you have a more dispersed network than if the average length is small. Diameter of graph is Max(i,j) d(i,j) — the biggest distance between any two vertices.

How local are connections? If you are connected to two neighbors, what is the chance they are connected? Take a point, look at all its neighbors, find max links, divide actual links by max. **Cluster coefficient** measure local neighborhood, C(v). If v has k neighbors, the ratio of actual edges to possible edges among neighbors: (k) = k!/[(k-2)! 2!] = k (k-1)/2. If your neighbors are closely linked we have a highly clustered neighborhood.

Three types of graphs

Regular Network — Lattice graph. Each vertex is connected to its k nearest neighbors. All vertices have same degree Long characteristic path length because you move slowly from your neighborhood to some other space. NO SHORTCUTS. Large clustering because everyone is connected in area. Know only people in your dorm/group. Get linear increase in characteristic path length as n grows.

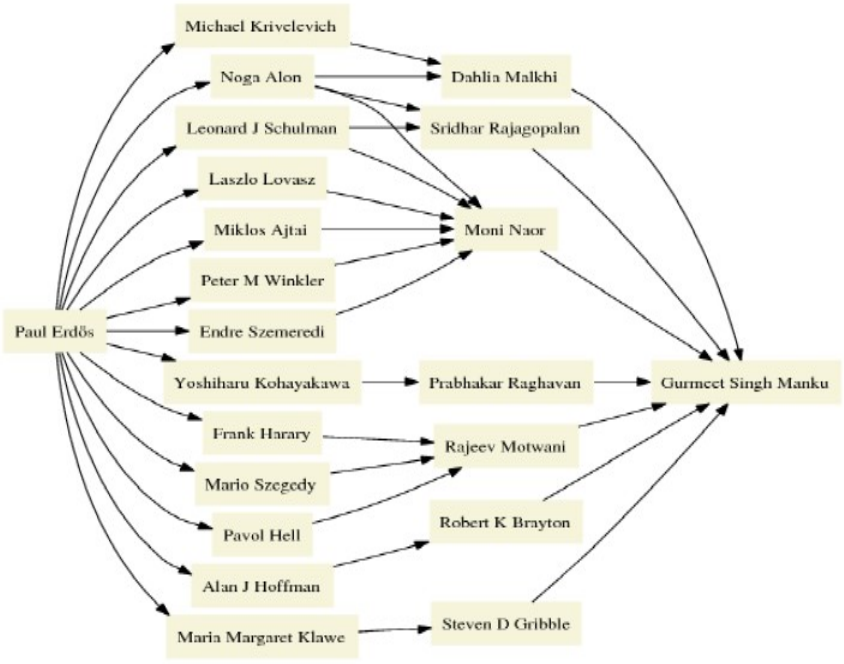
Random graph — In the 1950s Paul Erdos, the great wandering mathematician, and Alfred Renyi developed random graph model: G(n,p), where n = number of vertices of the graph; and p is the probability any two vertices are connected. There are n(n-1)/2 or —$n^2/2$. number of edges. The p that is the probability of two nodes linking does not depend on distance

Small World: (Watts-Strogatz) — start with a lattice graph and reconnect vertices with probability P by randomly shifting one edge to a randomly selected vertex (or just add link). The SW graph has path length comparable to the random graph but connectedness close to the lattice graph. On average, nodes can be connected through a short path in the network through one long shortcut; while probability that two nodes are linked is greater if they share a neighbor, the network has a large cluster coefficient. **Small characteristic path length and large cluster coefficient**. We can connect to people far away because we are in a cluster where someone has a long link.

Scientific productivity and networks Scientists write papers with other scientists. They can be treated as nodes and the papers as edges or links. So write with me and I write with Pierre and you are connected to Pierre so we have a connected gaph. This is co-authorship network

Very famous measure of graph are Erdos numbers – Erdős wrote around 1,500 mathematical articles in his lifetime, mostly co-written. He had 511 direct collaborators; these are the people with Erdős number 1. The people who have collaborated with them (but not with Erdős himself) have an Erdős number of 2 (8,162 people as of 2007), those who have collaborated with people who have an Erdős number of 2 (but not with Erdős or anyone with an Erdős number of 1) have an Erdős number of 3, and so forth.

Question: do people who write with Erdos have particular characteristics? Does writing with Erdos improve their mathematical knowledge/skills? In the production of science was Erdos a positive force only on the papers he co-authored? Could we track an Erdos effect on people with number 2?



Moe Howard (I)

was in

The Outlaws Is Coming (1965)

with

Henry Gibson (I)

was in

The Last Remake of Beau Geste (1977)

with

Michael York (I)

was in

The Mill and the Cross (2011)

with

Tomasz Luczak

was in

N Is a Number: A Portrait of Paul Erdös (1993)

with

Paul Erdös

Moe Howard (I) has a Paul Erdös number of 4.

Barabasi et al have a differential equation model for the evolution of networks that is mechanical based on preferential attachments where people co-author with people who have lots of co-authors. Not decision behavioral but descriptive differential equation which provides dynamic perspective.