Lecture 11- Insights into Power Laws and Network Models

Arithmetic of Power Laws

You are asked to present the view that unemployment is widely distributed in the population, with lots of people having short spells, and therefore is not so bad or needing much. Your clone is asked to present the view that unemployment is concentrated so there is group that suffers a lot and needs help.

Here is the data to help you decide whether you or clone have got it right.

Fact 1: The distribution of **unemployment** follows a power law. In population of 13 people, 1 person is unemployed for 12 months while 12 people are unemployed for one month each.

You say OF those unemployed OVER THE YEAR, just $1/13^{th}$ will be long-term so let's focus our effort to reduce unemployment on the short term unemployed workers.

Clone says Of those unemployed at ANY MONTH: ½ will be long-term so we ought to worry most about the long term unemployed, who make up half of the problem.

You say mean spell of unemployment is 24/13 ≈ 2 months; the median spell is 1 month; so too is the mode, so this talk about longterm being the root of the problem is nonsense

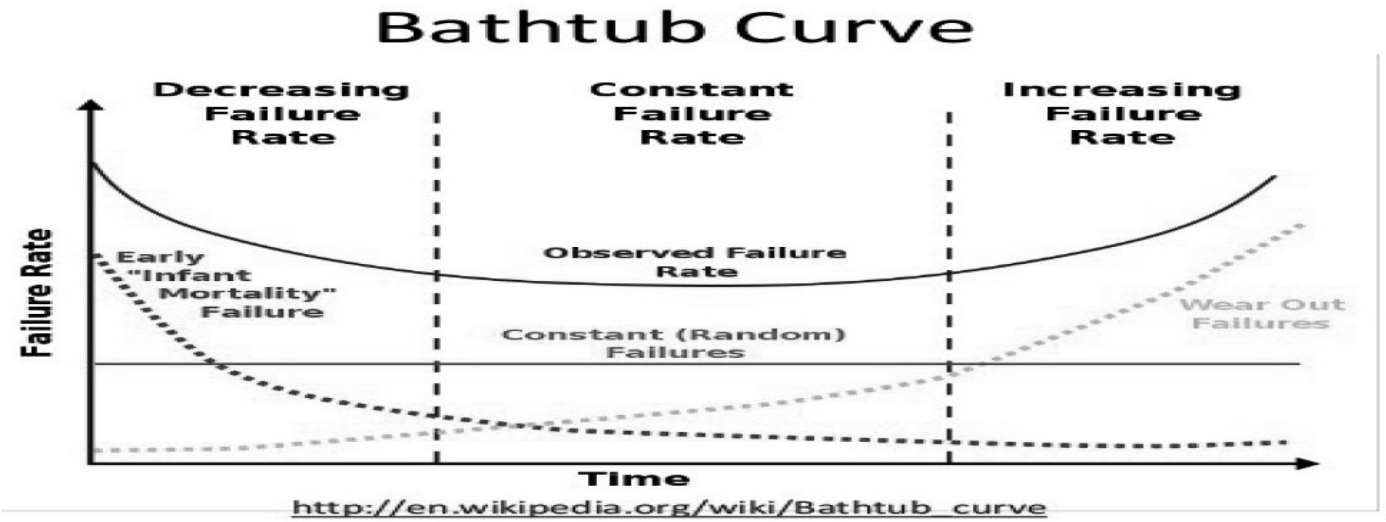
Clone says, But ½ of UI moneys will go to the long spell person. They are ½ of the problem and merit ½ of the effort to reduce unemployment.

These statements are all true for a power-law. Long spells contribute more to the stock of anything in a given period; so that small number of LONG TIMERS take up most resources. Short spells contribute to MEAN spell and dominate the MEDIAN so the average looks short. If you want to argue that most unemployment is short term give average length of spell. If you want to say most unemployment is long-term, give % long term in period.

Fact 2: The distribution of homelessness follows a power law. In population of 13 people, 1 person is homeless for 12 months while 12 people are homeless for one month each.

Fact 3: The distribution of crime follows a power law. In population of 13 people, 1 person commits lots of crimes while 12 people commit one crime.

These measures are **duration statistics** – how long a person is in a particular state. The key statistic is the Hazard Rate, the chance of leaving the state. It can be constant – same probability of leaving state over time. It can be decreasing – the longer you are married more likely you will stay married. It can be increasing – the longer you live, the more likely the Grim Reaper will catch up with you. In reliability of products they have bathtub curve, which mixes them: new products from decreasing rate of failure to constant rate to increasing rate.



Surveys ask people how long they have been in a state – for how many years have you been at Harvard? Unemployed? In Jail? Since they are in the state and have not left it, we measure of **INCOMPLETE** duration of spells. For the people in the state, the **completed spell must exceed the incomplete spell.**

If the sample is randomly drawn, the mean of a completed spell for **persons currently unemployed** will be twice the mean of the incomplete spell for the unemployed, assuming no changes in the world that generated the spells. Thus, when unemployed people reporting how long they have been unemployed have average UNE spell of

4 months, the average completed spell will be 8 months. Why? The same proportion will report a spell that has just begun as will report a spell that is close to ending so the mean will be halfway through a spell. If the hazard rate is constant, the mean for the incomplete **spell** equals the mean for ALL completed spells.

Finally, note difference in averages depending on whom we ask. We want to know the average size of employment in a population with 160 workers and 15 firms

size of firm	# firms with given size	Total in group
40	1	40
20	2	40
10	4	40
5	8	40

Average size of employment REPORTED AMONG FIRMS (40+40+40+40)/15 total firms = 160/15= 10.7

Average size of employment REPORTED BY THE 160 WORKERS with each group having 40 workers = 40 (40) +40(20) +40 (10) +40 (5)/160 = 75/4 = 18.75

Median worker is in firm with size of 2-3. But m edian firm has 5 workers.

Another paradoxical numerical relation that reflects power law type pattern is that the average number of friends of people is invariably less than than the average number of friends of their friends. Numerical example: List of friends of four people. **https://en.wikipedia.org/wiki/Friendship_paradox.** Consider the following possible patterns of friends, where each term gives number of friends of four people.

	Average of friends	Average friends of friends	Grand Avg of friends of friends
Most unequal – 3 edges (3,1,1,1)	(3+1+1+1+)/4=1.5	Person with 3 friends has 3 friends with just 1 friend (namely her). Persons with 1 friend all have sole friend with 3 friends	¼ 1 ++ ¾ x 3 = 10/4 = 2.5
Another friendship-- 4 edges (3, 2, 2, 1)	(3+2+2+1)/4 =2.0	Most pop avg 5/3; 2 avg 2.5; 1 avg 3	¼ 5/3 + 2/4 2.5 + ¼ 3 = (29/12) ~ 2.42
Another friendship – 5 edges (3, 3, 2, 2)	(3+3+2+2)/4= 2.5	2 most popular's friends avg 7/3; other 2's friends avg 3	½ 7/3 + ½ 3 = 8/3 ~2.7
Add another edge – 6 edges (3,3,3,3)	3	3	3

Why? Because it is more likely you are friends with someone with lots of friends.

These arguments/confusion/paradox are because we are used to thinking in normal distribution terms in a world where power laws often hold.

Small world fits lots of data: most nodes connected through short path L**arge cluster coefficient** to reflect clustered n-hood. BUT also lower **characteristic path length –** distance measured as # links along shortest path – than in random n-hood or regular n-hood, indicating existence of long edge short-cuts. Some s**cale free** measures – fraction of nodes with k neighbors decays by power law (after some point) -->critical nodes in hub-and-spoke network.

1.Marvel Comic Books Network with Power Law for Edges

DISNEY SHUTTING DOWN MARVEL COMICS? 03/01/2019

Rumor Disney Is Shutting Down Marvel Comics Causes Online Uproar March 2, 2019

Marvel Exec Slams Rumor Disney Might Shut Down Marvel Comics – Mar 02, 2019

How would you write out a network for Marvel Comics? What would be your nodes? What would be your edges? Node could be a comic book with the edges being books having the same characters in it: Comic book 6 connected to Comic book 7 by having 10 edges means?

Nodes could be characters with the edges being comic book in which characters appeared together. Hulk and Wolverine linked by being in comic book 9: War against aliens. If this was Batman and Robin would have more edges linking them than any other in Marvel comics except … ???

In 2002 Spanish mathematicians analyzed Marvel Comics database making nodes significant Marvel characters and an edge where characters appeared jointly in the same comic book "after Issue 1 of Fantastic Four (Nov 1961)." (http://arxiv.org/abs/cond-mat/0202174). Marvel Chronology Project (MCP) collected over 96 000 appearances by more than 6 500 characters in about 13 000 comic books, and thus yields extensive picture of the Marvel Universe. See Marvel Universe Social Graph http://aws.amazon.com/datasets/5621954952932508.

How can you set up Marvel comics as **graph** from basic data on

Comic book 1 --- Spiderman, Wolverine, Hulk :
Comic book 2-- Spider, Blackpanther, Hobgoblin:
Comic book 3--Spider, Wolverine, Aquaman, Howard-the-Duck
3 comics, 6 characters (Spider, Hulk, Hobgob ,Aqua, Howard, Panther)

Some crude statistics:

average characters per book = (3+3+ 4)/3 = 10/3 – not quite power law but fewer characters – more books
average books per character = (3 +2+ 1+ 1 + 1 +1)/ 6 = 9/6 which has "power law" flavor, with one character in many books and most characters in few

You can form a network with characters connected to characters using comics as edges (collaboration network in that the same characters work in comic that links them together) . Each person/vertex has degree distribution of # books character is in

You can also form a network of comics connected to comics with people as edges. Each book/vertex has degree distribution of # characters in a book.

Data on appearances of characters in comic books: Number of characters: 6 486 Number of books: 12 942
Books per character: 14.9 --- # of appearances from 1 (?) to 1625 (Spider-Man!); Mean characters per book: 7.47

The probability comic book has k characters in it is **$Pb(k) = k^{-3.12}$ – a power law.** Most comic books have a few characters but a few have a lot. About 50% have 5 or fewer characters and 90% have around 10 or fewer. A few have lot → uniform distribution from 1 to 10 and power law after 10. This is the standard heavy tailed distribution,

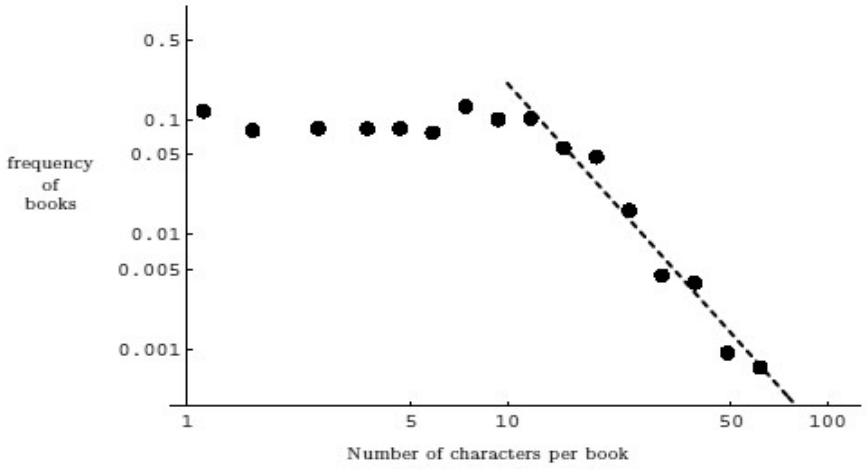


Fig. 1. Distribution of characters per comic books in the bipartite graph. The horizontal axis corresponds to the number of characters that appear in a comic book, while the vertical axis represent the frequency of books with those many characters. Note that the scales on both axis are logarithmic. The dashed line shows the tail probability distribution $P_b(k) \sim k^{-3.12}$.

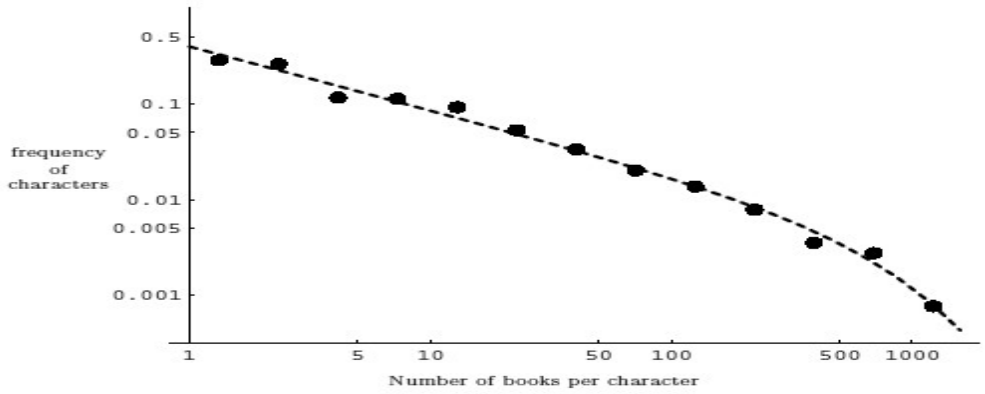


Fig. 2. Distribution of books per character in the bipartite graph. The horizontal axis corresponds to the number of comic books in which a character appears, while the vertical axis represents the frequency of characters that appear in those many books. Note that the scales on both axis are logarithmic. The dashed line shows the probability distribution $P_c(k) \sim k^{-0.66}10^{-k/1895}$.

The probability a character appears in k books is $P_c(k) = k^{-0.66}10^{-k/1895}$-- which is smoother and has relatively small power law coefficient. Most characters appear in a few books, but superstar characters appear in many.

Compare to a random network. Take 6486 characters as nodes & link characters randomly to a book to match the number of edges for character – ie if Hulk is in 2 comics, randomly choose which two. Marvel world diverges greatly from random. Average character collaborates with 52 others whereas if random would collaborate with 176 others.

Table 2 Summary of results of the MU network.

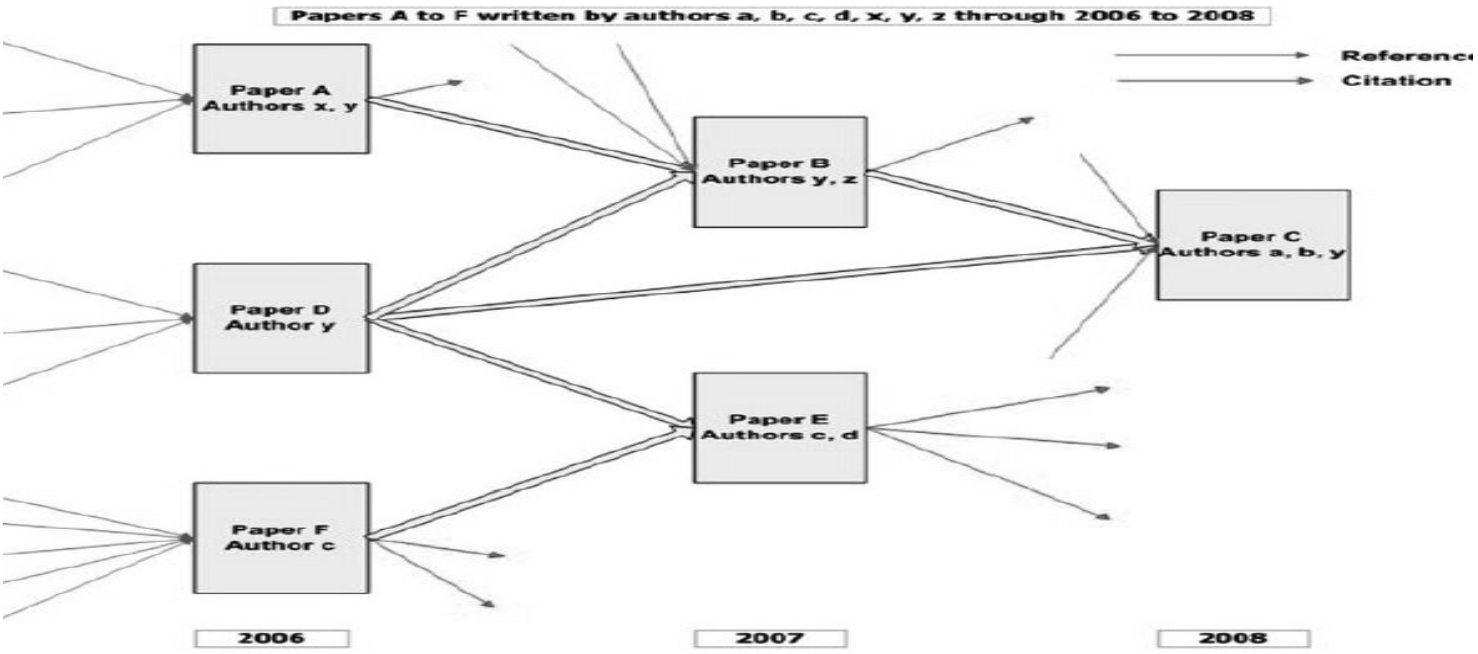
Mean partners per character: 51.88	Size of giant component: 6 449 characters (99.42%)
Mean distance: 2.63	Maximum distance: 5
Clustering coefficient: 0.012	Distribution of partners: $P(k)\ k^{-0.72}10^{-k/2167}$

Size of giant component is large at 99.42% because vertices with high degree connect to vertices with high-degree. Clustering coefficient is small, close to random graph because superheroes dominate tail. Paper? Look at another artificial network: Pro-wrestling shows or soap operas, where same characters interact on shows.

But same would be try of any real world social network.

2.Scientists and Science Paper Networks

Now look at scientific authors and papers where take scientists as nodes and edges as writing a paper together == per Erdos and co-authors. But also take paper or scientists as nodes and edges as citing other paper. But citations are uni-directional … future paper can cite older papers but older papers cannot cite future ones!



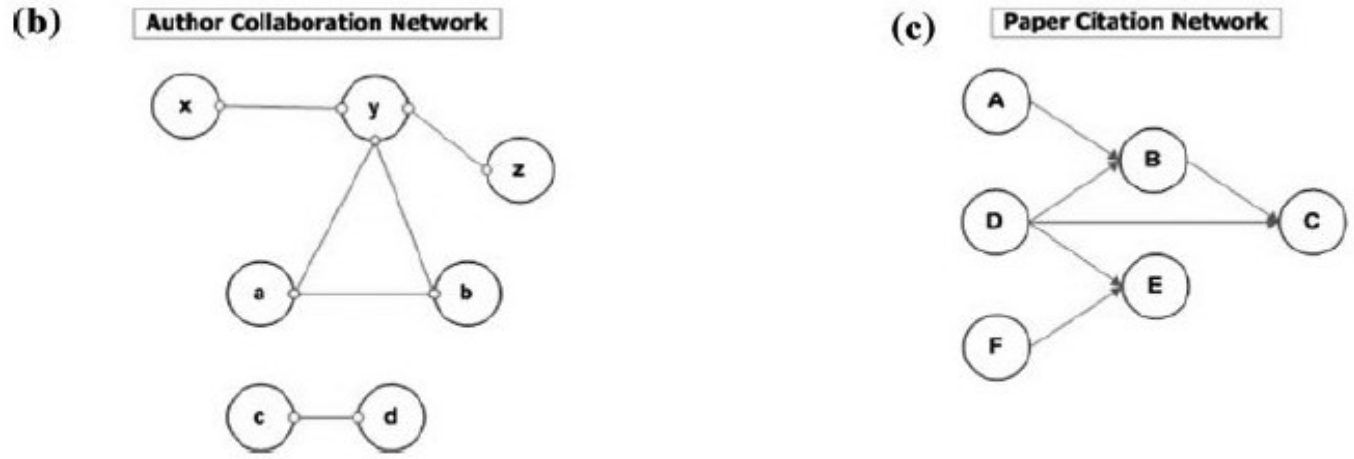


Fig. 1 a Illustration of a collection of six papers; (b) Co-author network, and (c) Paper citation network are extracted from the collection of six papers

How do these two networks connect? Do co-authors cite collaborators more than others, creating cliques of citers? Do people with given characteristics cite people like them more than others? YES!

1) There are power laws For # of papers Lotka (1926) law of inverse square of scientific productivity: Number of scientists (N) who publish P papers (in his case in Chemistry and Physics) = a P^{-2} : so that ln N = constant – 2 ln P

# papers	# of scientists
1	100
2	25
3	11
4	6
5	4

Get different coefficients for different fields but usually around 2 (Economics coefficient is 1.84). Lots of curve fitting in this business. Some use Tsallis statistic which are form with a parameter that adjusts so that distribution has properties intermediate to Gaussian and Levy statistics.
Some use stretched exponential

2)Citations to Papers: some papers get a lot and some get none:

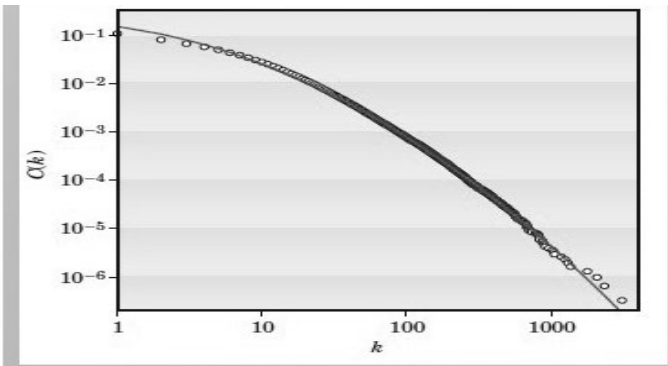


Figure 2. The cumulative citation distribution $C(k)$ versus the number of citations k for all papers published from July 1893 through June 2003 in the *Physical Review* journals. Circles indicate the data. The curve is the log-normal fit $C(k) = A \exp[-b \ln k - c(\ln k)^2]$, with $A = 0.15$, $b = 0.40$, and $c = 0.16$.

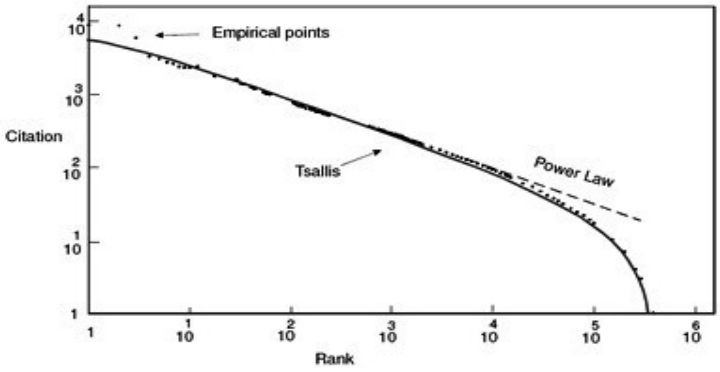


FIG. 2: Zipf plot of the number of citation of the *nth* ranked paper Y_n versus rank n on a double logarithmic scale of ISI data. The continuous line is a theoretical curve on the basis of Tsallis statistics, the dots are empirical points, and the dashes come from a power law.

3. Cites to Top 10 Physicists Most Cited Physicists, 1981-June 1997 Out of over 500,000 Examined

Author name		Institute	Country	Field	avg. cites	total art.	total cites	rank by total cit.
Witten	E	Princeton (U)	USA, NJ	High-energy (T)	168	138	23235	1
Gossard	AC	UCSB (U)	USA, CA	Semiconductors (E)	41	419	16994	2
Cava	RJ	Bell Labs (I)	USA, NJ	Superconductors (E)	65	223	14405	3
Batlogg	B	Bell Labs (I)	USA, NJ	Superconductors (E)	83	170	14164	4
Ploog	K	Max-Planck (NL)	Germany	Semiconductors (E)	19	712	13491	5
Ellis	J	Euro Nuclear Cent.	Switzerland	Astrophysics (E)	40	305	12255	6
Fisk	Z	Florida State (U)	USA, FL	Solid State (E)	23	520	12030	7
Cardona	M	Max Planck (NL)	Germany	Semicomductors (E)	20	571	11465	8
Nanopoulos	DV	Texas A&M (U)	USA, TX	High-energy (E)	39	293	11314	9
Heeger	AJ	UCSB (U)	USA, CA	Polymers (E)	34	320	10872	10

Some view science as giant collaborative network. Giant connected component – all scientists linked up except for some odd groups: Astrologers? Creationists?

Some facts about science network

Clustering coefficients differ –> different modes of research: biologists more smaller clusters than physicists;

Trend in co-authored papers;

Papers with more authors get more cites; Papers with more references get more cites;

Homophily in paper writing and in citations – National/ethnic/gender groups more likely to write together and to cite their group. If women and men tend to cite people of same gender more, who loses?

Science is a small world. Mathematicians separated from one another by 7.6 links, while the 1.6 million biomedical researchers in the analysis were separated by only four links.

Small number of extremely well-connected scientists serve as "brokers" for communications between others, with most connections among collaborators passing through them; –> Power to the connector.

Table 1. Summary of results of the analysis of seven scientific collaboration networks

		Los Alamos e-Print Archive					
	MEDLINE	Complete	astro-ph	cond-mat	hep-th	SPIRES	NCSTRL
Total papers	2,163,923	98,502	22,029	22,016	19,085	66,652	13,169
Total authors	1,520,251	52,909	16,706	16,726	8,361	56,627	11,994
First initial only	1,090,584	45,685	14,303	15,451	7,676	47,445	10,998
Mean papers per author	6.4 (6)	5.1 (2)	4.8 (2)	3.65 (7)	4.8 (1)	11.6 (5)	2.55 (5)
Mean authors per paper	3.754 (2)	2.530 (7)	3.35 (2)	2.66 (1)	1.99 (1)	8.96 (18)	2.22 (1)
Collaborators per author	18.1 (1.3)	9.7 (2)	15.1 (3)	5.86 (9)	3.87 (5)	173 (6)	3.59 (5)
Cutoff z_c	5,800 (1,800)	52.9 (4.7)	49.0 (4.3)	15.7 (2.4)	9.4 (1.3)	1,200 (300)	10.7 (1.6)
Exponent τ	2.5 (1)	1.3 (1)	0.91 (10)	1.1 (2)	1.1 (2)	1.03 (7)	1.3 (2)
Size of giant component	1,395,693	44,337	14,845	13,861	5,835	49,002	6,396
First initial only	1,019,418	39,709	12,874	13,324	5,593	43,089	6,706
As a percentage	92.6 (4)%	85.4 (8)%	89.4 (3)	84.6 (8)%	71.4 (8)%	88.7 (1.1)%	57.2 (1.9)%
Second largest component	49	18	19	16	24	69	42
Mean distance	4.6 (2)	5.9 (2)	4.66 (7)	6.4 (1)	6.91 (6)	4.0 (1)	9.7 (4)
Maximum distance	24	20	14	18	19	19	31
Clustering coefficient C	0.066 (7)	0.43 (1)	0.414 (6)	0.348 (6)	0.327 (2)	0.726 (8)	0.496 (6)

THE ECONOMICS in Networks relates to Decision to link to someone else. In papers, it is for collaborative research. There are two decisions: writing paper; – alone, with 1,2, … N others; Decision to cite other papers. #papers and #cites important in promotion decisions.

Why collaborate? Time is chief input. If writing paper with you takes ½ time and I get full credit for paper, wow! But more likely I get half credit. Then no reason to write with you unless paper actually has productive payoff. Comparative advantage is big reason; I do the blue sky, you do the heavy math, Jones does the writing. We mix biology and engineering to create Frankenstein Monster

SCIENTOMETRICS focuses on measuring scholarly communication. **ALTMETRICS** captures impact measures from a broader audience, including social media. Tweets and citations are two separate networks but … a tweet can cite an article and a scientist can tweet. If you tweet about my article, will more people read and cite it? What if someone in seminar tweets "XX just said the dumbest thing in history" and it gets retweeted and retweeted.

Haustein et al study "Tweeting Biomedicine: An Analysis of Tweets and Citations in the Biomedical Literature: 2014 JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY, 65(4):656–669,

gives "systematic evidence" about Twitter disseminaing information about journal articles in biomedical sciences based on 1.4 million documents covered by PubMed and Web of Science, 2010 to 2012. The number of tweets was compared to citations to evaluate the degree to which certain journals, disciplines, and specialties were represented on Twitter and how far tweets correlate with citation impact.

Find almost all journals had at least one tweeted publication. The mean number of tweets per journal is 88.7.The most frequently tweeted journal is Nature, with 13,430 tweets linking to its articles. A total of 9.4% (134,929 of the 1,431,576 documents) of PubMed/WoS articles were tweeted at least once: 2.4% of the papers published in 2010 were tweeted at least once, 10.9% in 2011; and 20.4% of the articles published in 2012 received at least one tweet. There were 340,751 tweets mentioning 134,929 unique articles, providing a global Twitter citation

rate of 2.5 (0.2 including untweeted documents). The distribution of tweets per document is positively skewed, with 63.0% of documents only mentioned once. Majority of journals had less than 20% of content tweeted. Those with high Twitter coverage had designated twitter handles for the journal or the associated publisher or asso

Correlations between tweets and citations are low, implying that impact metrics based on tweets are different from those based on citations. But long delay in cites suggests the tweet impact could be delayed. What might you do to get better measure? Look at downloads? Could you devise an experiment to nail down causality? Pick out a random article, tweet about it to connected person on network where tweets may get attention of large group, see if this generates attention. Must be studies of social media on such phenomena .

: 22173204

Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact J Med Internet Res. 2011 Oct-Dec; 13(4): e123.

Monitoring Editor: Anne Federer

Reviewed by Mike Thelwall and Jason Priem

Gunther Eysenbach, MD, MPH, FACMI[1,2,3]

A total of 4208 tweets cited 286 distinct JMIR articles. The distribution of tweets over the first 30 days after article publication followed a power law (Zipf, Bradford, or Pareto distribution), with most tweets sent on the day when an article was published (1458/3318, 43.94% of all tweets in a 60-day period) or on the following day (528/3318, 15.9%), followed by a rapid decay. The Pearson correlations between tweetations and citations were moderate and statistically significant, with correlation coefficients ranging from .42 to .72 for the log-transformed Google Scholar citations, but were less clear for Scopus citations and rank correlations.

A linear multivariate model with time and tweets as significant predictors ($P < .001$) could explain 27% of the variation of citations. Highly tweeted articles were 11 times more likely to be highly cited than less-tweeted articles (9/12 or 75%of highly tweeted article were highly cited, while only 3/43 or 7% of less-tweeted articles were highly cited; rate ratio 0.75/0.07 =10.75, 95% confidence interval, 3.4–33.6). Top-cited articles can be predicted from top-tweeted articles with 93% specificity and **Social media activity either increases citations or reflects the underlying qualities of the article that also predict citations, but the true use of these metrics is to measure the distinct concept of social impact. Social impact measures based on tweets are proposed to complement traditional citation metrics.**

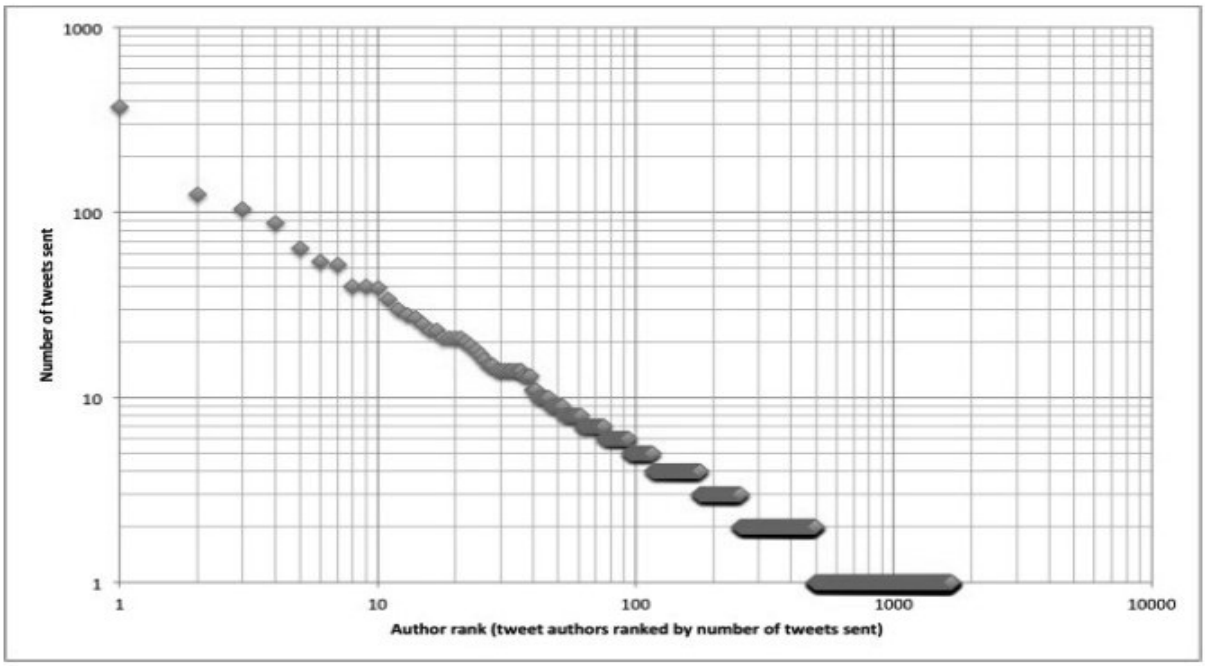


Figure 6

Tweetation density by account. Each Twitter account is ranked by the number of tweetations sent and plotted by rank on the x-axis. The y-axis shows how many tweetations were sent by each ranked account. For example, the top Twitter account ranked number 1 (@JMedInternetRes) sent 370 tweetations. Note the linear pattern on a log-log scale, implying a power law.