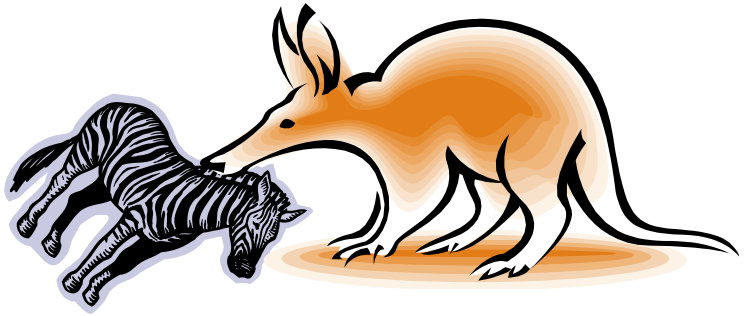
Tim Evans

Imperial College London

Scaling and Citations



Paper A

Aardvarks Eat Zebras

D.Yan, D.Tan, P.Tethera

Abstract

Sed accumsan sollicitudin feugiat. Aenean nec est vehicula magna egestas commodo. Sed commodo lorem in sem ullamcorper eu pulvinar magna auctor.

Bibliography

1. P.Tethera, "More about Aardvarks", J. Orycteropus **10** (2005) 26-975.
2. P.Tethera, M.Mether, "Zebras as Prey", Equus Studies 87 (2010) 78-79
3. ...

Paper Z

Zebras as Prey

P.Tethera, M.Mether

Abstract

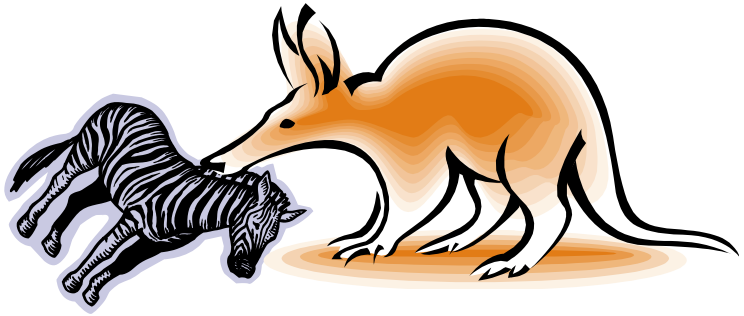
Ut nibh lacus, bibendum vel tempus ut, suscipit vitae magna. Donec massa nisl, aliquam vel imperdiet ut, pharetra at lectus.

Bibliography

1. P.Tethera, "Interesting facts about Zebras", Equus Studies 72 (2002) 1-979.
2. ...

What is a Citation?

If paper A refers to paper Z in its bibliography then we say this is a citation from A to Z



Paper A

Aardvarks Eat Zebras

D.Yan, D.Tan, P.Tethera

Abstract

Sed accumsan sollicitudin feugiat. Aenean nec est vehicula magna egestas commodo. Sed commodo lorem in sem ullamcorper eu pulvinar magna auctor.

Bibliography

1. P.Tethera, “More about Aardvarks”, J. Orycteropus **10** (2005) 26-975.
2. P.Tethera, M.Mether, “Zebras as Prey”, Equus Studies 87 (2010) 78-79
3. ...

Paper Z

Zebras as Prey

P.Tethera, M.Mether

Abstract

Ut nibh lacus, bibendum vel tempus ut, suscipit vitae magna. Donec massa nisl, aliquam vel imperdiet ut, pharetra at lectus.

Bibliography

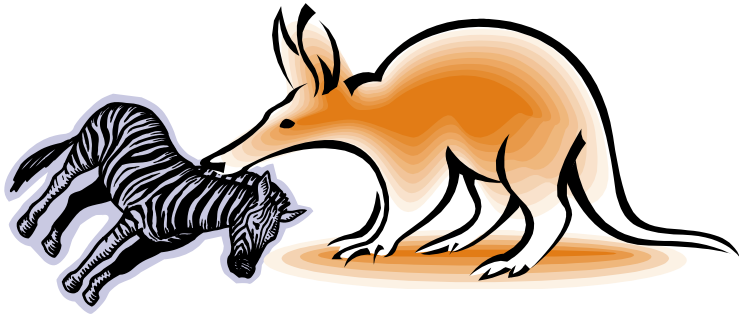
1. P.Tethera, “Interesting facts about Zebras”, Equus Studies 72 (2002) 1-979.
2. ...

There are Citations and Citations

- Different types of ‘article’
  - Academic papers, books, patents, chapters in books, editorials, commentaries, essays, newspaper articles, web pages, blogs, tweets, ...
- Different sources
  - Journals, books, patent agencies, web sites, newspapers, social networking sites, ...
  - Peer review?
  - One item, several copies/editions/translations
  - Sources in different languages
- Information in items imperfect
  - Author copied already incorrect bibliography entry

What is a Citation really?

A citation from A to Z
is just a reference to Z
identified in an item A
within a dataset D



Paper A

Aardvarks Eat Zebras

D.Yan, D.Tan, P.Tethera

Abstract

Sed accumsan sollicitudin feugiat. Aenean nec est vehicula magna egestas commodo. Sed commodo lorem in sem ullamcorper eu pulvinar magna auctor.

Bibliography

1. P.Tethera, “More about Aardvarks”, J. Orycteropus **10** (2005) 26-975.
2. P.Tethera, M.Mether, “Zebras as Prey”, Equus Studies 87 (2010) 78-79
3. ...

Paper Z

Zebras as Prey

P.Tethera, M.Mether

Abstract

Ut nibh lacus, bibendum vel tempus ut, suscipit vitae magna. Donec massa nisl, aliquam vel imperdiet ut, pharetra at lectus.

Bibliography

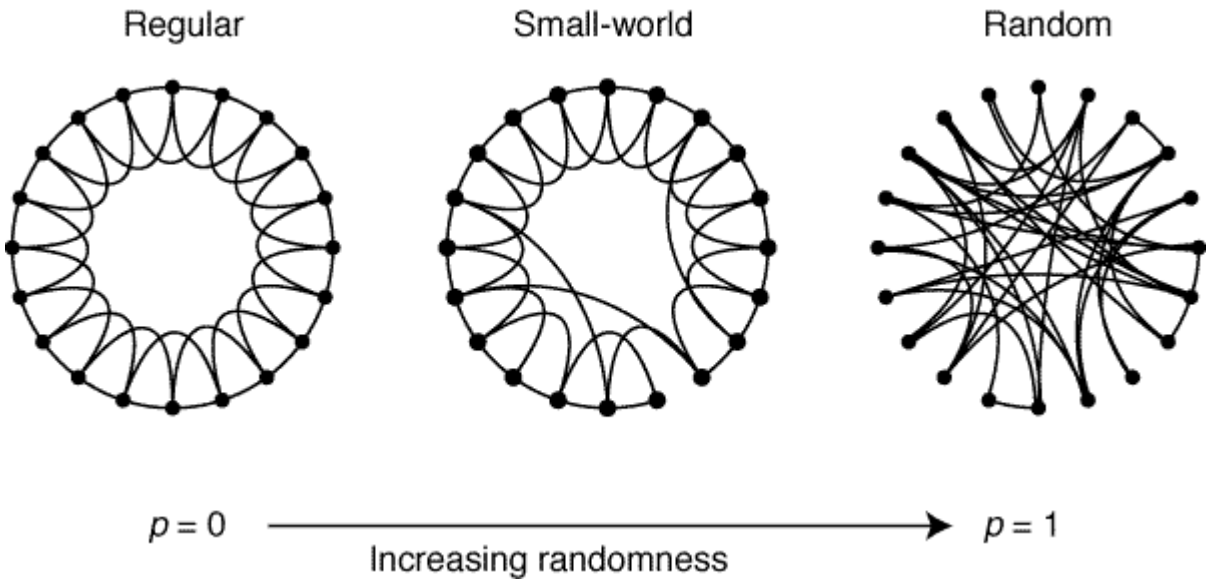
1. P.Tethera, “Interesting facts about Zebras”, Equus Studies 72 (2002) 1-979.
2. ...

What does a citation mean? ***Good Things***

Authors add references to paper Z because:-

- Authors learnt something useful from paper Z

e.g. Watts & Strogatz (1998) told us how a few short cuts can completely change some properties of a network



What does a citation mean? ***Less useful***

Authors add references to paper Z because:-

- Authors were aware of paper Z
- Authors could access paper Z
- Authors could read and understand paper Z

e.g. Barabási-Albert paper is recent (1999), in English, placed on **`arXiv`**, and published in Science. More accessible to statistical physicists than work of Yule (1925,1944), Simon (1955), Price (1965,1976) etc.

What does a citation mean? ***Not so useful***

Authors add references to paper Z because:-

- Authors learnt something from paper Y that cited paper Z though they never read paper Z
- Some of the current authors wrote paper Z

e.g. How many people have ***read*** original Erdős-Réyni (1959,1960) papers on random graphs? One is in Hungarian, they are in journals hard to find, fifty years old, yet work has 1000's citations

Signal to Noise in Vitations

Citations carry useful information but its a very noisy signal

Paper A

Aardvarks Eat Zebras

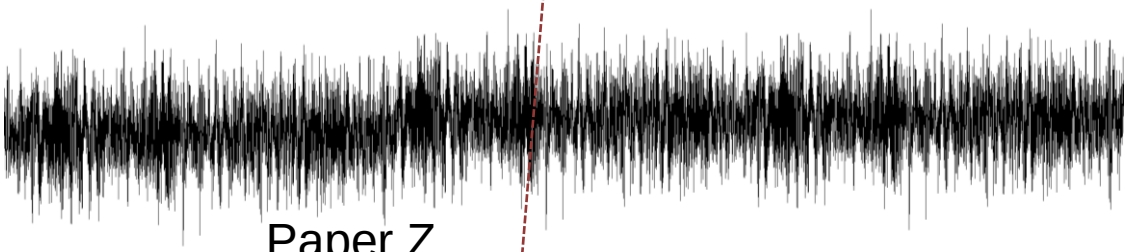
D.Yan, D.Tan, P.Tethera

Abstract

Sed accumsan sollicitudin feugiat. Aenean nec est vehicula magna egestas commodo. Sed commodo lorem in sem ullamcorper eu pulvinar magna auctor.

Bibliography

1. P.Tethera, “More about Aardvarks”, J. Orycteropus **10** (2005) 26-975.
2. P.Tethera, M.Mether, “Zebras as Prey”, Equus Studies 87 (2010) 78-79
3. ...



Paper Z

Zebras as Prey

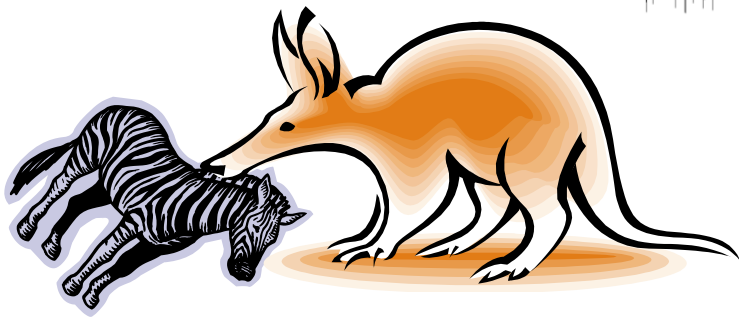
P.Tethera, M.Mether

Abstract

Ut nibh lacus, bibendum vel tempus ut, suscipit vitae magna. Donec massa nisl, aliquam vel imperdiet ut, pharetra at lectus.

Bibliography

1. P.Tethera, “Interesting facts about Zebras”, Equus Studies 72 (2002) 1-979.
2. ...



What can we learn from bibliographies?

- Their information allows us to find the best existing work
  $\Rightarrow$ measures of quality and relevance
- They trace the innovation process

Modern large data sets allow statistical methods to overcome problems and inherent fluctuations in data for individual papers

Measures of Quality

Count the citations of a paper

- Simple
  = Cheap – only one to be used in REF

- Varies with age of paper
- Varies with field

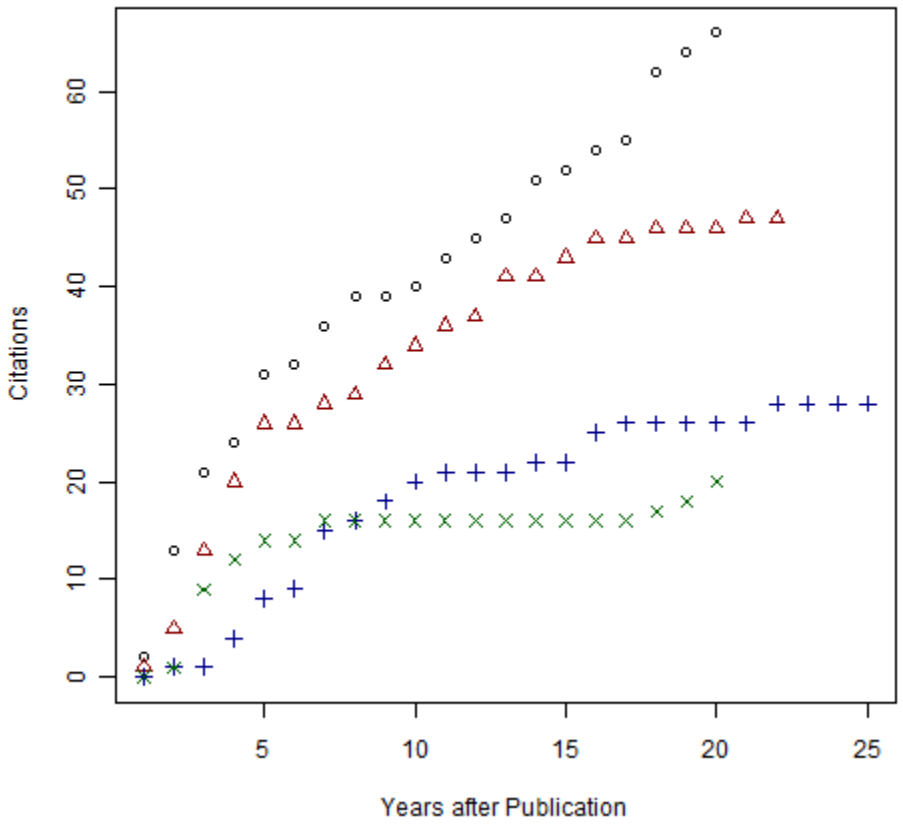
Citation Time Scales

Citations to papers typically rise steadily then plateau

Exceptions are the rule

- Large variations in behaviour even in papers by one author
- Sleepers, papers whose worth is recognised only after a significant delay

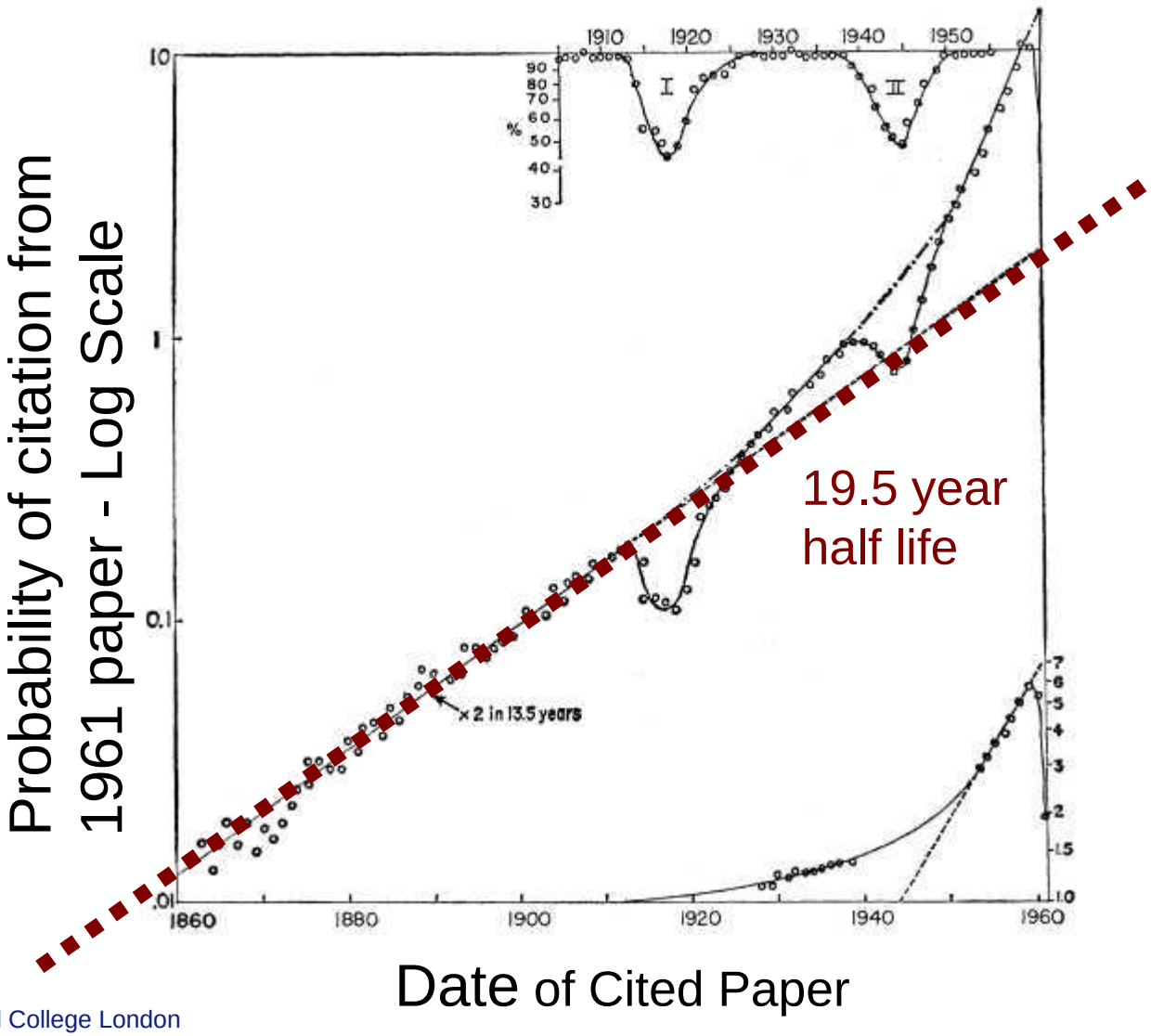
Papers from one author



Citation Time Scales

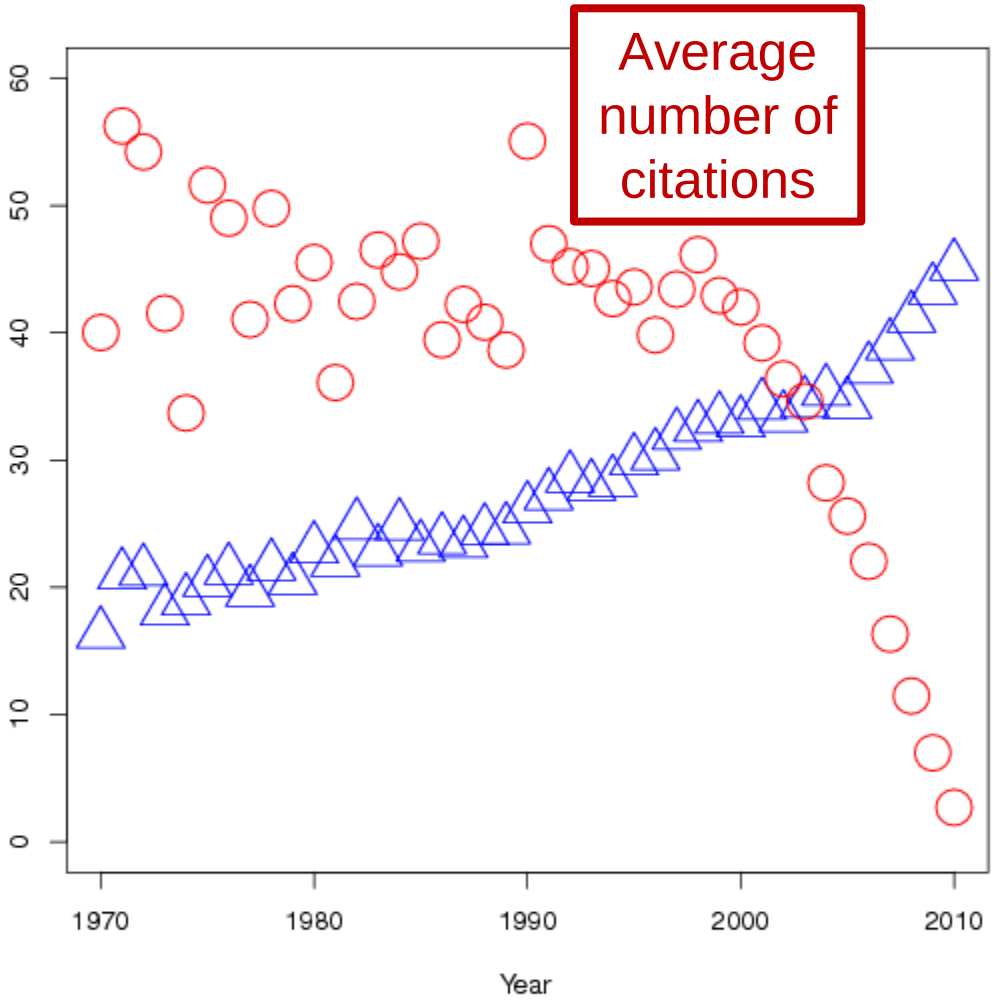
Probability of citing an earlier paper falls off fast

Exponential fall off for cites from 1961 papers (Price 1965)



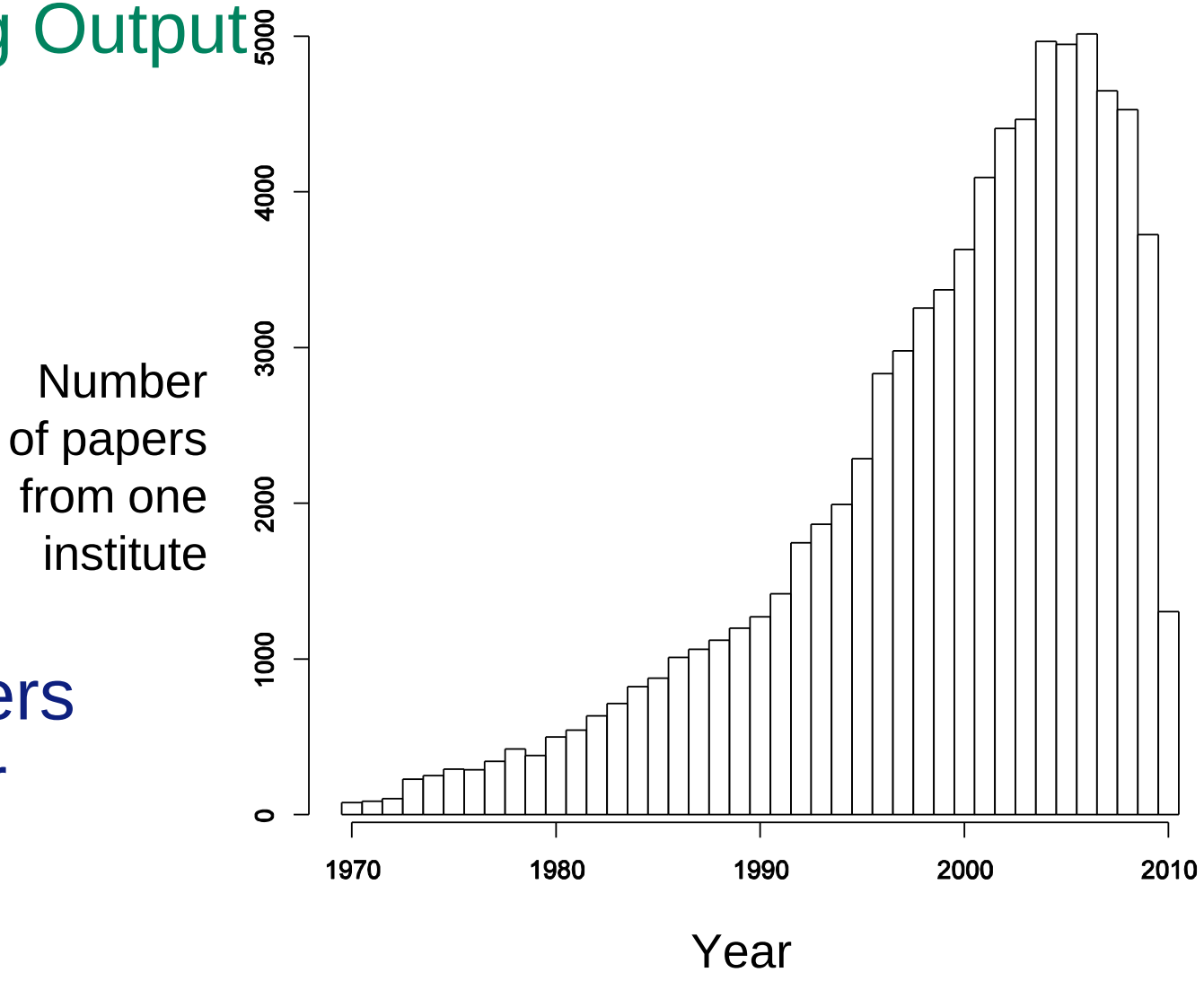
Citation Time Scales

Recent data show an average 10 year time scale



WoS output from one Institute [Evans et al, 2011]

Increasing Output

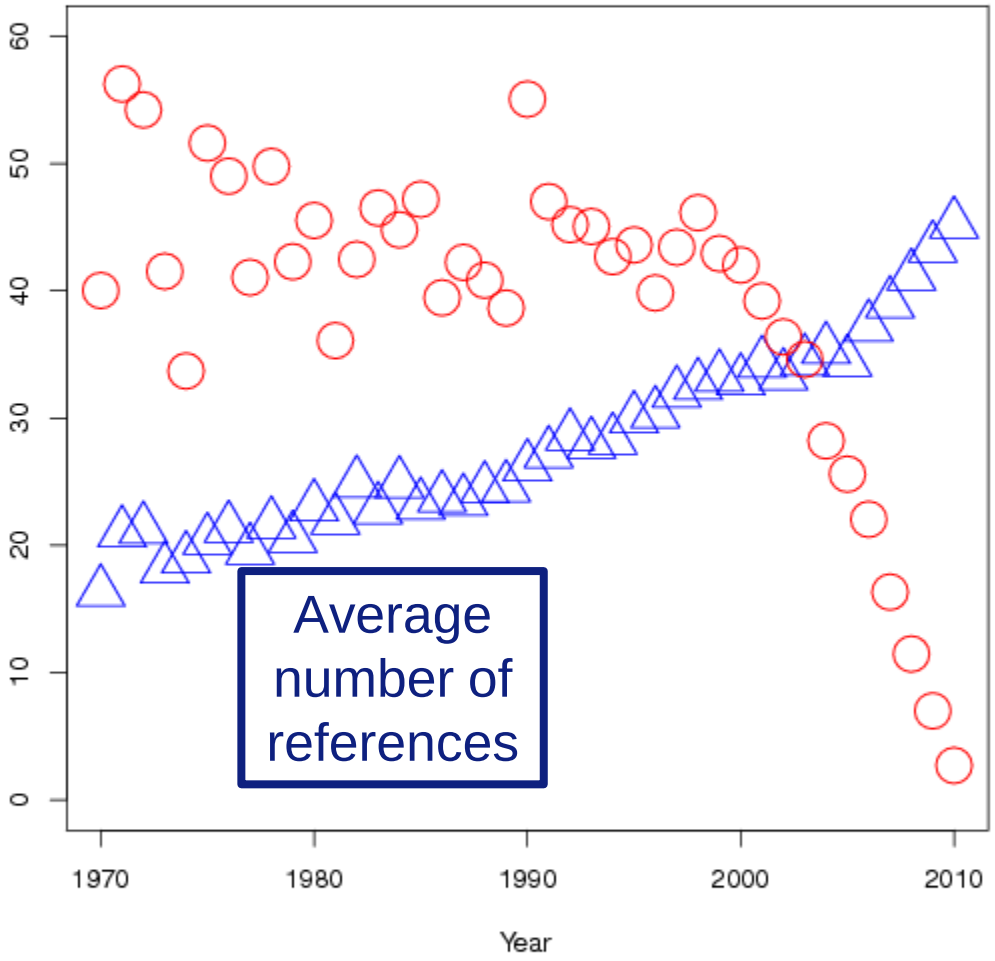


More papers every year

WoS output from one Institute [Evans et al, 2011]

Increasing Output

Longer bibliographies every year

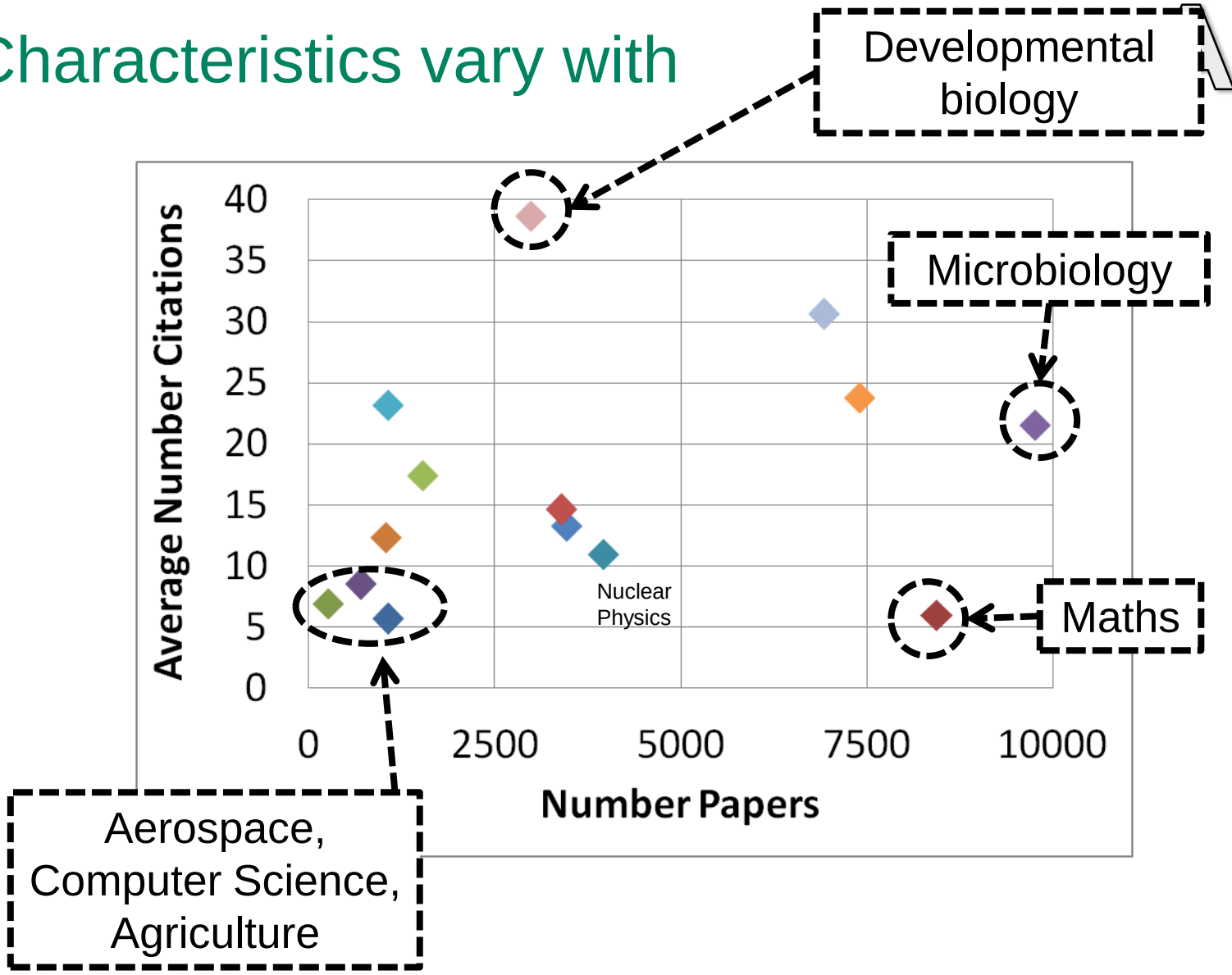


WoS output from one Institute [Evans et al, 2011]

Citation Characteristics vary with Field

Data for various fields in same year

(Radicchi et al, 2008)



WoS data for articles and letters, 1999, field defined by Journal of Citation Report

Citation Characteristics vary with Field

In same year number of papers and average number of citations vary by more than an order of magnitude

Field	Number papers
Agricultural economics and policy	266
Microbiology	9,761

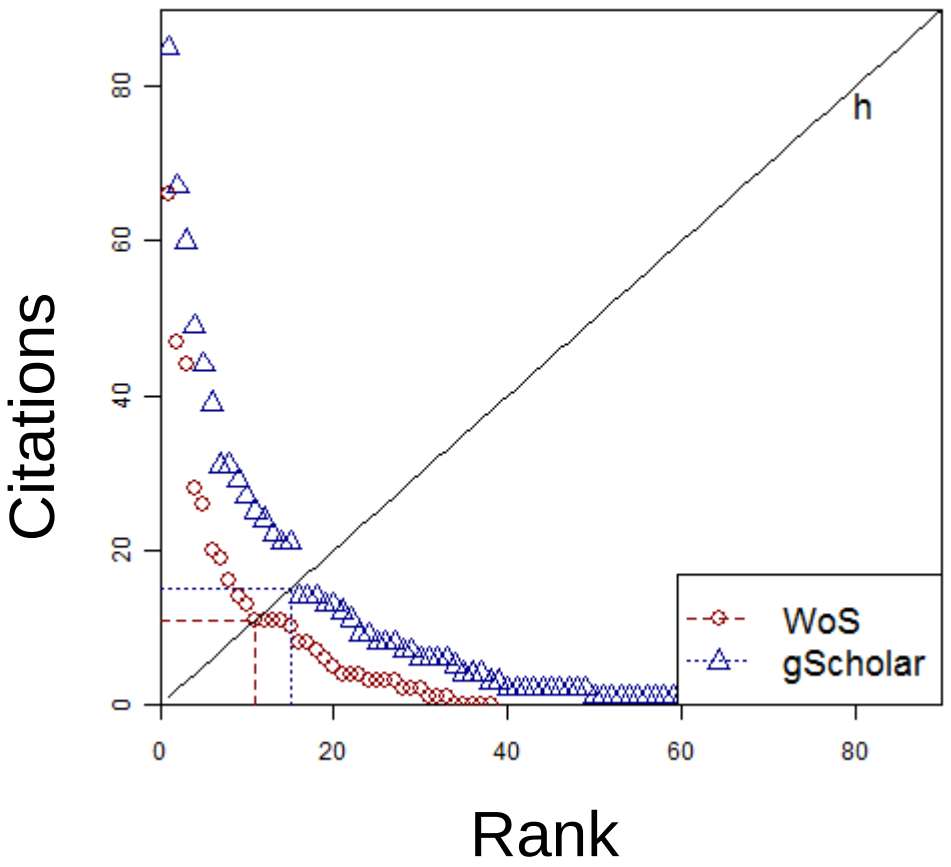
Field	Average Citations
Aerospace Engineering	5.65
Developmental biology	38.67

WoS data for articles and letters, 1999, field defined by Journal of Citation Report [Radicchi et al, 2008]

Long Tails - individuals

For individuals
typically
many papers with
few citations,
a few papers
have many

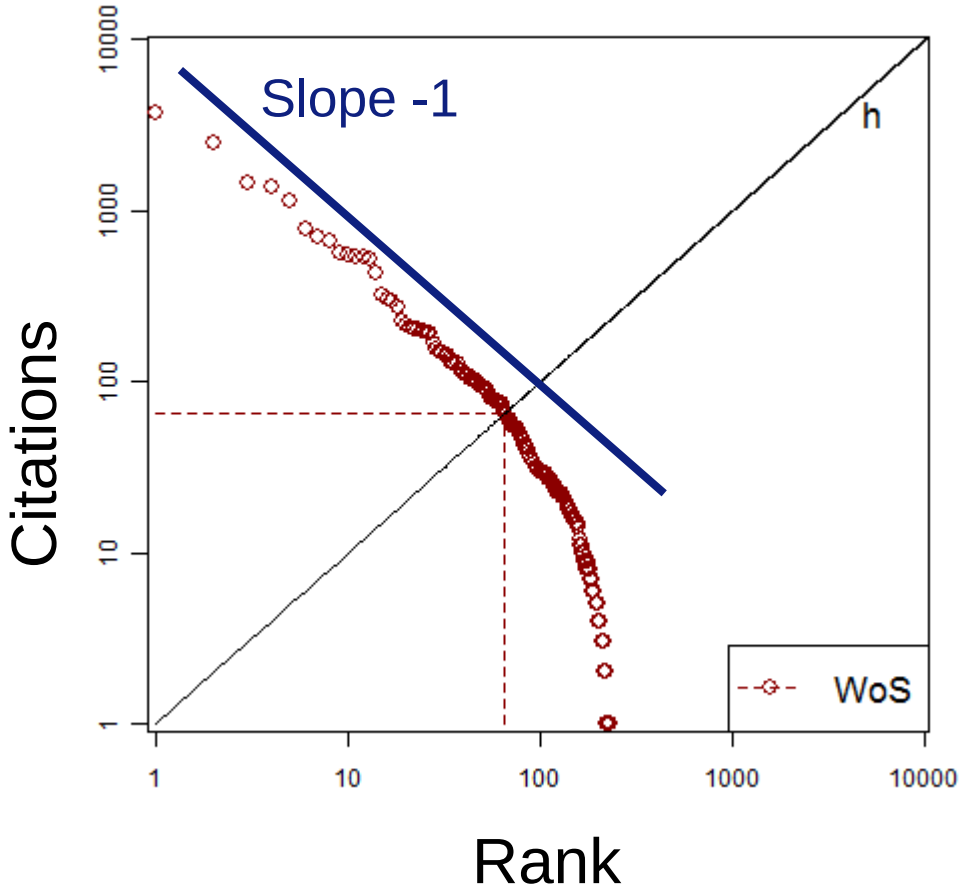
Single author, normal scaling,
same papers two data sources



Long Tails - individuals

For individuals
typically
many papers with
few citations,
a few papers
have many

Single author, log-log scales



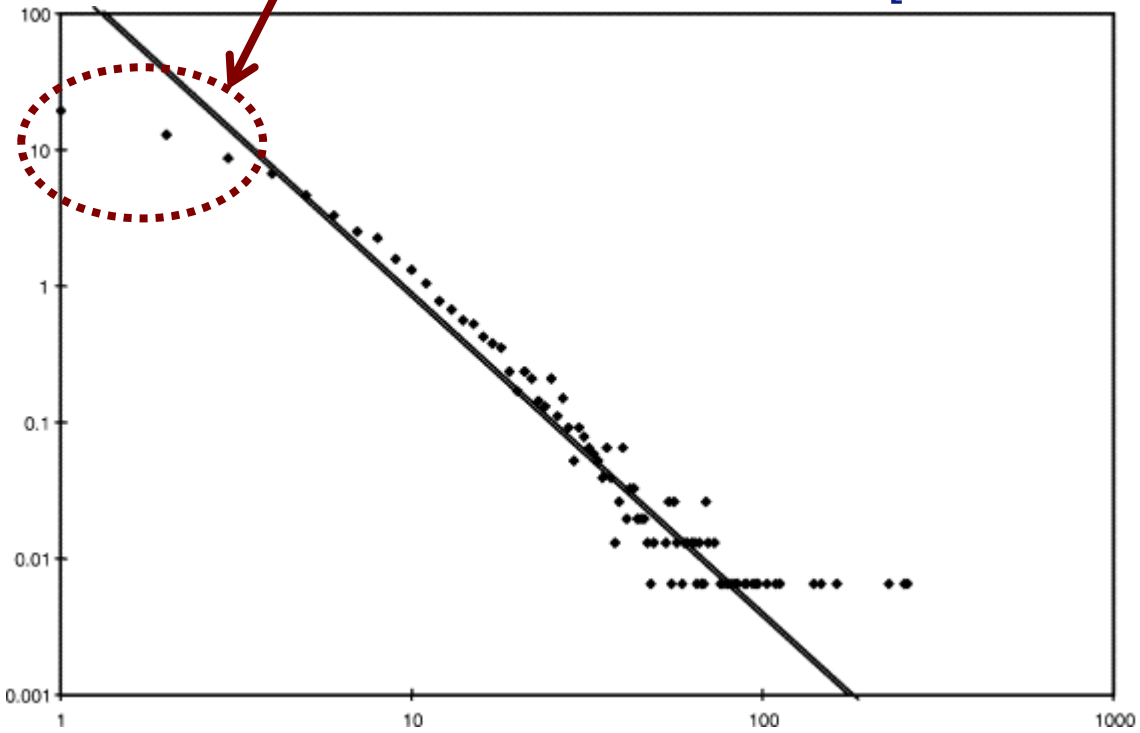
Possible future Physics Nobel prize winner?

Long Tails by Field

Papers with 1, 2 or 3 citations off trend

[van Raan 2001]

Log Number of Chemistry Papers 1985–1993



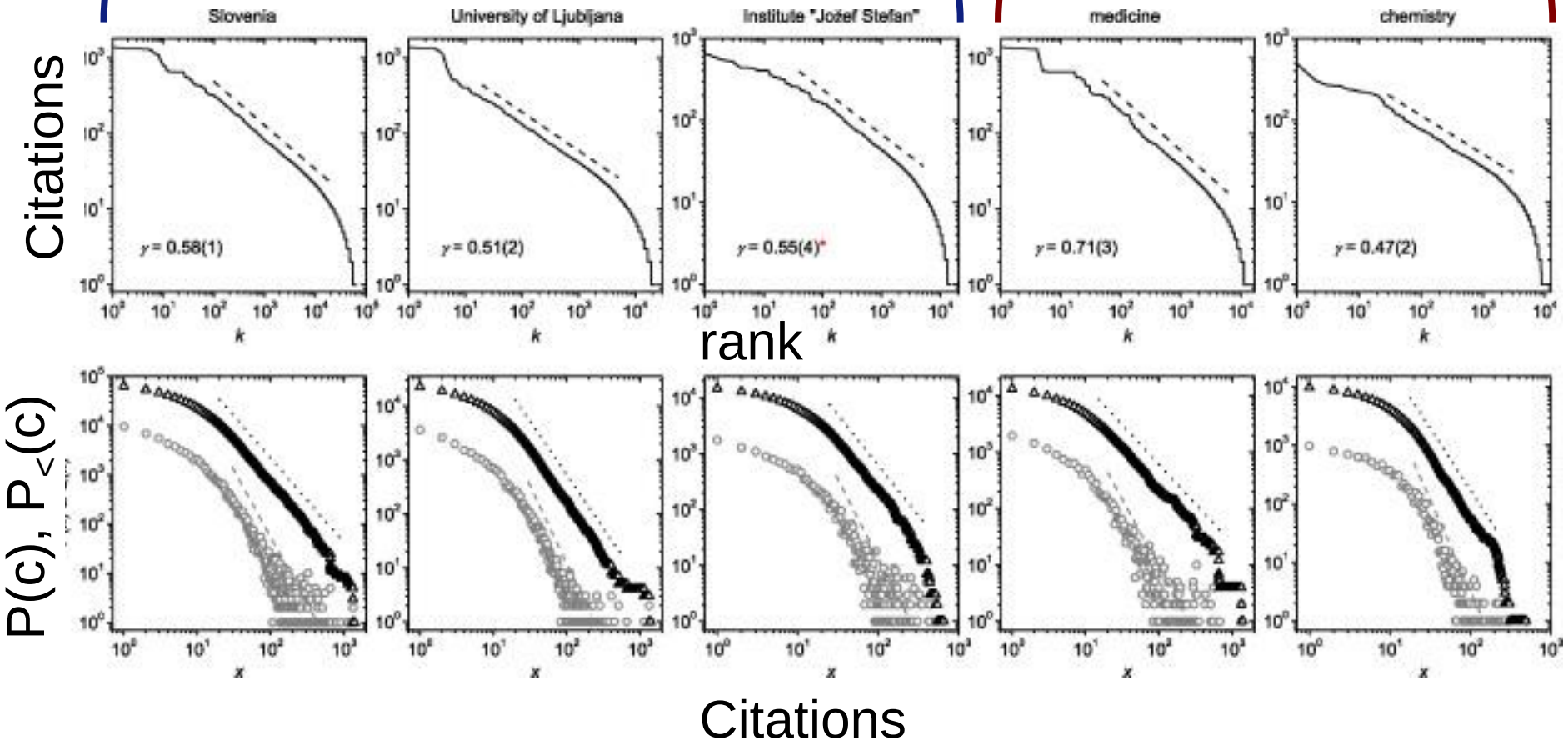
Log Number of Citations after 3 years
(no self-citations)

Long Tails by Institution/Field/Country

[Perc 2010]

Solovenian Universities

Solvenian outputs in single fields



Finding Universal features

- Citations vary with age
- Features differ between fields

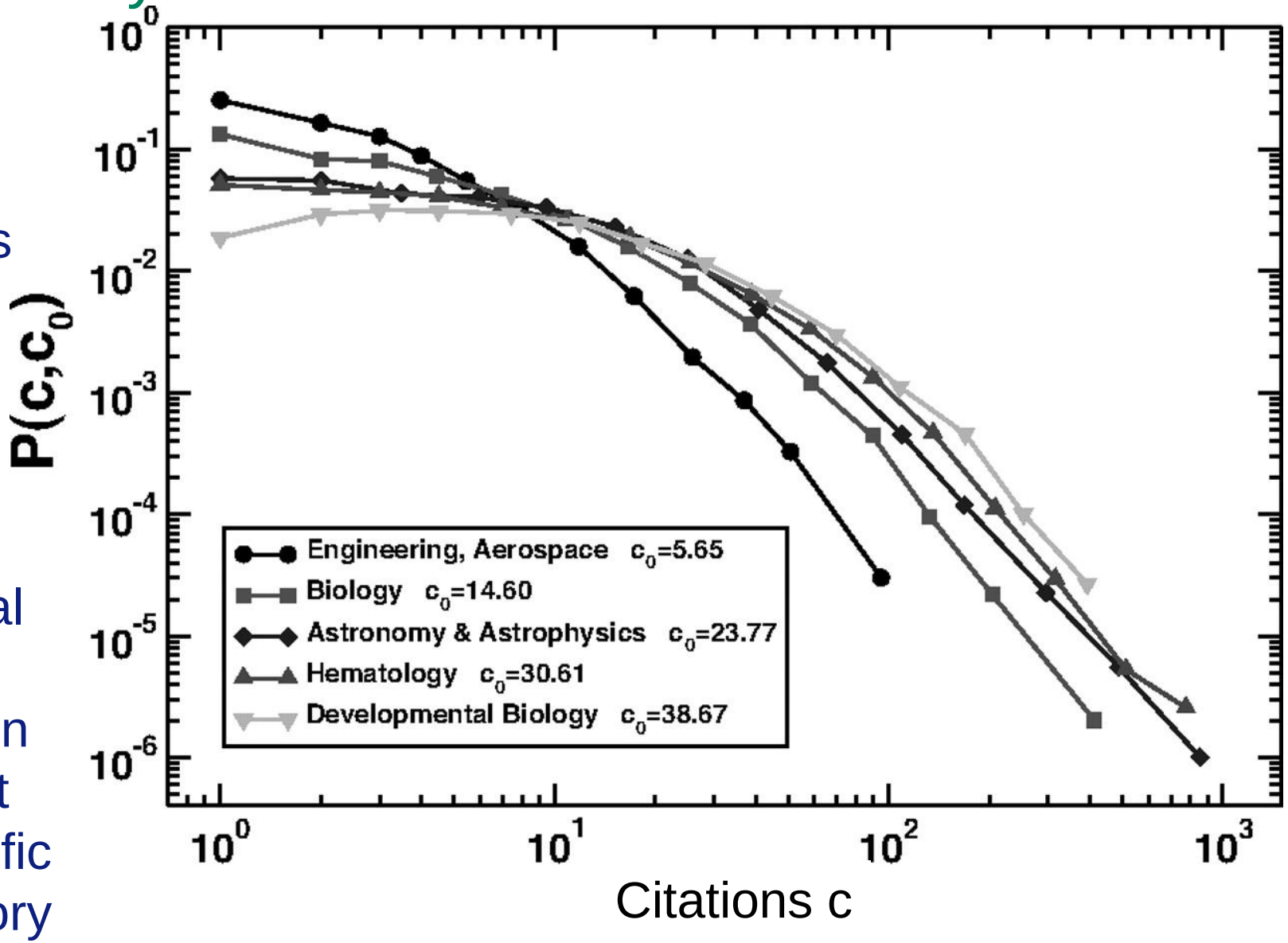
$\Rightarrow$ Focus on papers published in one year in one `field'

Citations by Year and Field

[Radicchi et al, 2008]

1999 WoS papers

Journal of Citation Report scientific category



Finding Universal features - $c_f(t)$

- Define the average number of citations for papers ***p*** in field ***f*** and year ***t*** to be $\langle c(t) \rangle_f$

$$\langle c(t) \rangle_f = \frac{\sum_{p \in f} c(t)}{N_f(t)}$$

$$N_f(t) = \sum_{p \in f} 1$$

$N_f(t)$ Number of papers in field ***f*** at year ***t***

- Define relative citation count of paper $c_f(t)$

$$c_f(t) = \frac{c(t)}{\langle c(t) \rangle_f}$$

c(t) is number of citations to a paper in year ***t***

Finding Universal features – Crown Indicator

Relative citation counts of paper $c_f(t)$ are also used in the '***Crown Indicator***' [van Leewun et al, 1995] as used in assessment of whole institutes, countries etc.

$$c_f(t) = \frac{c(t)}{\langle c(t) \rangle_f}$$

Finding Universal features - c_f distribution

- Find distribution $p(c_f)$ of $c_f(t)$ for each field f and year t

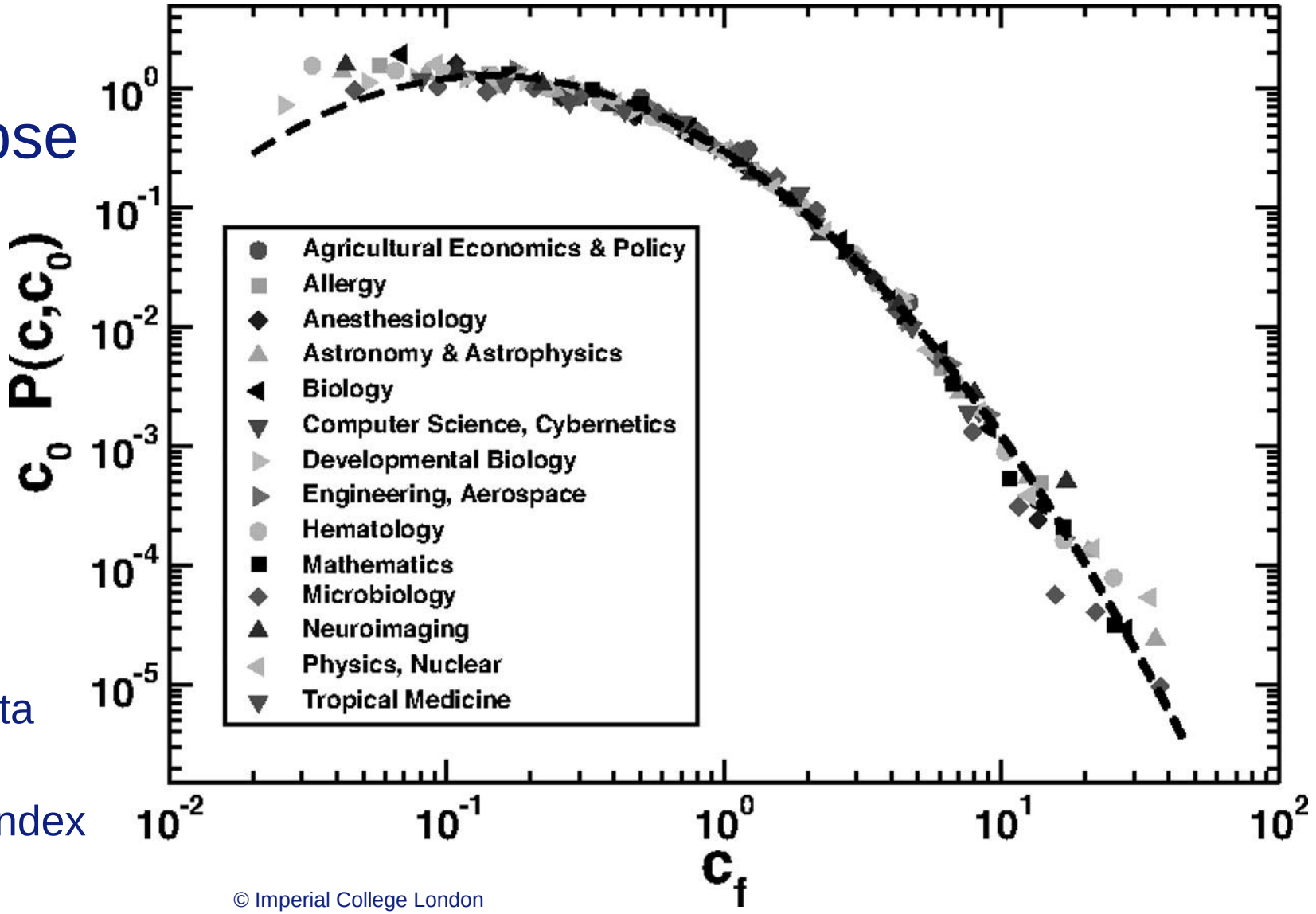
$$p(c_f) = p\left(c(t) \Big/ \langle c(t) \rangle_f \right)$$

- Average is trivially 1
- Standard deviation still free

[Radicchi et al, 2008]

Universal form for $p(c_f)$ – world output

Data Collapse

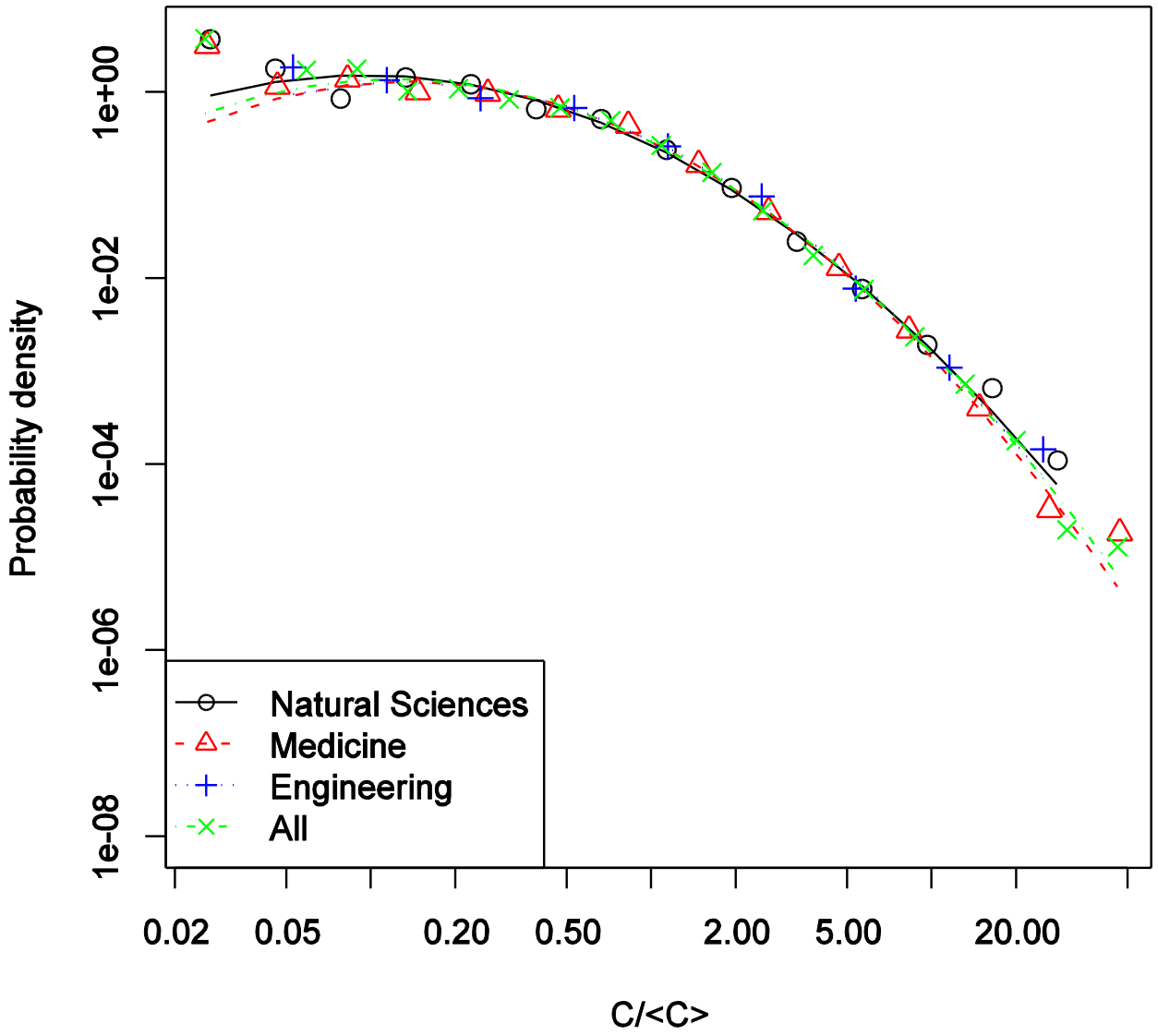


WoS data 1999, f=J.Cit.Index

[Evans et al, 2011]

Universal form for $p(c_f)$ - faculties

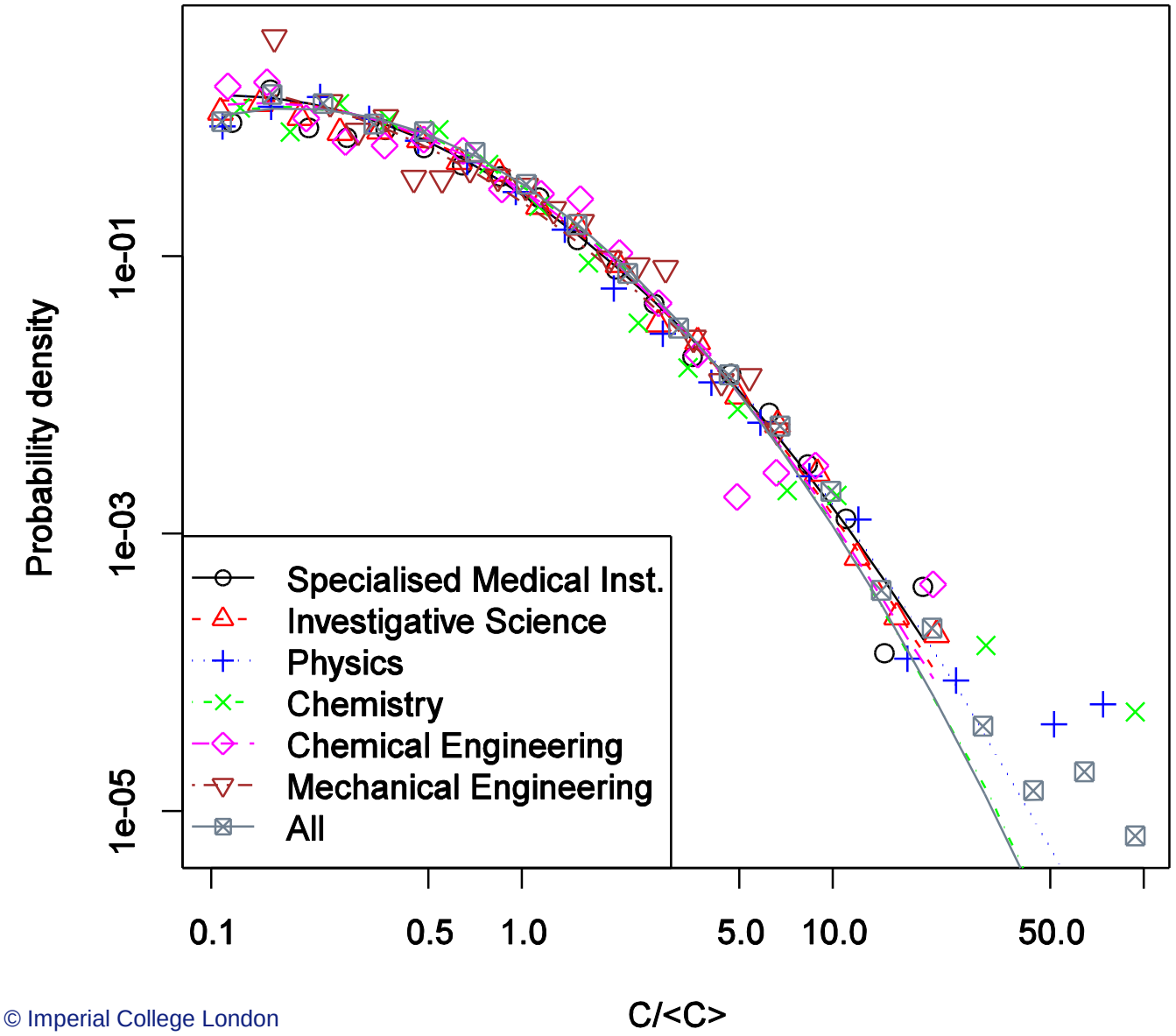
Data collapse for papers from one faculty in one year



[Evans et al, 2011]

Universal form for $p(c_f)$ - departments

Data collapse for papers from one department in 3 years 1999-2001



[Evans et al, 2011]

Universal form for $p(c_f)$

In all cases our best representation of the data (lines on plots) for $c_f > 0.1$ are **log normals**

$$p(c_f) = \frac{1}{\sigma c_f \sqrt{2\pi}} \exp\left\{ \frac{-(\ln(c_f) - \mu)^2}{2\sigma^2} \right\}$$

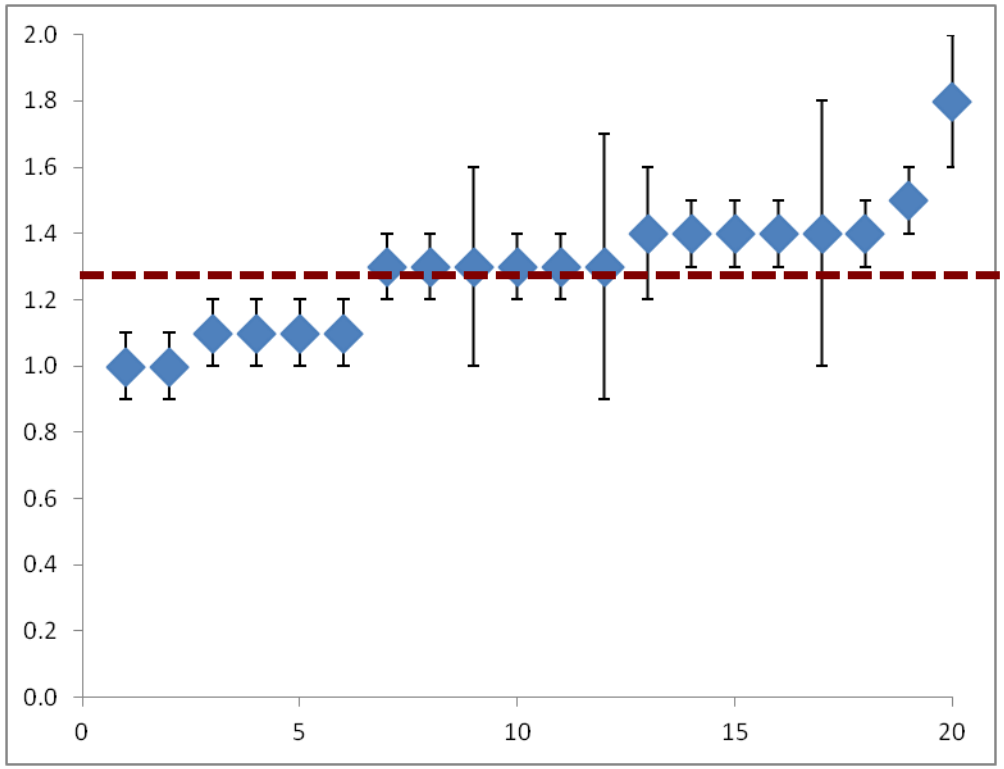
N.B.
As $\langle c_f \rangle = 1$
$\mu = -\sigma^2/2$

- Very common form [Limpert et al, 2001]
- We tried many other forms:
  - Power laws, Shifted power laws, Stretched Exponentials

Universal Variance – World Data

The variance from log normal fits to the data is roughly constant and universal at

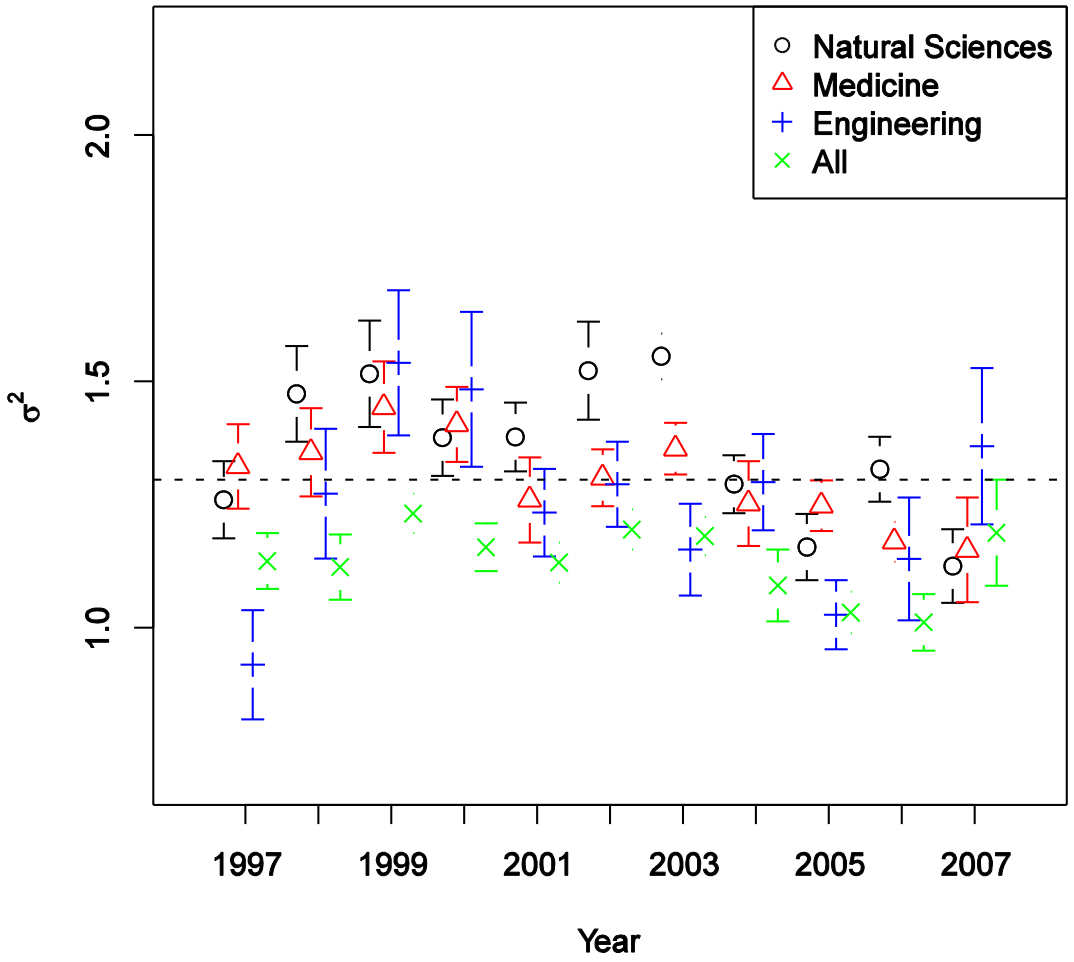
$\sigma^2 \approx 1.3$ [Radicchi et al, 2008]



Universal Variance – Faculty Data

Faculty data also close to $\sigma^2 \approx 1.3$

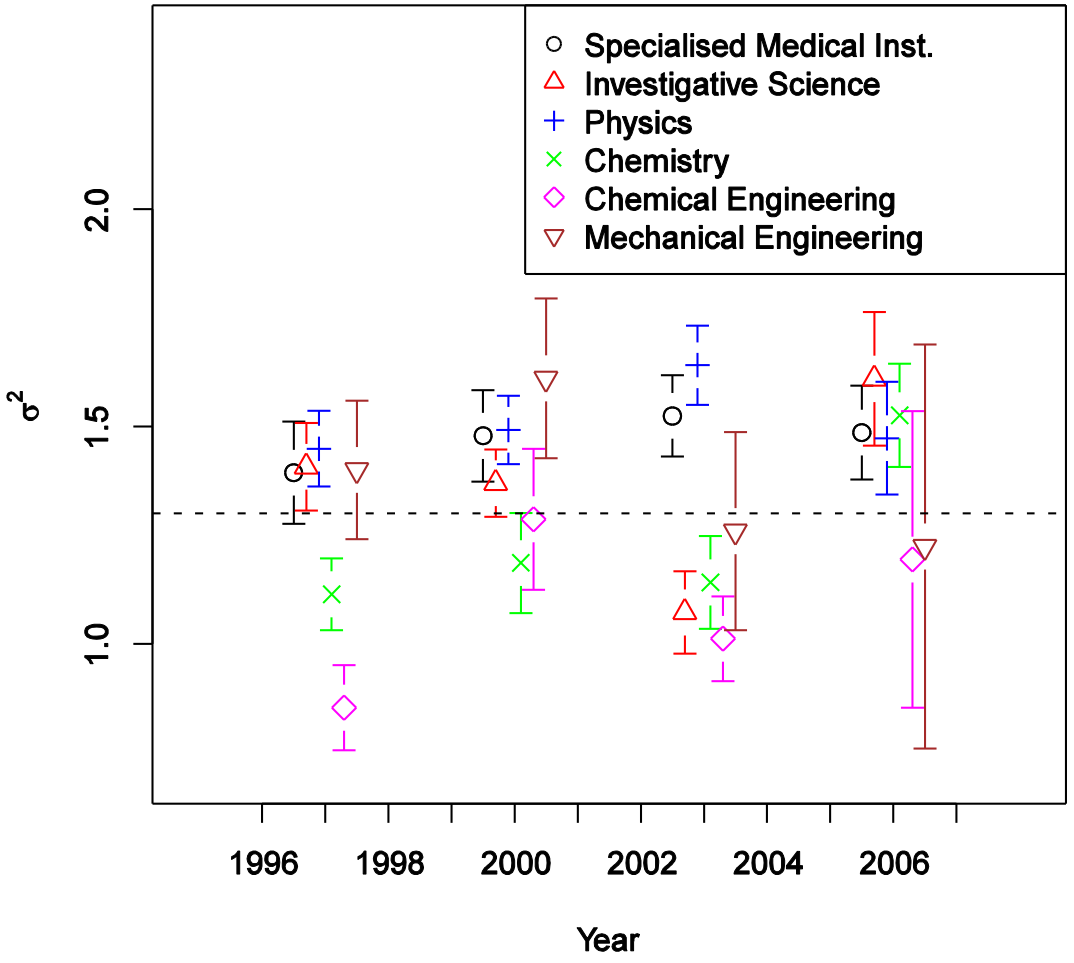
[Evans et al, 2011]



Universal Variance – Department Data

Department data also close to $\sigma^2 \approx 1.3$

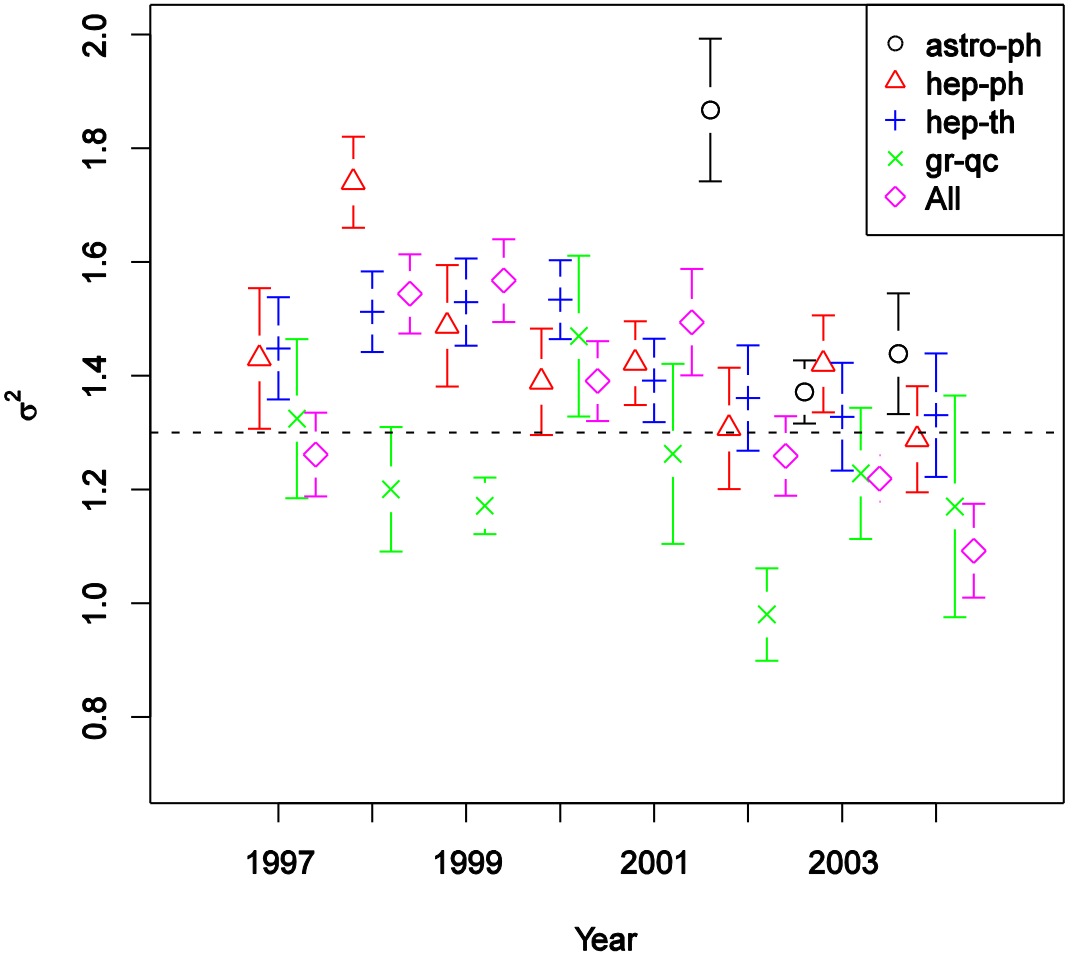
[Evans et al, 2011]



Universal Variance – arXiv categories

arXiv data close to $\sigma^2 \approx 1.3$

[Evans et al, 2011]



Using a universal form to compare papers

The rescaling by average number of citations in the same field and year produces a good data collapse.

$\Rightarrow$ The field/year normalised index $\boldsymbol{c_f}$ (Crown index)

$$c_f = \frac{c}{\langle c(t) \rangle_f}$$

allows useful comparisons of papers across different fields and times

Universal indices to compare papers

Useful indices are:-

- Straight use of index $c_f = c / \langle c \rangle_f$
  - ***Exploits fact variance σ^2 is roughly constant***
- Use logarithm $\ln(c_f) = \ln(c / \langle c \rangle_f)$
  - ***Deals with long tail***
- Use $z = [\ln(c_f) - \mu] / \sigma$
  - ***Allows for variations in mean and variance***

Universal indices to compare papers

Results show that these indices can be used

- On global data set
  - Expensive
- Local data sets
  - Useful for local comparisons at University level
  - Useful at Country level
    e.g. REF (Research Excellence Framework) will supply only citation count ***c*** to expert reviewers

Open Questions

- Is universal form a log normal?
  - Many other forms in literature, we excluded them
  - Universality more useful than precise form
- What is the origin of the log normal with $\sigma^2=1.3$?
  - Simple models we tried failed e.g. Price
- Other processes for low citation papers?
  - We found best to exclude papers with $c_f < 0.1$
  - Errors in bibliographies, self-citation only significant for low cited papers

Thanks

- Nicola Hopkins
- Ben Kaube
- O.Kibaroglu and D.Hook for help with data
- Nuffield Foundation
- Imperial College London UROP scheme

T.S.Evans, N.Hopkins, B.Kaube, "Universality of Performance Indicators based on Citation and Reference Counts" **arXiv:1110.3271**

Bibliography

Evans, T.; Hopkins, N. & Kaube, B. “Universality of Performance Indicators based on Citation and Reference Counts” **2011,** arXiv:1110.3271.

T. Van Leewun, H. Moed, R. de Bruin. “New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications” *Scientometrics, **1995**, 33, 381–422.*

Limpert E.; STAHEL, W. A. & ABBT, M., “Log-normal Distributions across the Sciences: Keys and Clues” *BioScience,* **2001***, 51*, 341.

Radicchi, F.; Fortunato, S. & Castellano, C., “Universality of citation distributions: Toward an objective measure of scientific impact” PNAS*,* **2008**, 17268—17272.

van Raan, A. F. J., “Two-step competition process leads to quasi power-law income distributions - Application to scientific publication and citation distributions”, *Physica A,* **2001***, 298*, 530-536

Perc, M. “Zipf's law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenia's research as an example” *Journal of Informetrics,* **2010***, 4*, 358-364

Price, D. S. “Networks of Scientific Papers”, *Science,* **1965***, 149*, 510-515.

Watts, D. J. & Strogatz, S. H. “Collective dynamics of 'small-world' networks” *Nature,* **1998***, 393*, 440-442

Tim S. Evans – Mini Biography

Tim studied the mixture of quantum field theory and statistical physics in his PhD at Imperial College London. He was supervised by Prof. Ray Rivers who also supervised another speaker, Prof. Luis Bettencourt. Tim then spent time as a researcher at the University of Alberta in Edmonton Canada, before returning to research positions back here at Imperial, latterly as a Royal Society University Research Fellow. He was appointed to a lectureship at Imperial in1997.

Around 2003 he expanded his work on statistical physics to cover at problems in complexity, with a particular interest in network methods. This has included participation in an EU collaboration with social scientists on innovation, “ISCOM”, run in part by Prof. Geoff West (another speaker today). This fuelled his interest in social science applications and started an on going collaboration with an archaeologist.