Entropic selection of concepts unveils hidden topics in documents corpora

Alessio Cardillo
Department of Engineering Mathematics, University of Bristol, Bristol (UK)

@a_cardillo

Friday 29th March 2019, MRC IEU Data Science Group Meeting,
Bristol, United Kingdom



University of BRISTOL

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

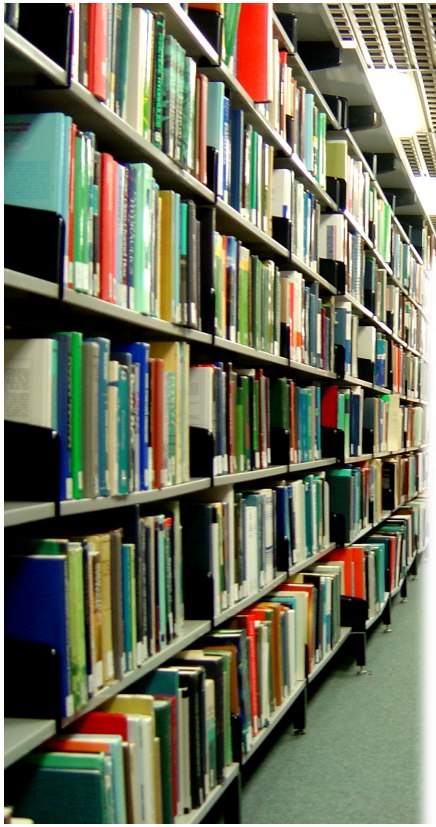
Once upon a time . . .

Once upon a time . . .



Once upon a time . . .







Nowdays . . .



Nowdays . . .




CHICAGOLAND SKYDIVING CENTER
ARE YOU READY FOR THE THRILL OF YOUR LIFE?!?

Flood of information



newsblog

Nature brings you breaking news from the world of science

News & Comment › News blog Archive › Post

Previous post
Climate change is present danger, US warns

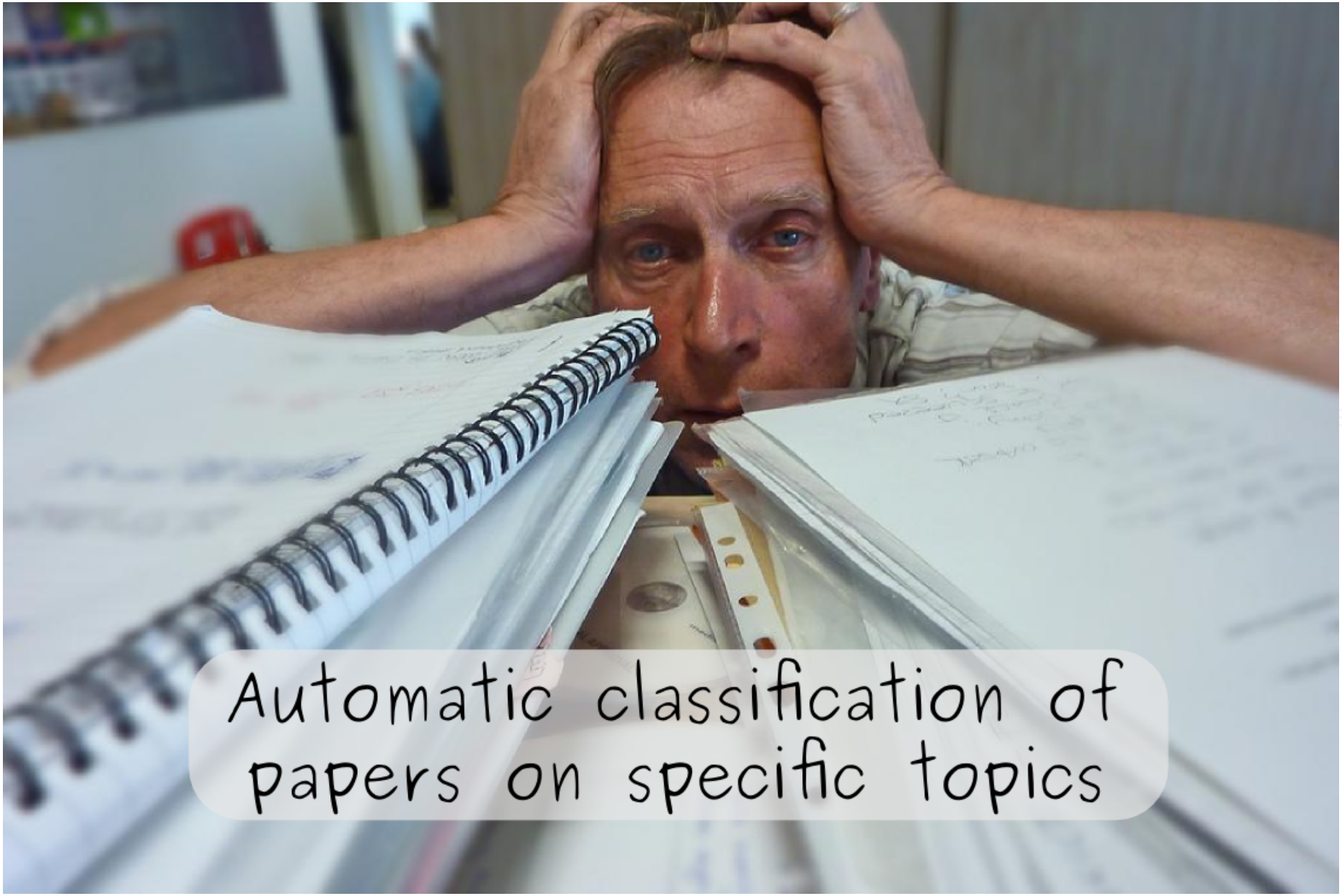
Next post
German research agencies condemn animal-rights attack on neuroscientist

NEWS BLOG

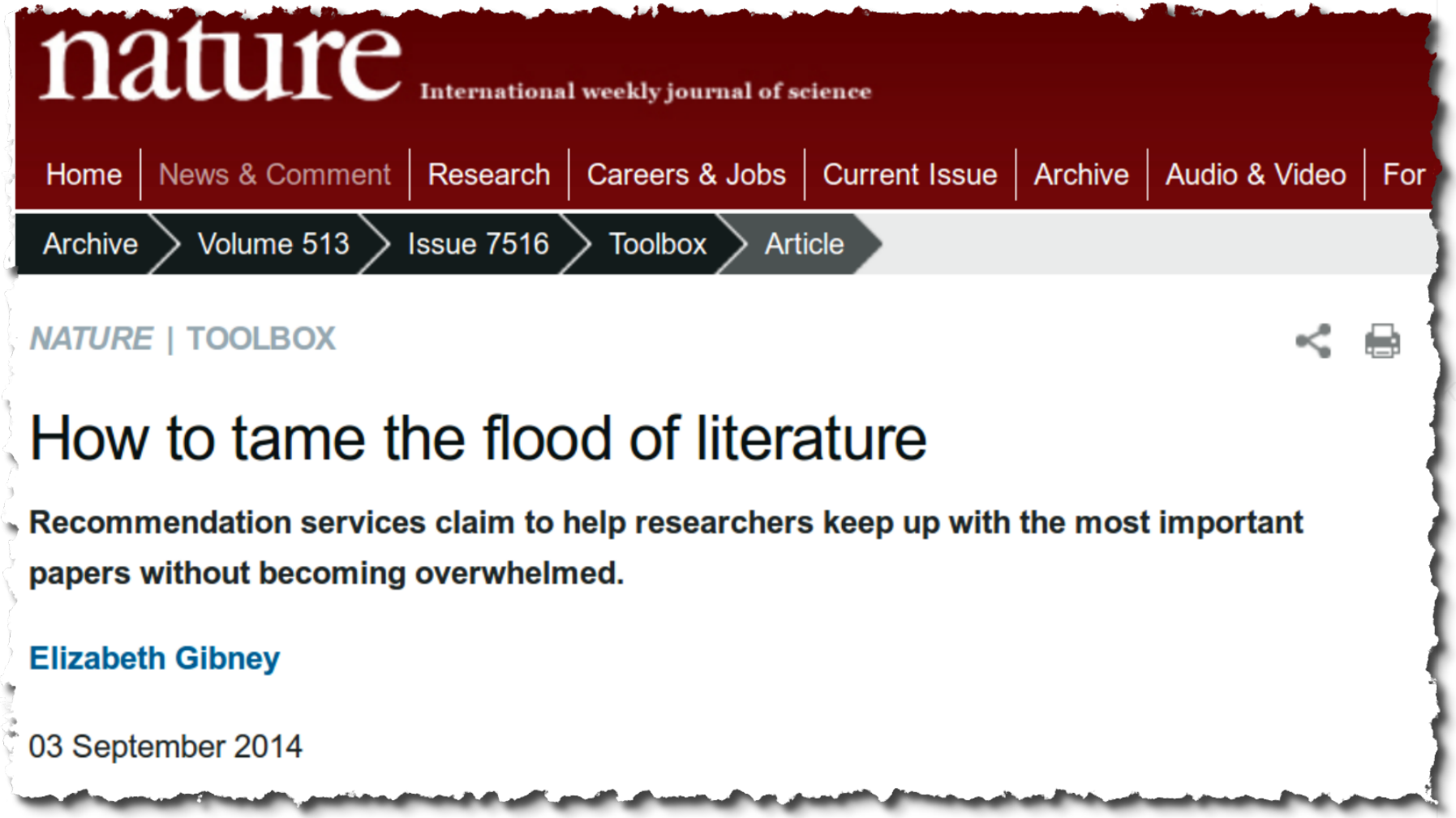
Global scientific output doubles every nine years

07 May 2014 | 16:46 GMT | Posted by Richard Van Noorden | Category: Policy, Publishing

Flood of information



Flood of information



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For

Archive > Volume 513 > Issue 7516 > Toolbox > Article

NATURE | TOOLBOX

How to tame the flood of literature

Recommendation services claim to help researchers keep up with the most important papers without becoming overwhelmed.

Elizabeth Gibney

03 September 2014

What do we need?

There is an inherent problem to giving you information that you weren't actively searching for. ***It has to be relevant*** *– so that we are not wasting your time –* ***but not too relevant****, because you already know about those articles.*

Anurag Acharya
Google Scholar co-creator

What do we need?

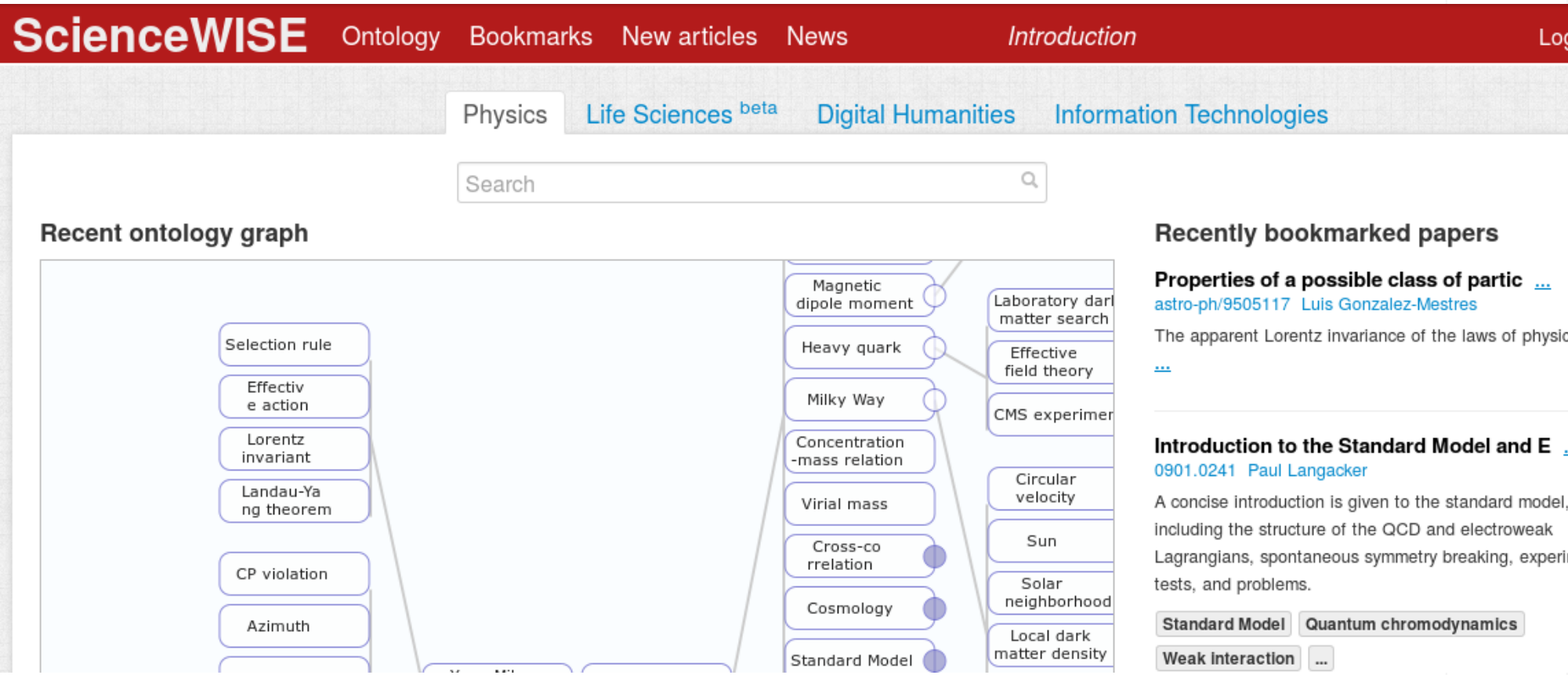
There is an inherent problem to giving you information that you weren't actively searching for. ***It has to be relevant*** *– so that we are not wasting your time –* ***but not too relevant****, because you already know about those articles.*

Anurag Acharya
Google Scholar co-creator

Semantic Scholar offers a few innovative features, including picking out the ***most important keywords and phrases*** *from the text without relying on an author or publisher to key them in. "****It's surprisingly difficult for a system to do this****,"*

Oren Etzioni
CEO of AI2 (Semantic Scholar)

What do we need?



http://sciencewise.info

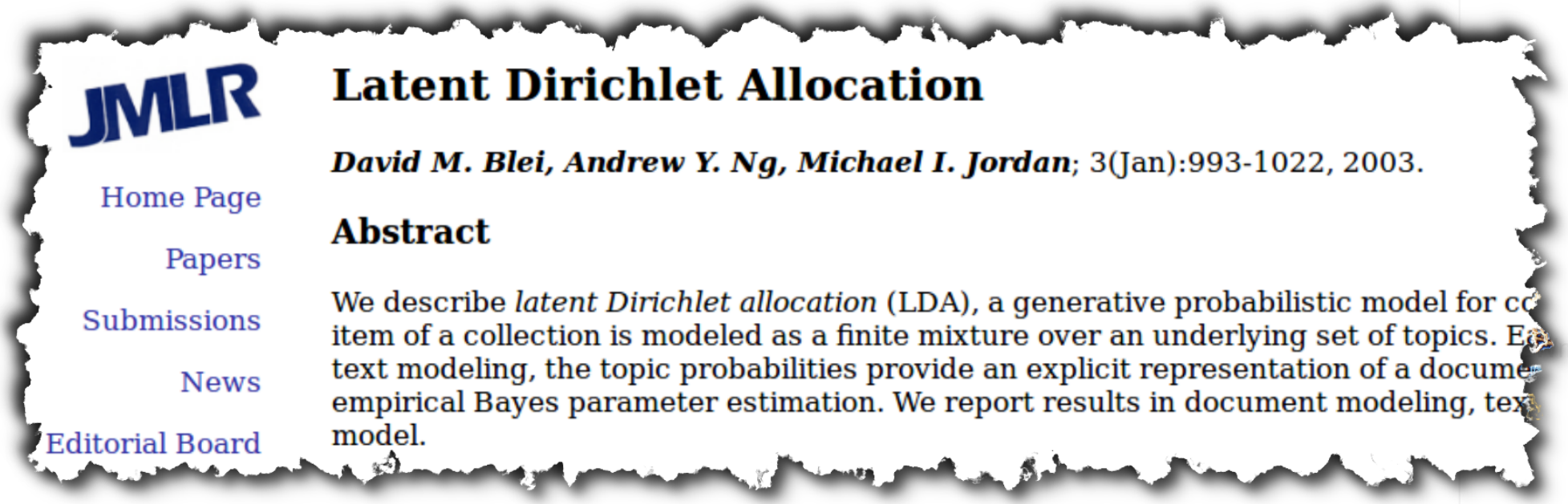
Outline

- Introduction on topic modeling & LDA.
- ★ Filtering of concepts.
- ★ **Entropic** filtering of concepts.
- ★ Results with "*Special Effects*".
- Take home messages
- Questions

Section 1

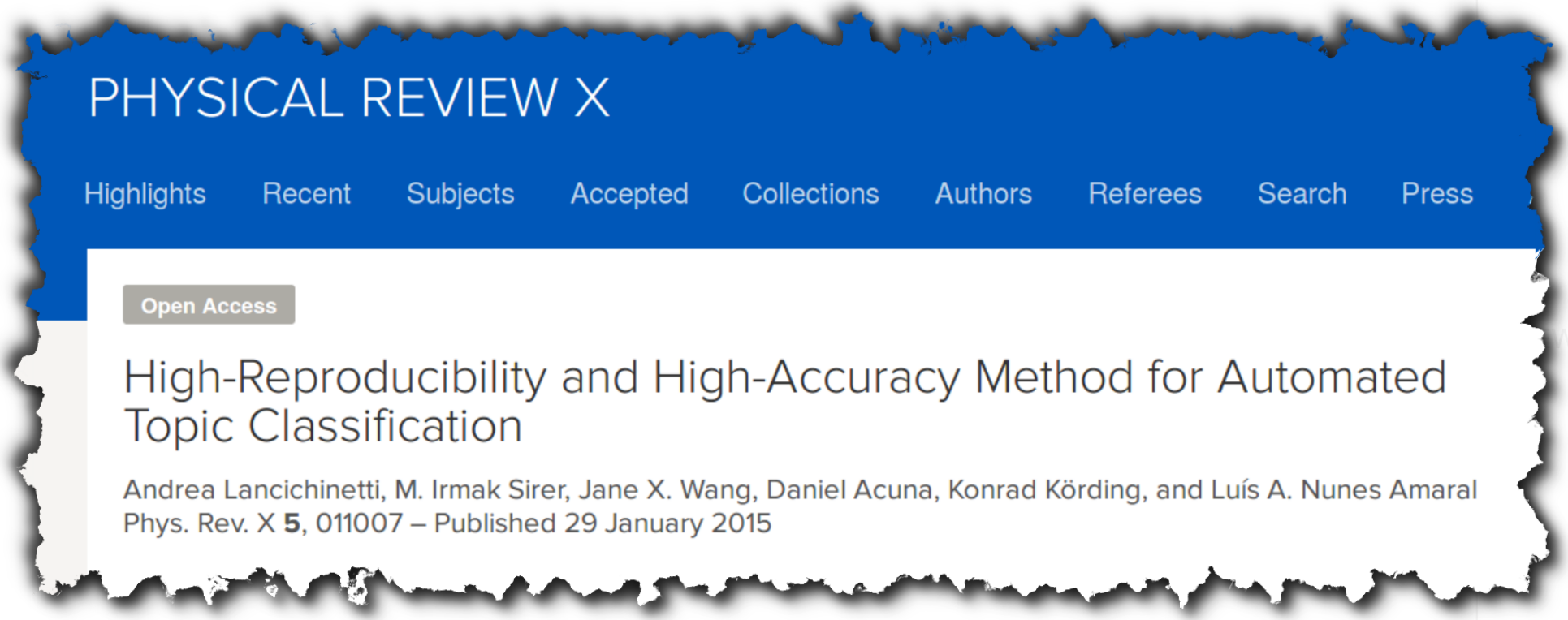
Topic modeling

Topic Modeling & LDA in a nutshell



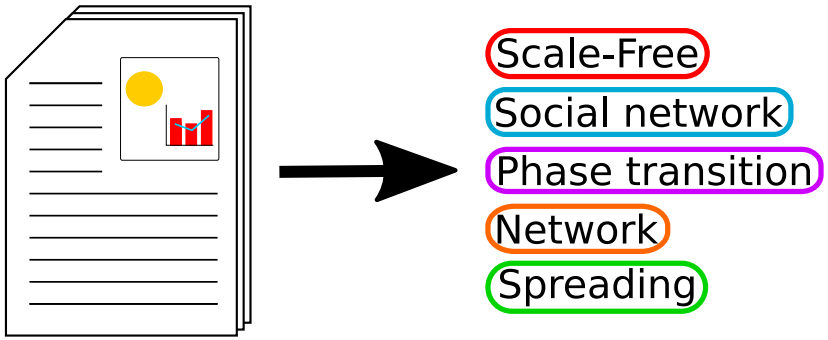
- D.M. Blei *et al.* "*Latent dirichlet allocation*". Journal of Machine learning Research **3** 993 (2003).
- C.D. Manning *et al.* "*Introduction to Information Retrieval*". Cambridge University Press, (2008)

Topic Modeling & LDA in a nutshell

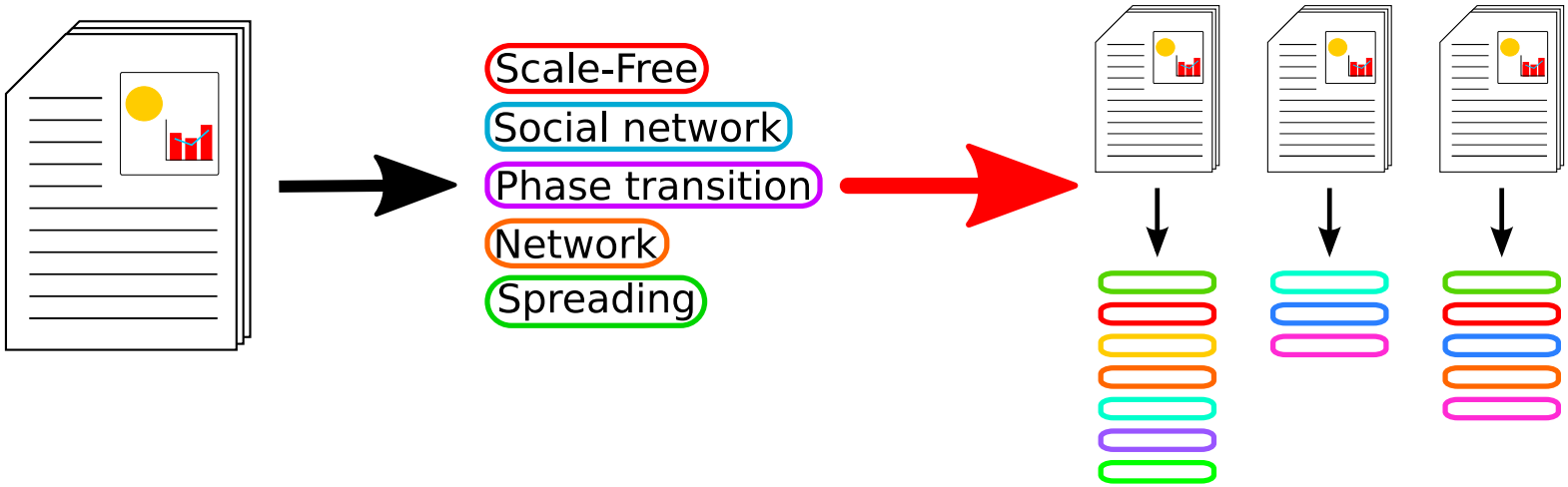


- A. Lancichinetti *et al.* "*High-Reproducibility and High-Accuracy Method for Automated Topic Classification*". Phys. Rev. X **5** 011007 (2015)

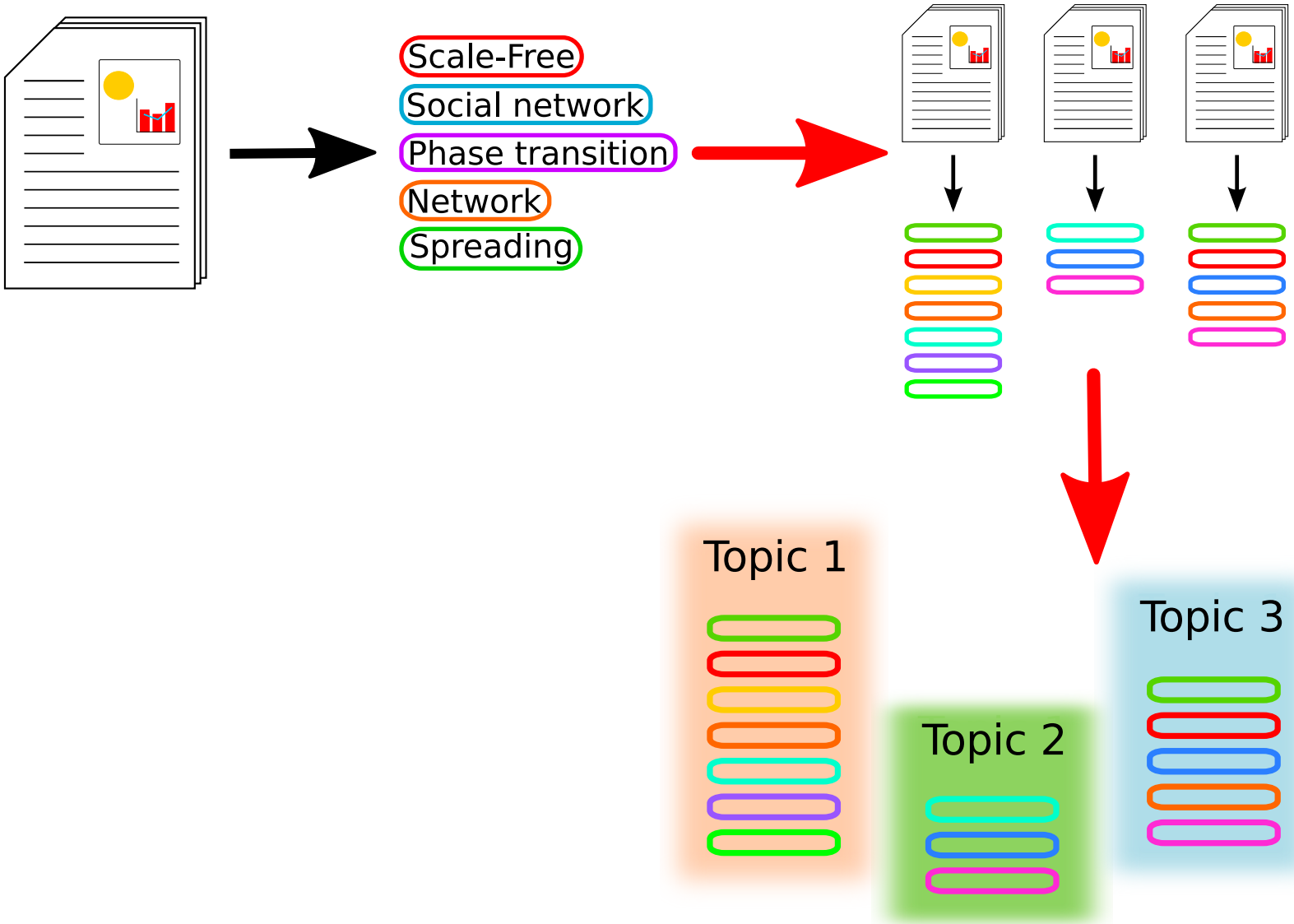
Topic Modeling & LDA in a nutshell



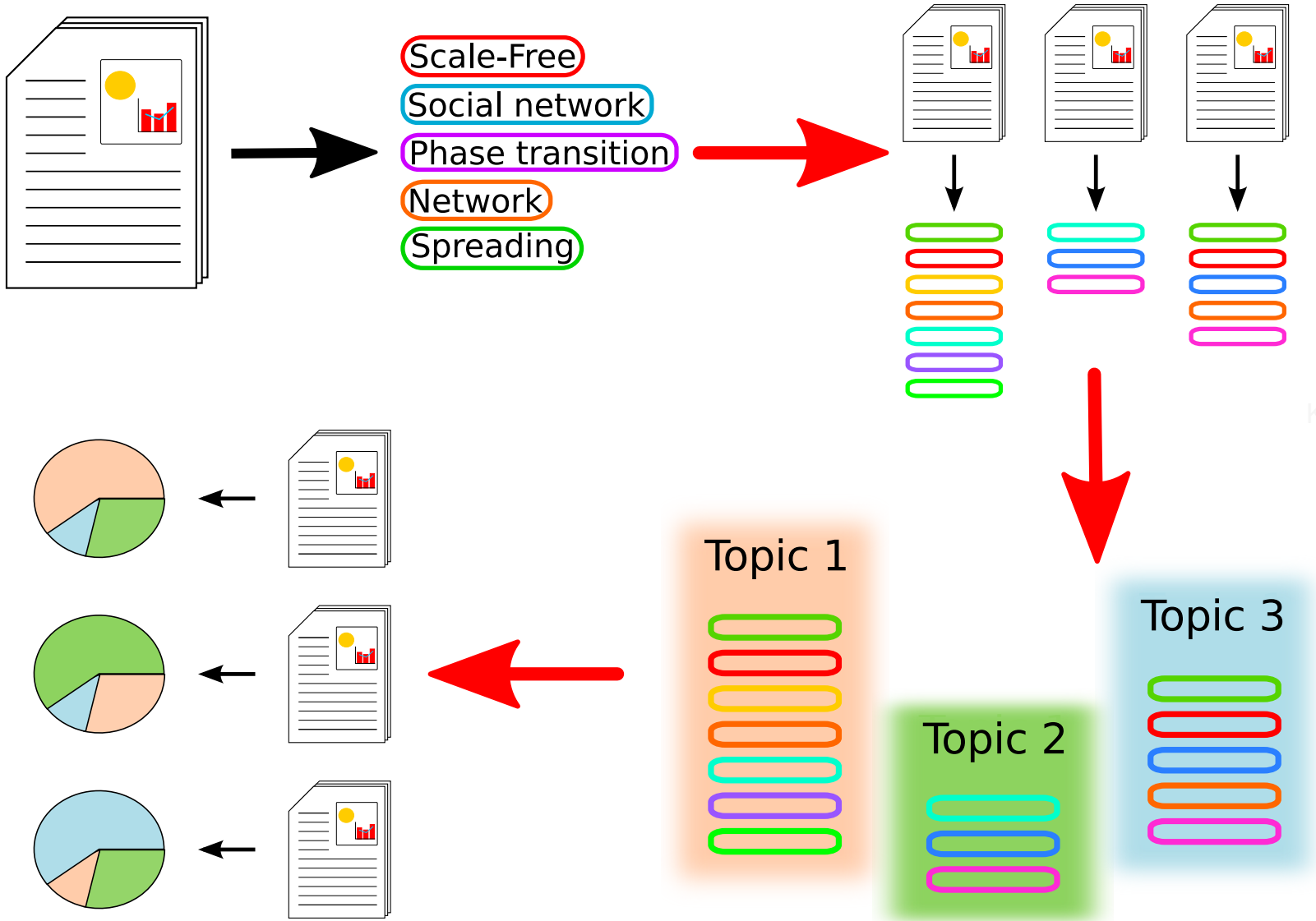
Topic Modeling & LDA in a nutshell



Topic Modeling & LDA in a nutshell



Topic Modeling & LDA in a nutshell

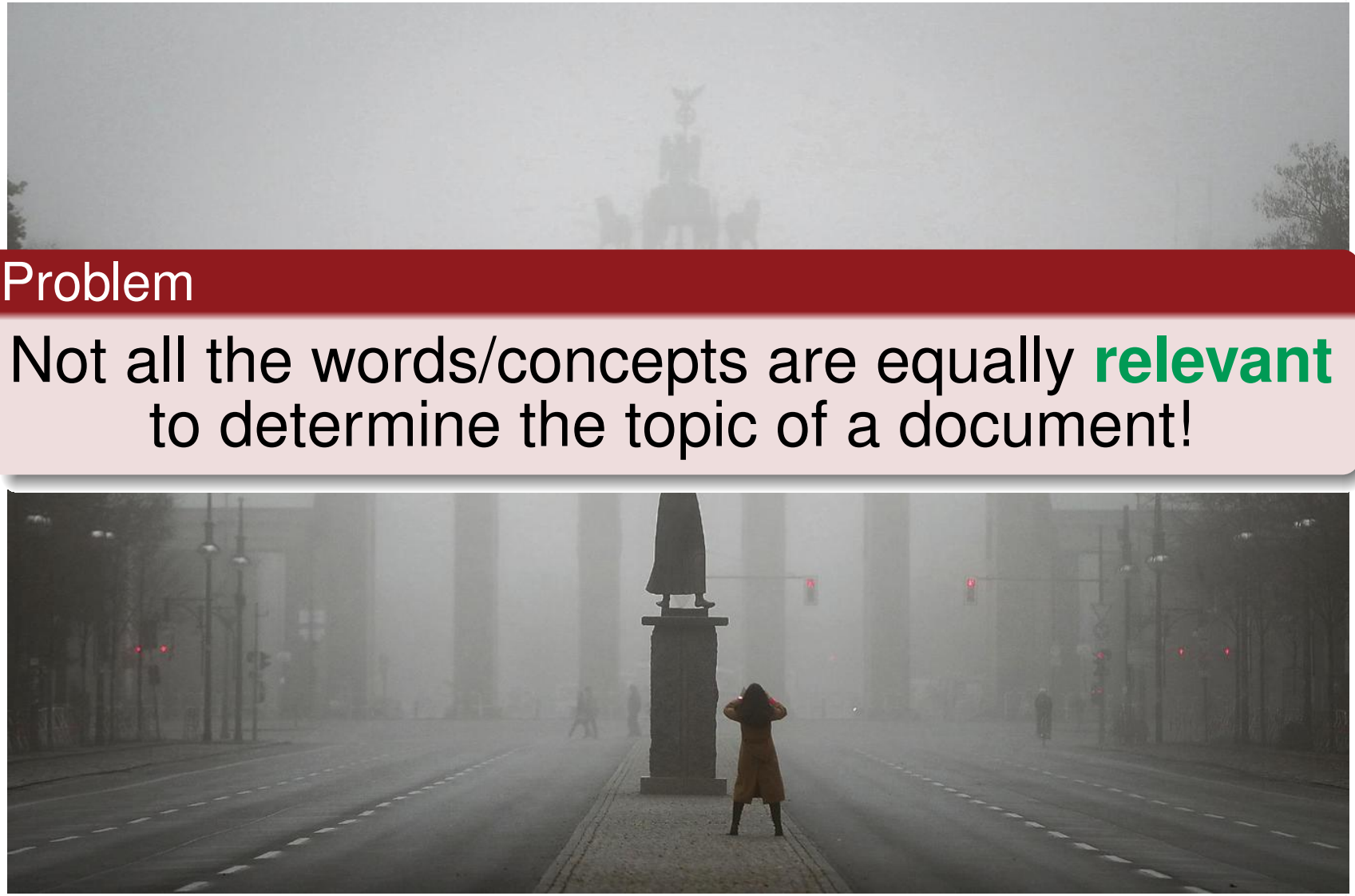


Topic Modeling & LDA in a nutshell

Topic Modeling & LDA in a nutshell



Topic Modeling & LDA in a nutshell



Problem

Not all the words/concepts are equally **relevant** to determine the topic of a document!

Topic Modeling & LDA in a nutshell



Section 2

Filtering

Relevant concepts



Relevant concepts

Key features

- # of documents a concept appears in

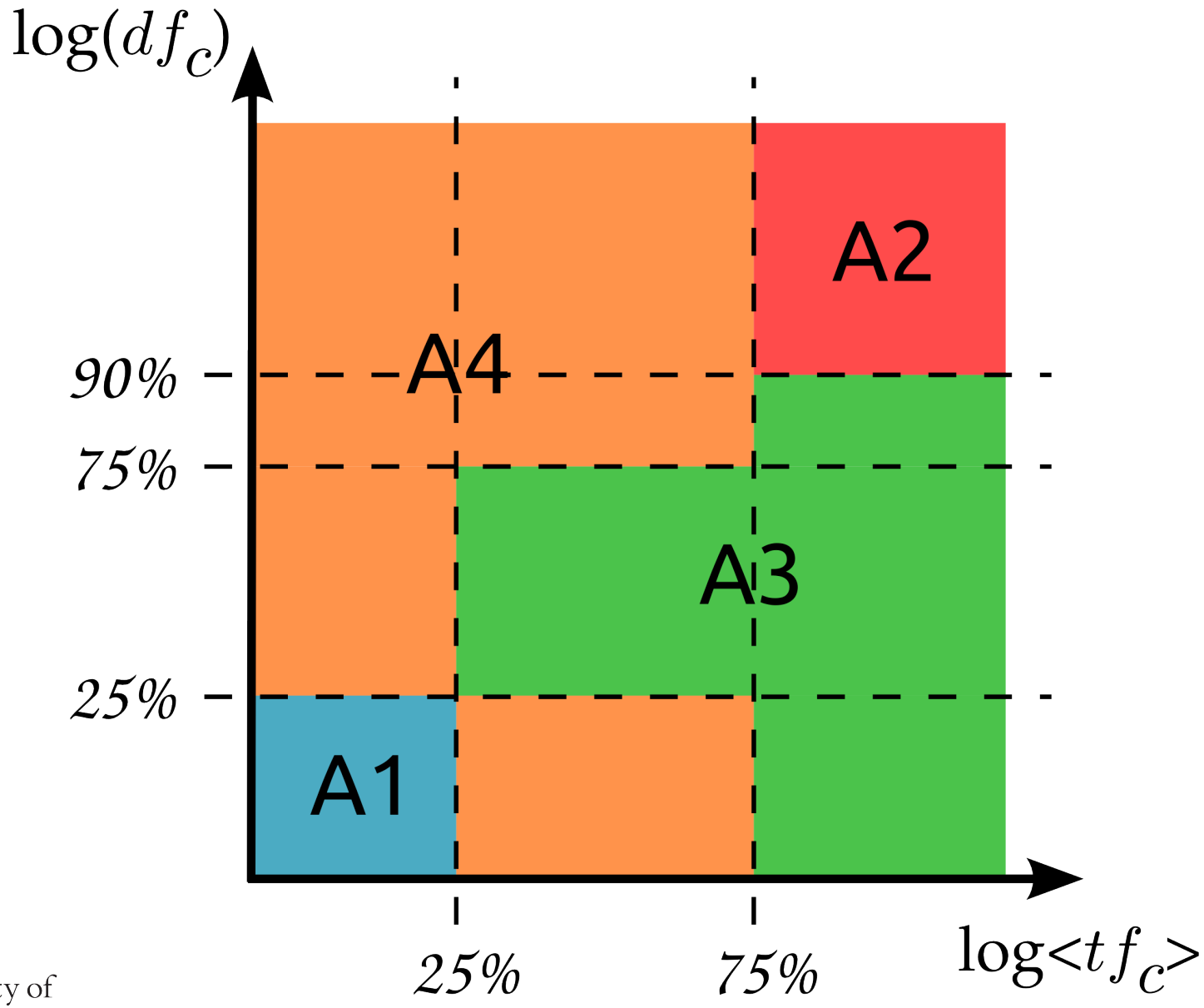
$$df_c \rightarrow \text{ document frequency}$$

- average # of times a concept appears inside a document

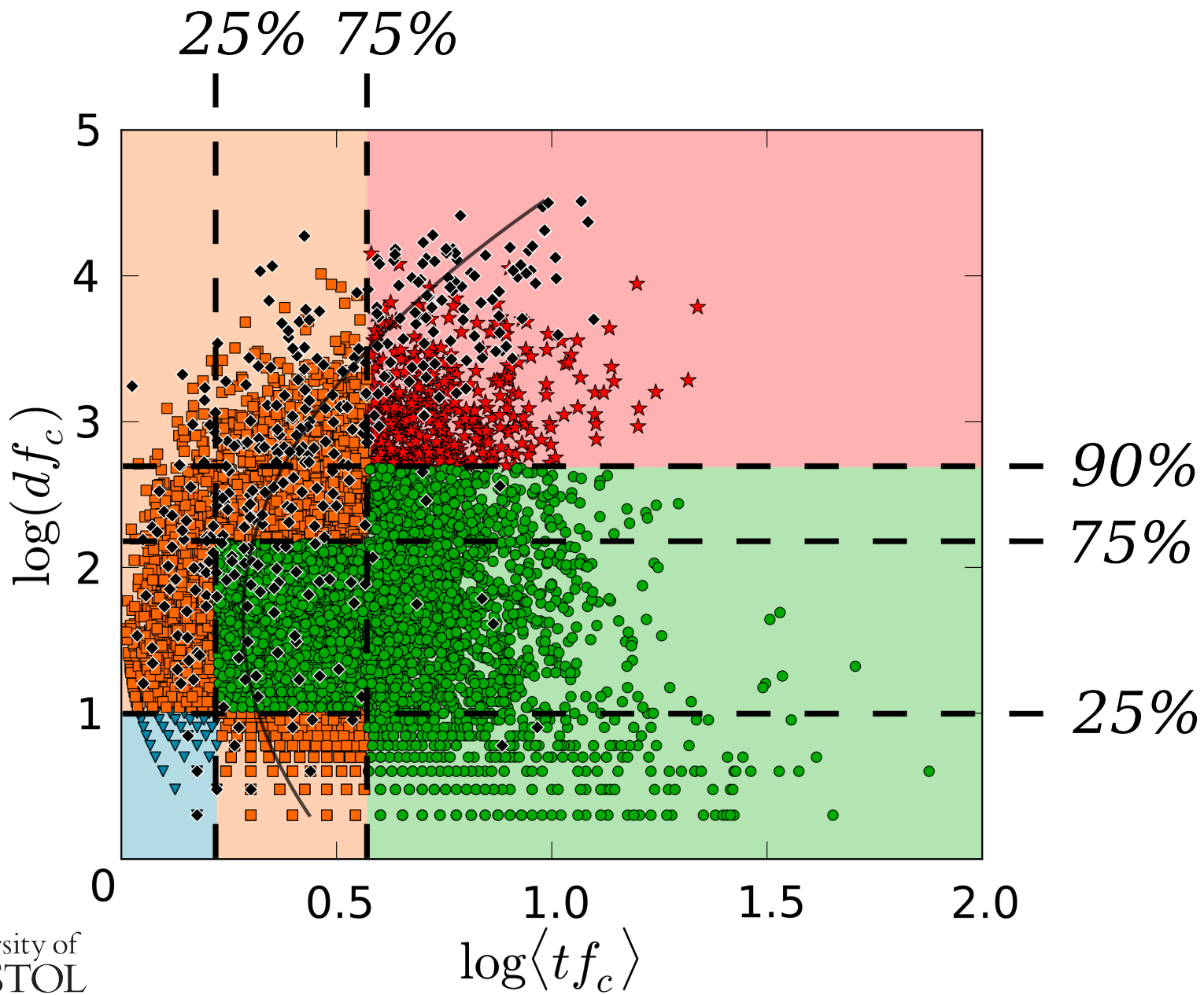
$$\langle tf_c \rangle \rightarrow \text{ average term frequency}$$

• D. Jurafsky and J. Martin "*Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*" Prentice Hall (2000).

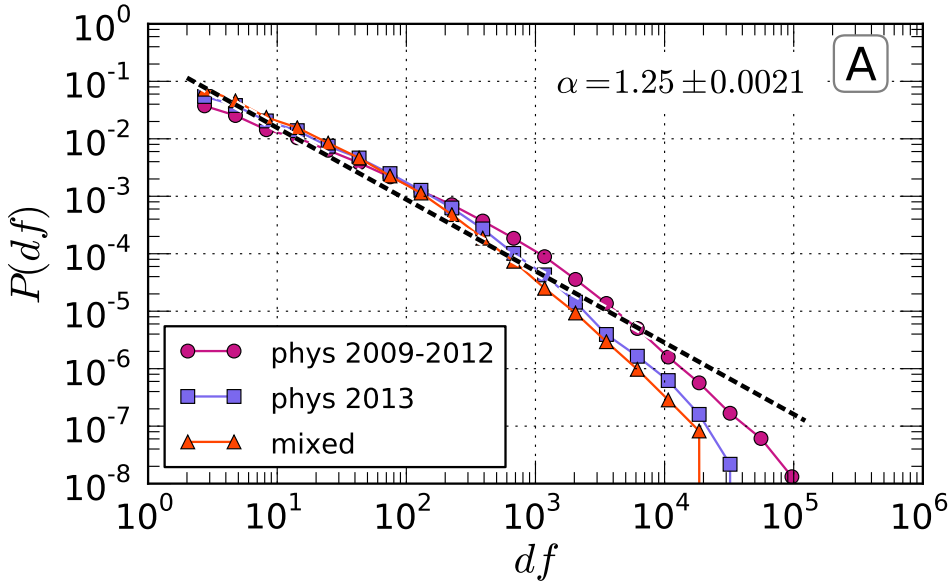
Bidimensional tessellation

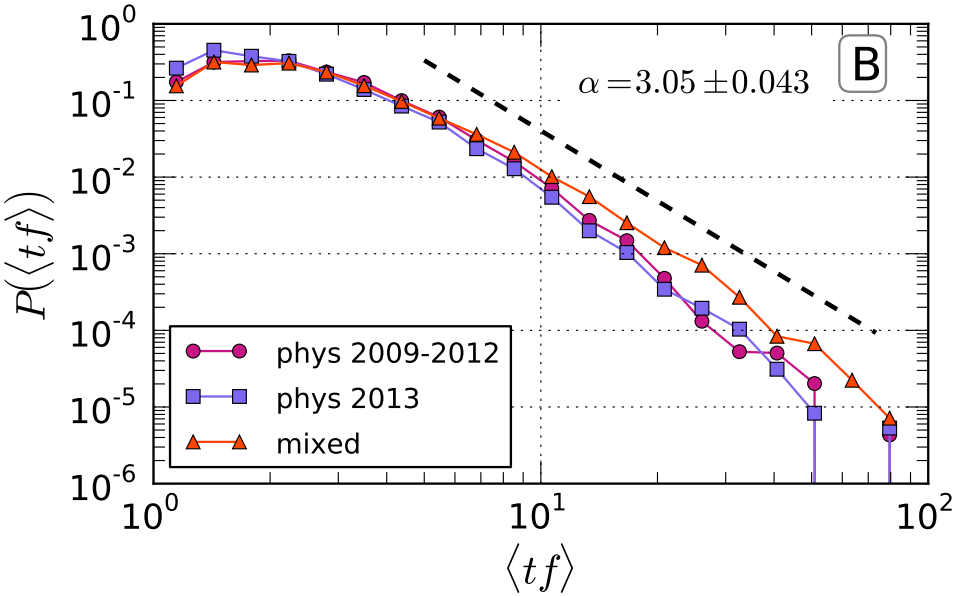


Bidimensional tessellation



Bidimensional tessellation



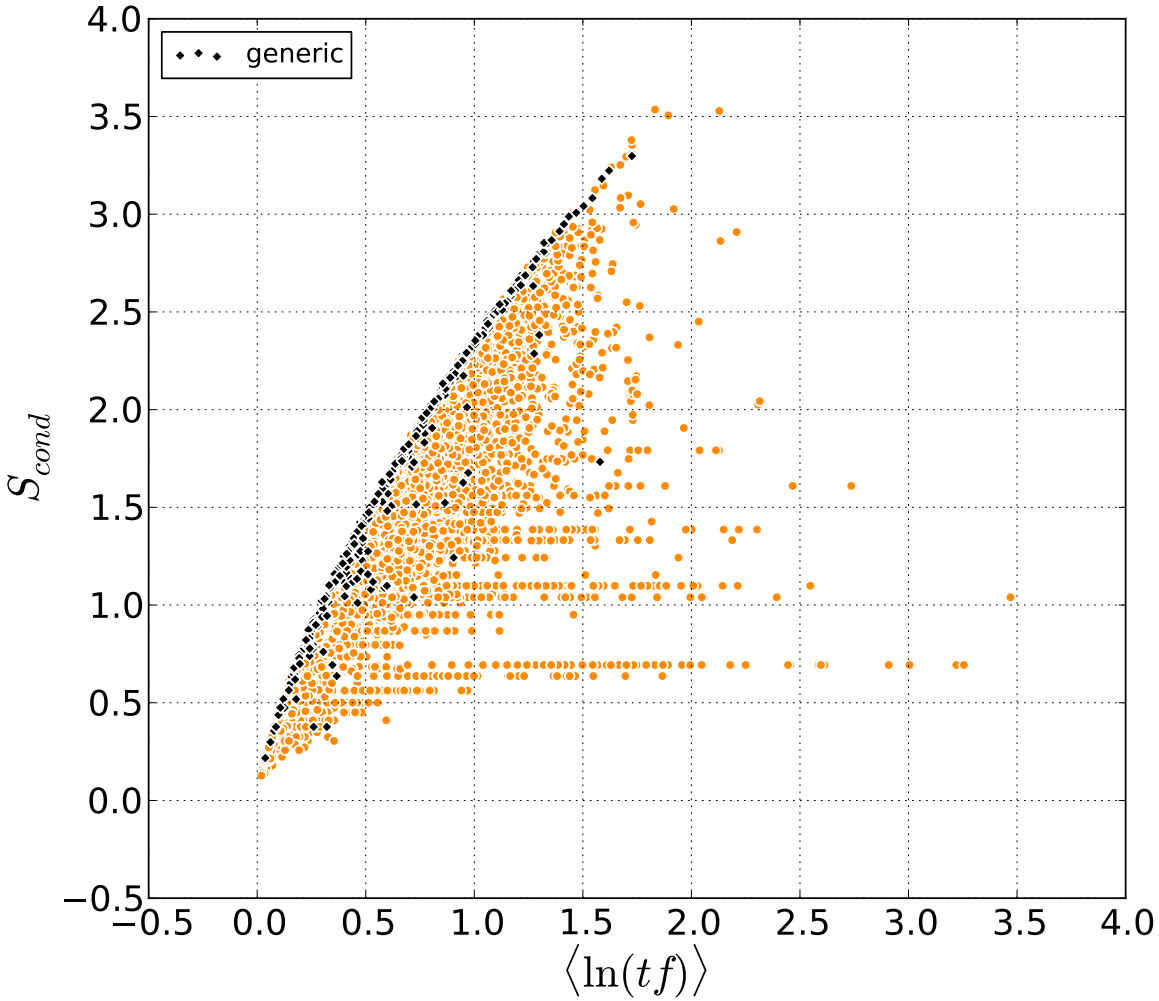


- F. Font-Clos *et al.* "*A scaling law beyond Zipf's law and its relation to Heaps' law*". New J. Phys. **15** 093033 (2013).
- M. Gerlach *et al.* "*Scaling laws and fluctuations in the statistics of . . .*". New J. Phys. **16** 113010 (2014).

Section 3

Entropic Filtering

Maximum entropy



$$S = -\sum_{j=0}^{\infty} p_c(j) \ln p_c(j)$$

- A. Berger *et al.* "*A Maximum Entropy Approach to Natural Language . . .*". Computational Linguistics **22** 39 (1996).

Maximum entropy

$$\sum_n p_n = 1$$

$$\sum_n p_n \, n = \langle n \rangle$$

$$\sum_n p_n \, \ln n = \langle \ln n \rangle$$

$$\ln p_n + \lambda n + \mu \ln n = 0$$

$$p_n = \frac{1}{Z} e^{-\lambda n} n^{-\mu}$$

- M. Gerlach *et al.* "*Scaling laws and fluctuations in the statistics of . . .*". New J. Phys. **16** 113010 (2014)
- R. Ferrer i Cancho *et al.* "*Least effort and the origins of scaling in . . .*". Proc. Nat. Acad. Sci. USA **100** 788 (2003).

Maximum entropy

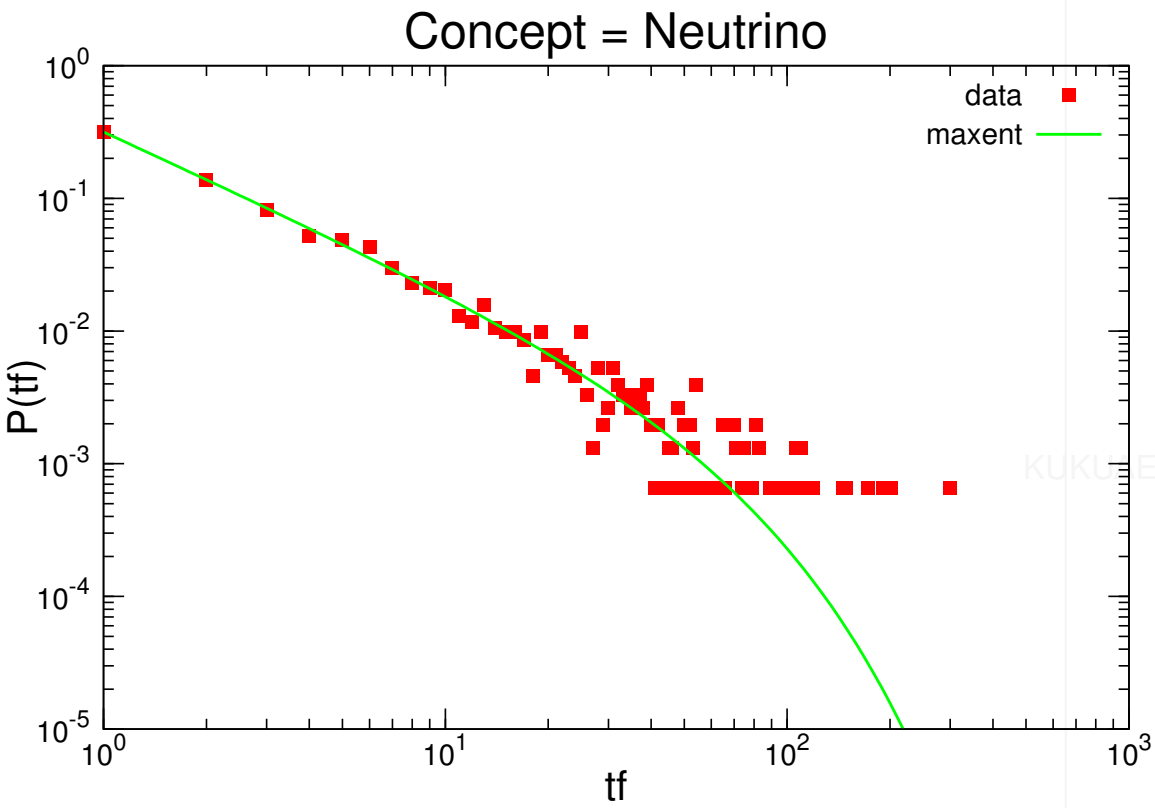
$$\sum_n p_n = 1$$

$$\sum_n p_n\, n = \langle n \rangle$$

$$\sum_n p_n \ln n = \langle \ln n \rangle$$

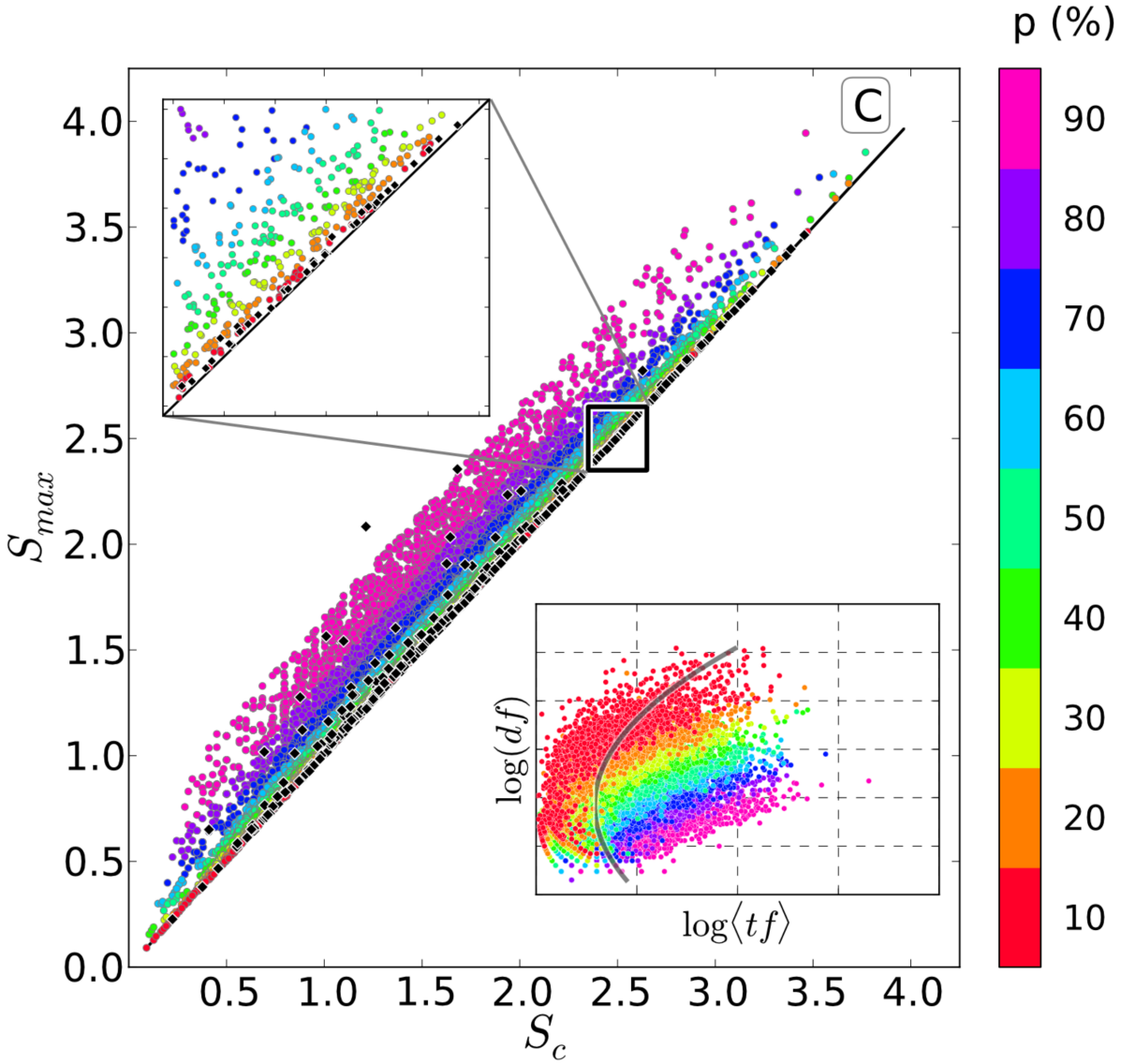
$$\ln p_n + \lambda n + \mu \ln n = 0$$

$$p_n = \frac{1}{Z} e^{-\lambda n} n^{-\mu}$$



- M. Gerlach *et al.* "*Scaling laws and fluctuations in the statistics of . . .*". New J. Phys. **16** 113010 (2014)
- R. Ferrer i Cancho *et al.* "*Least effort and the origins of scaling in . . .*". Proc. Nat. Acad. Sci. USA **100** 788 (2003).

Maximum entropy



Residual Entropy

$$S_d(c) = S_{\max}(c) - S_c(c)$$

Section 4

Results

Data

Documents collections

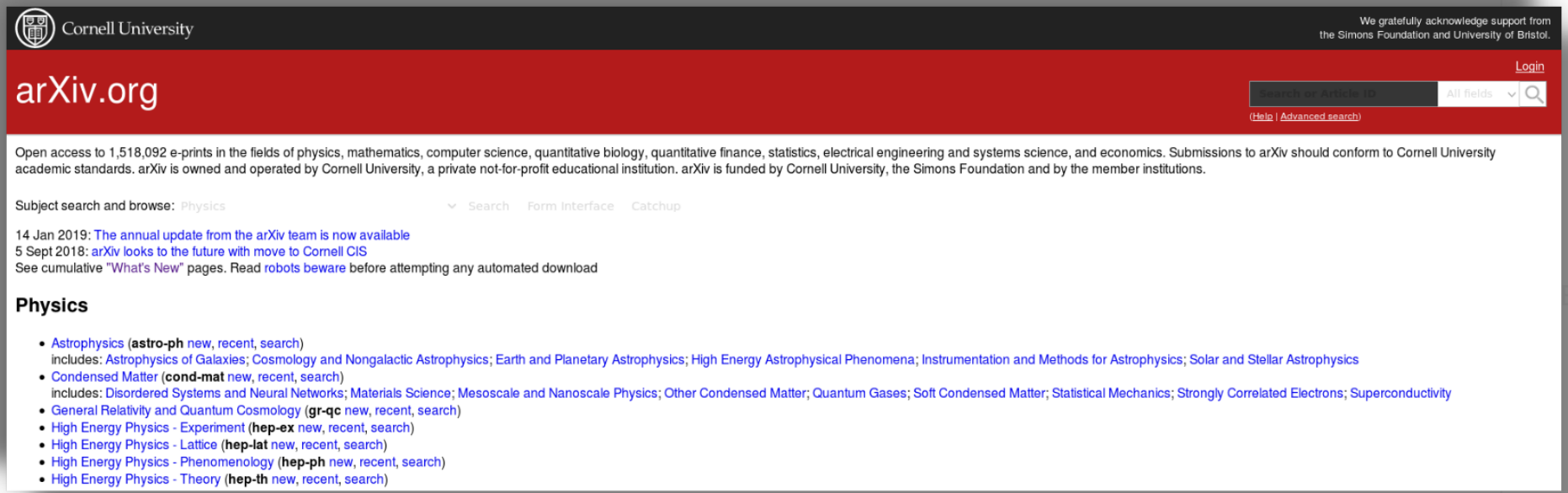
	N_a	N_{con}	T	T^*	$\langle N_a \rangle_{T^*}$	$\langle N_{con} \rangle_{T^*}$
`arXiv` Physics 2013	52979	13173	10	10	5298	4212
`arXiv` Physics 2009 – 2012	189759	15040	10	10	18976	6185
`arXiv` Mixed	50408	19843	14	12	4155	3994
Climate change webdocs	18770	822545	201	22	432	26004

Data

Documents collections

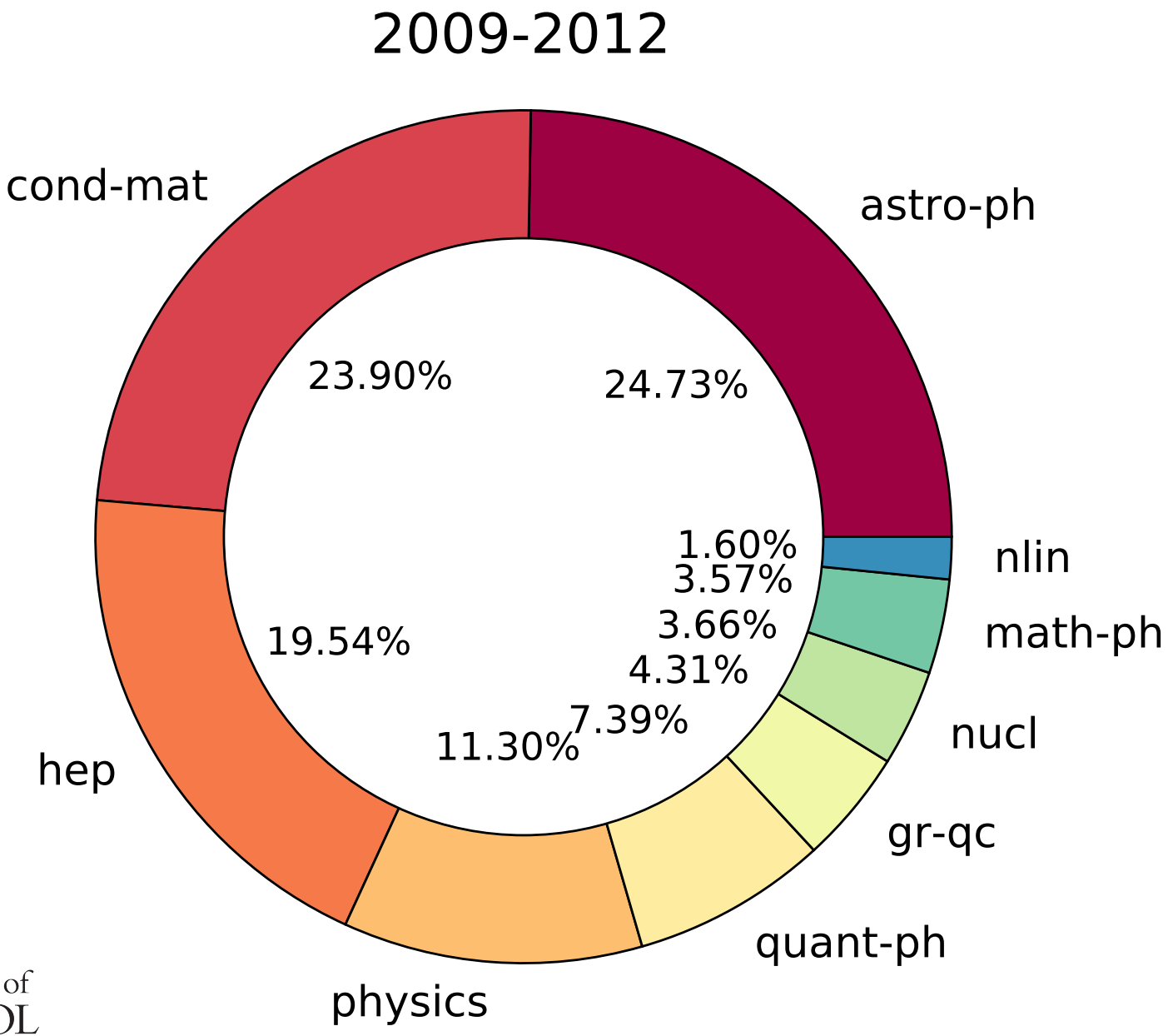
	N_a	N_{con}	T	T^*	$\langle N_a \rangle_{T^*}$	$\langle N_{con} \rangle_{T^*}$
`arXiv` Physics 2013	52979	13173	10	10	5298	4212
`arXiv` Physics 2009 – 2012	189759	15040	10	10	18976	6185
`arXiv` Mixed	50408	19843	14	12	4155	3994
Climate change webdocs	18770	822545	201	22	432	26004

Data

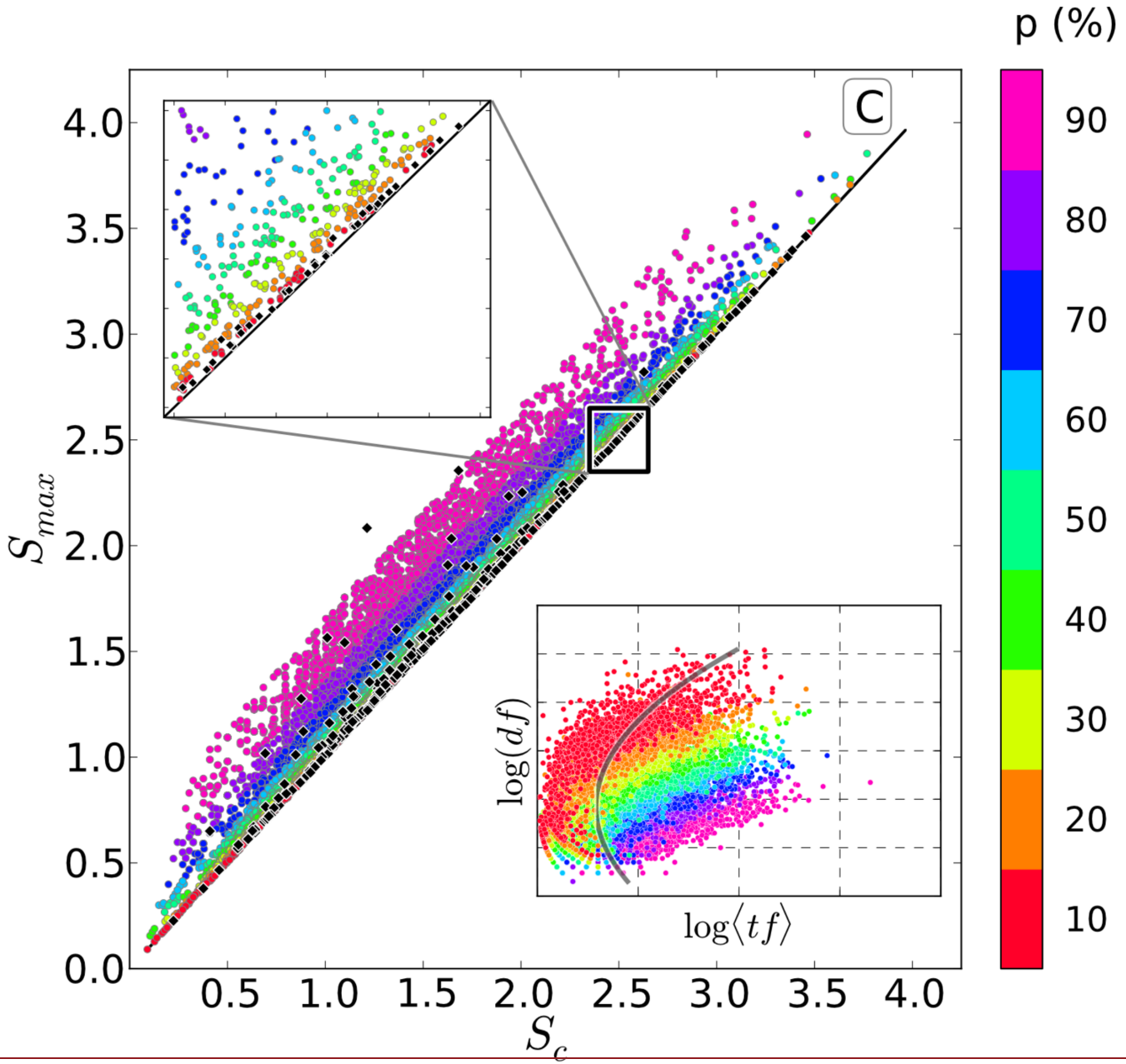


https://arxiv.org/

Data


2009-2012
cond-mat
astro-ph
23.90%
24.73%
1.60%
3.57%
3.66%
4.31%
7.39%
11.30%
19.54%
nlin
math-ph
nucl
gr-qc
quant-ph
physics
hep

What is a “generic concept”?

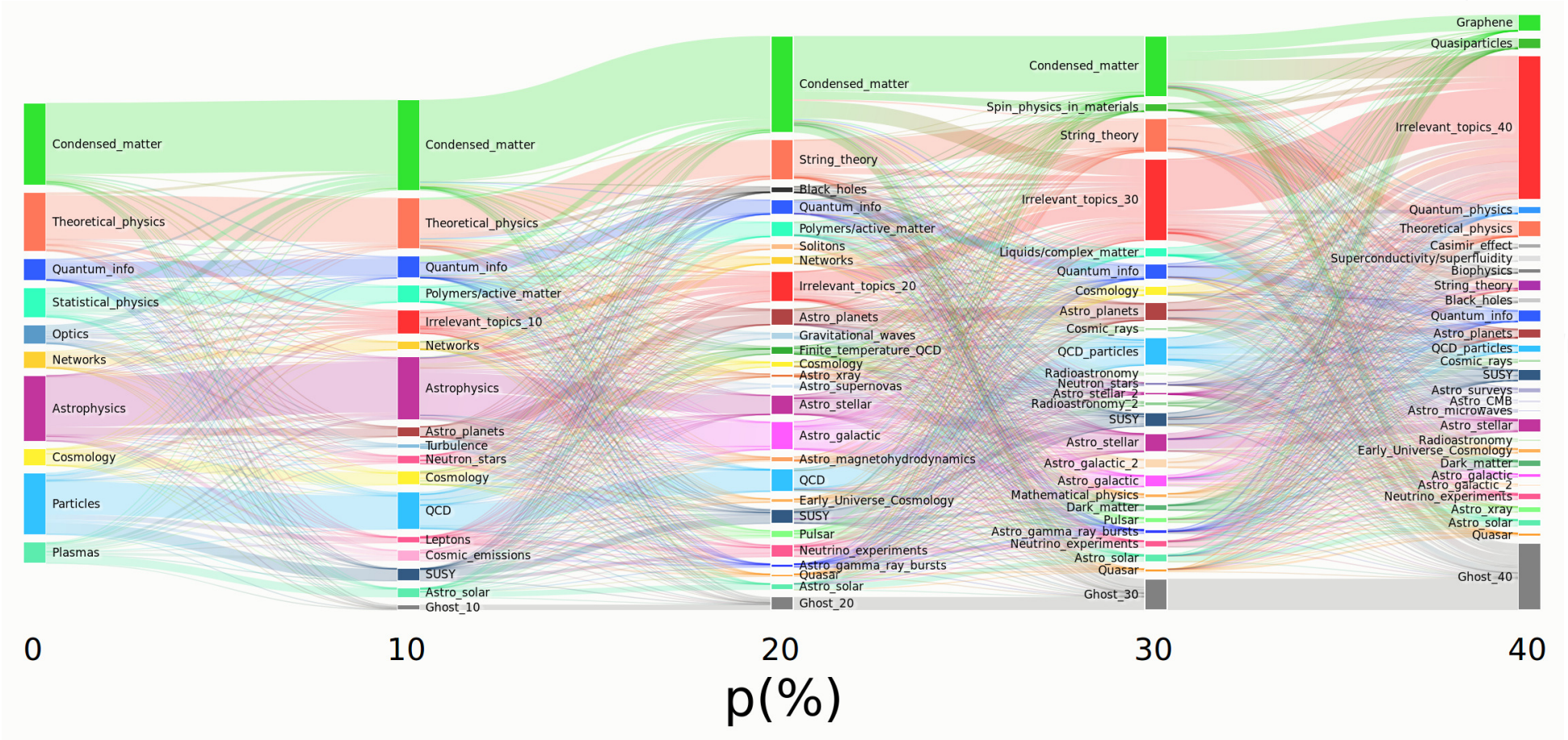


Effects of pruning concepts



▶ Link

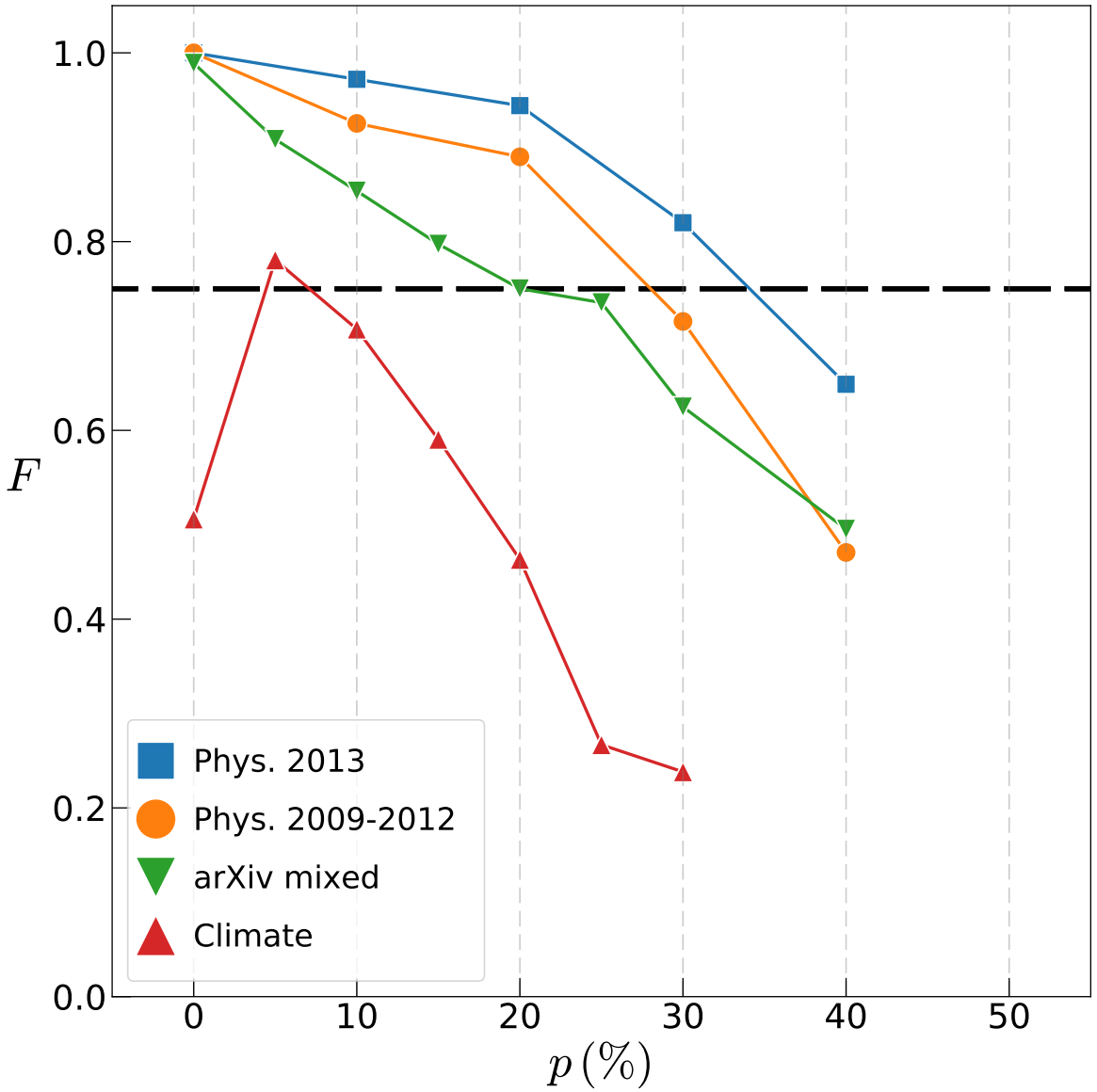
Effects of pruning concepts



Optimal filtering

$$F(p) = \frac{N_{good}(p)}{N_{tot}} =$$

$$= 1 - \frac{N_a^G(p) + N_a^I(p)}{N_{tot}}$$

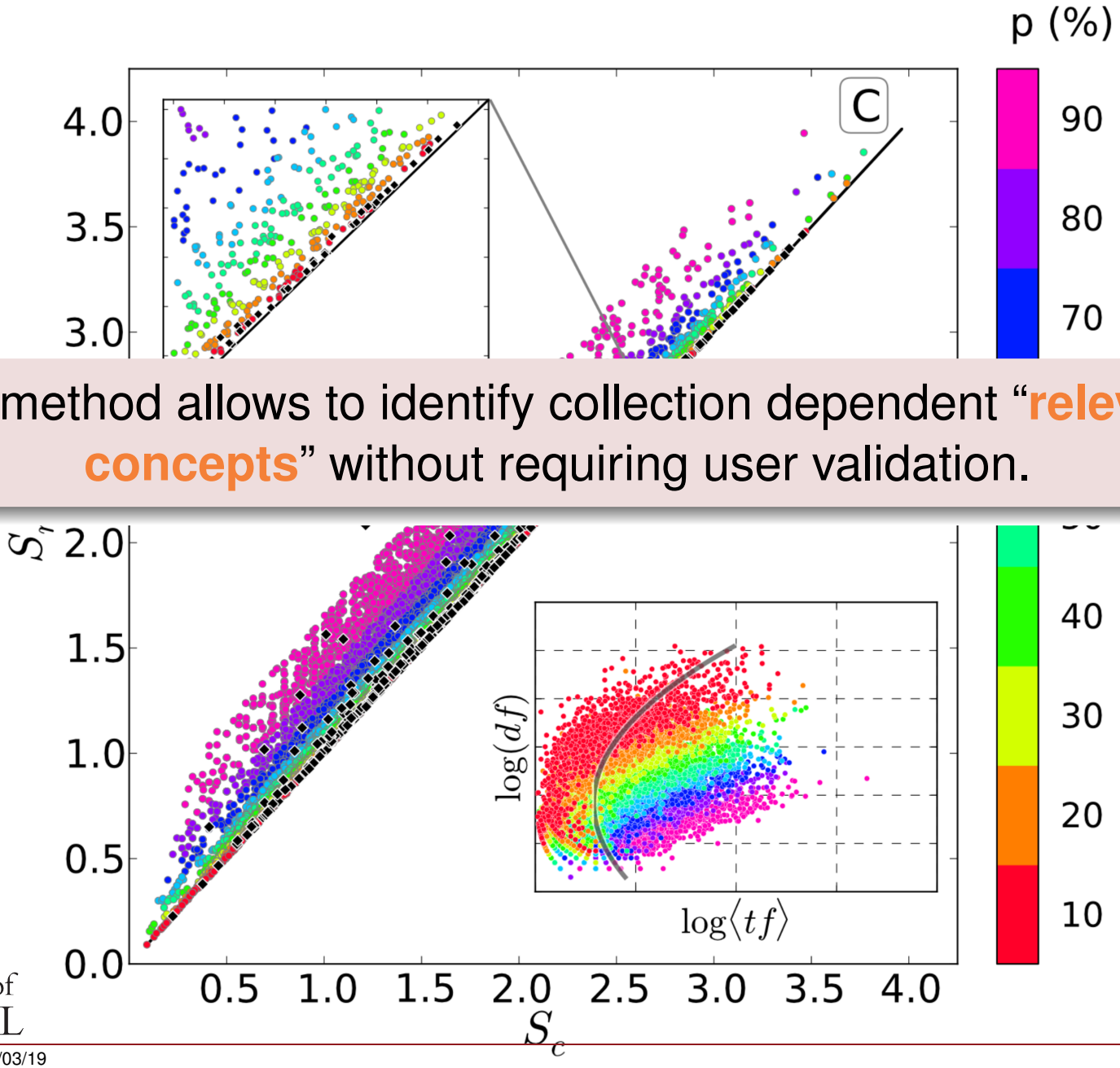


Section 5

Conclusions

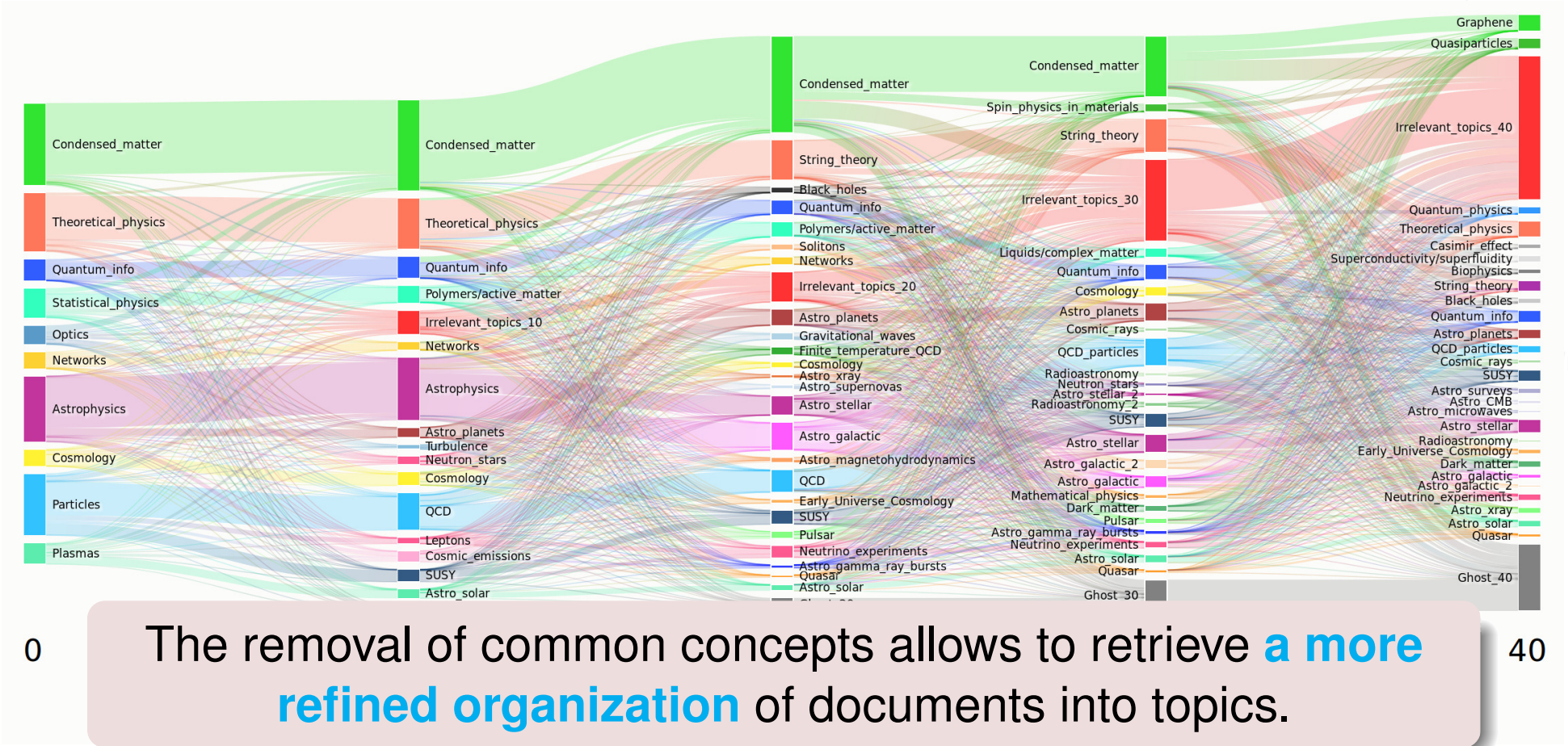
Summing up . . .

Summing up . . .



The method allows to identify collection dependent "**relevant concepts**" without requiring user validation.

Summing up . . .



The removal of common concepts allows to retrieve **a more refined organization** of documents into topics.

Summing up . . .



The removal of common concepts allows to retrieve **a more refined organization** of documents into topics.

Summing up . . .

What's next? Open questions

- Study the evolution in time of "generality".
- Study the relation between concepts.
- Use the method to build ontologies.

Acknowledgements



Paolo De Los Rios

Andrea Martini





Alex Constantin

&

ScienceWISE team

Acknowledgements



FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

Acknowledgements

arXiv.org > physics > arXiv:1705.06510

Physics > Physics and Society

Entropic selection of concepts unveils hidden topics in documents corpora

Andrea Martini, Alessio Cardillo, Paolo De Los Rios

(Submitted on 18 May 2017 (v1), last revised 11 May 2018 (this version, v2))

The organization and evolution of science has recently become itself an object of scientific quantitative investigation, thanks to the wealth of in documents, such as citations between papers and co-authorship between researchers. However, only few studies have focused on the conce can be extracted and analyzed, revealing the deeper organization of scientific knowledge. Unfortunately, several concepts can be so common emergence of the underlying topical structure of the document corpus, because they give rise to a large amount of spurious and trivial relation common concepts, we introduce a method to gauge their relevance according to an objective information-theoretic measure related to the sta corpus. After progressively removing concepts that, according to this metric, can be considered as generic, we find that the topic organization structure.

alessio.cardillo@bristol.ac.uk

http://www.bifi.es/∼cardillo/

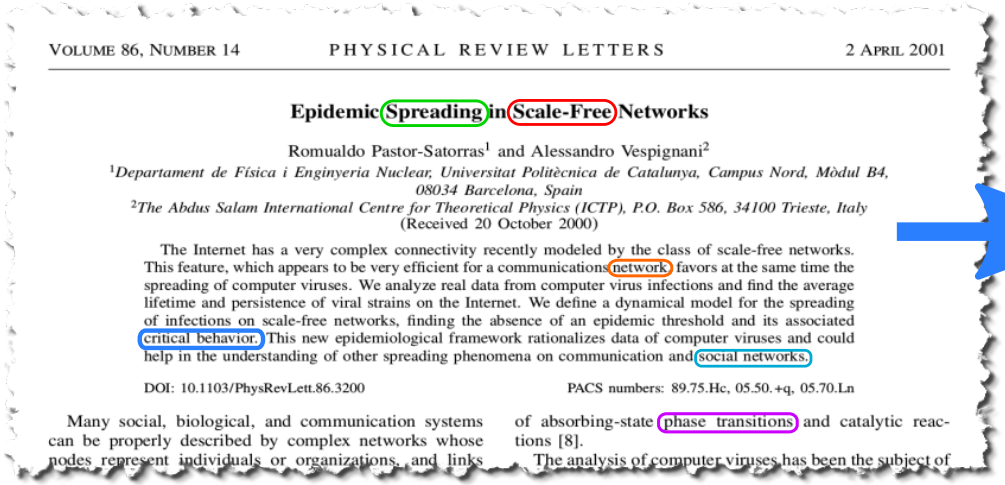
@a_cardillo

University of BRISTOL

Part II

Appendix

Networks of documents



VOLUME 86, NUMBER 14 PHYSICAL REVIEW LETTERS 2 APRIL 2001

Epidemic Spreading in Scale-Free Networks

Romualdo Pastor-Satorras[1] and Alessandro Vespignani[2]

[1]*Departament de Física i Enginyeria Nuclear, Universitat Politècnica de Catalunya, Campus Nord, Mòdul B4, 08034 Barcelona, Spain*

[2]*The Abdus Salam International Centre for Theoretical Physics (ICTP), P.O. Box 586, 34100 Trieste, Italy*

(Received 20 October 2000)

The Internet has a very complex connectivity recently modeled by the class of scale-free networks. This feature, which appears to be very efficient for a communications network, favors at the same time the spreading of computer viruses. We analyze real data from computer virus infections and find the average lifetime and persistence of viral strains on the Internet. We define a dynamical model for the spreading of infections on scale-free networks, finding the absence of an epidemic threshold and its associated critical behavior. This new epidemiological framework rationalizes data of computer viruses and could help in the understanding of other spreading phenomena on communication and social networks.

DOI: 10.1103/PhysRevLett.86.3200 PACS numbers: 89.75.Hc, 05.50.+q, 05.70.Ln

Many social, biological, and communication systems can be properly described by complex networks whose nodes represent individuals or organizations, and links of absorbing-state phase transitions and catalytic reactions [8].

The analysis of computer viruses has been the subject of

(a)

Scale-Free

Social network

Phase transition

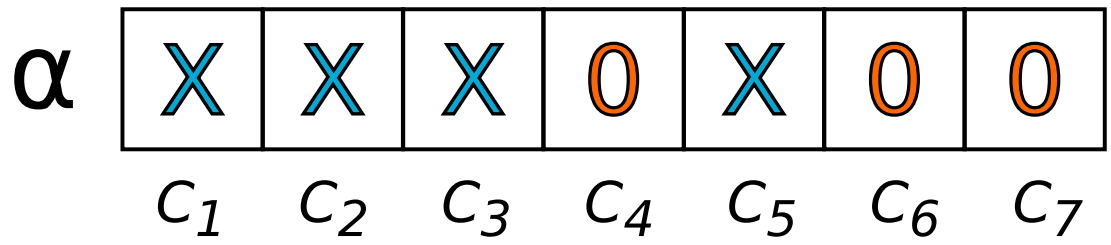
Network

Spreading

(b)

(c)

TF-IDF and similarity



$$TF\text{-}IDF_{\alpha c} = u_{\alpha c} = \underbrace{tf_{\alpha c}}_{\text{TF}} \underbrace{\log\left(\frac{1}{df_c}\right)}_{\text{IDF}} = tf_{\alpha c} \log\left(\frac{N}{N_c}\right) .$$

TF-IDF and similarity

Edge weight

$$w_{\alpha\beta} = \frac{\vec{u_\alpha} \cdot \vec{u_\beta}}{\left\|\vec{u_\alpha}\right\| \left\|\vec{u_\beta}\right\|},$$

$$w_{\alpha\beta} \in [0, 1],$$

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
β	2	43	0	18	0	11	27
α	13	5	9	0	30	0	0

TF-IDF and similarity

Edge weight

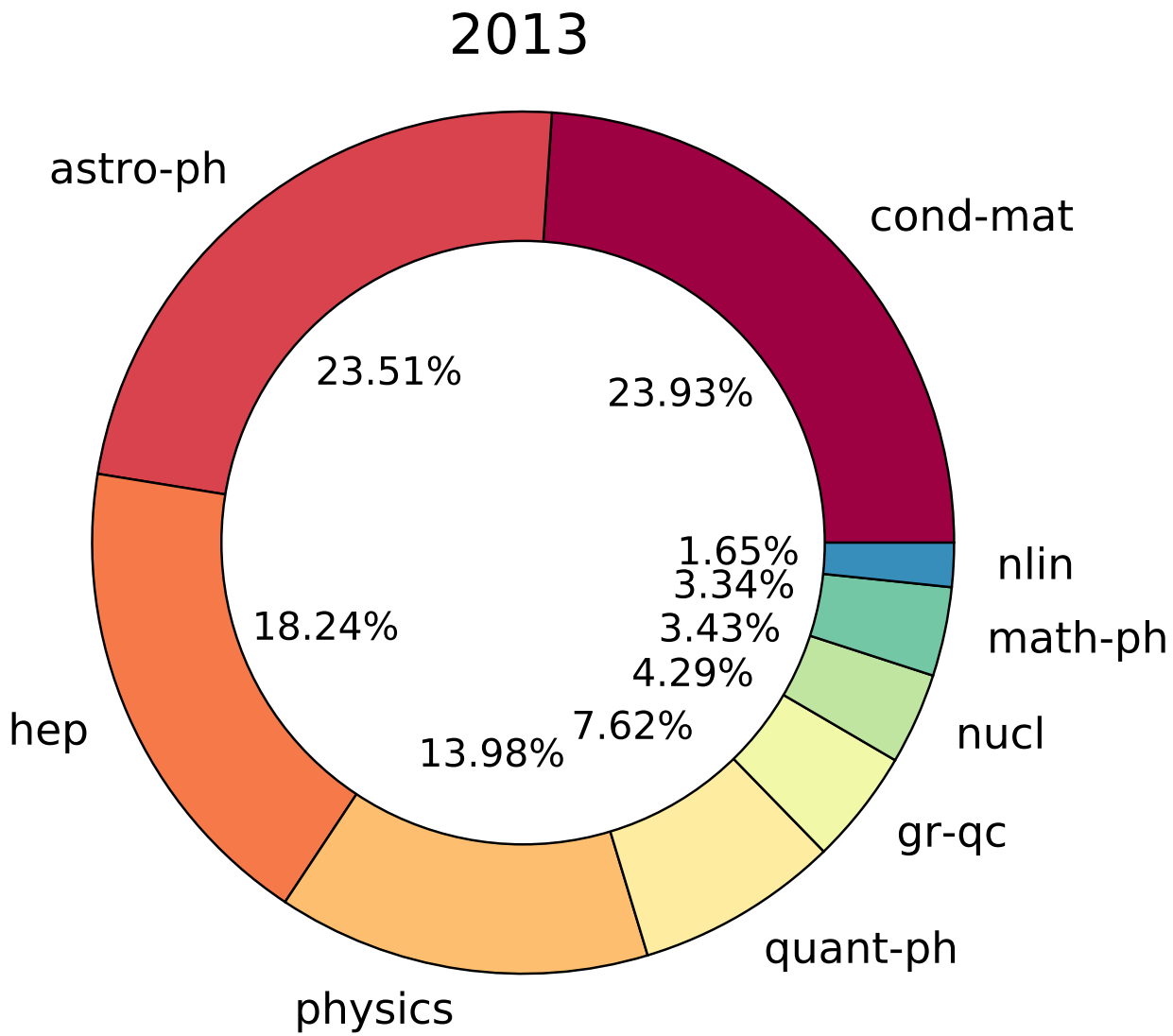
$$w_{\alpha\beta} = \frac{\vec{u_\alpha} \cdot \vec{u_\beta}}{\|\vec{u_\alpha}\| \, \|\vec{u_\beta}\|} ,$$

$$w_{\alpha\beta} \in [0, 1] ,$$

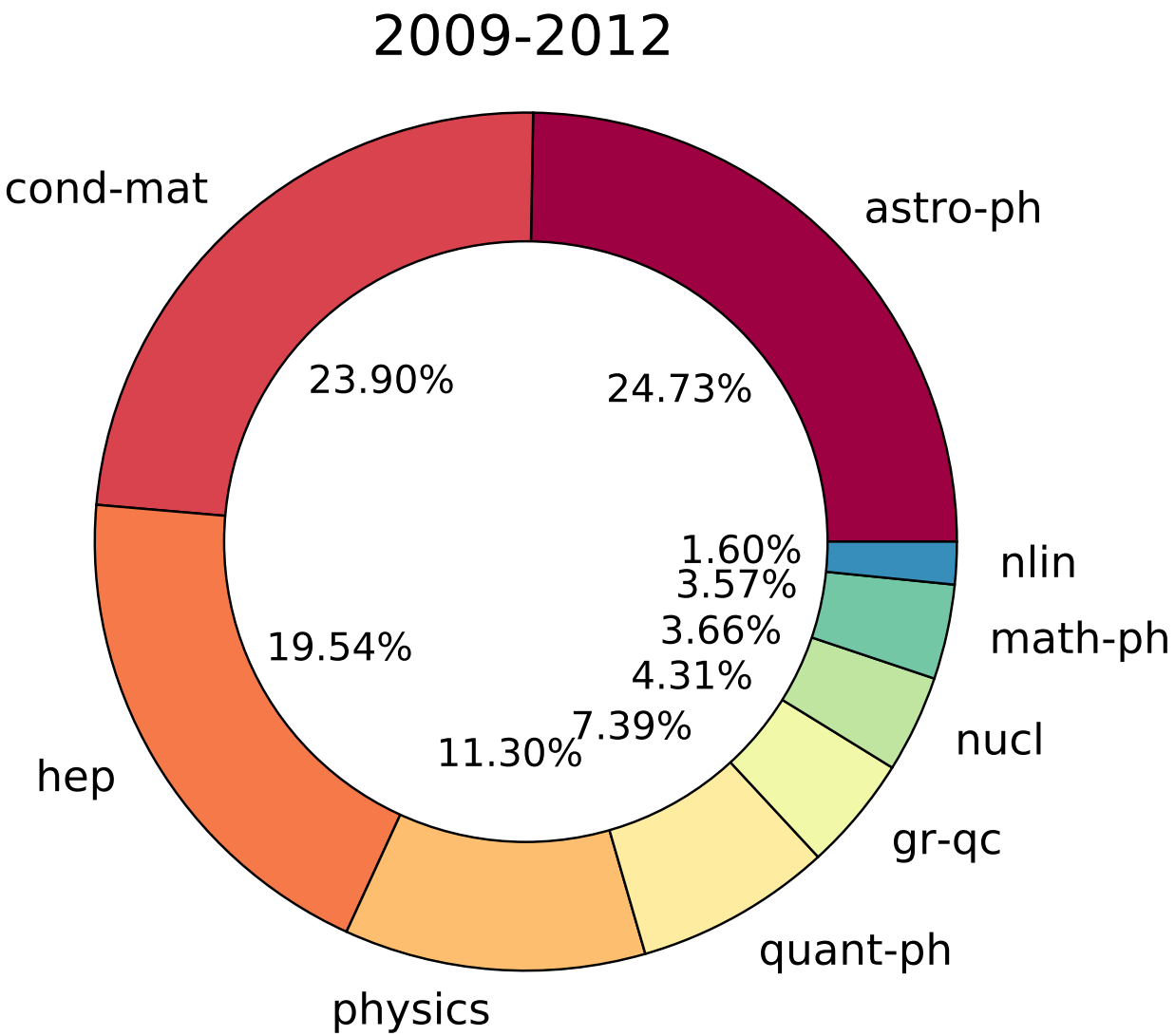
$$w_{\alpha\beta} = \frac{(13 \times 2) + (43 \times 5)}{55.02 \times 34.28} = \\ = \frac{241}{1886.09} \simeq 0.13 .$$

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
β	2	43	0	18	0	11	27
α	13	5	9	0	30	0	0

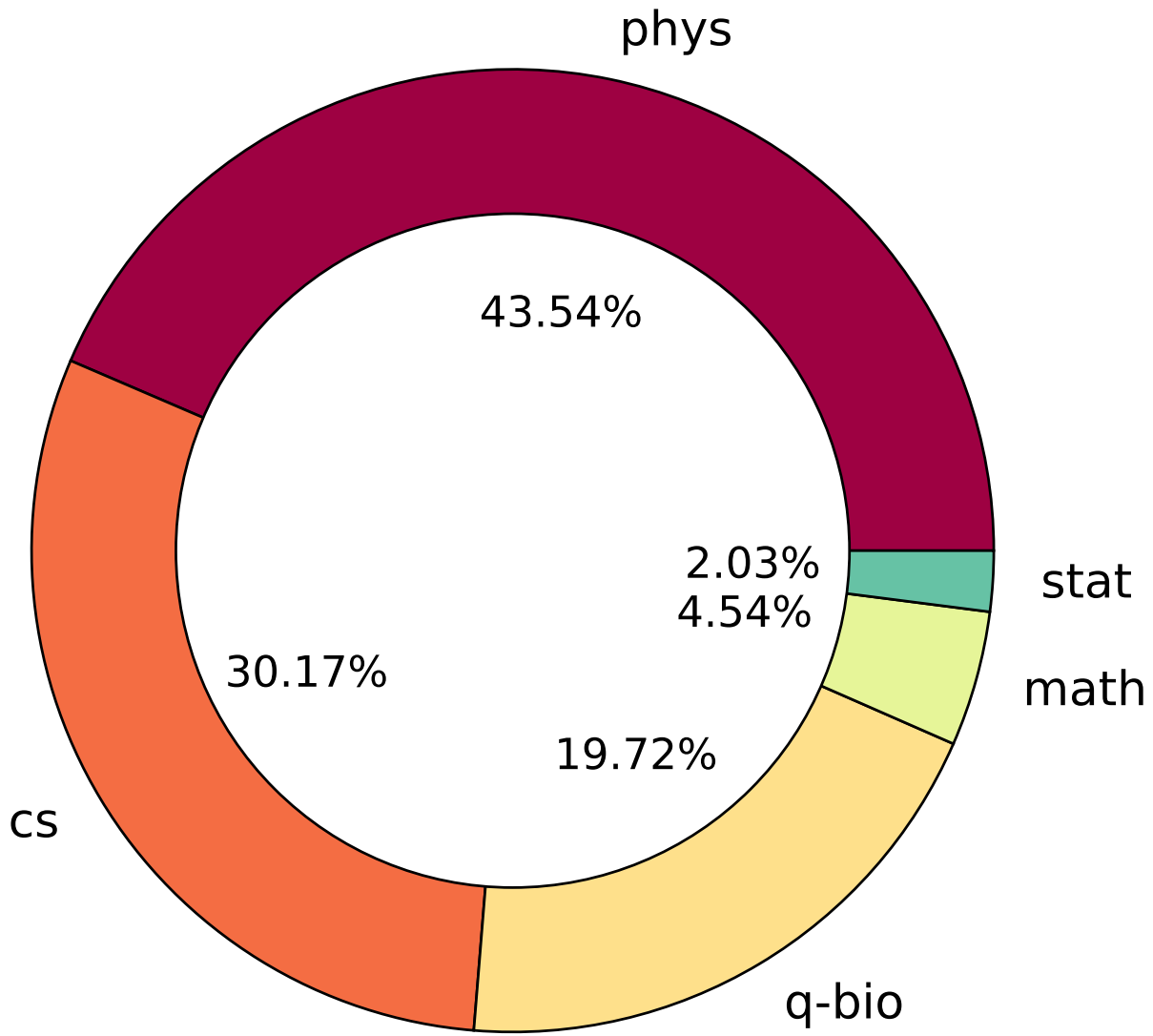
arXiv Collections



arXiv Collections



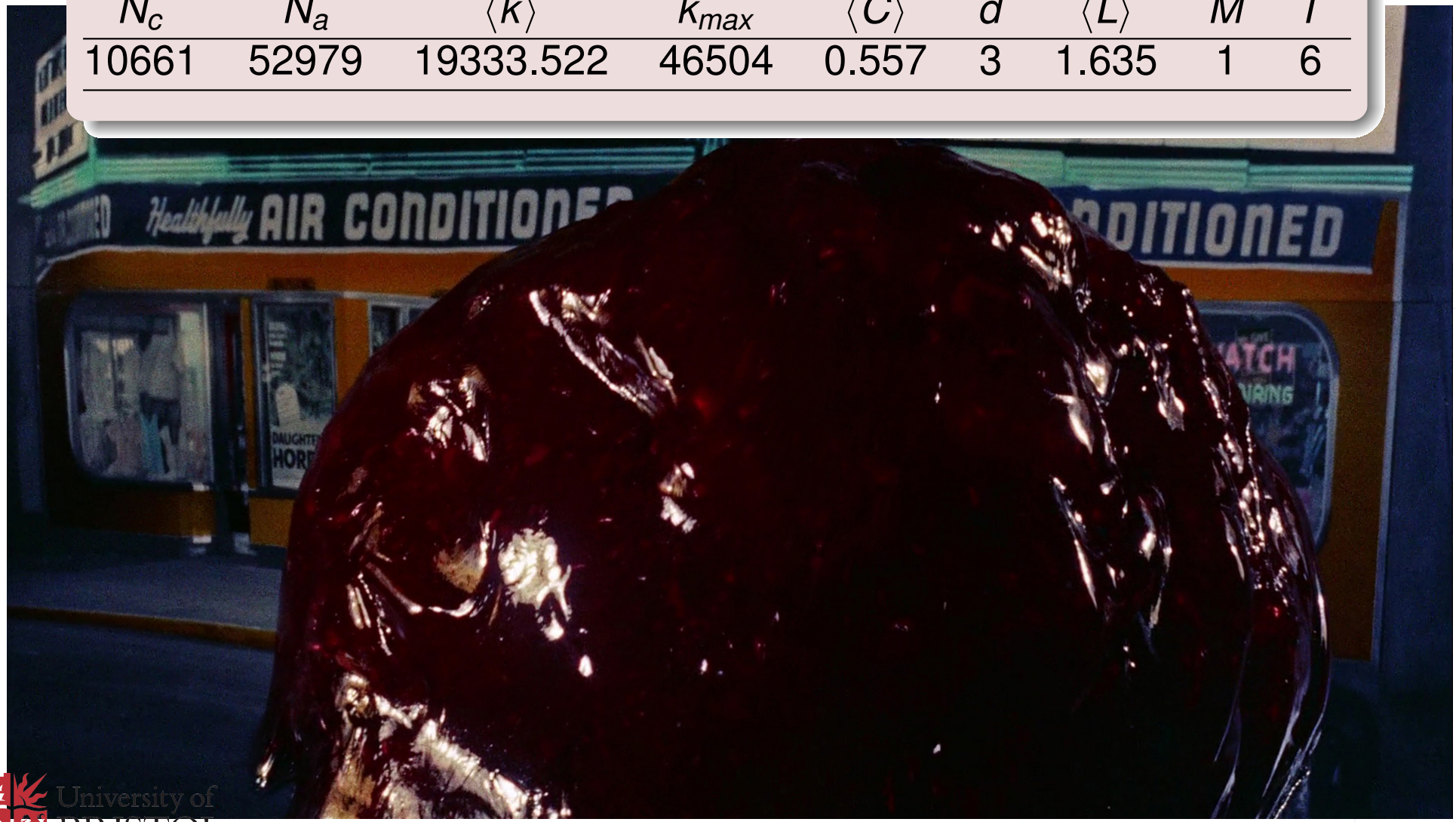
arXiv Collections



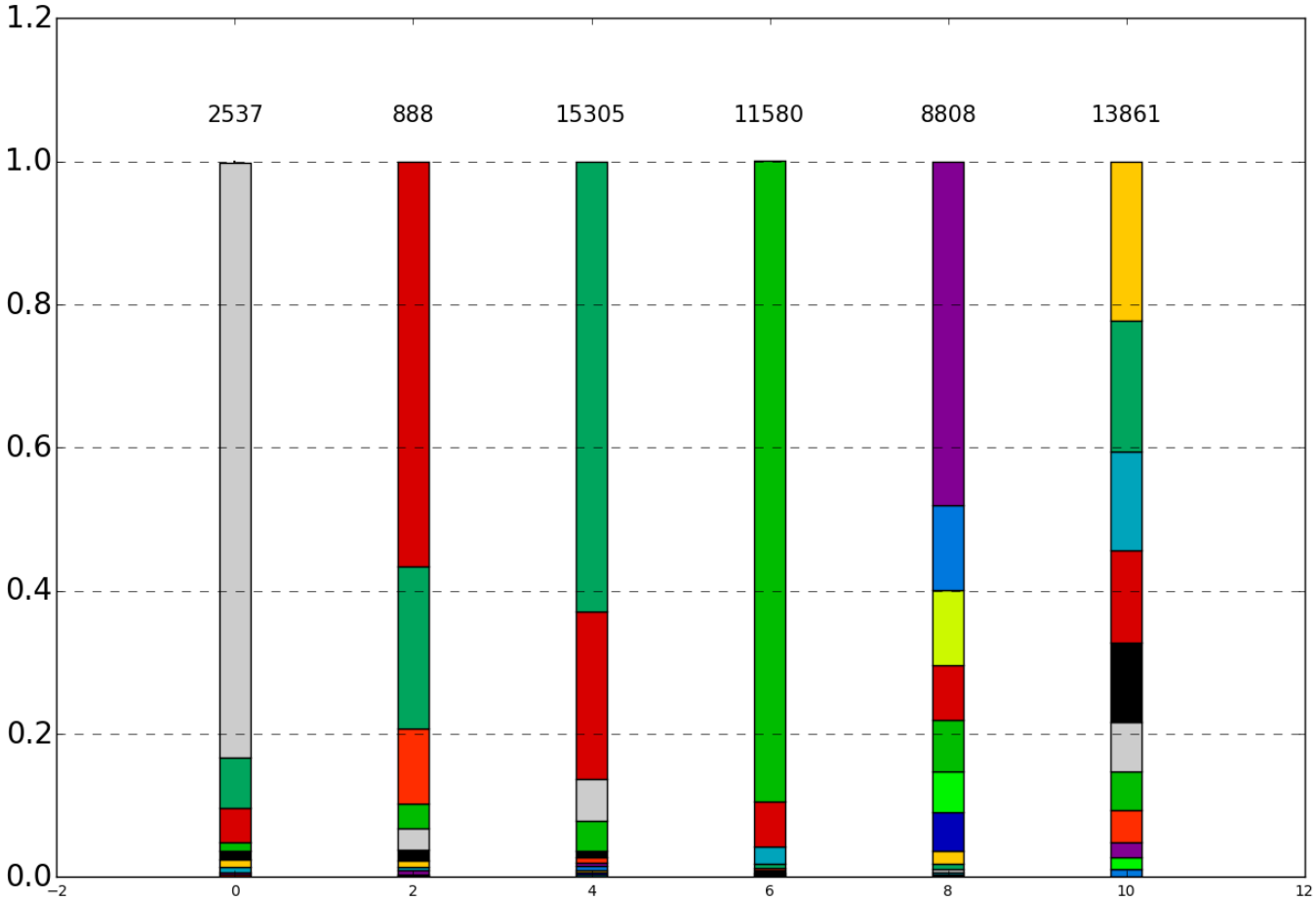
The data: 2013 Physics arXiv

Network properties

N_c	N_a	$\langle k \rangle$	k_{max}	$\langle C \rangle$	d	$\langle L \rangle$	M	T
10661	52979	19333.522	46504	0.557	3	1.635	1	6

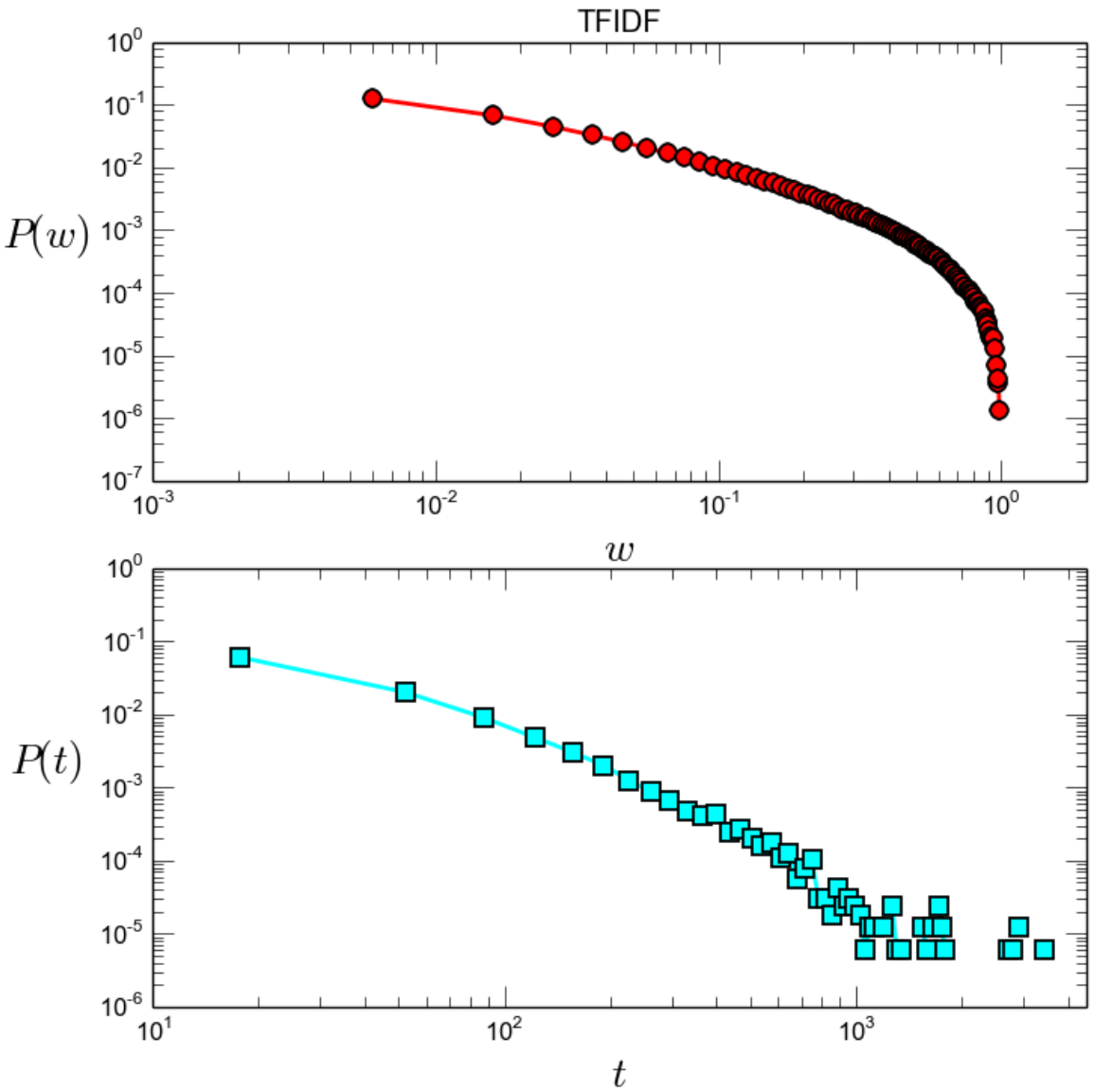


The data: 2013 Physics arXiv



$$\rho = \frac{K}{K_{\max}} \simeq 36\%$$

Edge pruning/sparsification methods



Edge pruning/sparsification methods



- Serrano M.A., *et al.* *Extracting the multiscale backbone of complex weighted networks*. Proc. Natl. Acad. Sci. (USA) **106** 6483 (2009).

Edge pruning/sparsification methods



PHYSICAL REVIEW E
statistical, nonlinear, and soft matter physics

Highlights Recent Accepted Authors Referees Search About

Information filtering in complex weighted networks

Filippo Radicchi, José J. Ramasco, and Santo Fortunato
Phys. Rev. E **83**, 046101 – Published 1 April 2011

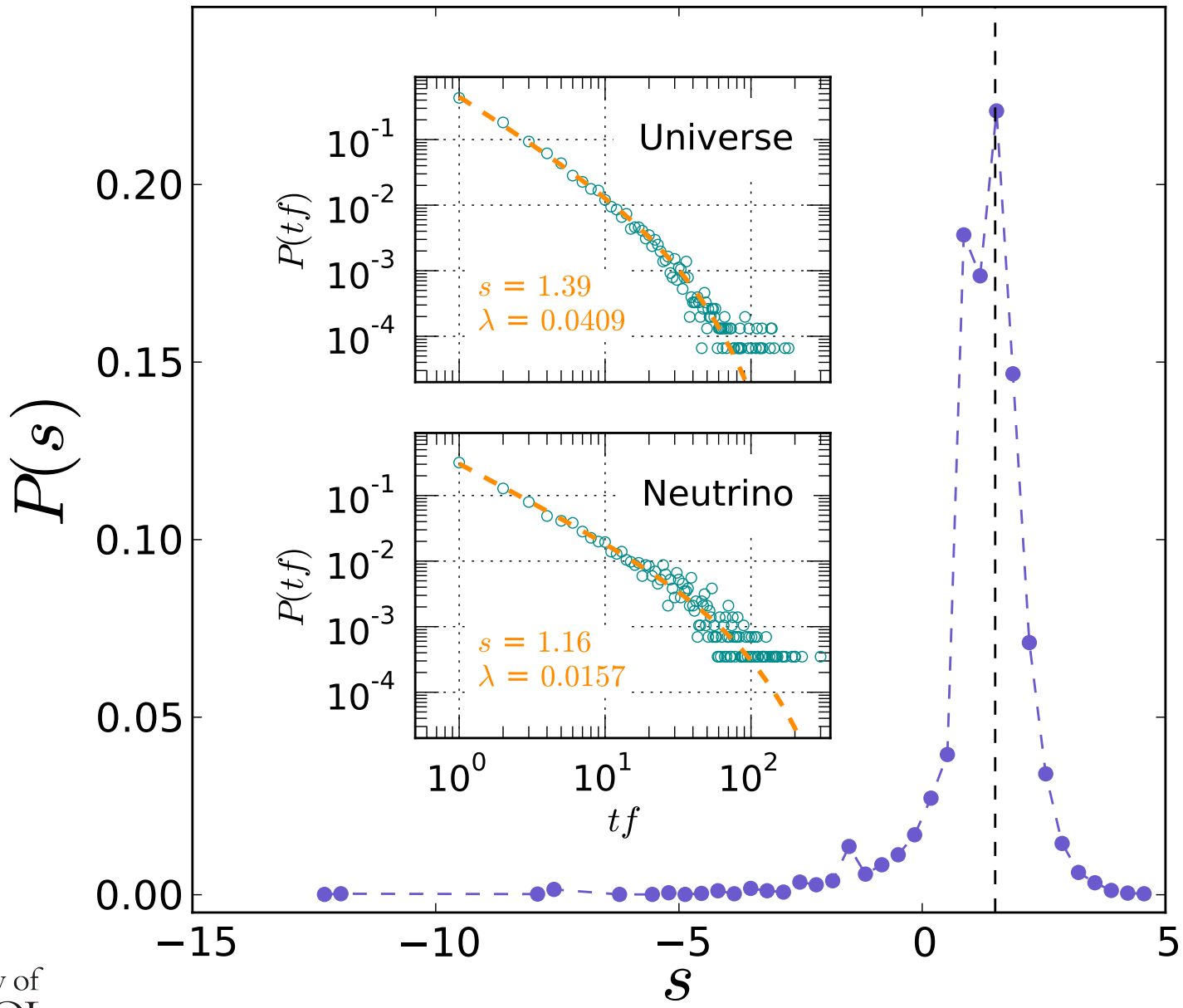
Article References Citing Articles (8) PDF HTML Export Citation

ABSTRACT

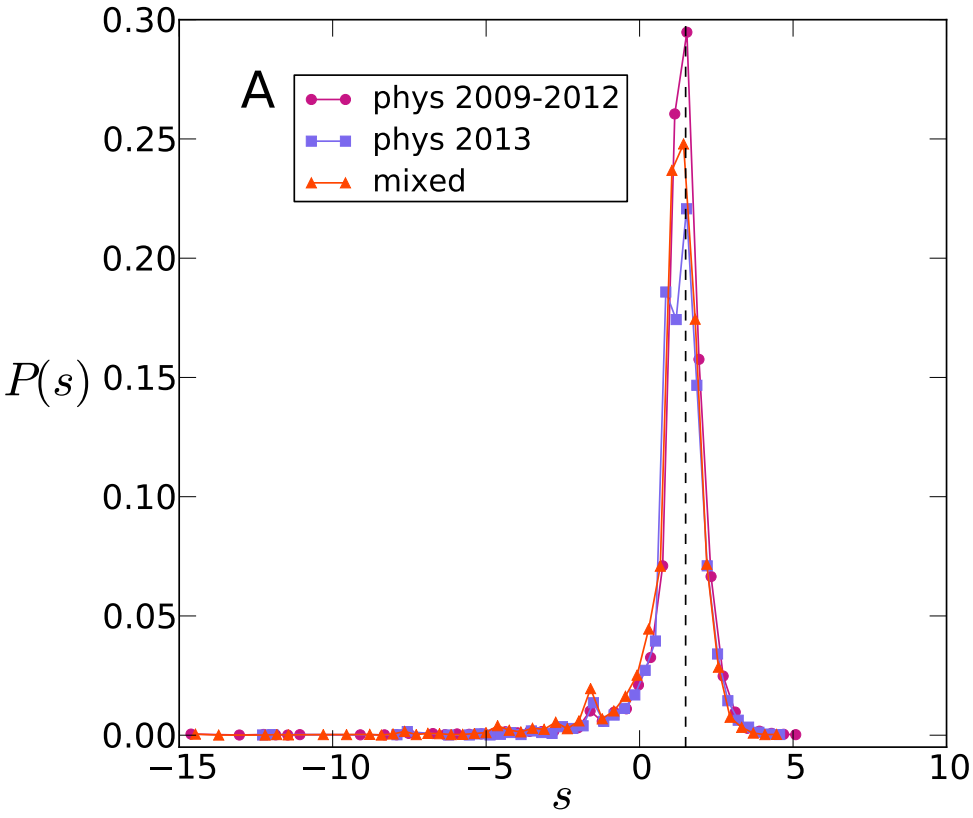
Many systems in nature, society, and technology can be described as networks, where the vertices are the system's elements, and edges between vertices indicate the interactions between the corresponding elements. Edges may be weighted if the interaction strength is measurable. However, the full network information is often redundant because tools and techniques from network analysis

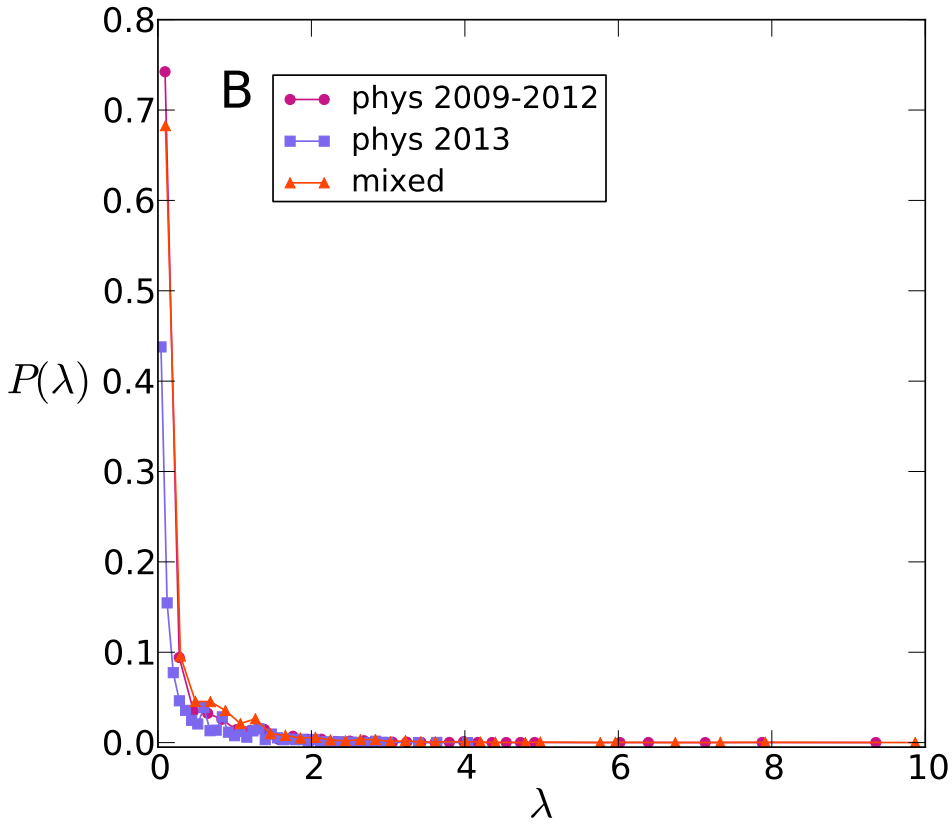
- Radicchi, F., *et al. Information filtering in complex weighted networks*. Physical Review E, **83** 046101. (2011).

Maximum entropy – why power law?

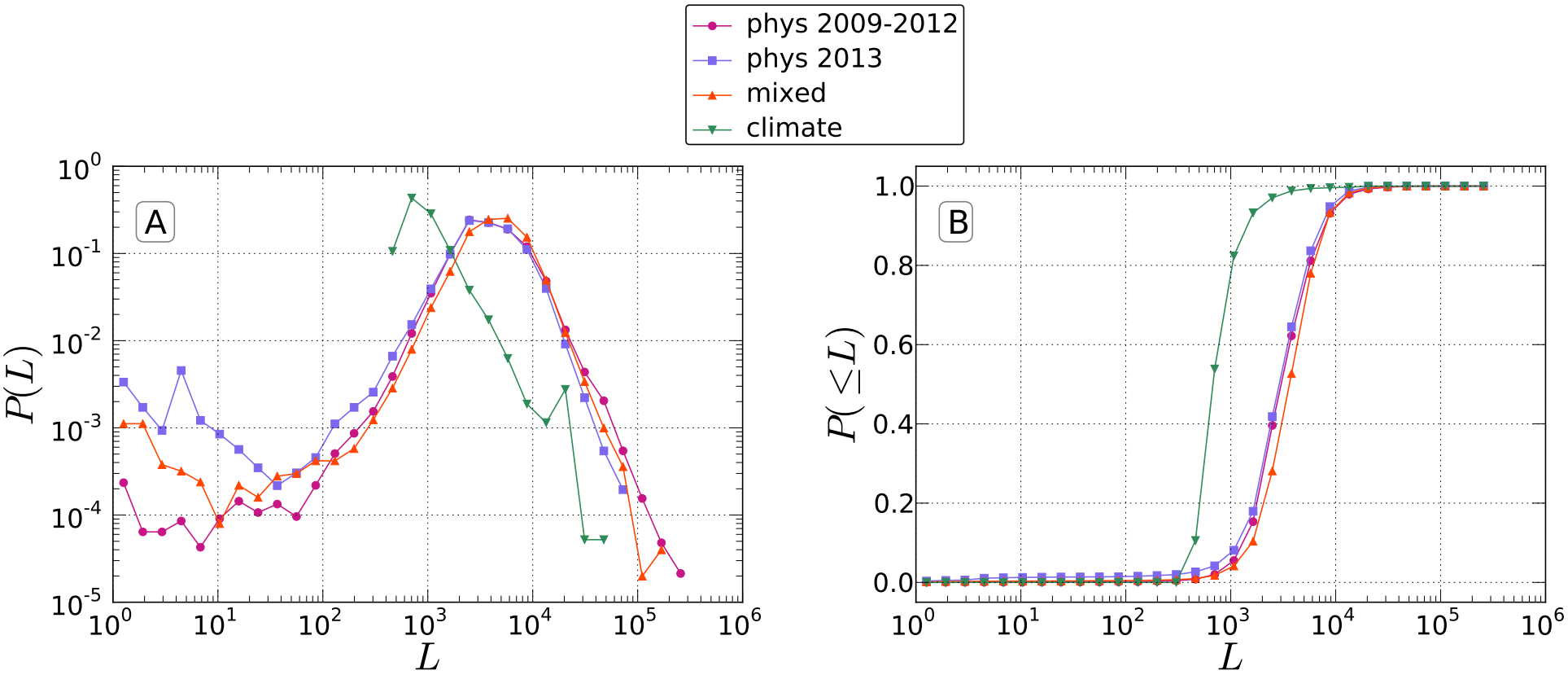


Maximum entropy – why power law?





Maximum entropy – TF density

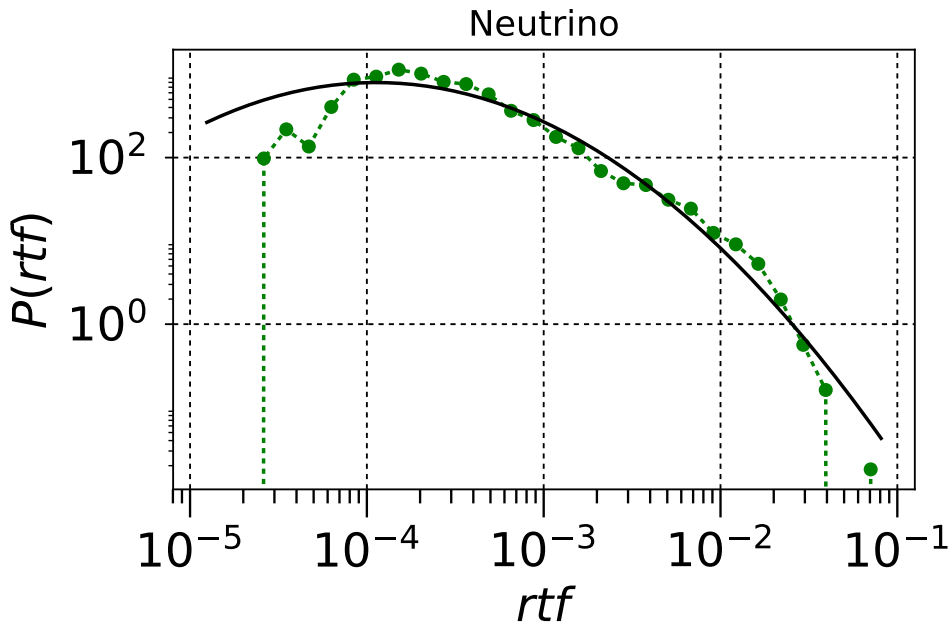


Maximum entropy – TF density

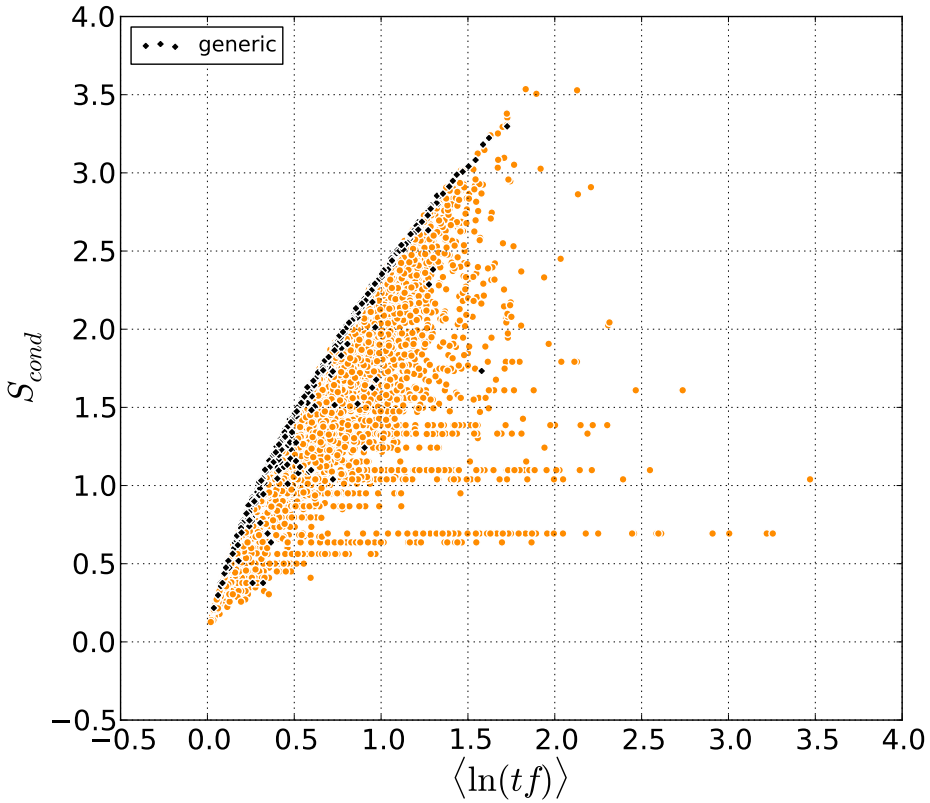
$$\sum_n p_n = 1$$

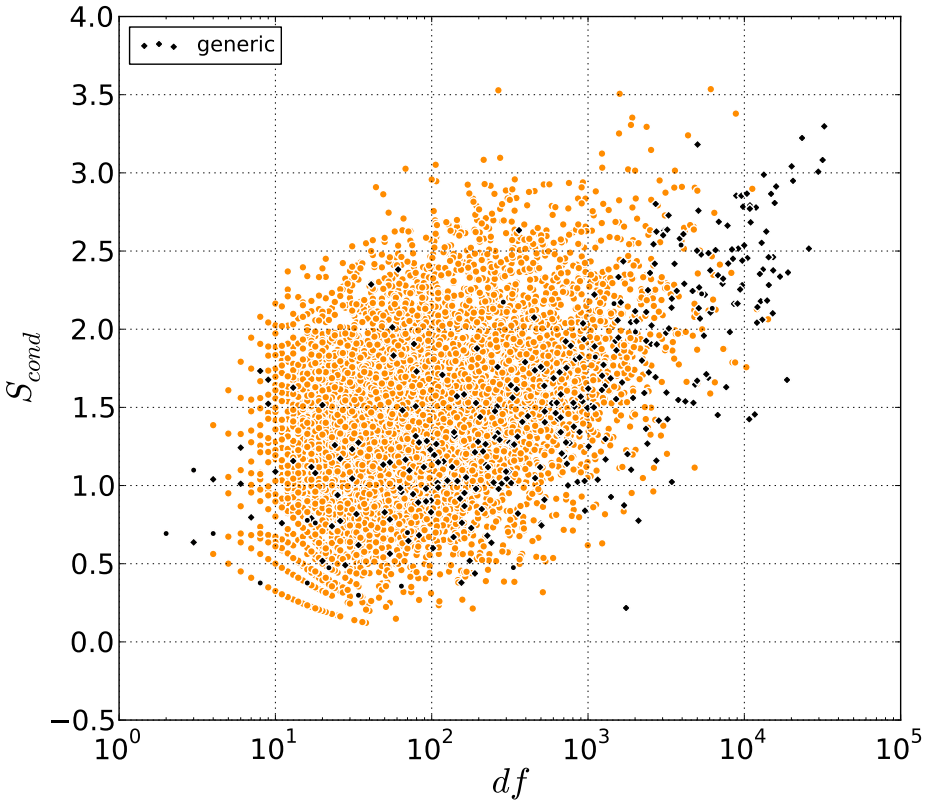
$$\sum_n p_n \ln n = \langle \ln n \rangle$$

$$\sqrt{\sum_n p_n (\ln n - \langle \ln n \rangle)^2} = \sigma_{\ln n}$$



Why not *df*?

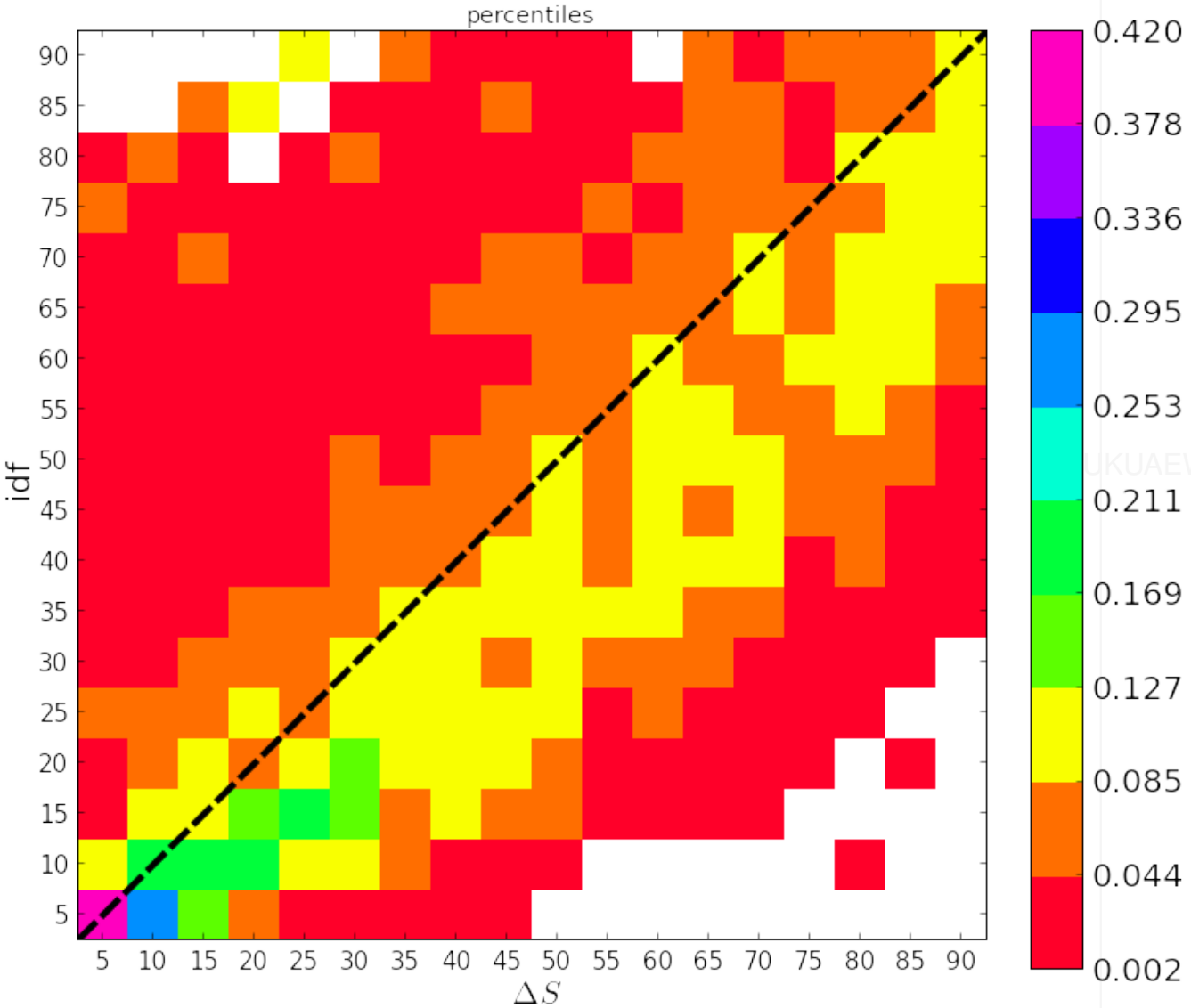




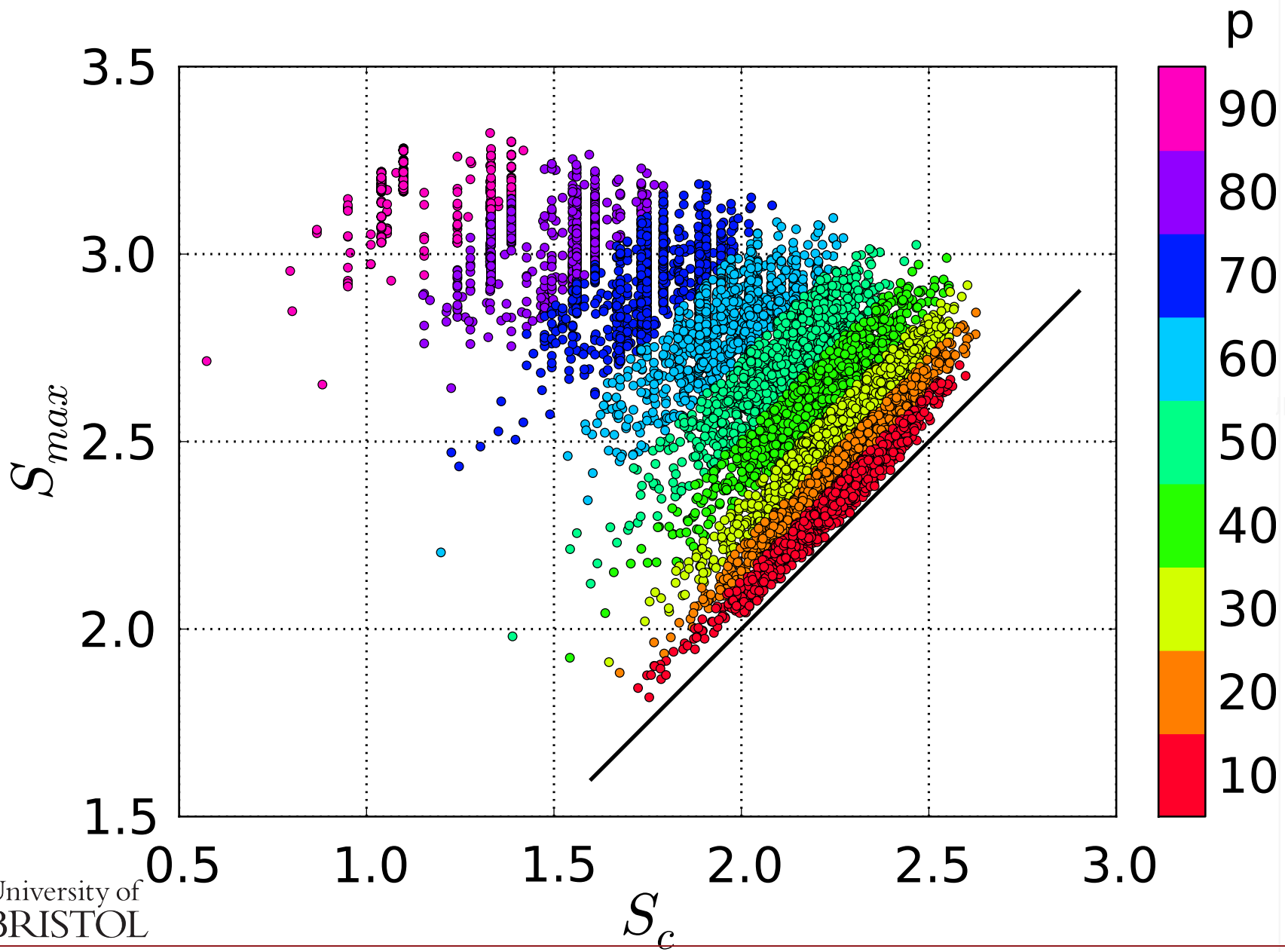
Why not *df*?

Jaccard Score

$$J = \frac{|A \cap B|}{|A \cup B|}.$$



Climate dataset



Climate dataset

