Automatic identification of relevant concepts in scientific publications

Alessio Cardillo
Laboratory for Statistical Biophysics (LBS)
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

http://bifi.es/~cardillo/

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

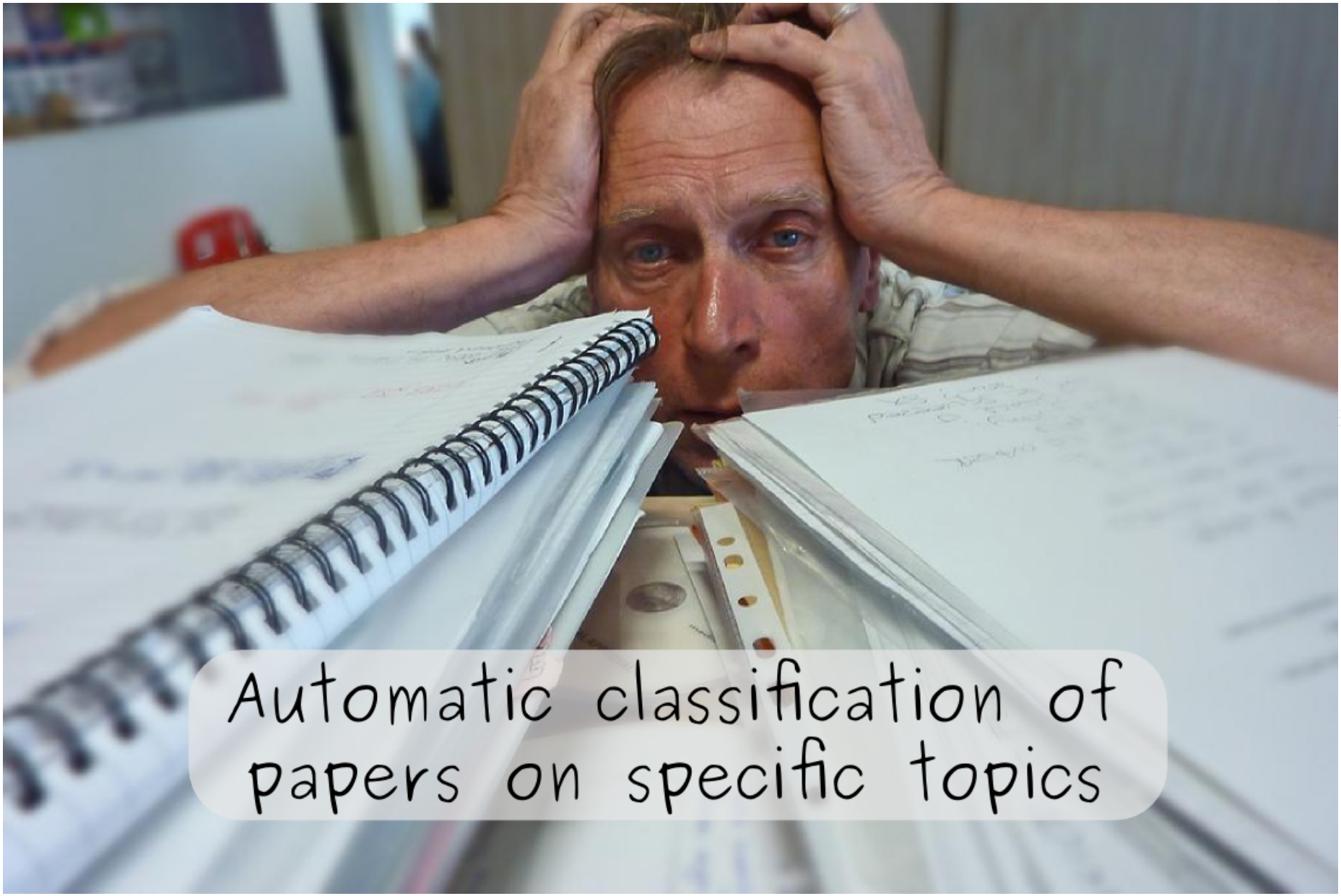
Next post
German research agencies condemn animal-rights attack on neuroscientist

NEWS BLOG

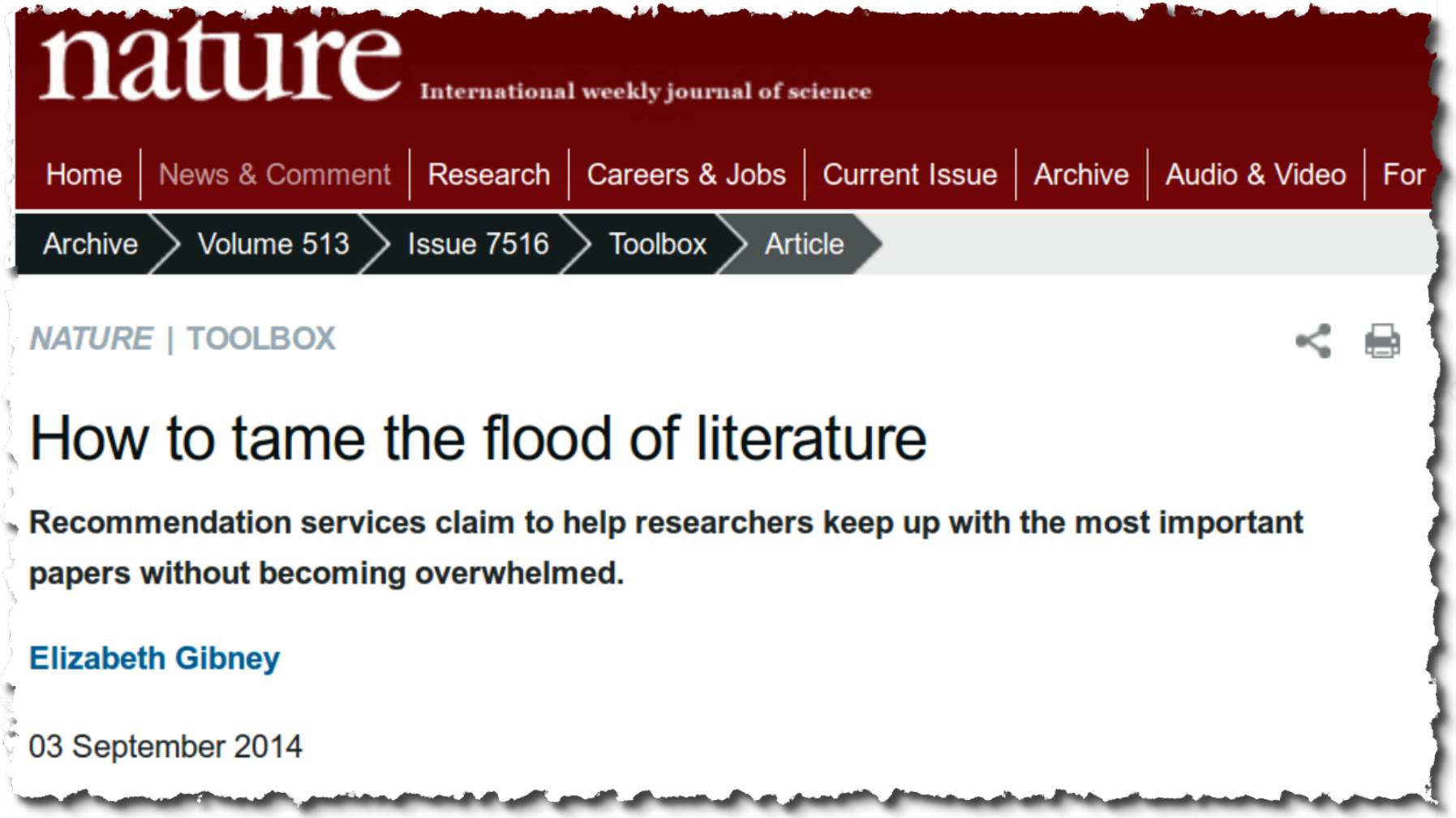
Global scientific output doubles every nine years

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?

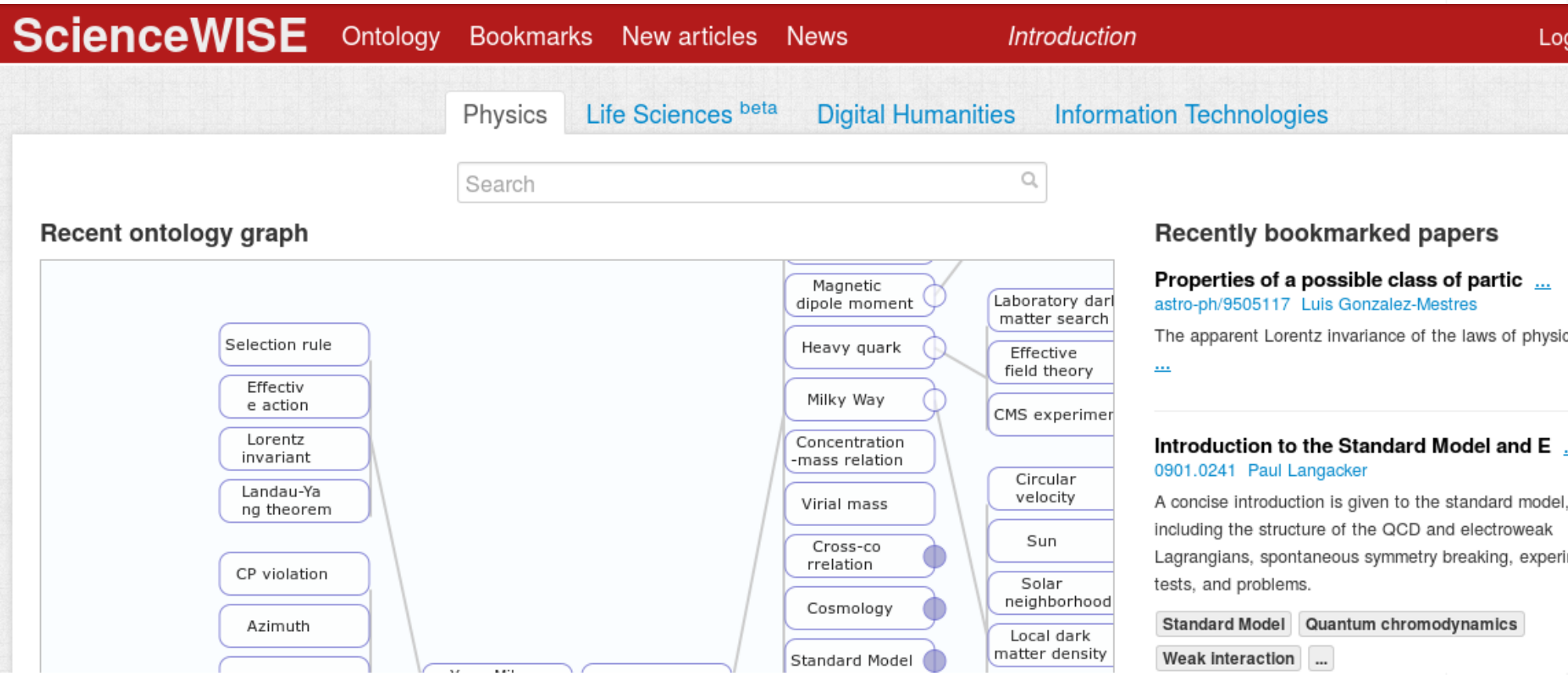
There is an inherent problem to giving you information that you weren't actively searching for. ***It has to be relevant*** *– so that we are not wasting your time –* ***but not too relevant****, because you already know about those articles.*

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

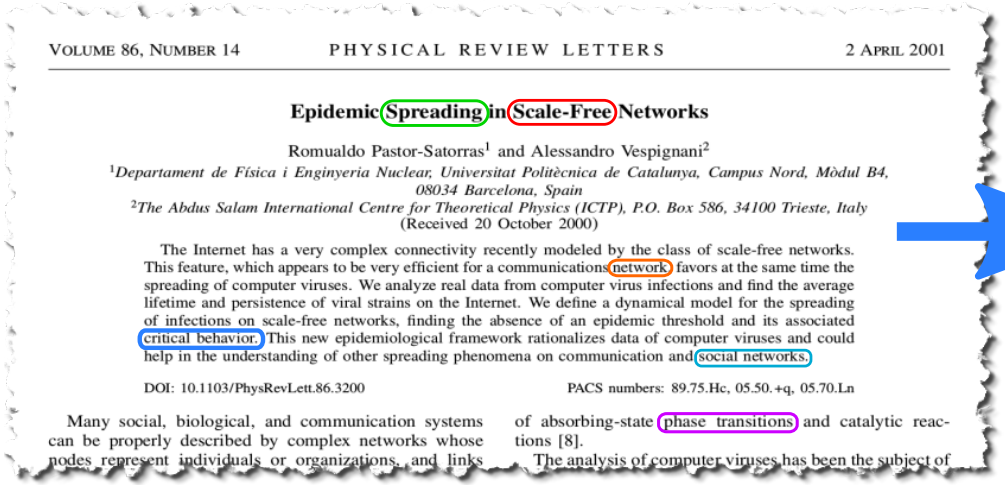
Outline

- Introduction on similarity networks.
- ★ Filtering of weighted networks.
- ★ Entropic filtering of concepts.
- ★ Results with "*Special Effects*".
- Take home messages
- Questions

Section 1

Similarity networks

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

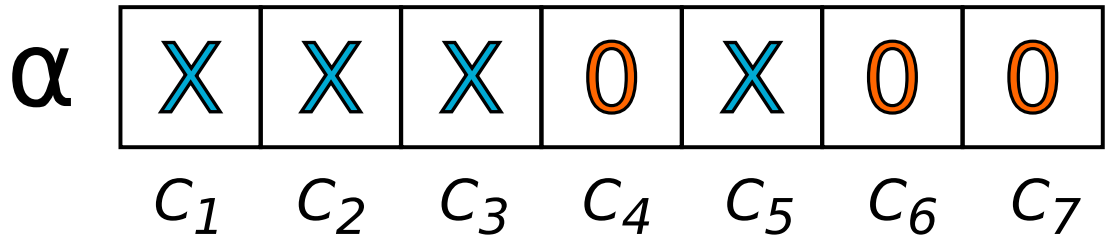
Network

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

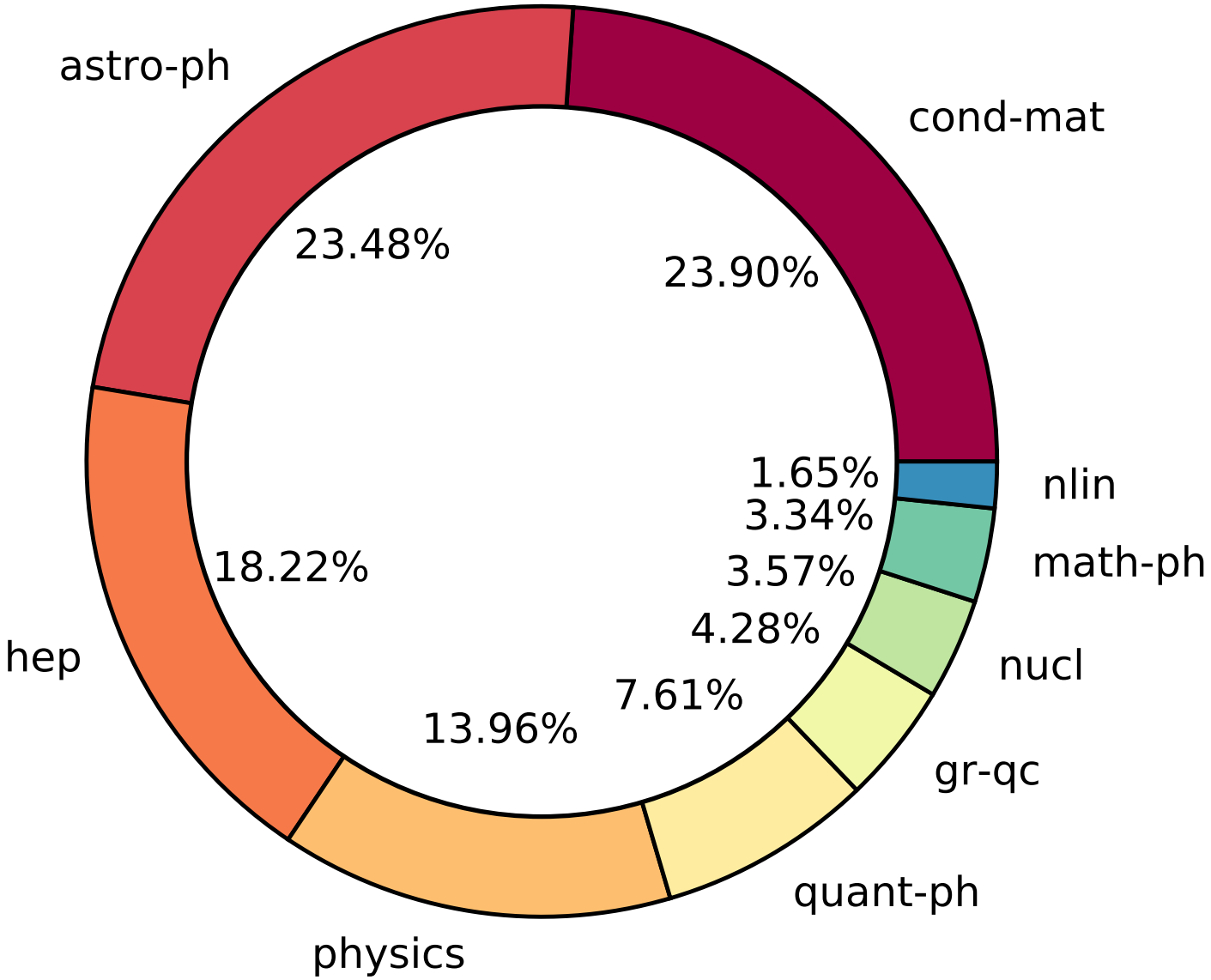
$$w_{\alpha\beta} = \frac{\vec{u_\alpha} \cdot \vec{u_\beta}}{\|\vec{u_\alpha}\| \, \|\vec{u_\beta}\|} \,,$$

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

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

	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

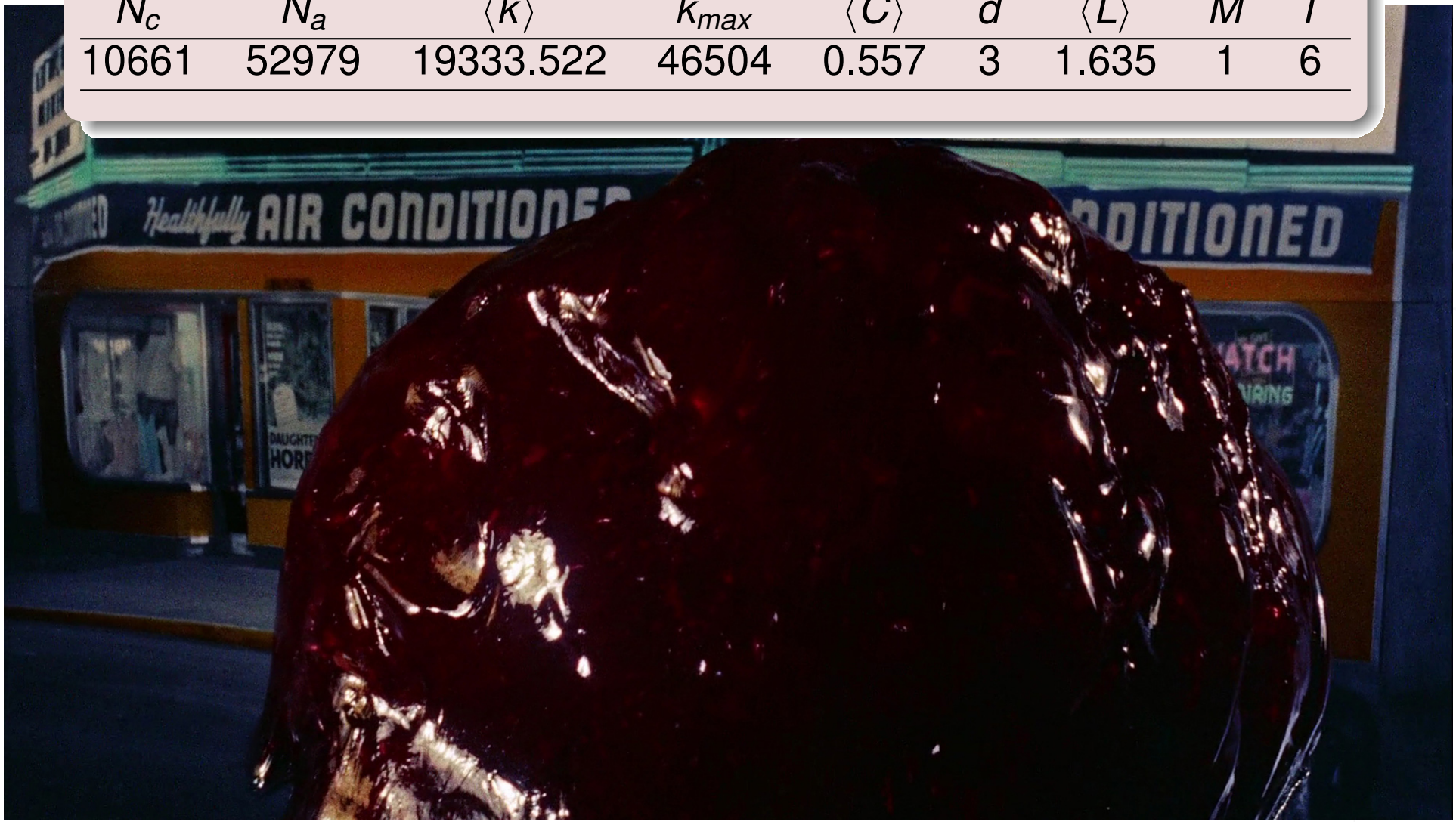
The data: 2013 Physics arXiv



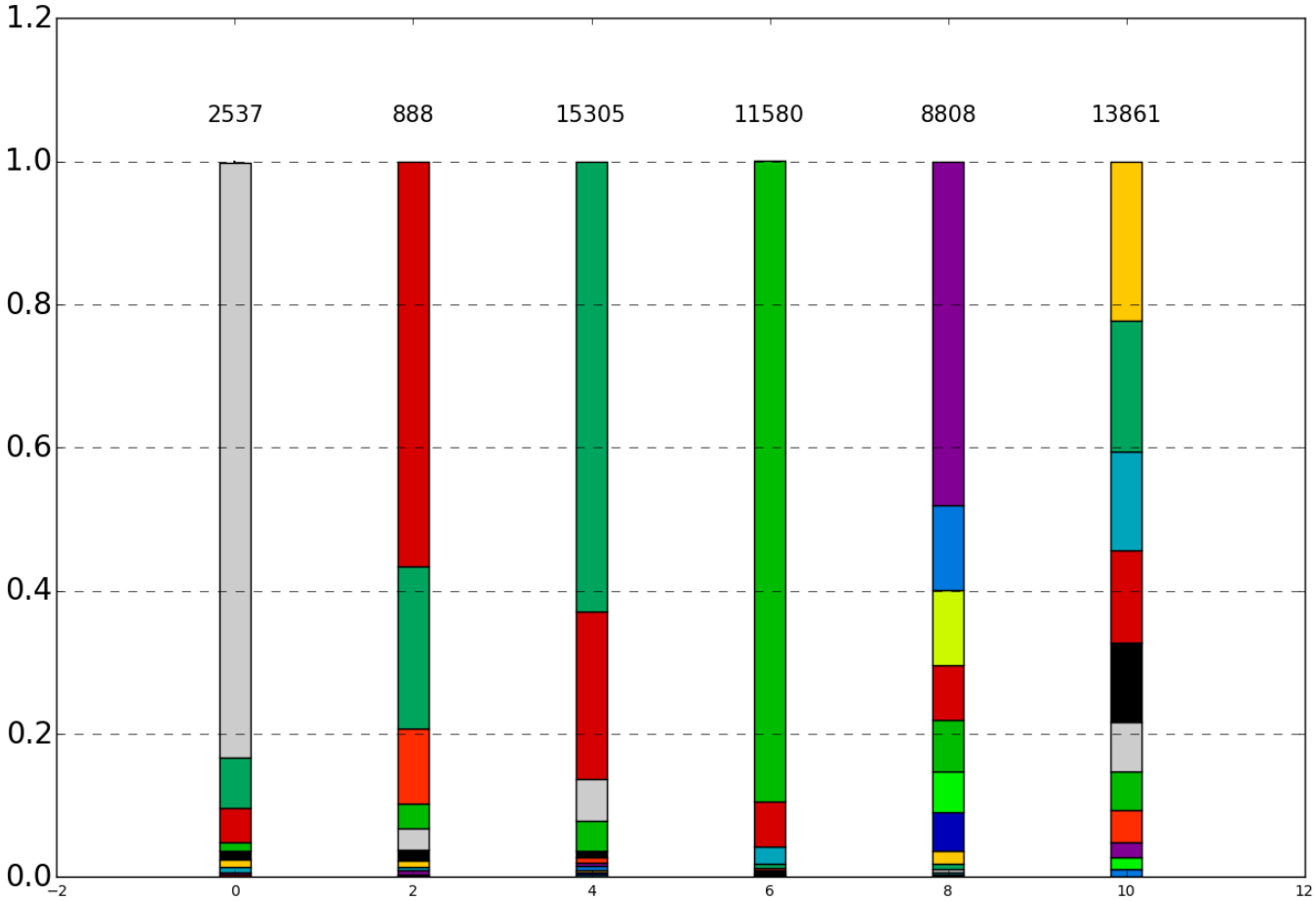
The data: 2013 Physics arXiv

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



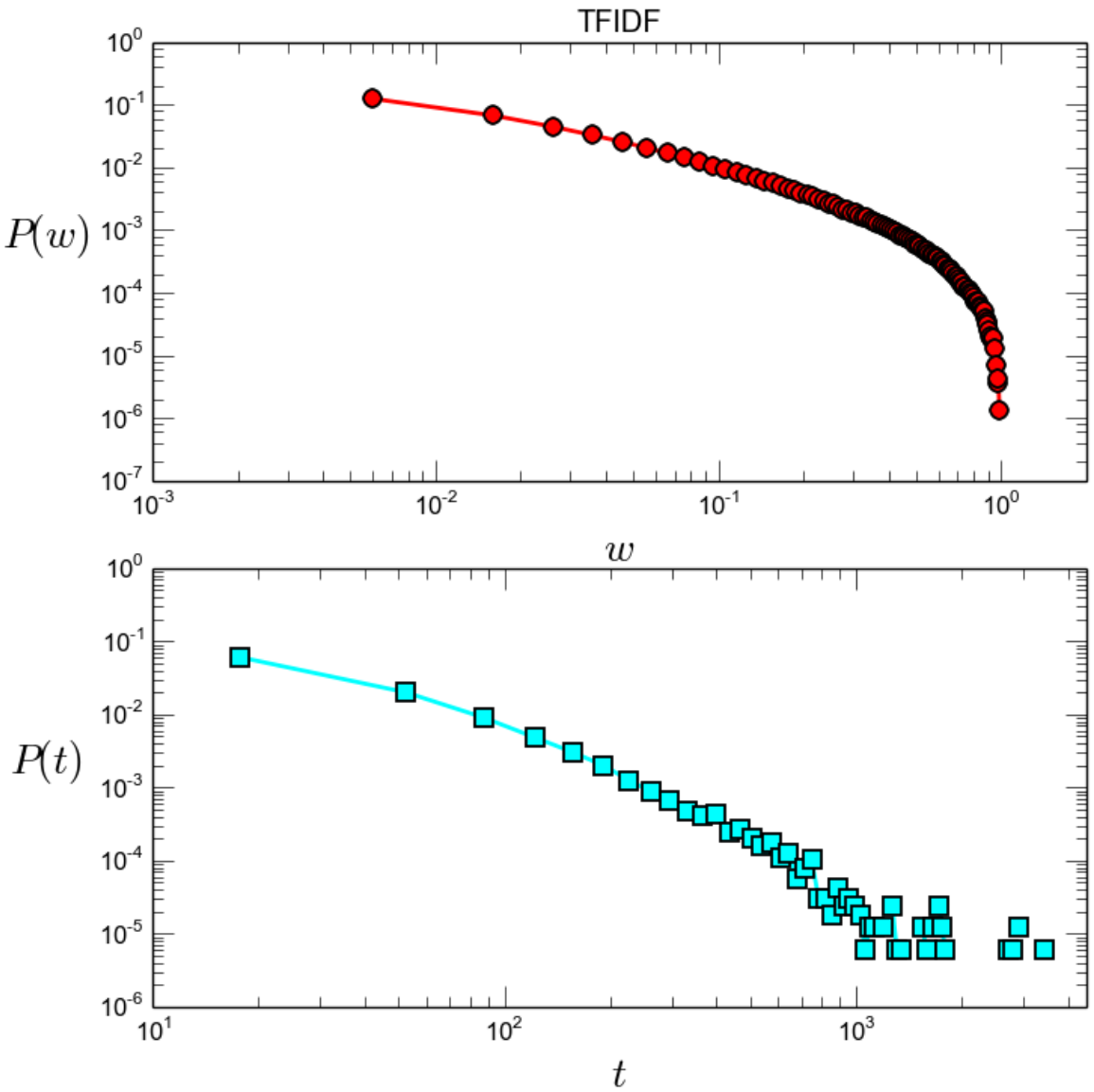
$$\rho = \frac{K}{K_{\max}} \simeq 36\%$$

The data: 2013 Physics arXiv

Section 2

Filtering

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

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

Relevant concepts

Relevant concepts

Key features

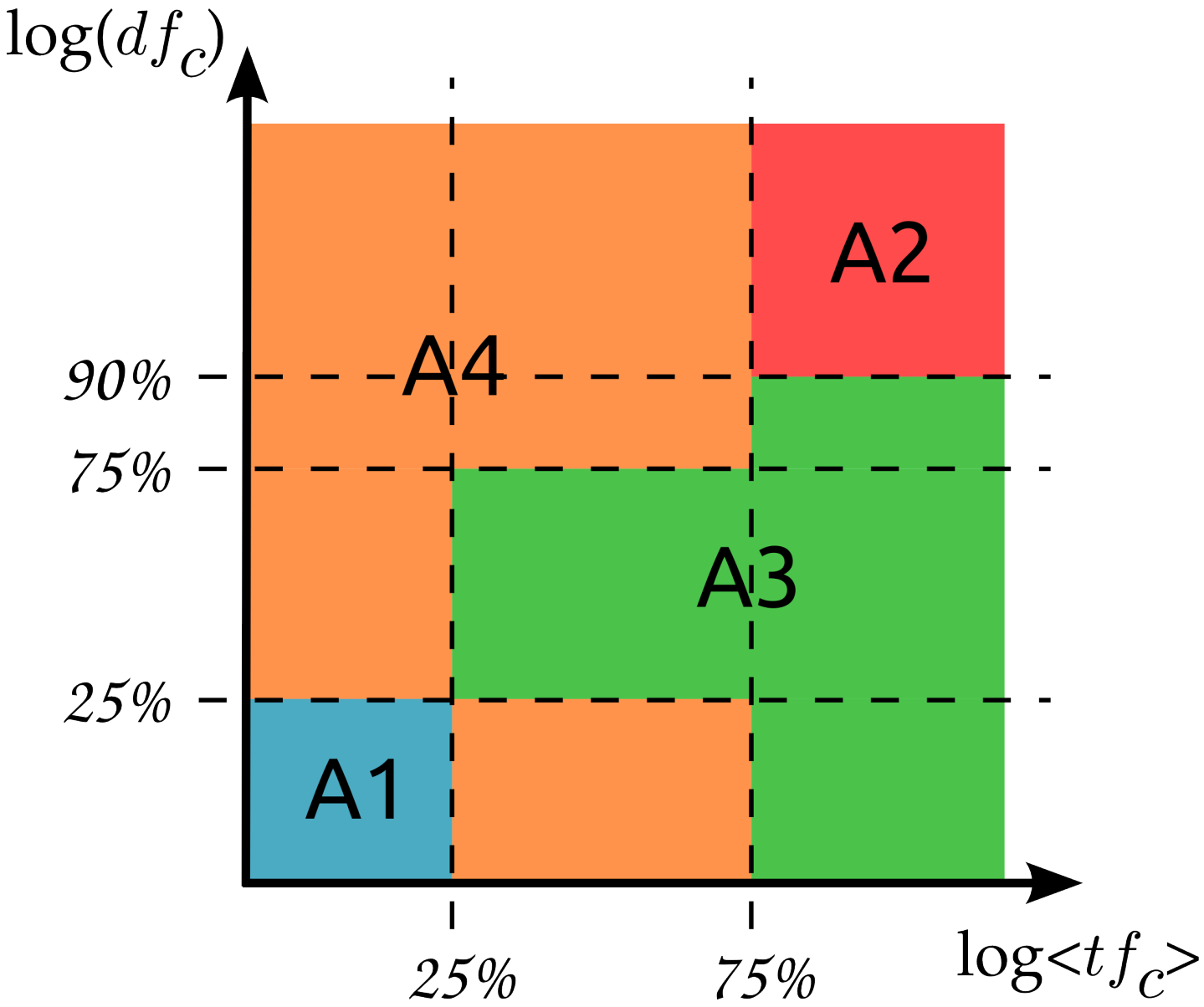
- # of papers a concept appears in

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

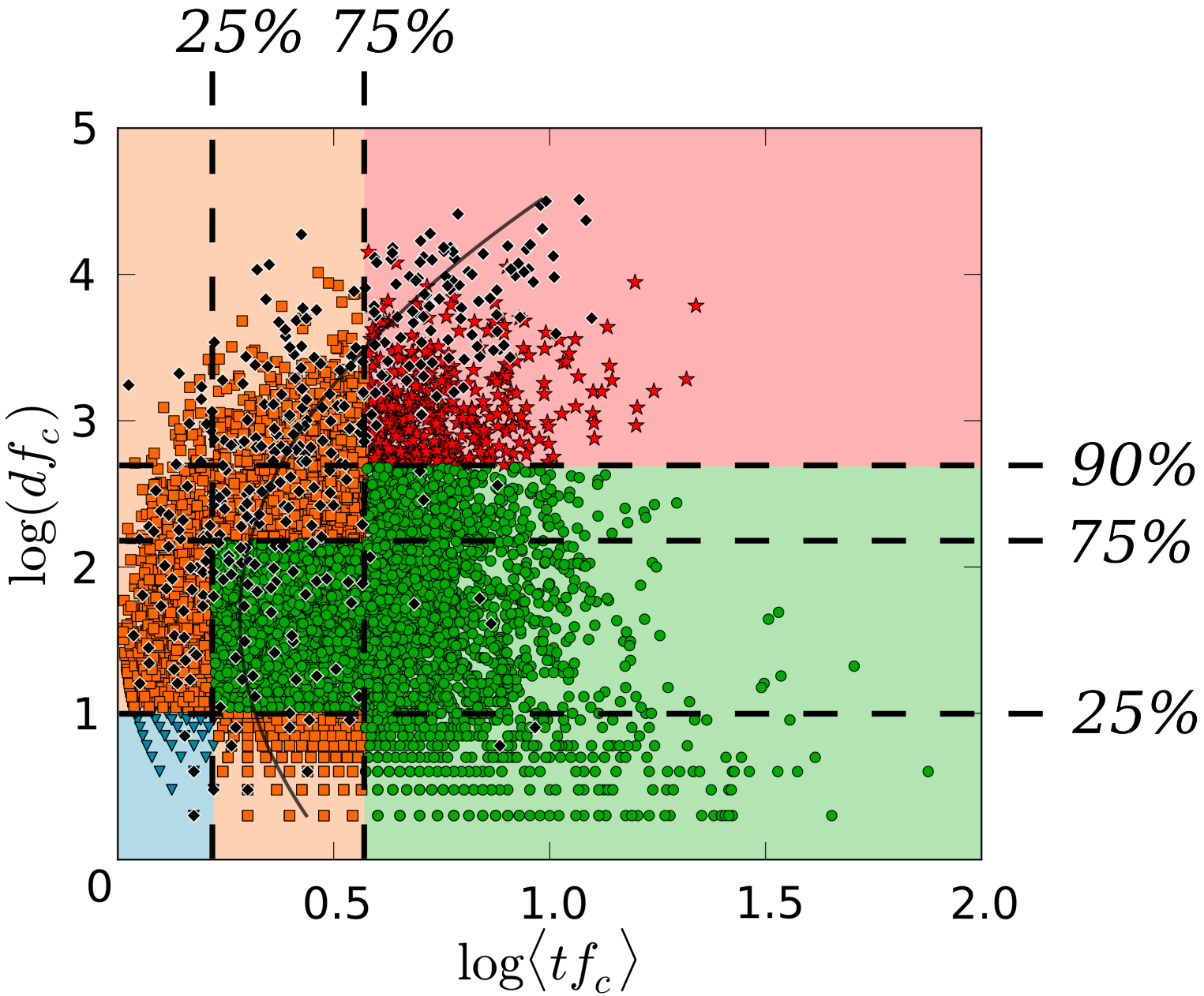
- average # of times a concept appears inside a paper

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

Bidimensional tessellation



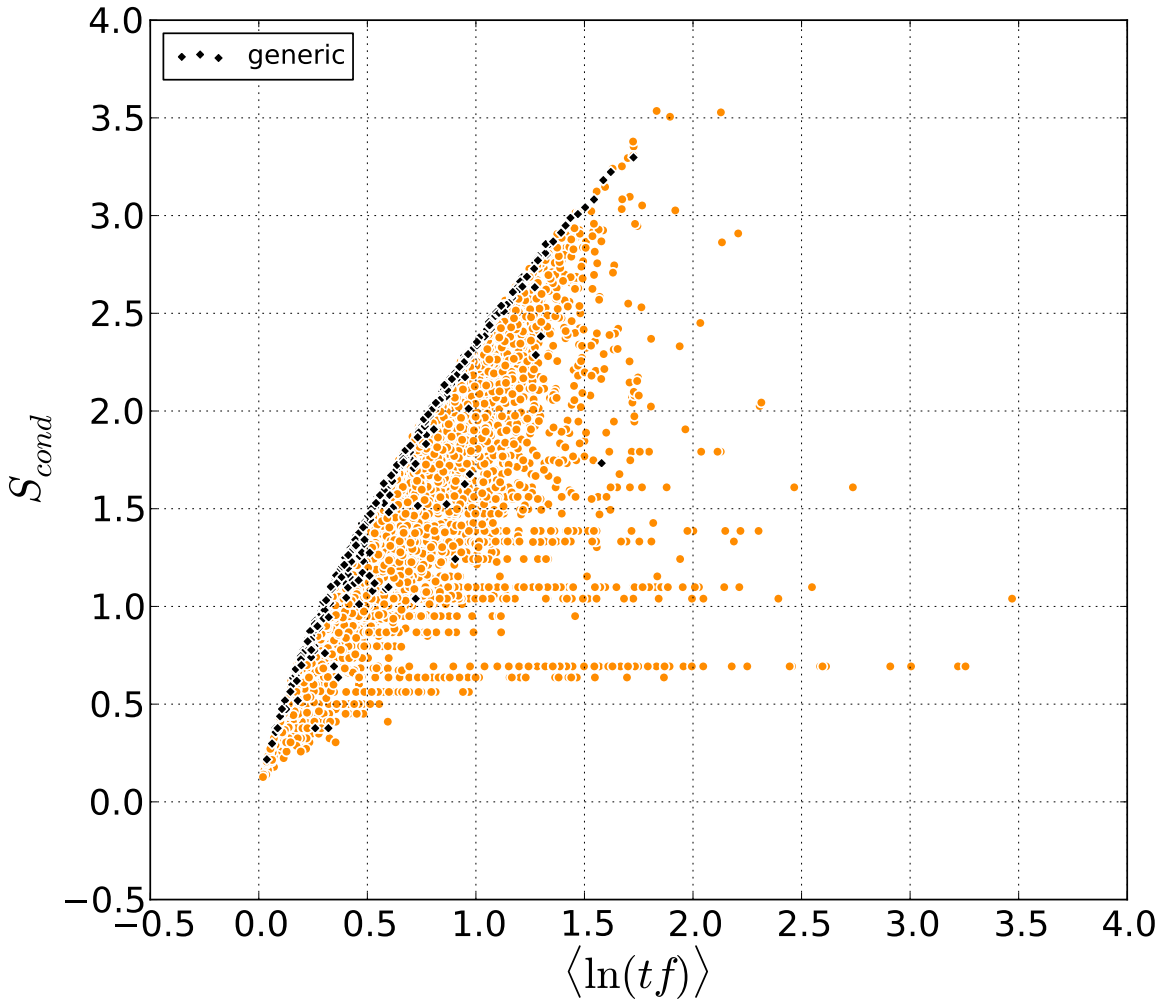
Bidimensional tessellation



Section 3

Entropic Filtering

Maximum entropy



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

Maximum entropy

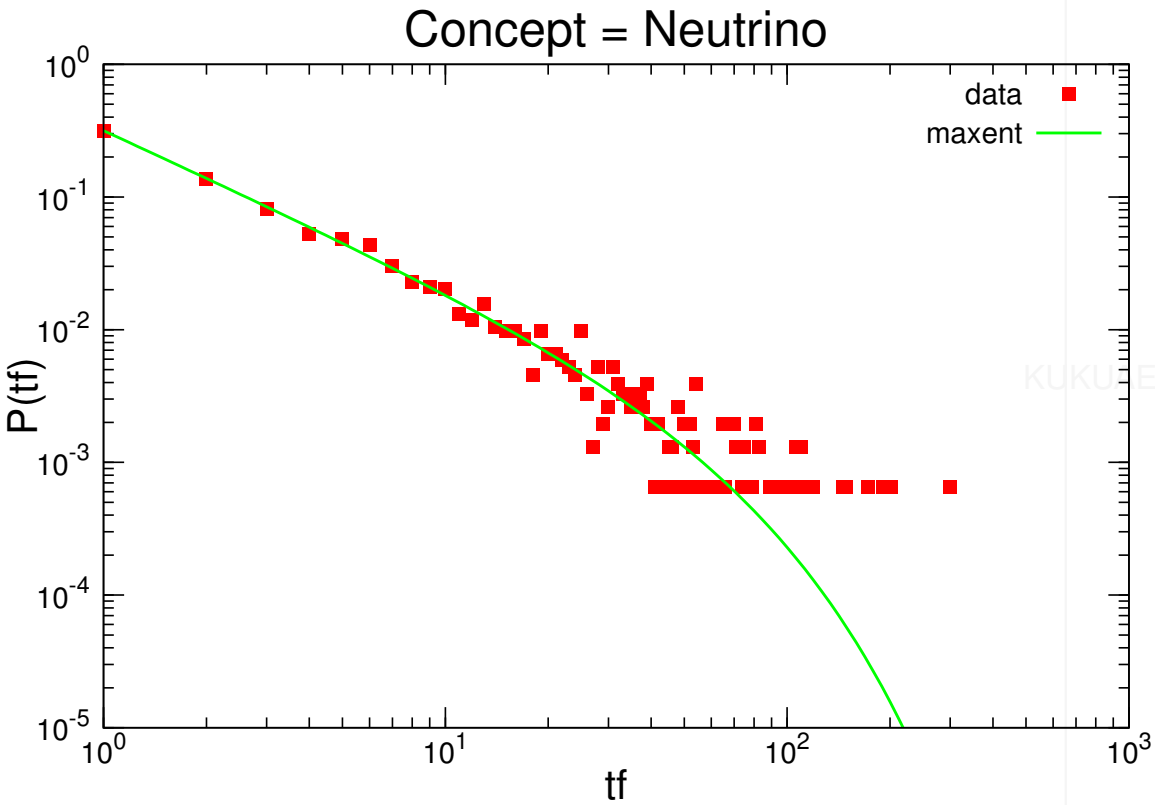
$$\sum_n p_n = 1$$

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

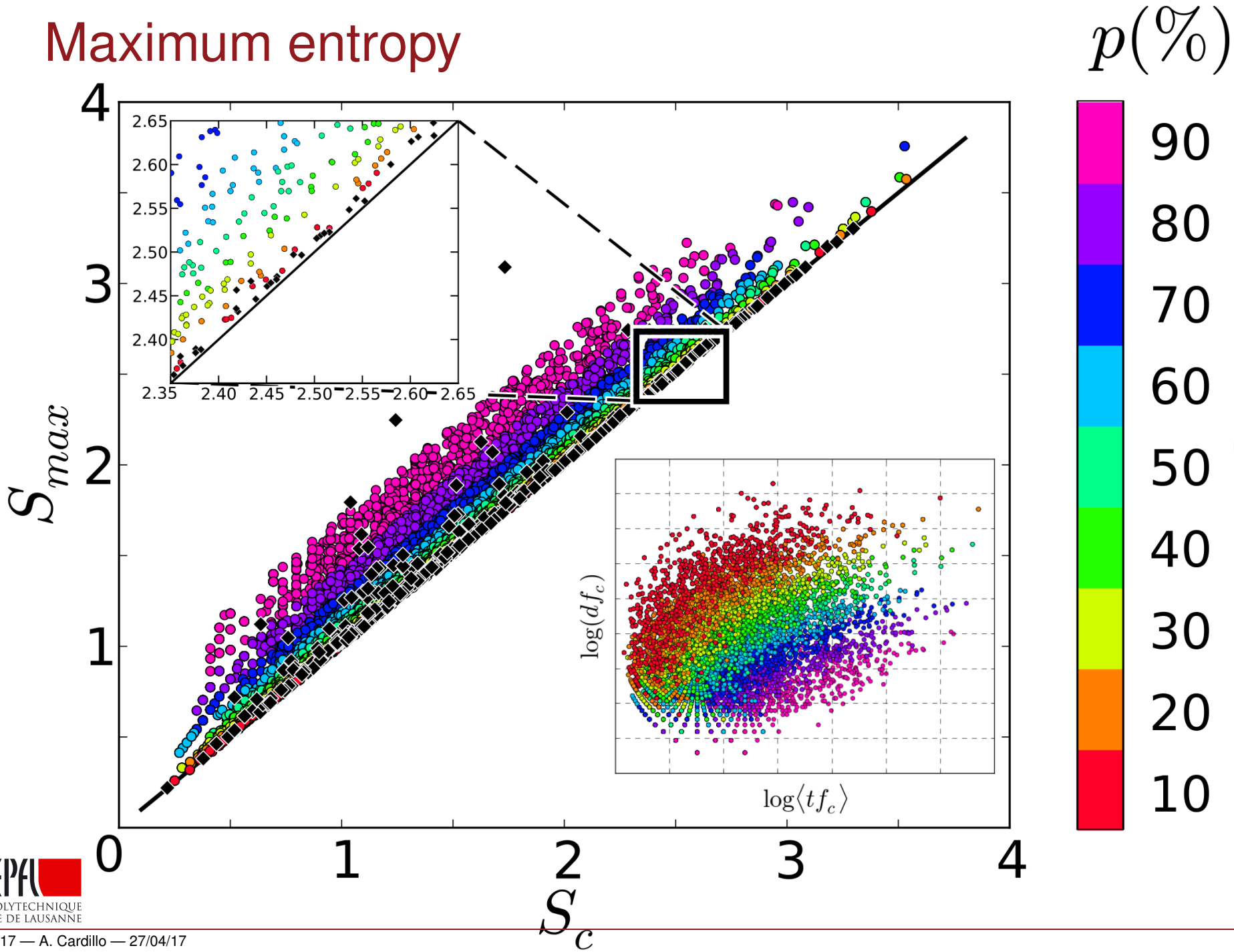
$$\sum_n p_n \ln n = \langle \ln n \rangle$$

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

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



Maximum entropy



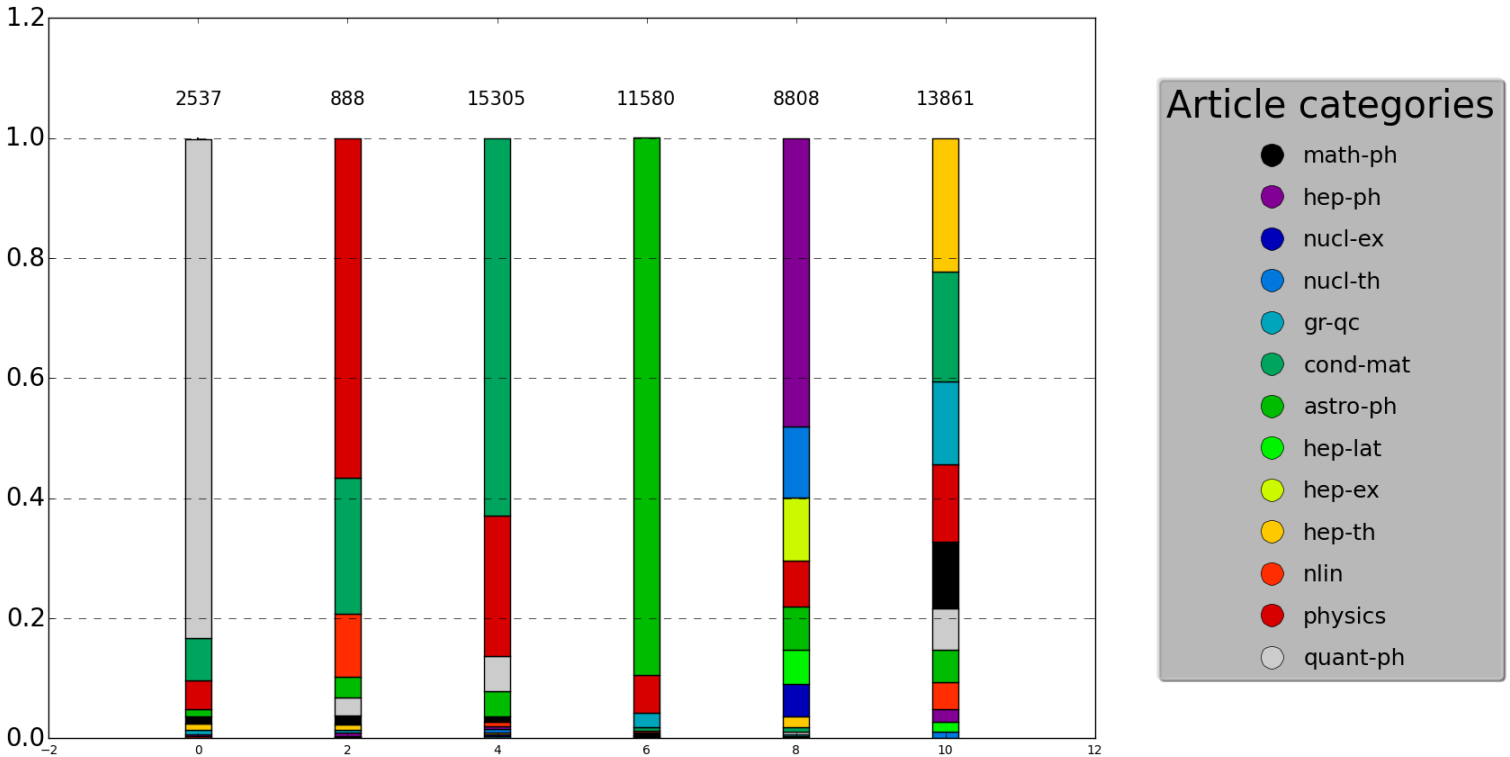
Section 4

Results

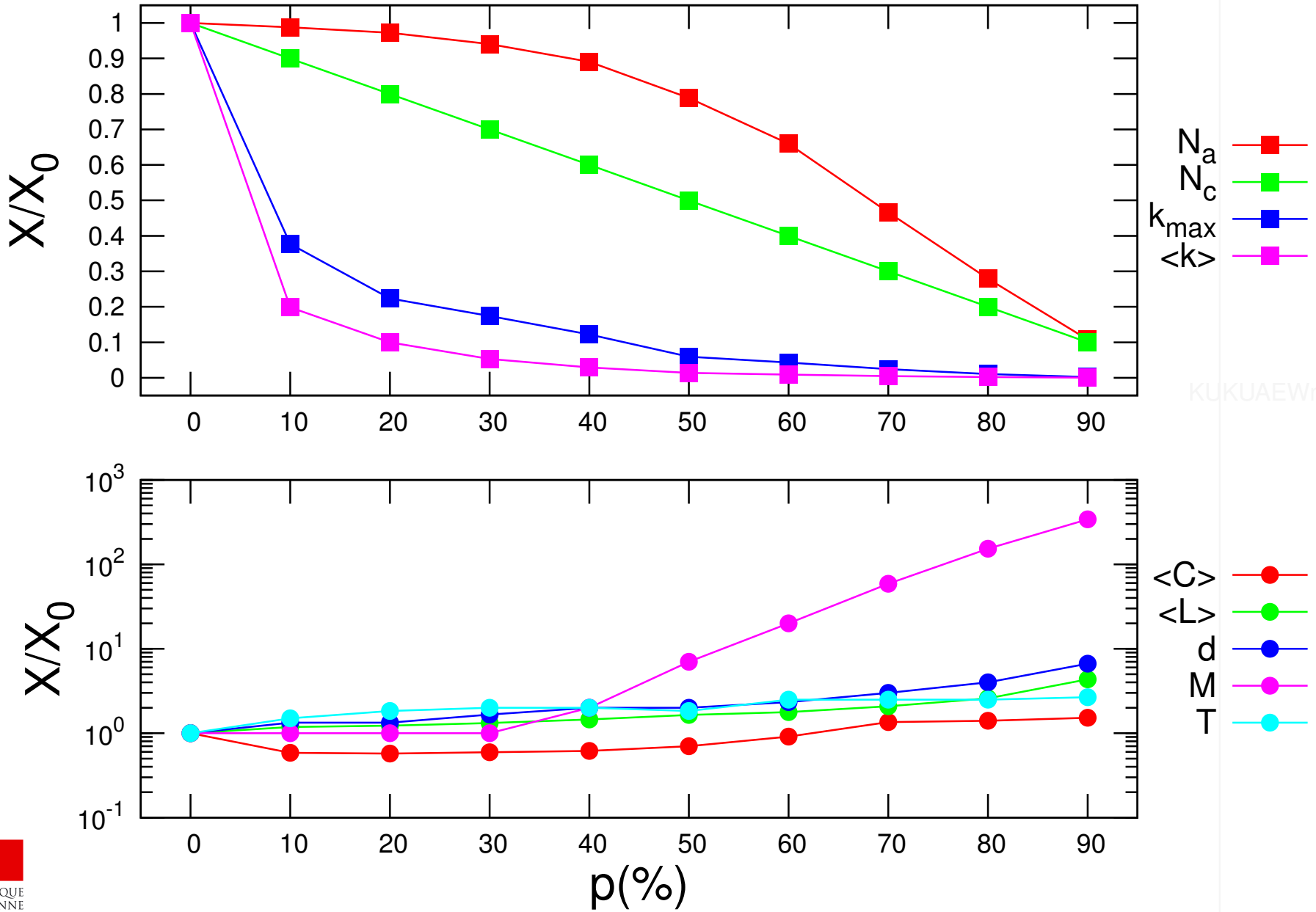
Topological properties

Original network

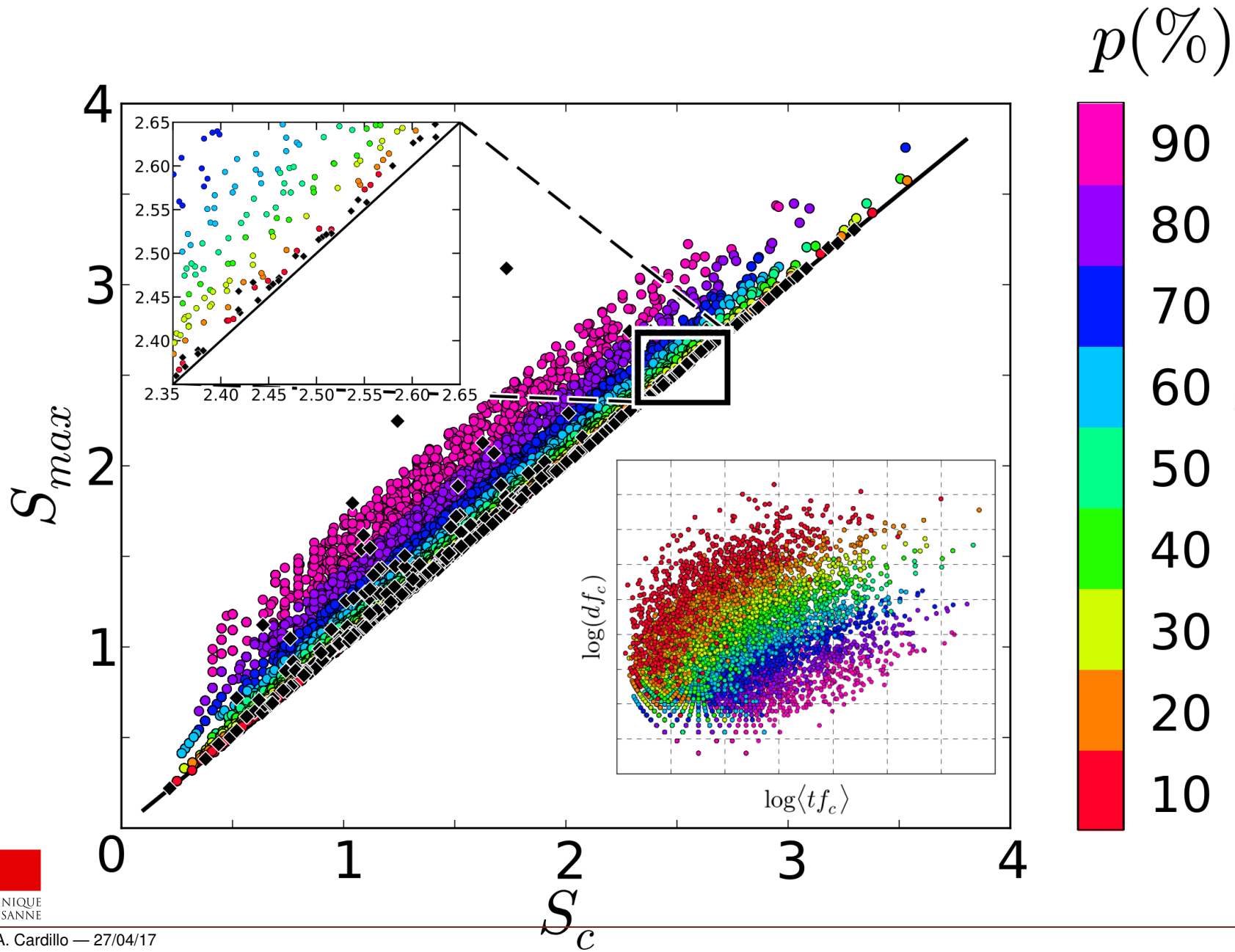
N_c	N_a	$\langle k \rangle$	k_{max}	$\langle C \rangle$	d	$\langle L \rangle$	M	T
10661	52979	19333.522	46504	0.557	3	1.635	1	6



Topological properties



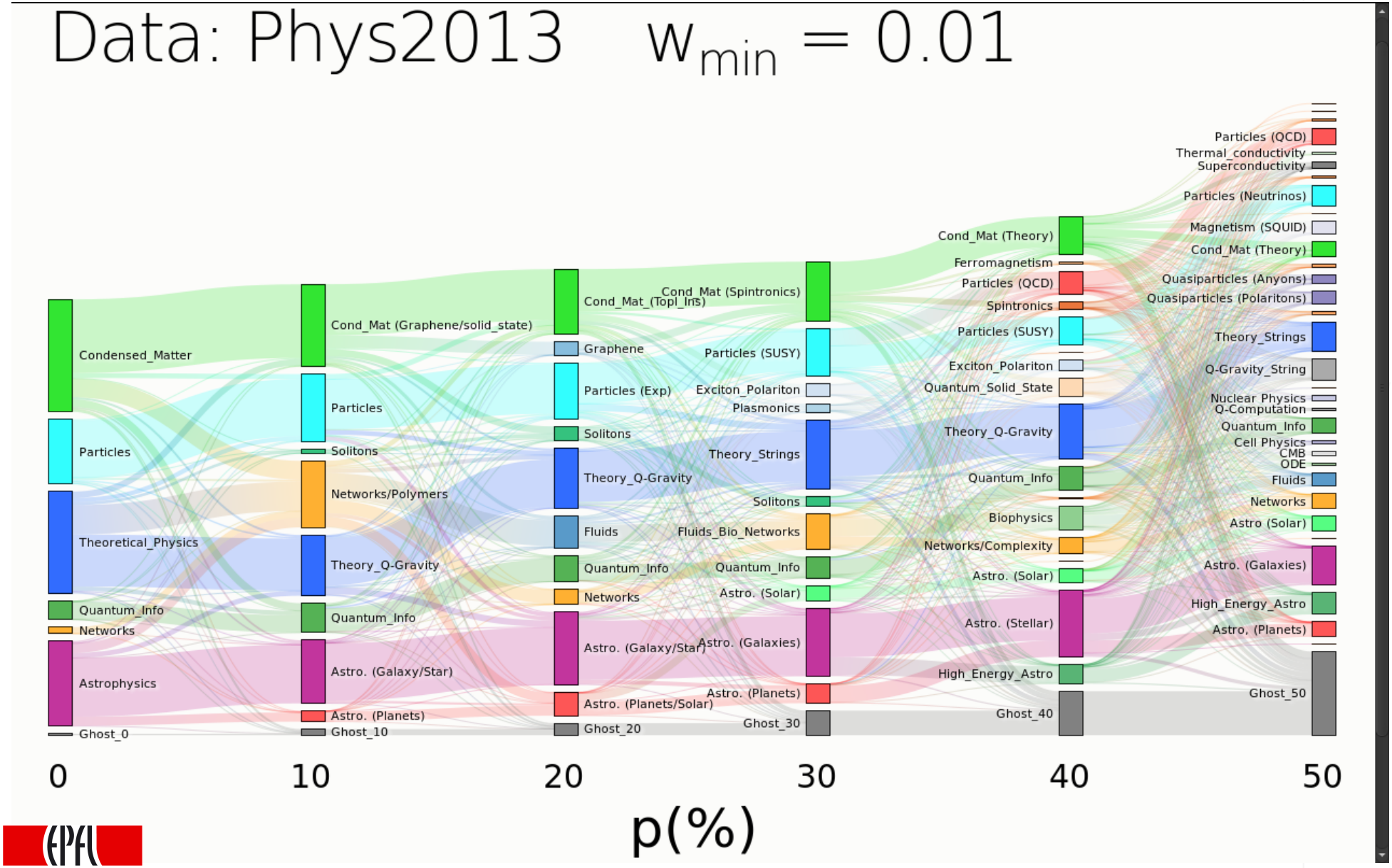
What is a “generic concept”?



Community detection

▶ Link

Community detection



Section 5

Conclusions

Summing up . . .

Take home messages

- We have used the maximum entropy principle to build a method to filter networks of similarity between documents.

Summing up . . .

Take home messages

- We have used the maximum entropy principle to build a method to filter networks of similarity between documents.
- The method allows to identify collection dependent "*relevant concepts*" without requiring user validation.

Summing up . . .

Take home messages

- We have used the maximum entropy principle to build a method to filter networks of similarity between documents.
- The method allows to identify collection dependent "*relevant concepts*" without requiring user validation.
- 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".

Summing up . . .

What's next? Open questions

- Study the evolution in time of "generality".
- Study the relation between concepts.

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.

Summing up . . .

Reference

A. Martini *et al.* , *Automatic selection of relevant concepts in scientific publications* – to be submitted

http://bifi.es/~cardillo/

alessio.cardillo@epfl.ch

Acknowledgements

Paolo De Los Rios

Andrea Martini

Alex Constantin

&

ScienceWISE team

Acknowledgements

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

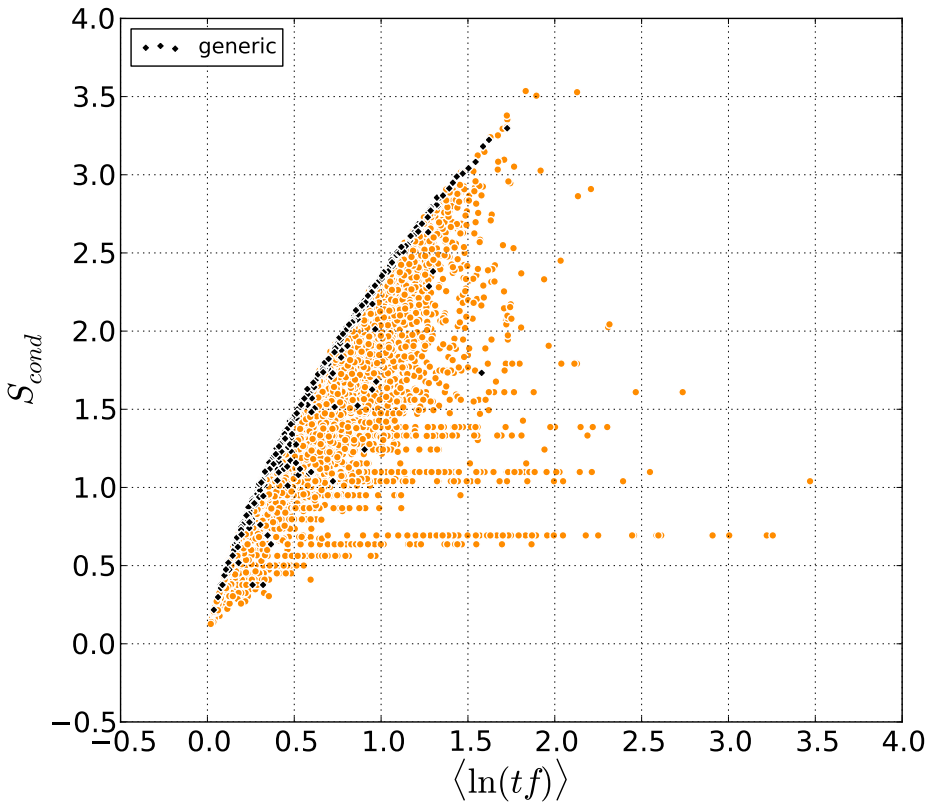
Academic freedom matters.

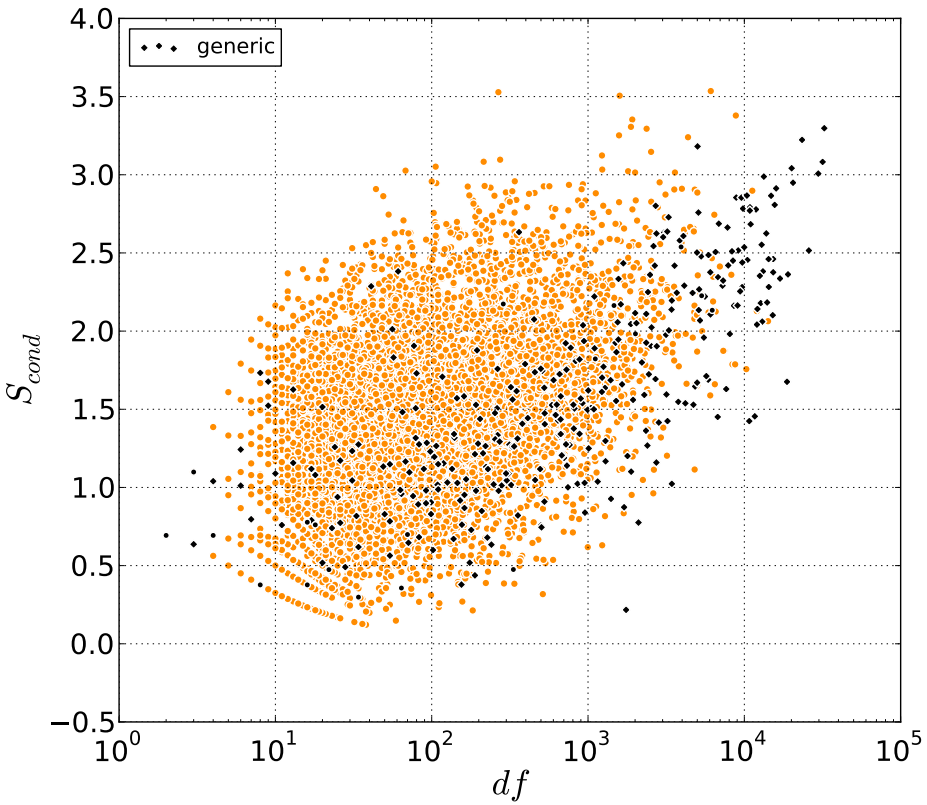
#IStandWithCEU

Part II

Appendix

Why not *df*?

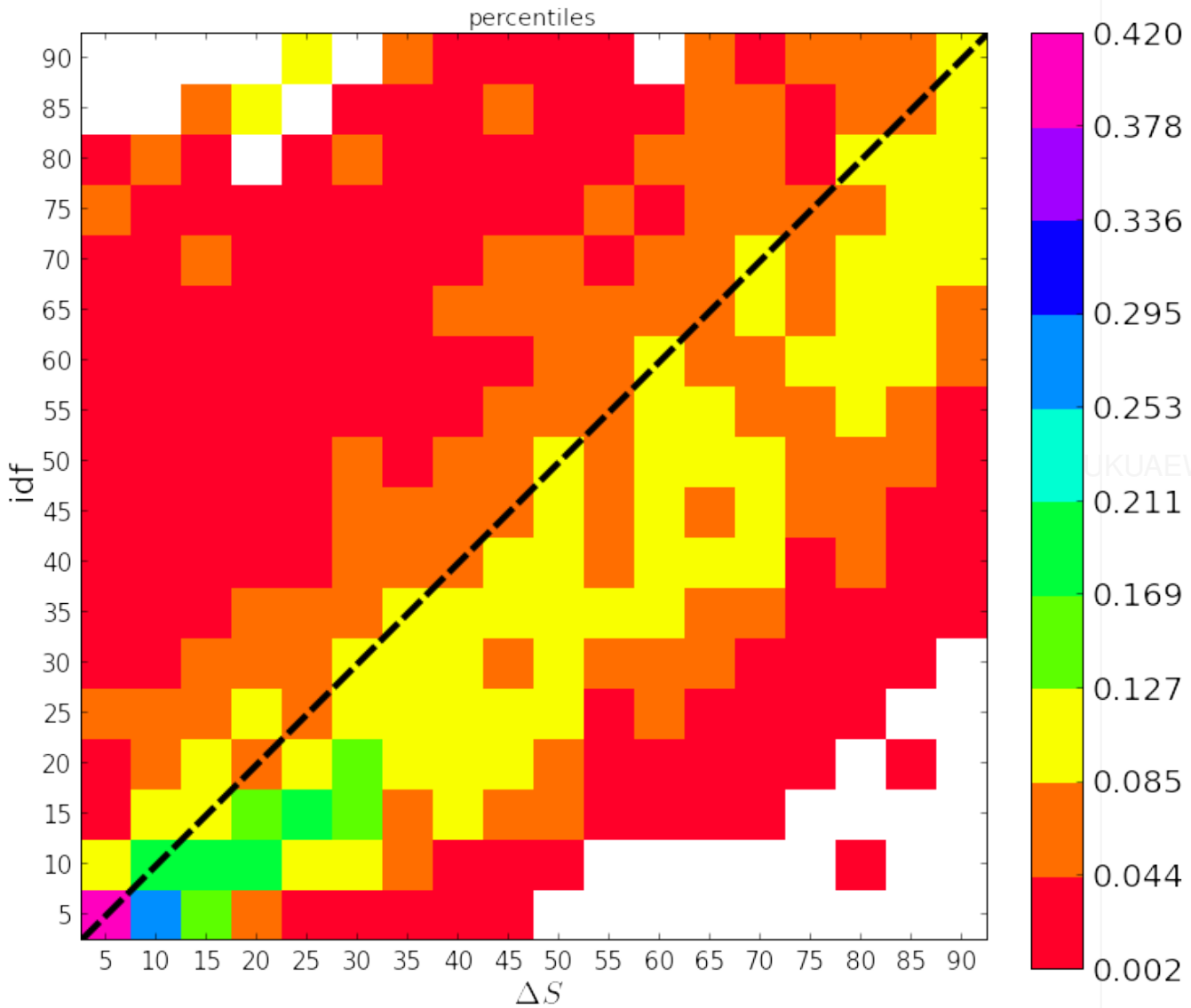




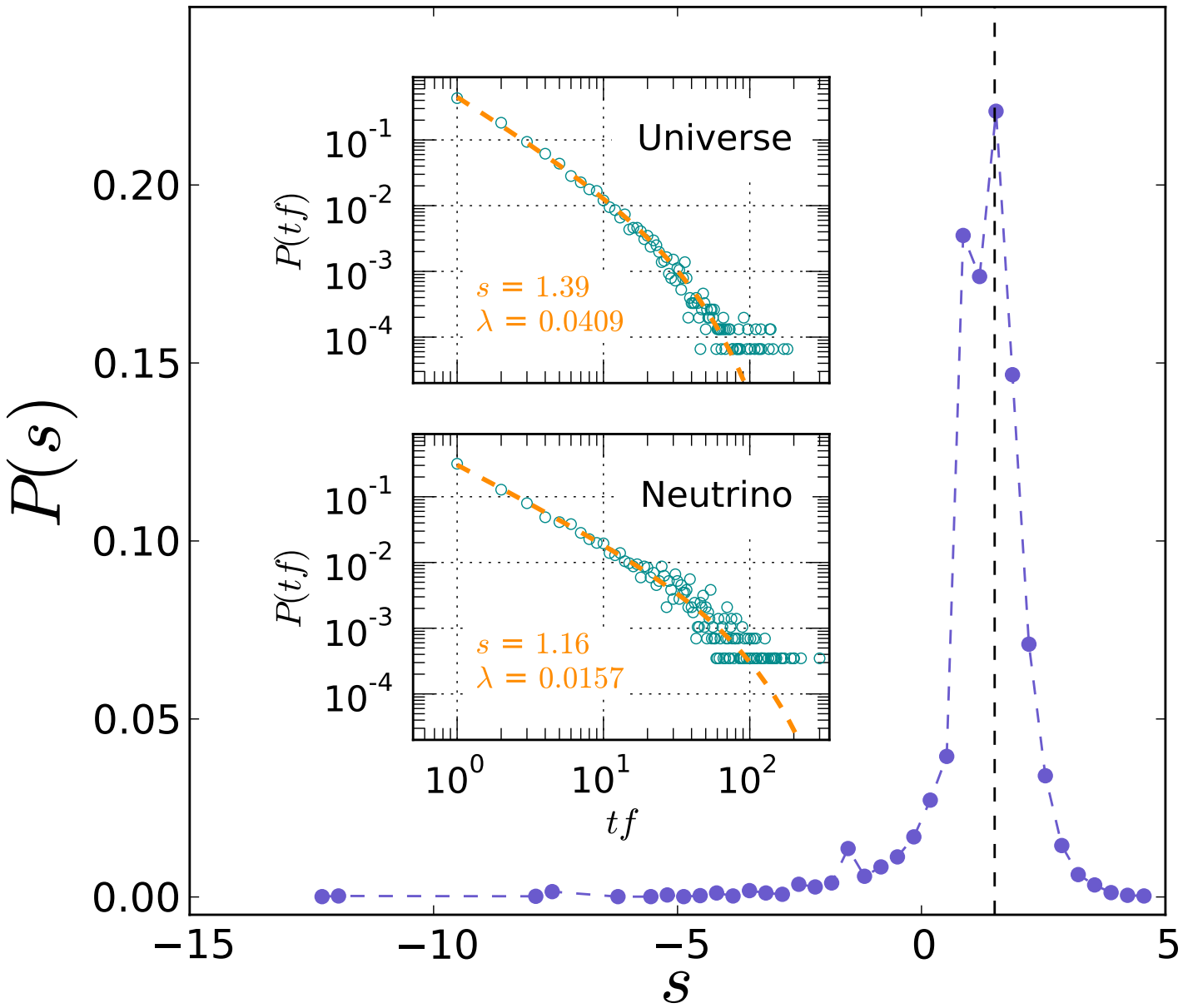
Why not *df*?

Jaccard Score

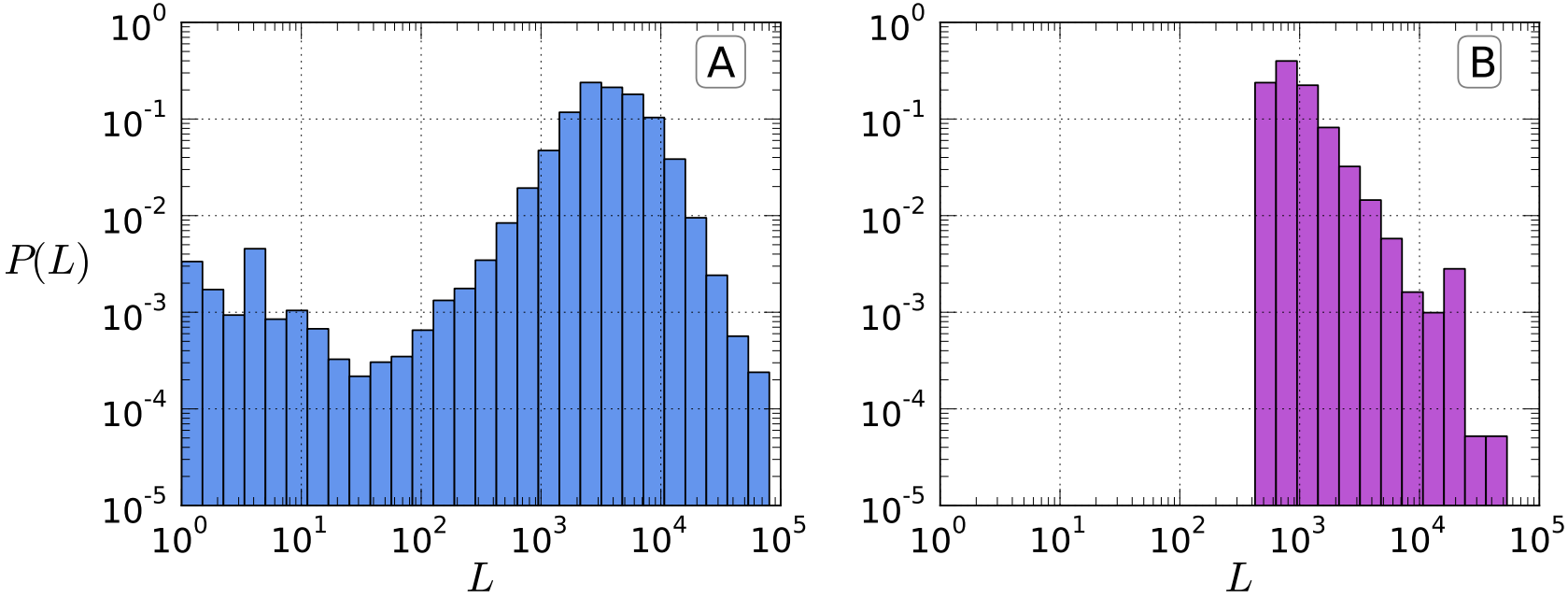
$$J = \frac{|A \cap B|}{|A \cup B|}.$$



Maximum entropy – why power law?



Maximum entropy – TF density

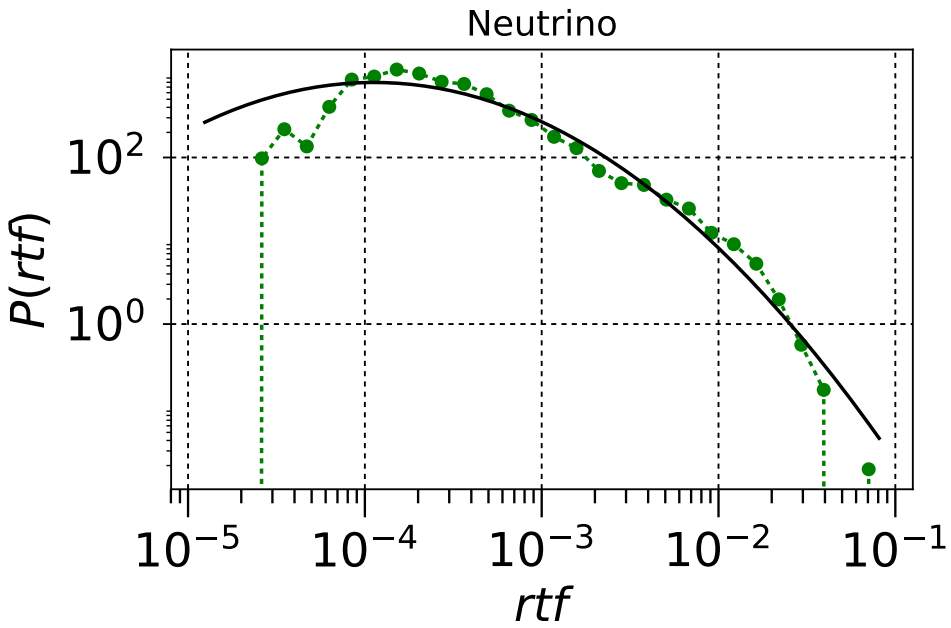


Maximum entropy – TF density

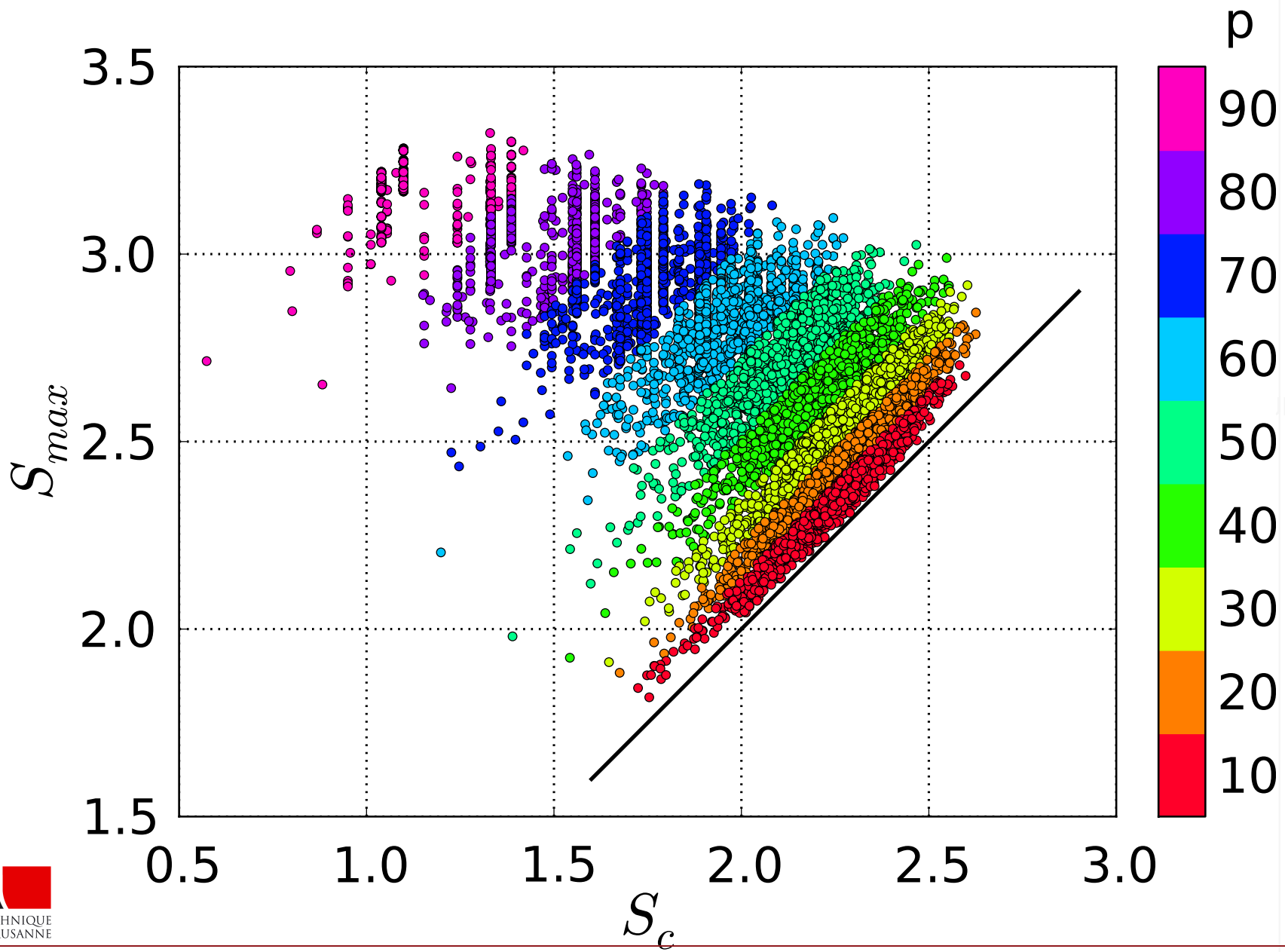
$$\sum_n p_n = 1$$

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

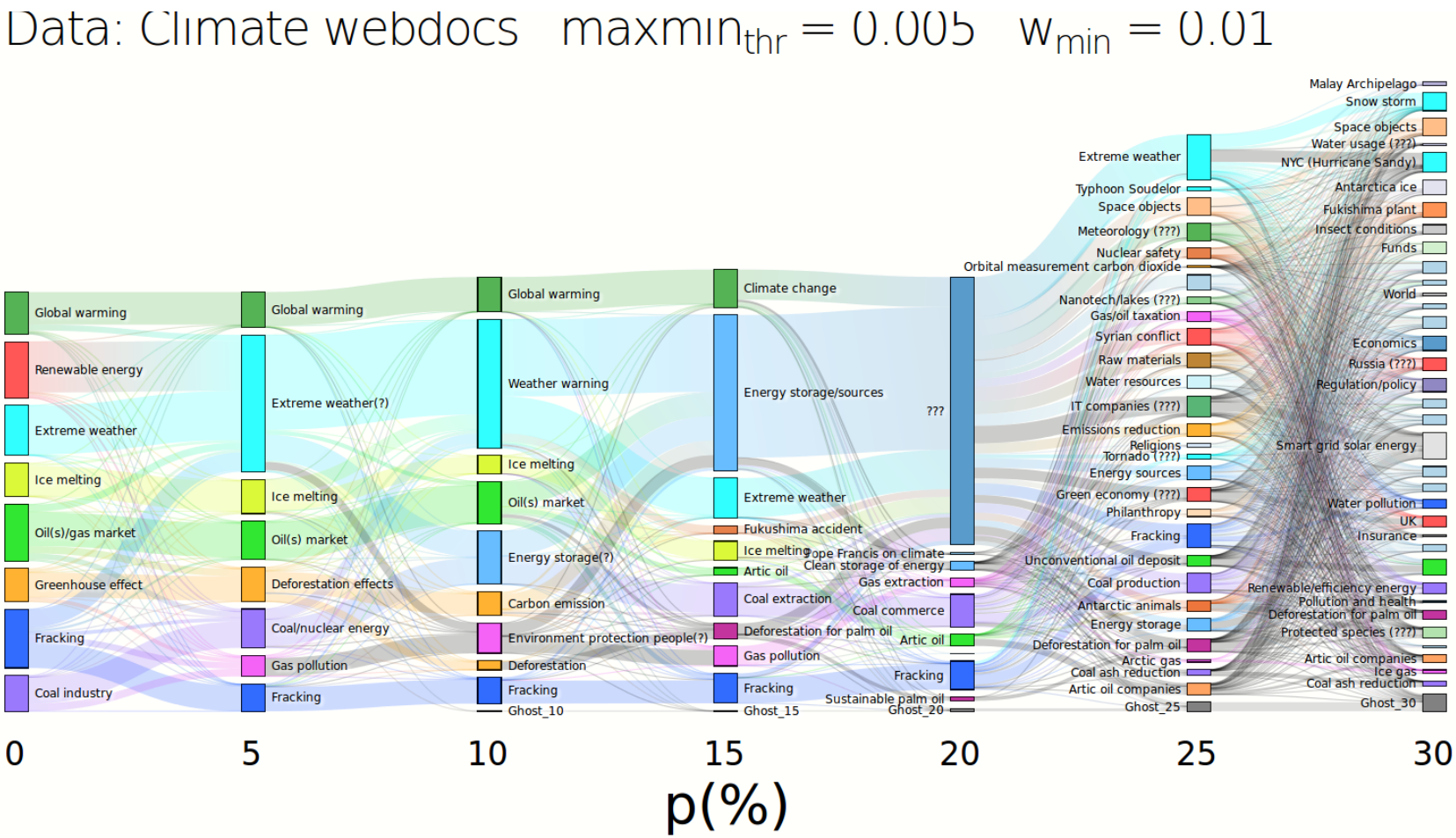
$$\sqrt{\sum_n p_n (\ln n - \langle \ln n \rangle)^2} = \sigma_{\ln n}$$



Climate dataset



Climate dataset



Optimal filtering