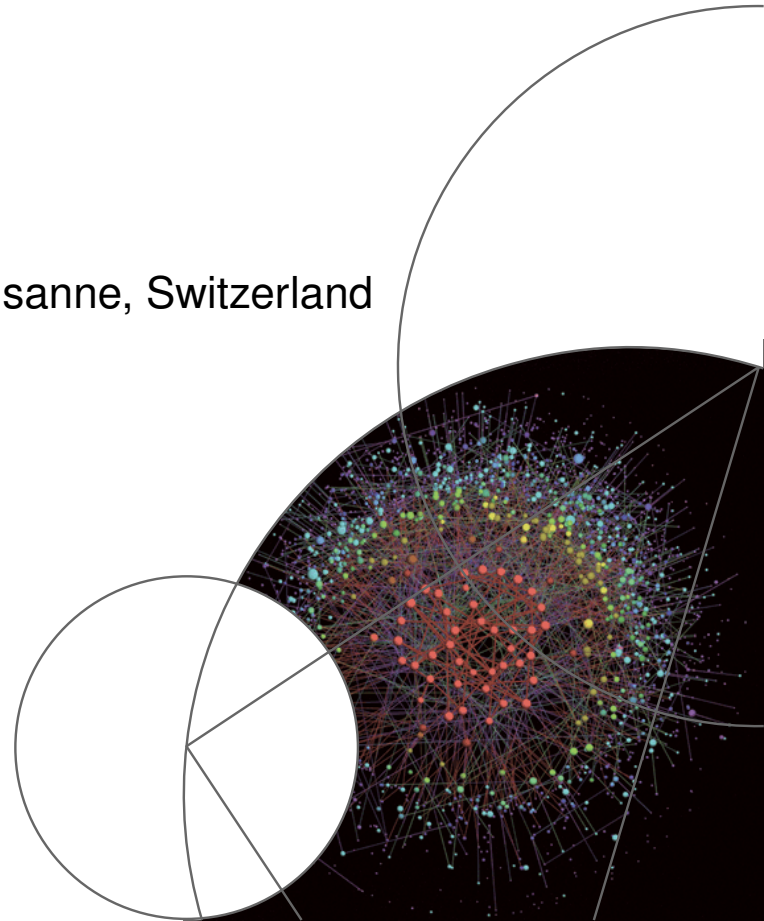
Automatic identification of relevant concepts in scientific publications

Alessio Cardillo
Laboratory for Statistical Biophysics (LBS)
École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland

http://bifi.es/~cardillo/



Acknowledgements

Main Collaborators

- Paolo De Los Rios (EPFL)
- Andrea Martini (EPFL)

Other Collaborators

- Alex Constantin (EPFL)
- Vasyl Palchykov, Valerio Gemmetto, Diego Garlaschelli (Leiden - The Netherlands)
- ScienceWISE Team (Lausanne)

Previous post
Climate change is present danger, US warns

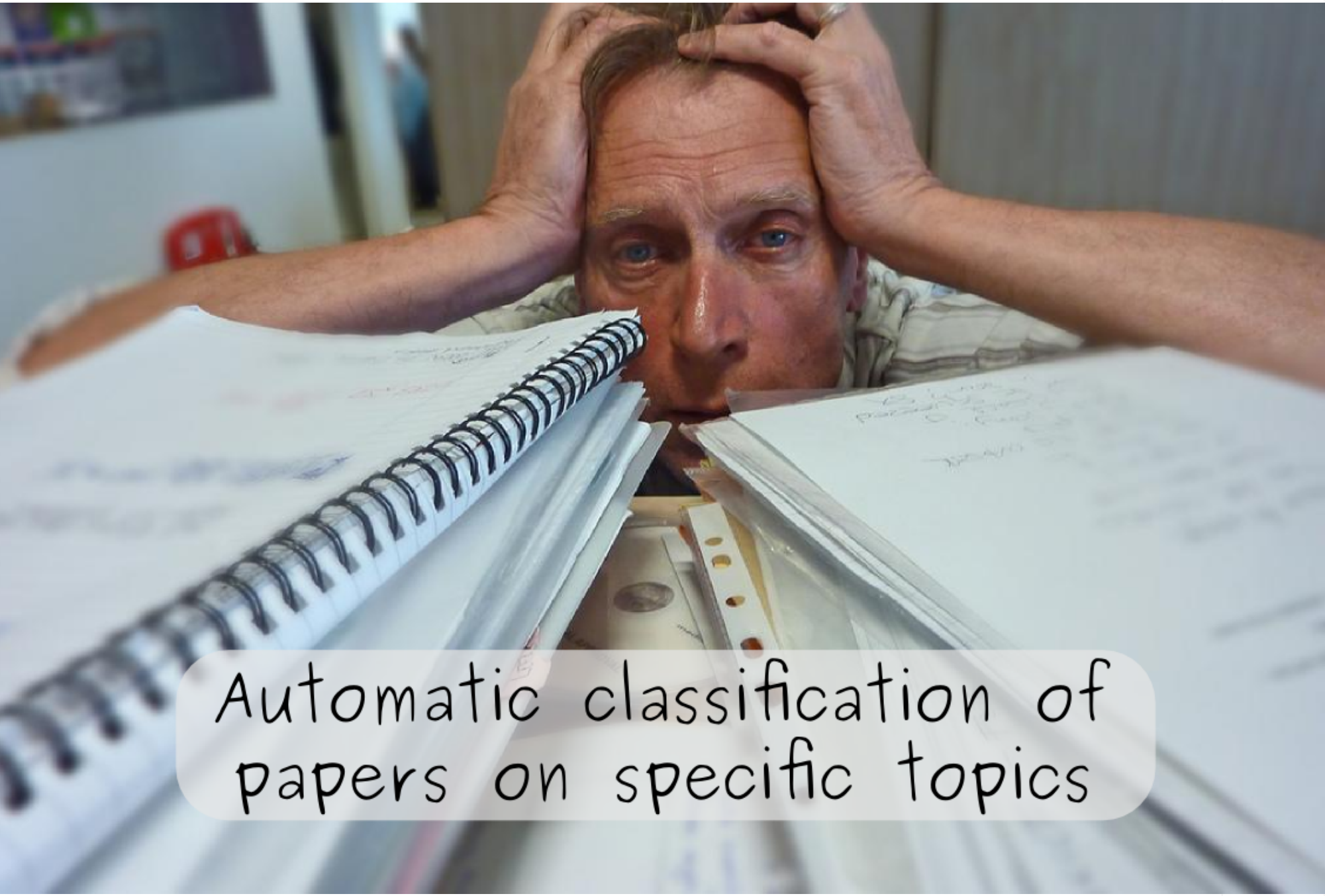
Next post
German research agencies condemn animal-rights attack on neuroscientist

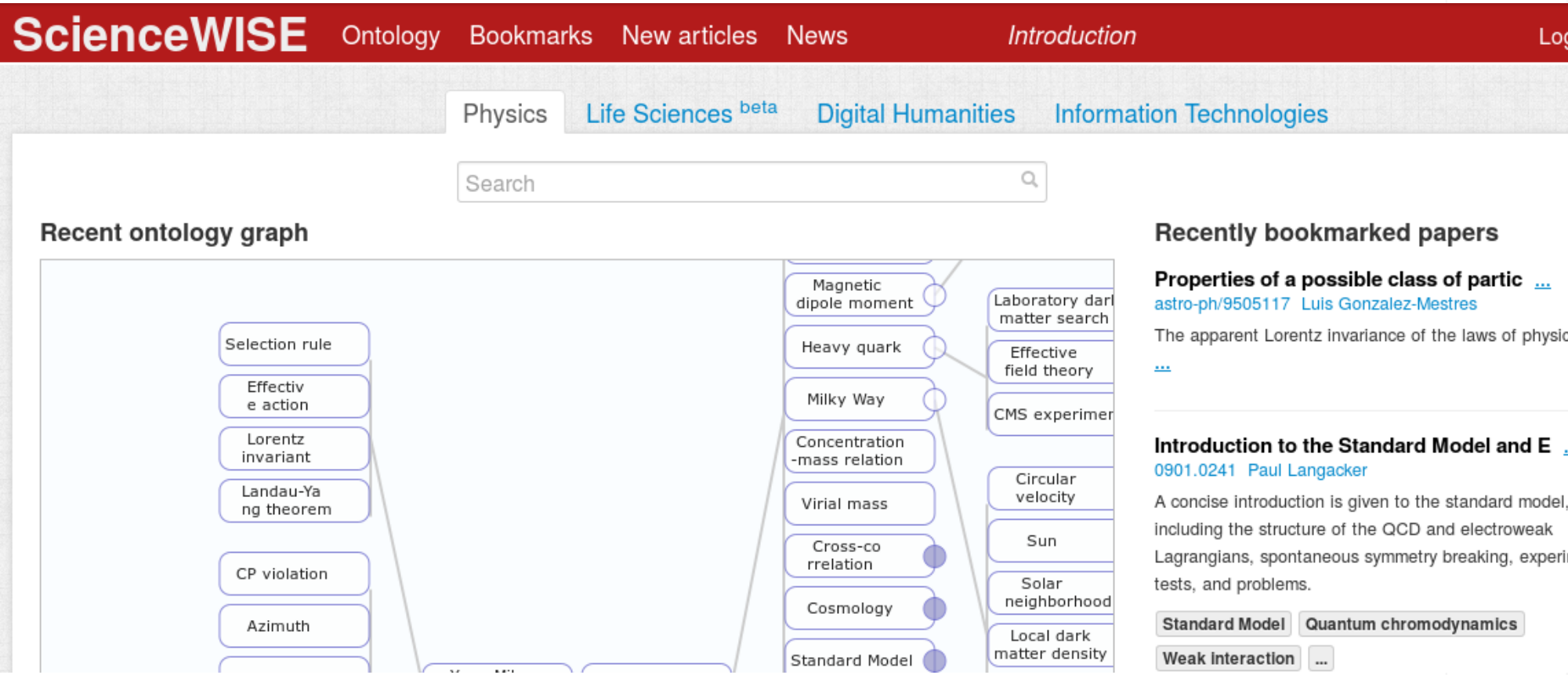
NEWS BLOG

Global scientific output doubles every nine years

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

It's a common complaint among academics: today's researchers are publishing too much, too fast. But jus


Automatic classification of
papers on specific topics



http://sciencewise.info

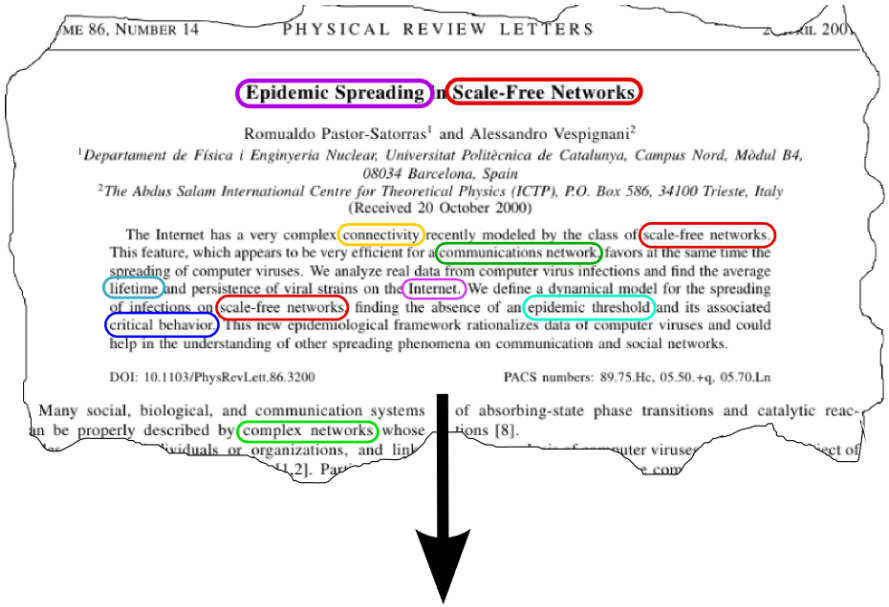
Outline

- 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 similarity between papers



VOLUME 86, NUMBER 14 PHYSICAL REVIEW LETTERS 2 APRIL 200

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 ... of absorbing-state phase transitions and catalytic reac- ... an be properly described by complex networks whose ... tions [8]. ... viduals or organizations, and link... computer viruses ... ect of ... [1,2]. Par... com...

Epidemics spreading

Scale-free networks

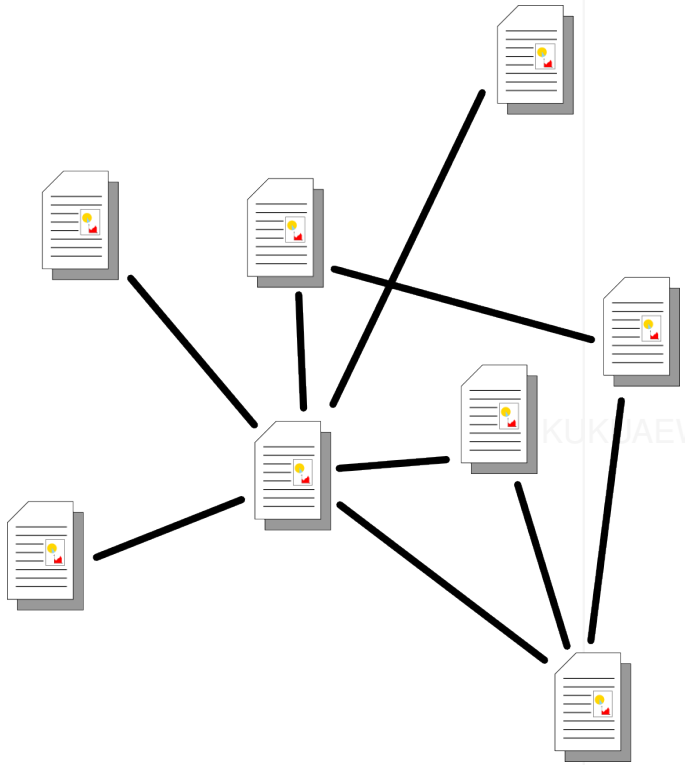
Connectivity

Communications nets

Lifetime

...

...



Similarity network

TF-IDF and similarity

α	1	1	1	0	1	0	0
	c_1	c_2	c_3	c_4	c_5	c_6	c_7

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

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
α	2	43	0	18	0	11	27
β	13	5	9	0	30	0	0

TF-IDF and similarity

Edge weight

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

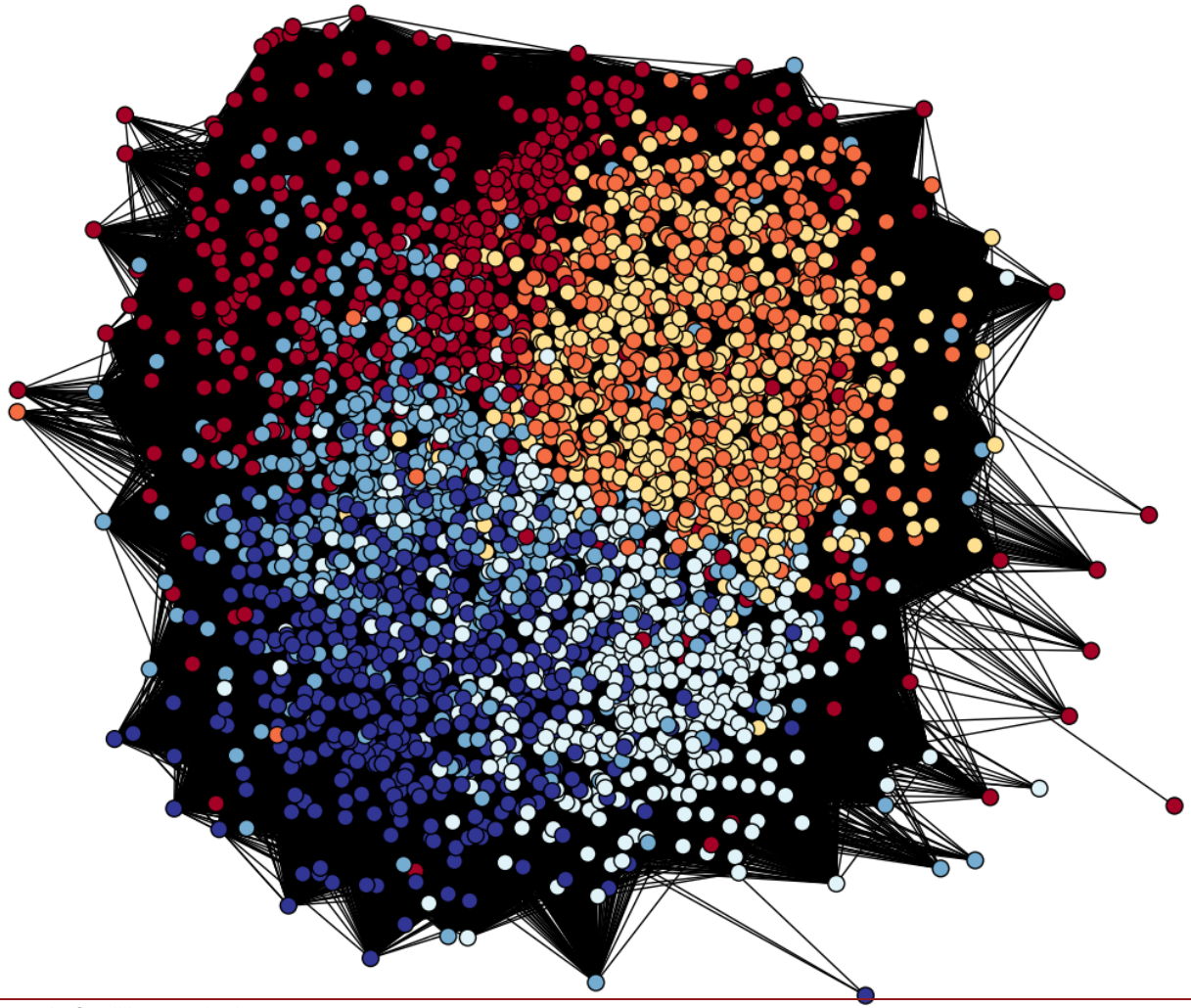
$$w_{\alpha\beta} \in [0, 1] ,$$

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

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

Network properties (Phys2013_pc)

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

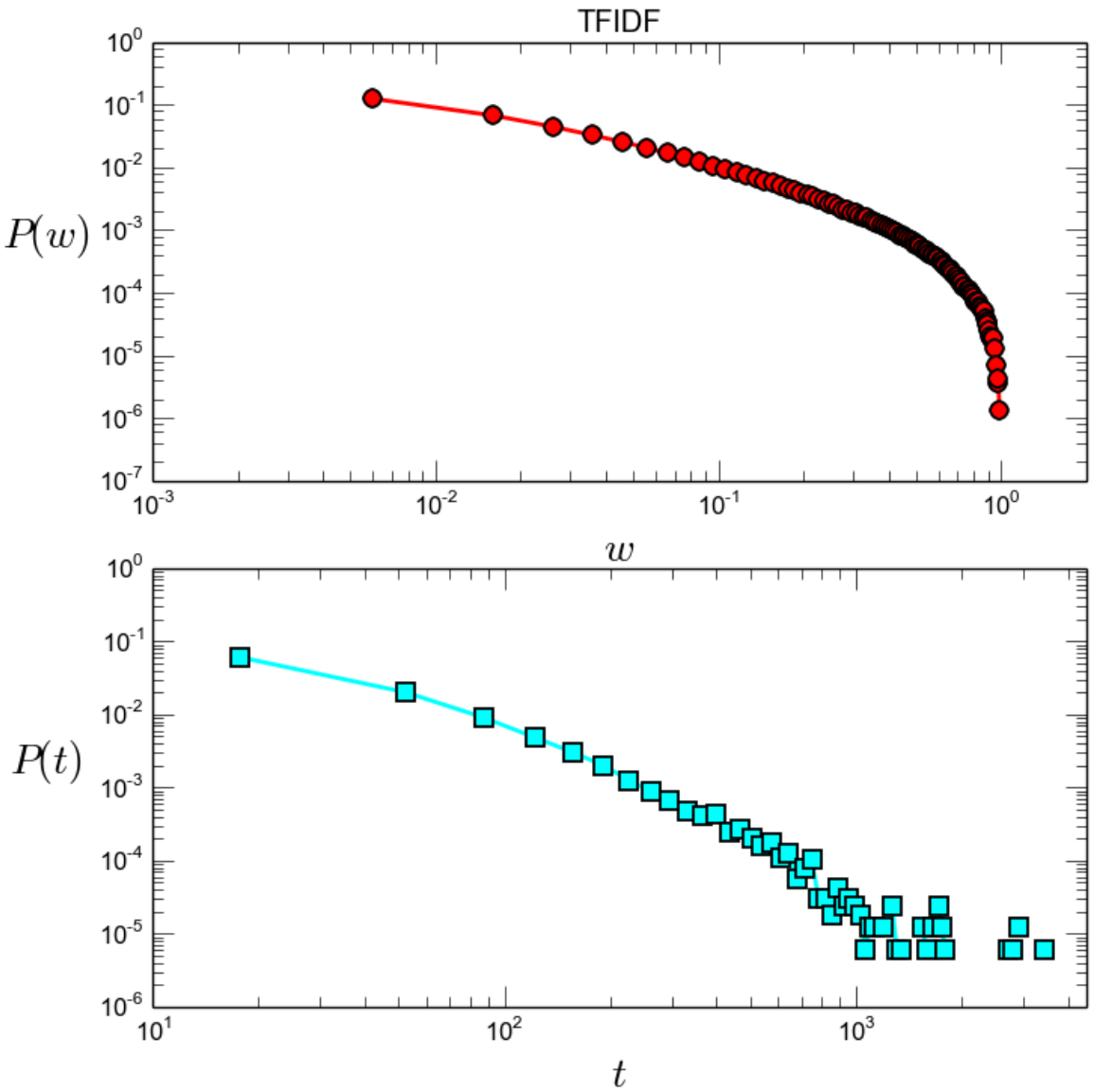



KEEP
CALM
AND
HOUSTON, WE
HAVE A PROBLEM
KeepCalmAndPosters.com

Section 2

Filtering

Edge sparsification methods



Edge 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).

Edge sparsification methods

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

Relevant concepts

Relevant concepts

Key features

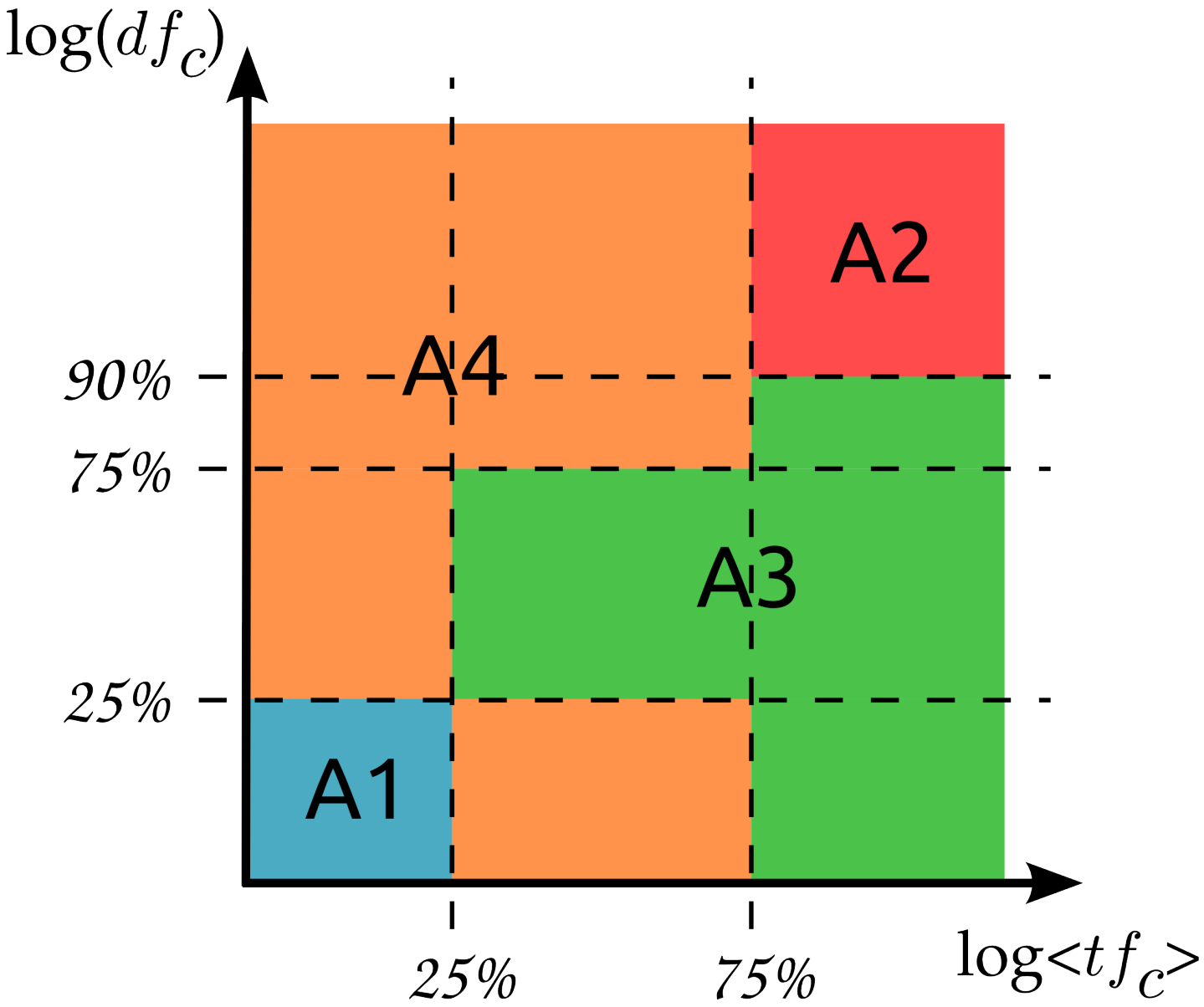
- # of papers a concept appears in

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

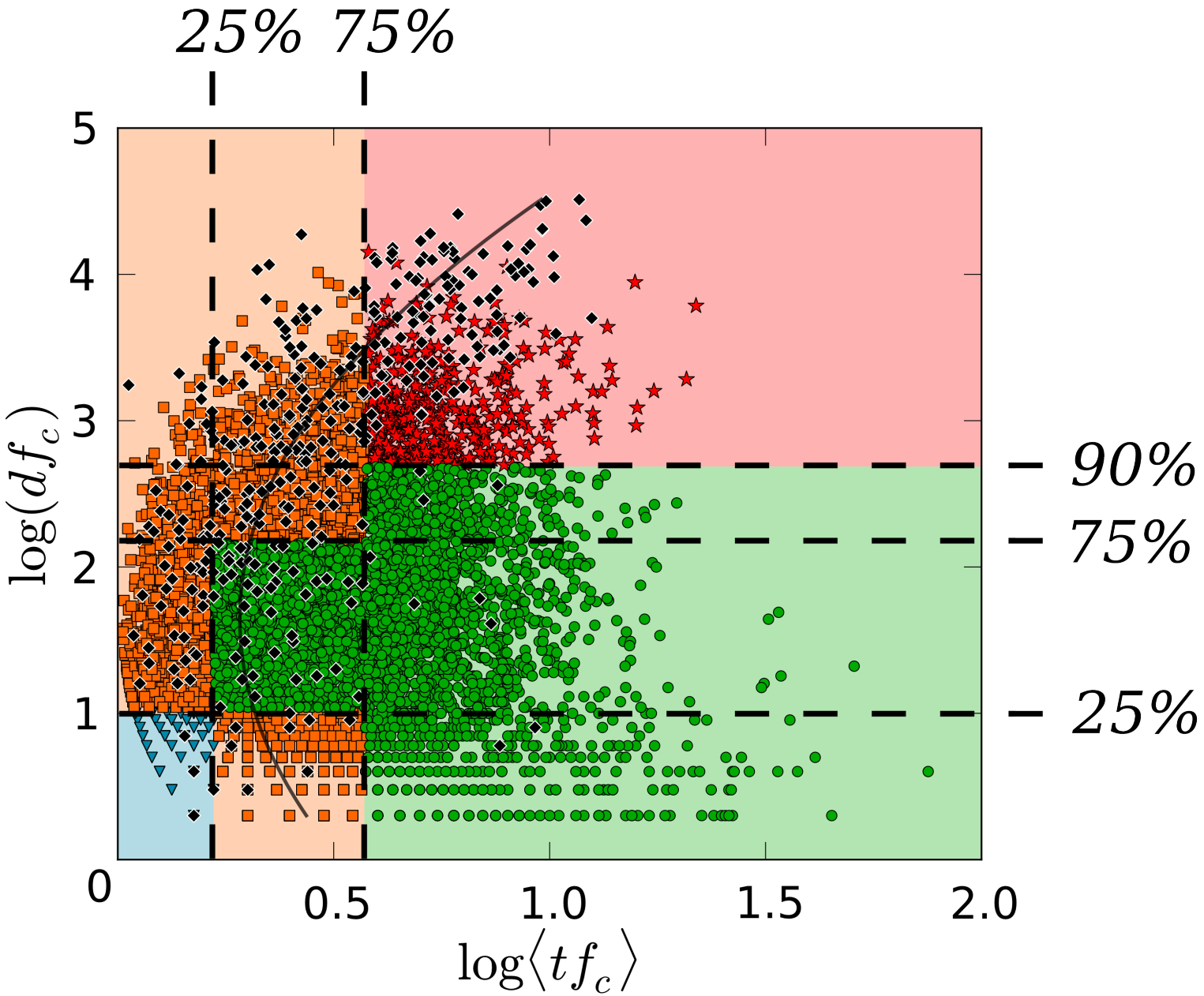
- average # of times a concept appears inside a paper

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

Bidimensional tessellation



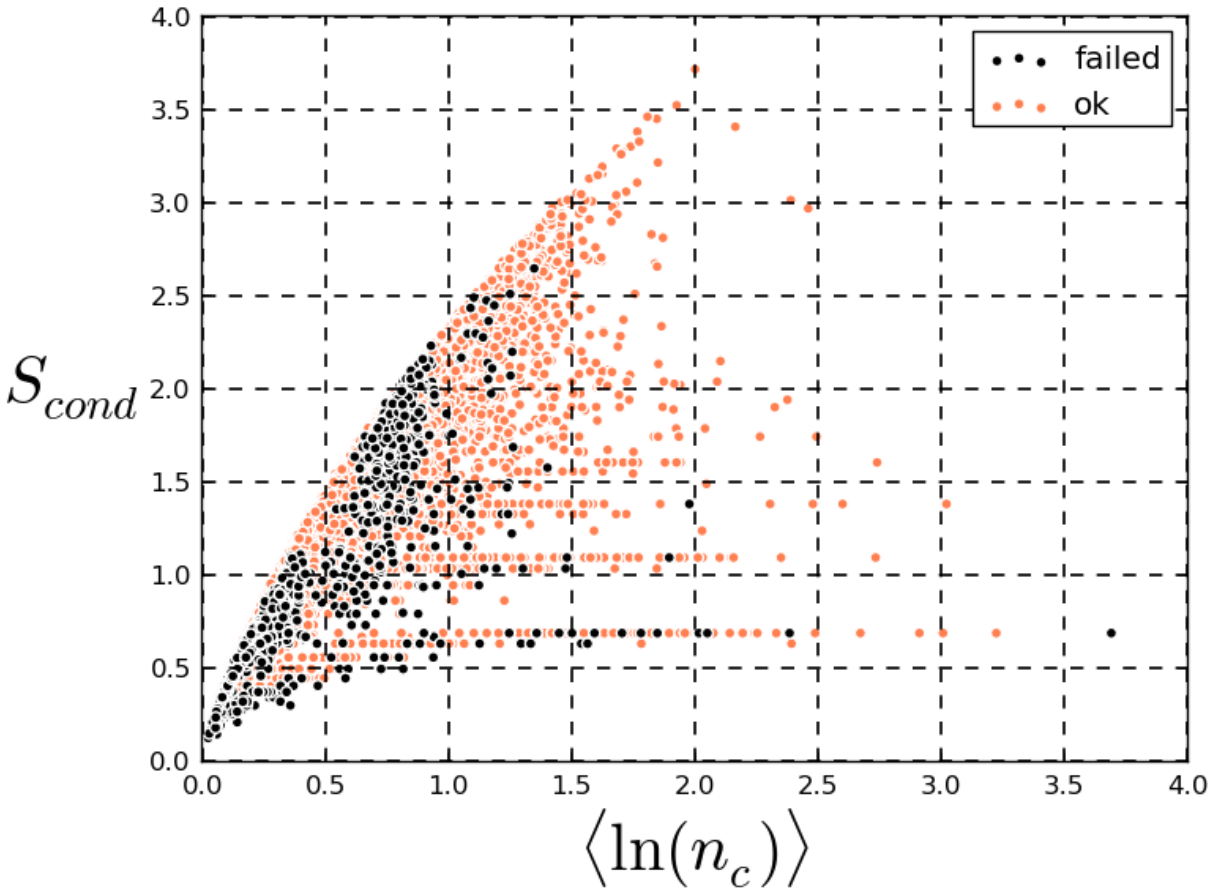
Bidimensional tessellation



Section 3

Entropic Filtering

Maximum entropy



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

Maximum entropy

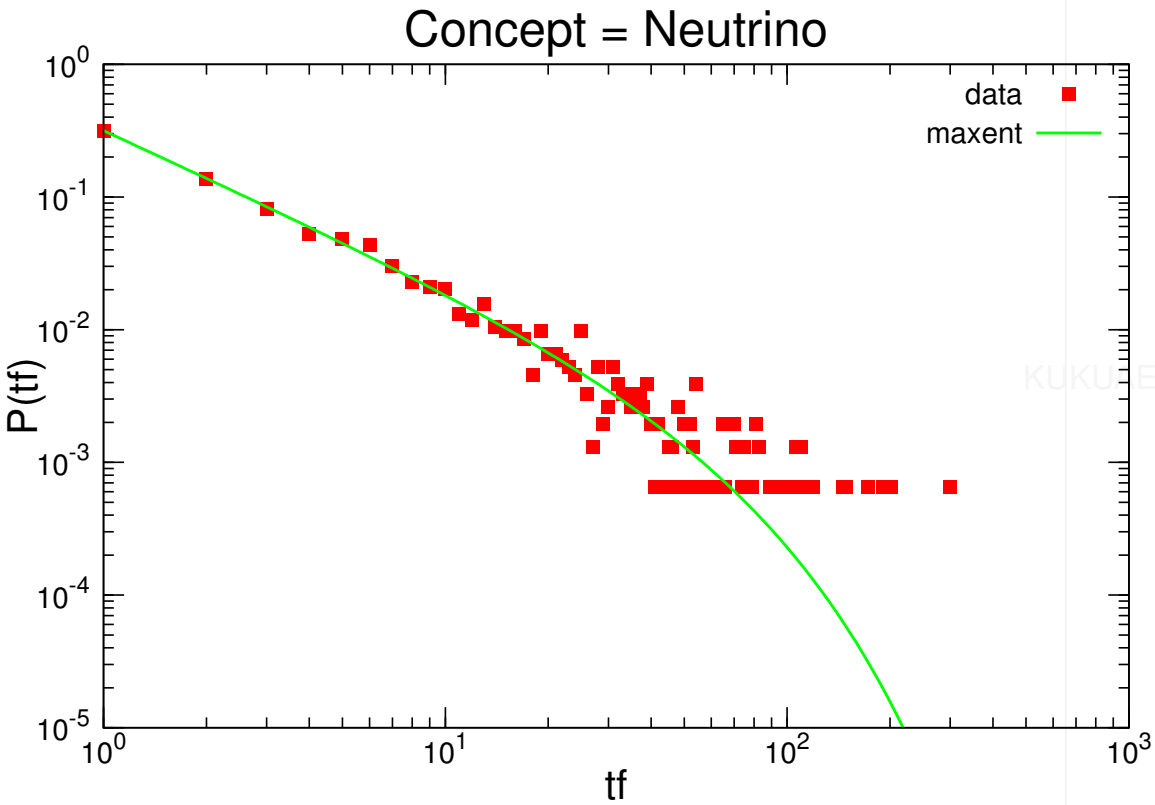
$$\sum_n p_n = 1$$

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

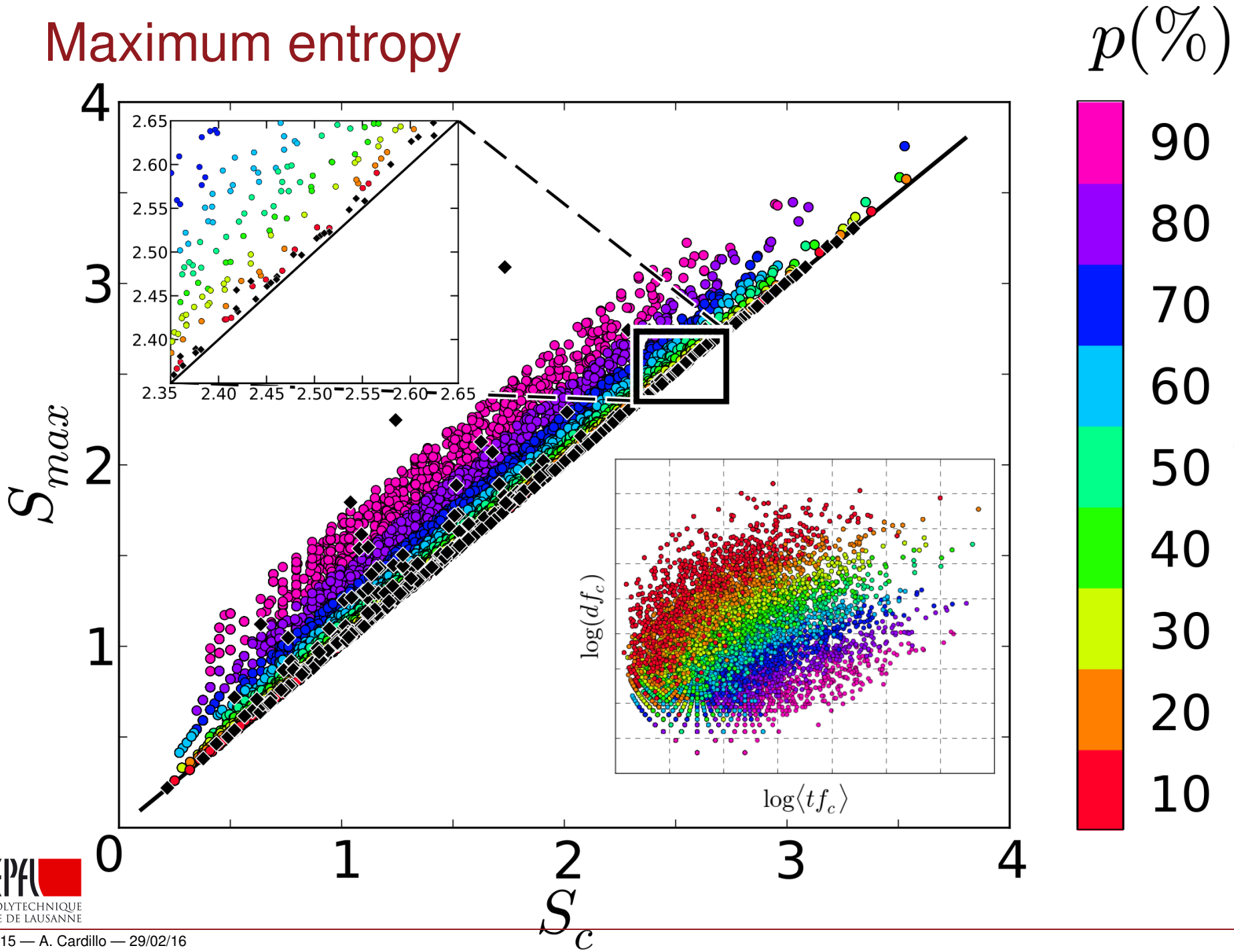
$$\sum_n p_n \ln n = \langle \ln n \rangle$$

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

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



Maximum entropy



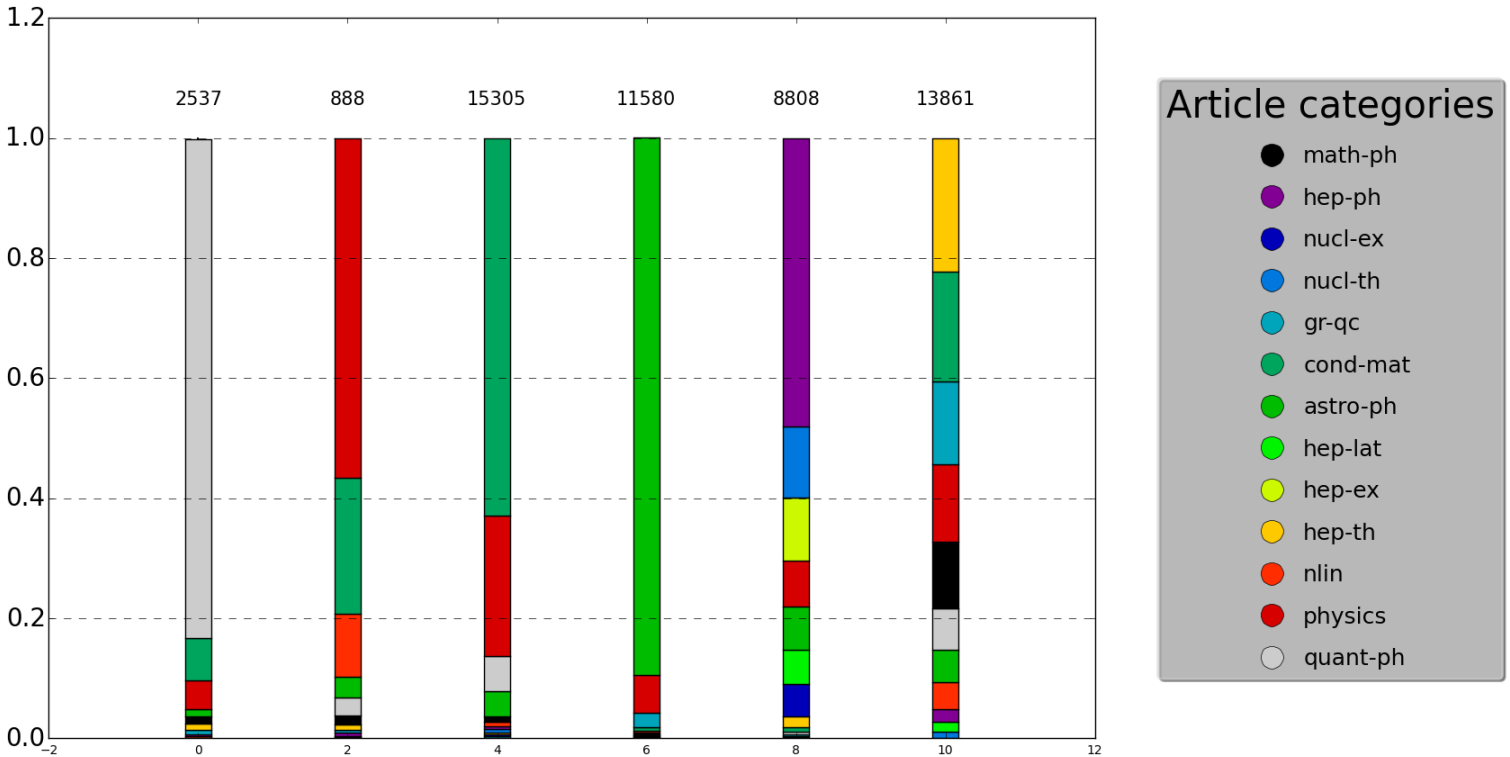
Section 4

Results

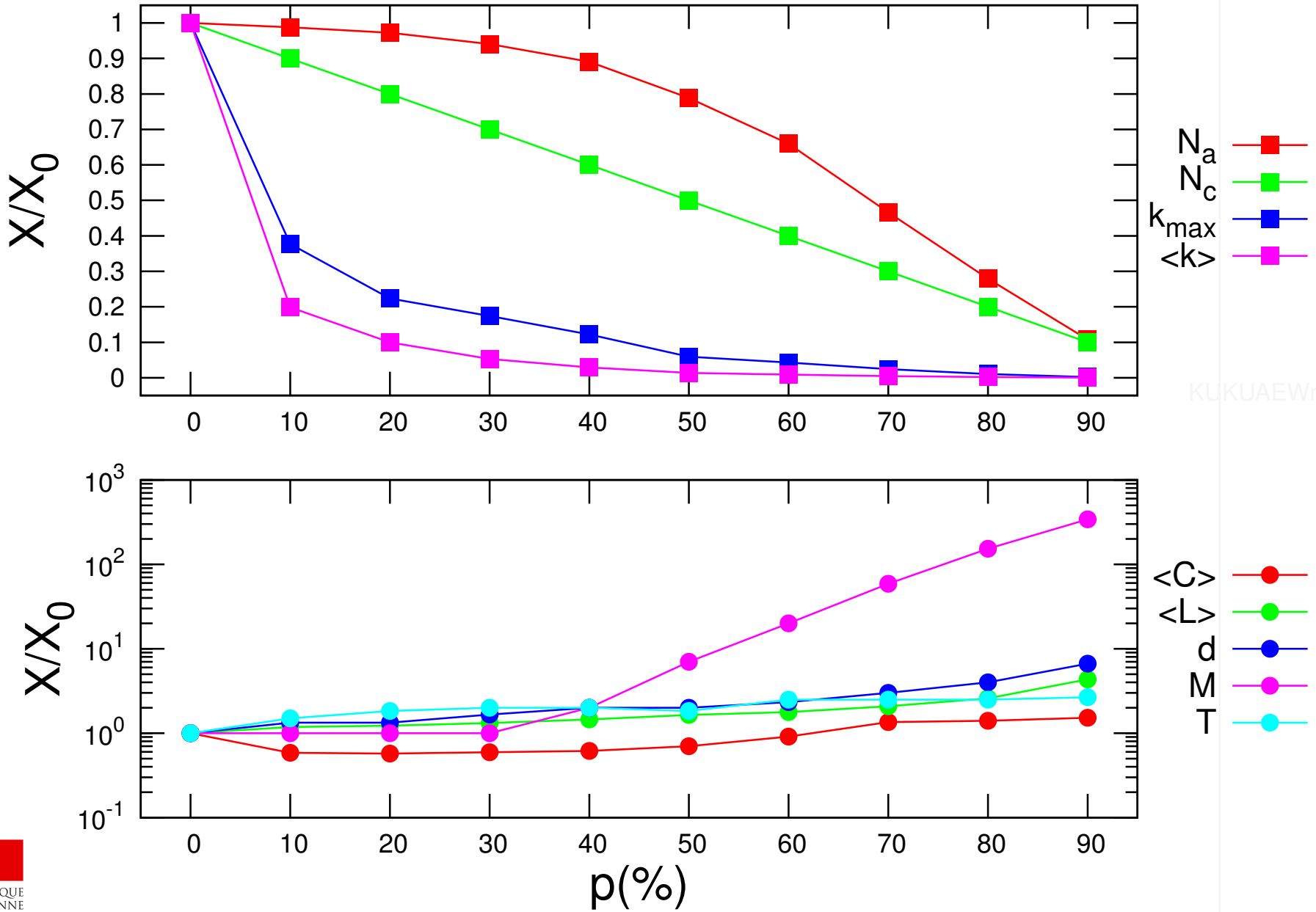
Topological properties

Original network

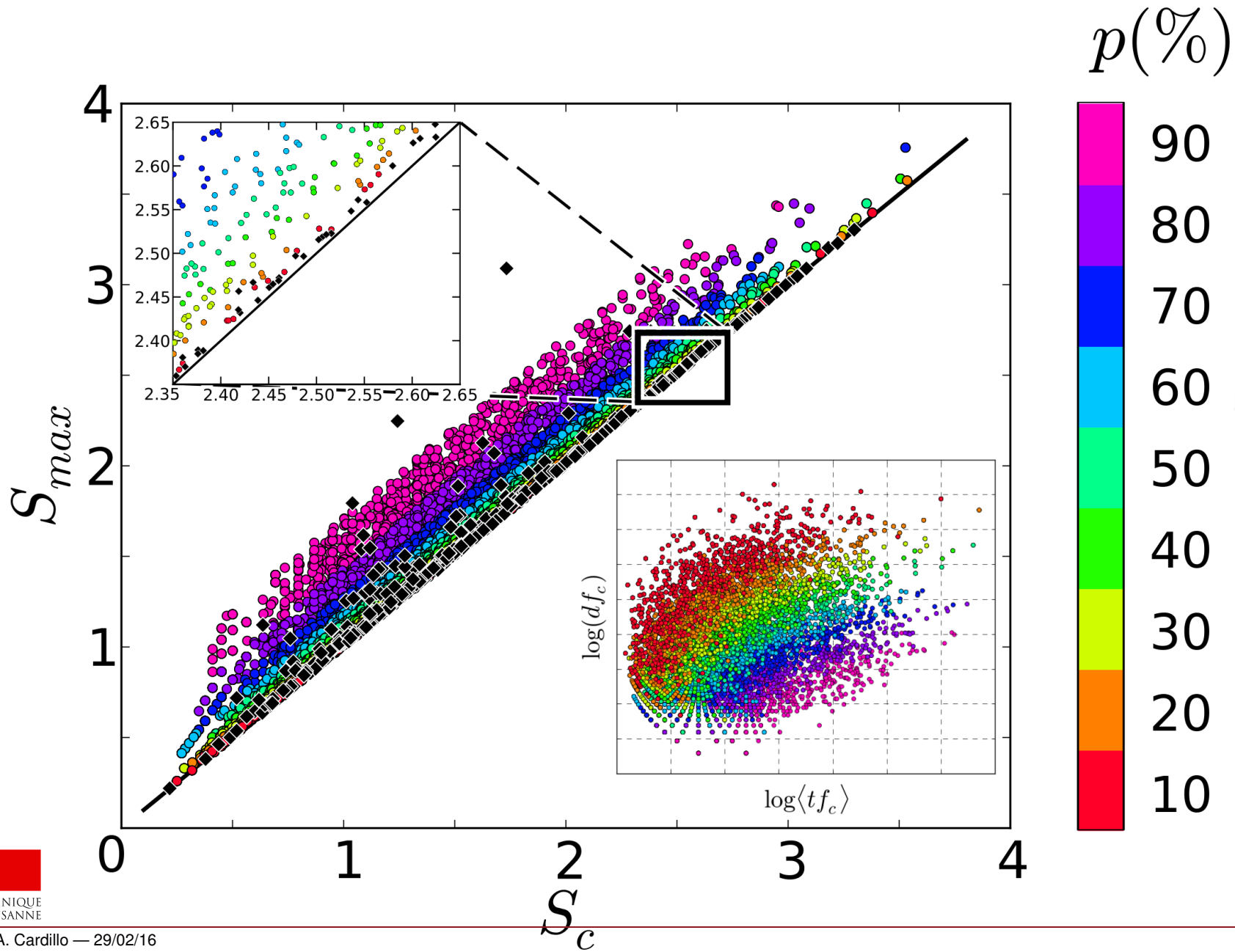
N_c	N_a	$\langle k \rangle$	k_{max}	$\langle C \rangle$	d	$\langle L \rangle$	M	T
10661	52979	19333.522	46504	0.557	3	1.635	1	6



Topological properties



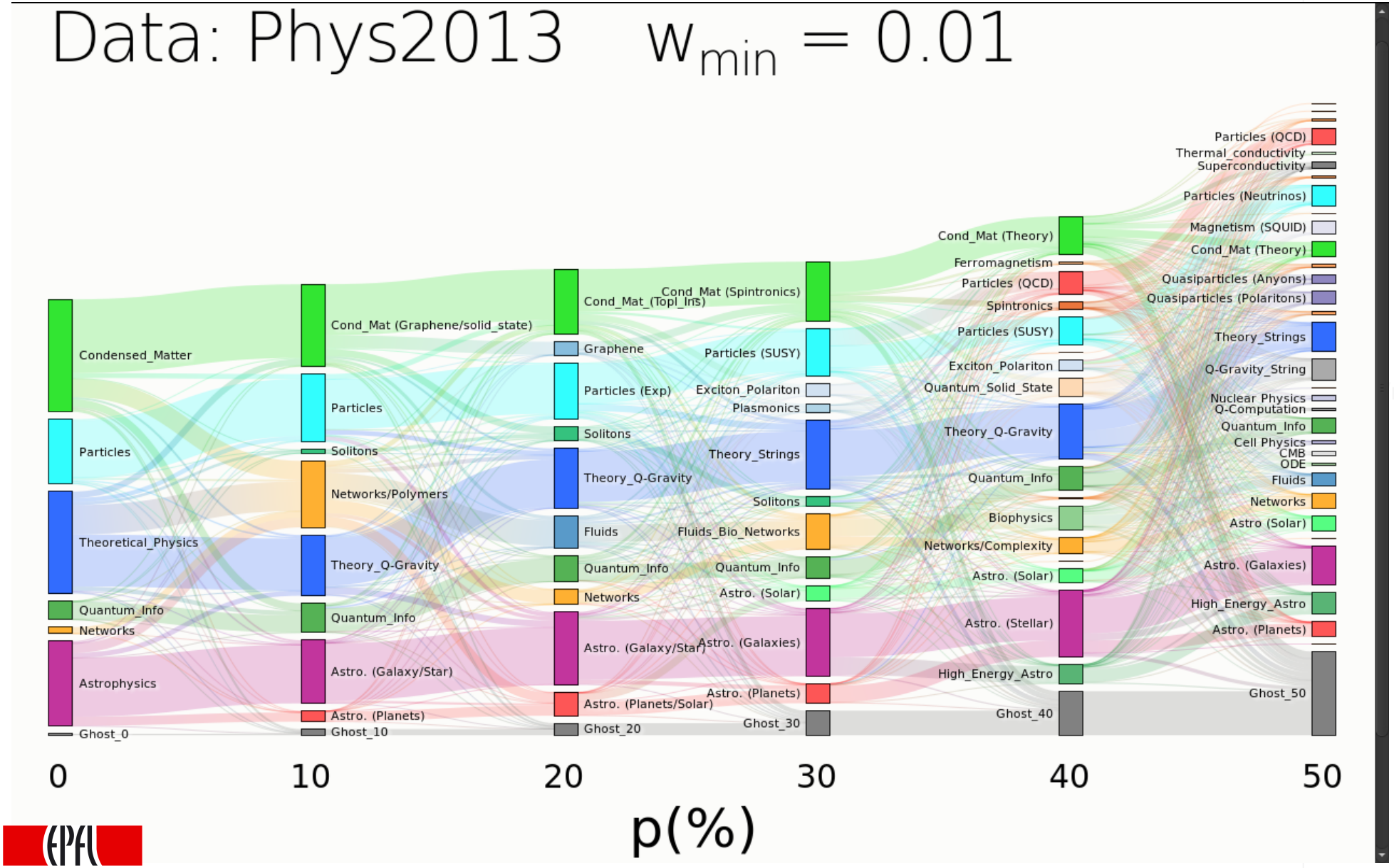
What is a “generic concept”?



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 removal of common concepts allows to retrieve a more well defined structure of documents into topics.

Summing up . . .

Take home messages

- We have used the maximum entropy principle to build a method to filter networks of similarity between documents.
- The removal of common concepts allows to retrieve a more well defined structure of documents into topics.
- The method allows to identify collection dependent "*relevant concepts*" without requiring user validation.

Summing up . . .

What's next? Open quesions

- Apply the methodology recursively.

Summing up . . .

What's next? Open quesions

- Apply the methodology recursively.
- Validate the method applying it on collection of documents different from scientific papers.

Summing up . . .

What's next? Open quesions

- Apply the methodology recursively.
- Validate the method applying it on collection of documents different from scientific papers.
- Study the evolution in time of "generality".

Summing up . . .

Reference

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

http://bifi.es/~cardillo/

alessio.cardillo@epfl.ch