Uncluttering document networks

Paolo De Los Rios, Alessio Cardillo and
Andrea Martini

Ecole Polytechnique Fédérale de Lausanne

Information in the old times…

Simple to access…

but limited (only documents physically in the library)...

and time consuming...

Information in the old times…

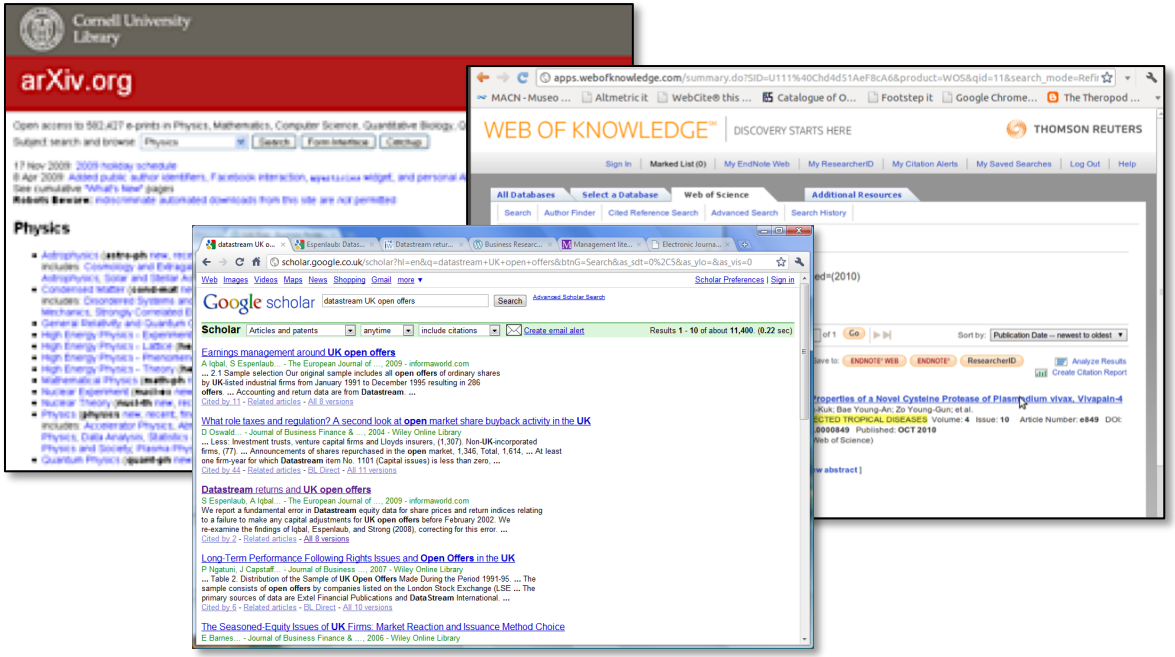
Simple to access…

but limited (only documents physically in the library)...

and time consuming...

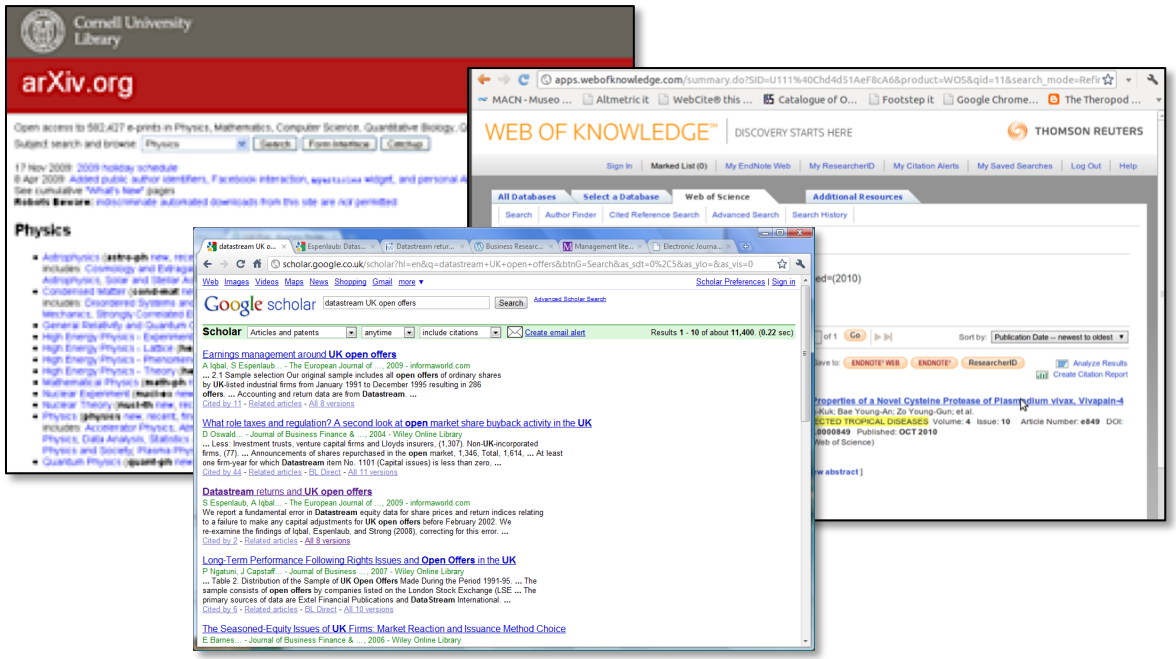
Although with some advantages…

… and in present times.



Navigate from document to document by a click

… and in present times.



Although some things never change…

Too much information can be a problem…

"Global scientific output doubles every nine years."
Richard van Noorden, Nature, 5/2014

At some point, one gets lost in the "information deluge".

We need smarter ways to navigate…

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 creator

Richard van Noorden, Nature, 11/2014

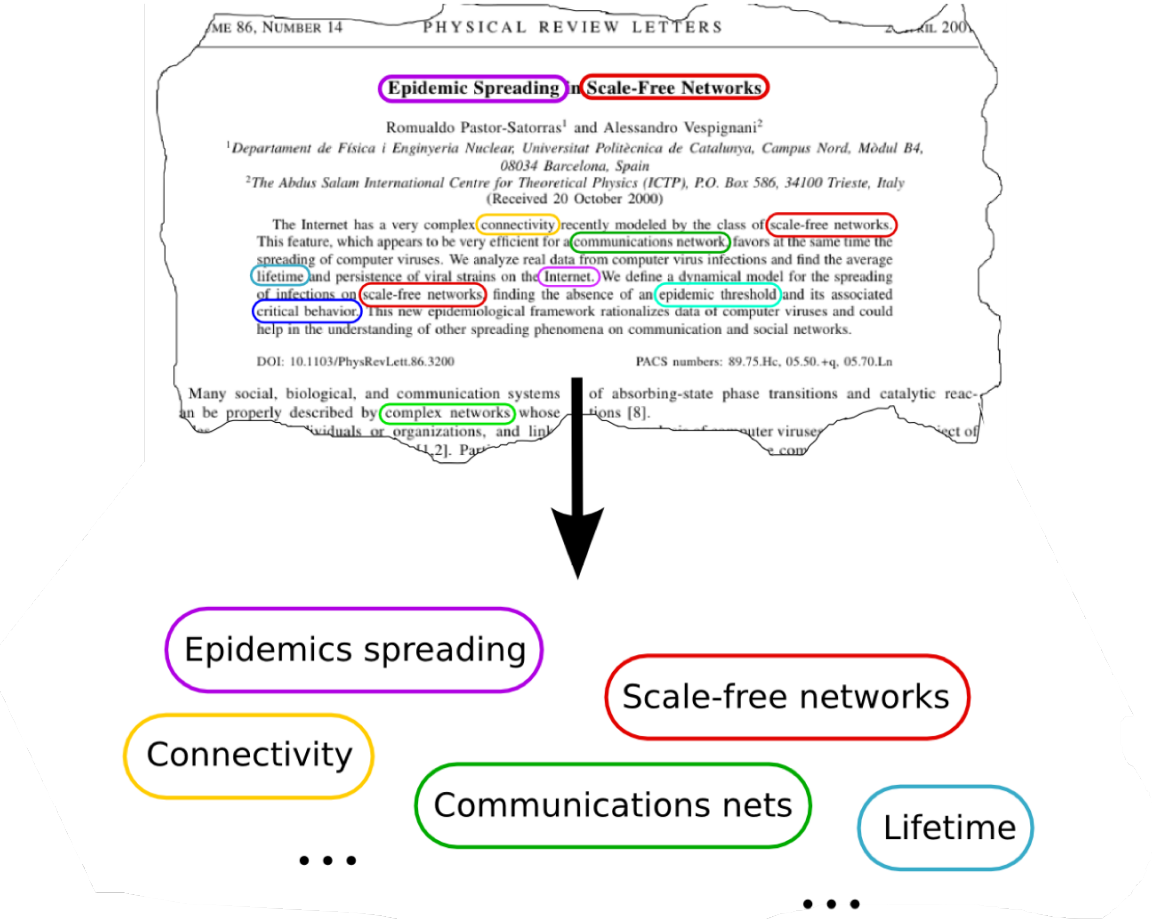
Semantic Scholar offers a few innovative features, including picking out the most important keywords and phrases from the text without relying on an author or publisher to key them in. "It's surprisingly difficult for a system to do this."

Oren Etzioni, CEO of AI2 (Semantic Scholar)

Nicola Jones, Nature 11/2015

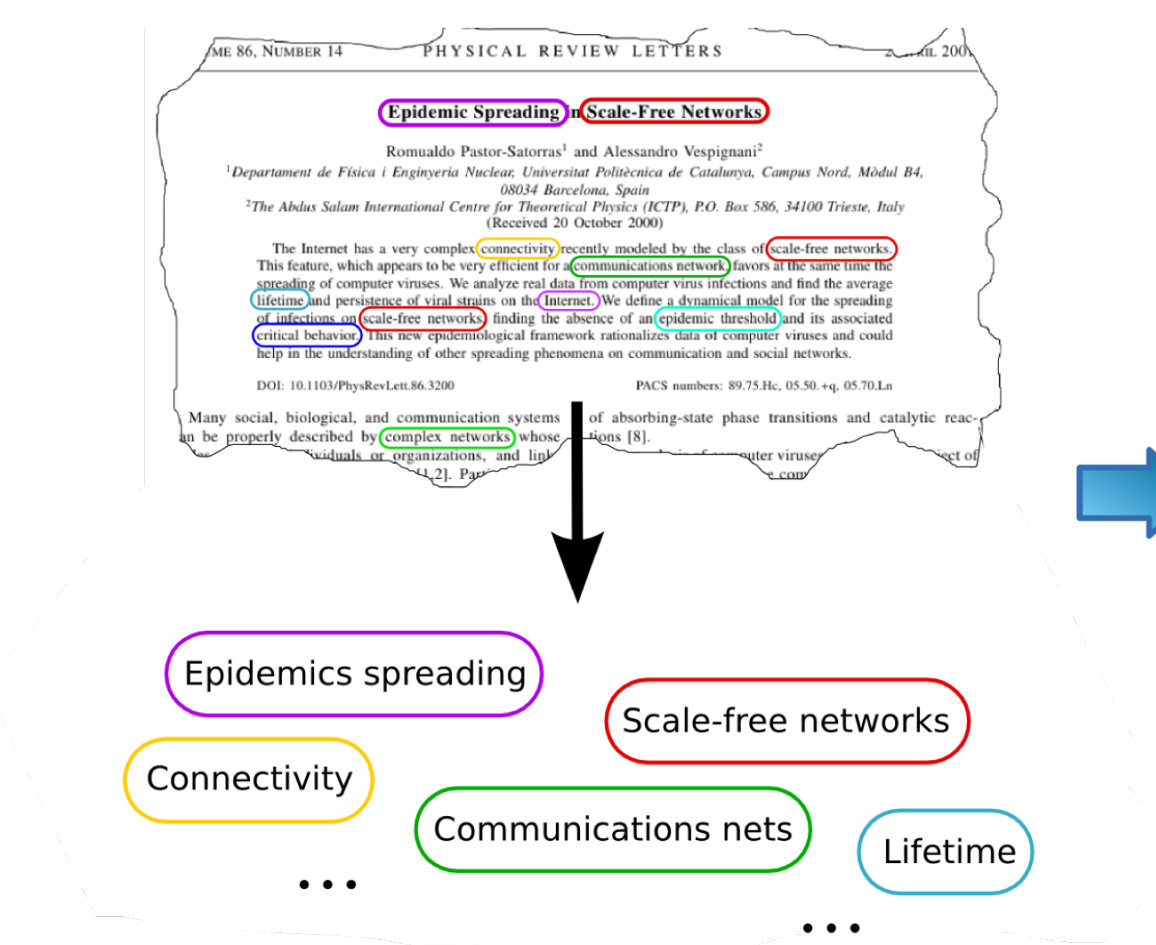
Of course no single route…

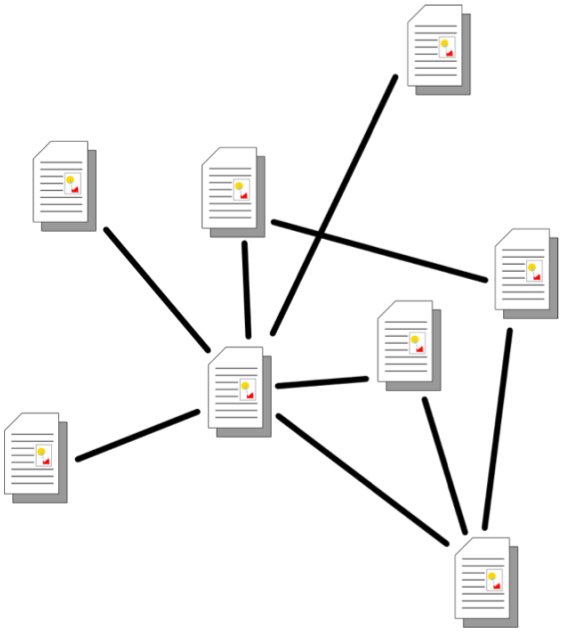
Here we focus on concepts…



Of course no single route…

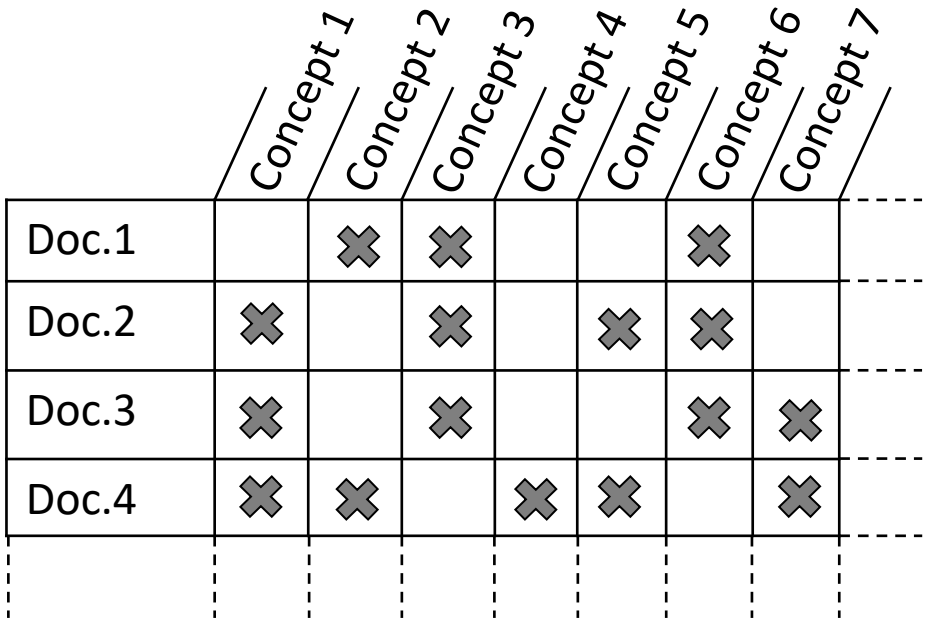
Here we focus on concepts…





Similarity network

How do we build the network?



Concepts are either present or absent in a document

The number of common concepts marks the similarity of two documents

Although a concept might be present in two documents, it might have different relevance in each of them

How do we build the network?

	Concept 1	Concept 2	Concept 3	Concept 4	Concept 5	Concept 6	Concept 7	
Doc.1		3	12			7		
Doc.2	2		99		3	10		
Doc.3	1		63			57	6	
Doc.4	45	34		3	2		7	

The number of times a concept appears in a document is the term frequency: $TF_\alpha(i)$

$TF_\alpha(i)$ = # of times concept i appears in document α

We might use this more detailed quantity…

The TF-IDF

What most scholars in the field of document science use is the TF-IDF

$$TF\text{-}IDF_{\alpha}(i) = TF_{\alpha}(i) * Log(N_{doc}/N_i)$$

where the second term is the IDF of concept i:

the (logarithm of the) inverse document frequency of concept i

TF increases the relevance of a concept in papers where it is heavily used (H.P. Luhn, 1957)

IDF decreases the relevance of concepts that appear in too many papers
(K. Spärck Jones, 1972)

The TF-IDF vector

$\underline{v}_\alpha$ = (TF-IDF(1), TF-IDF(2), …)

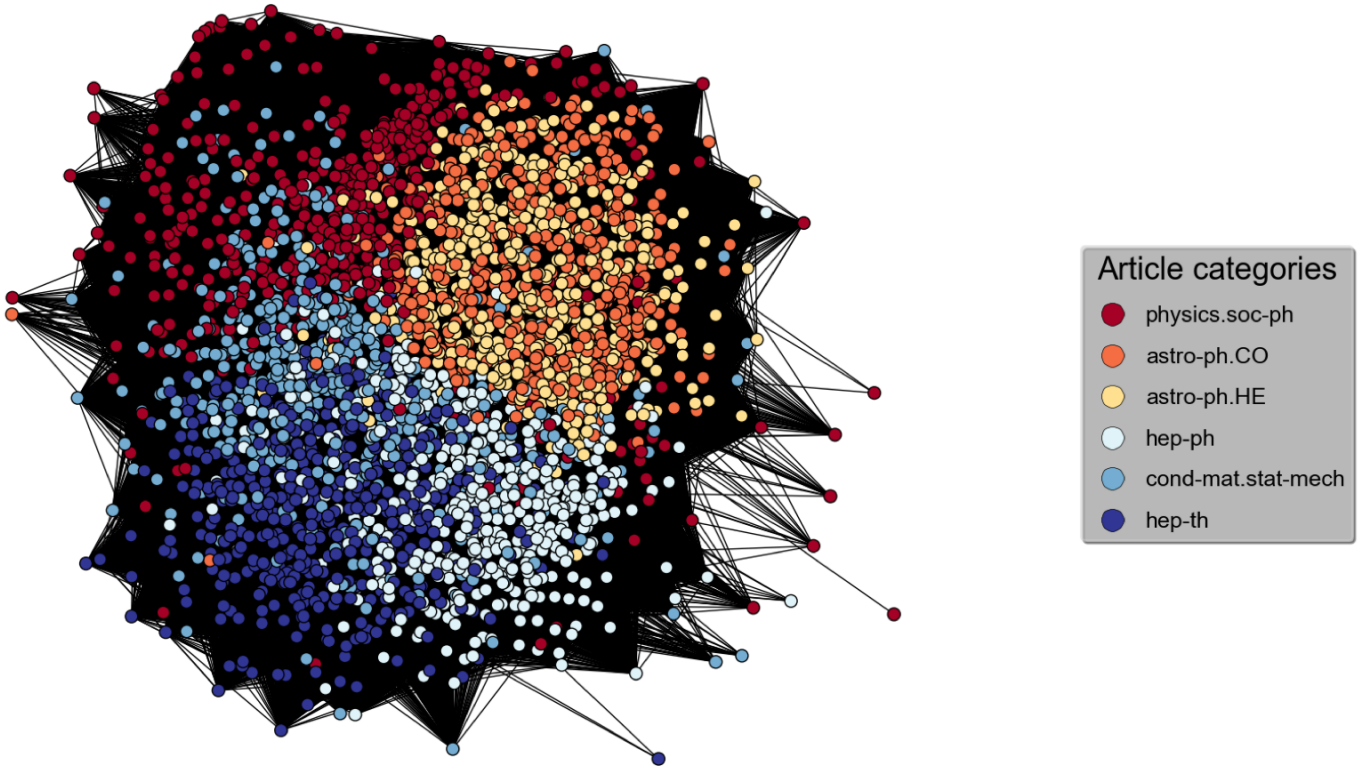
It contains the TF-IDF of the concepts belonging to document α

The similarity $W_{\alpha\beta}$ between two documents is computed by taking the scalar product of their normalized TF-IDF vectors ("cosine similarity")

$$W_{\alpha\beta} = \underline{V}_\alpha \cdot \underline{V}_\beta / (||\underline{V}_\alpha|| \, ||\underline{V}_\beta||)$$

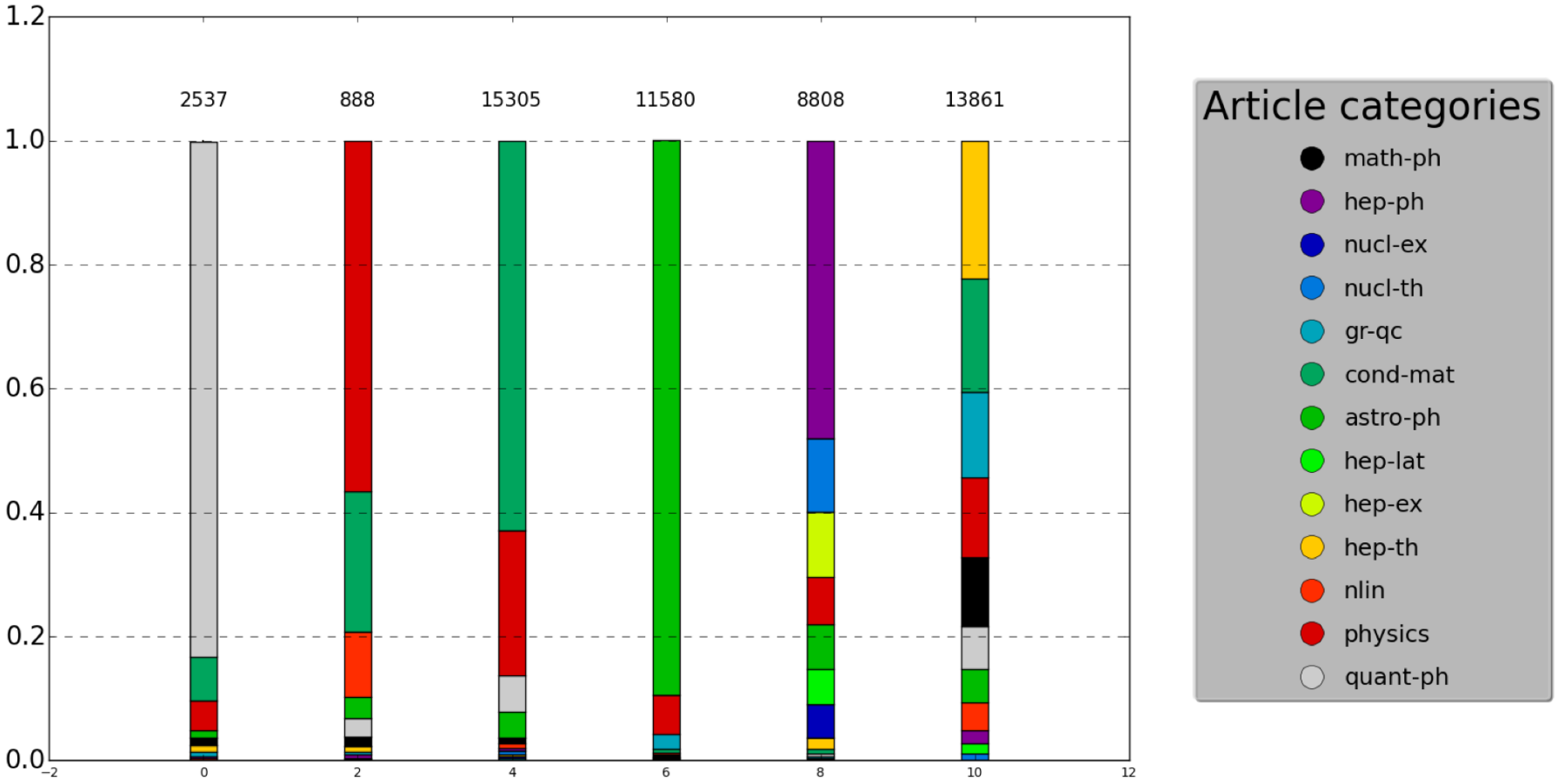
The network and its features: Physics 2013 from ArXiv

$N_{Concepts}$	$N_{Articles}$	<k>	k_{max}
10661	52979	19333.522	46504



Community detection

(modularity à la Louvain algorithm)



It ain’t that bad… but… Why the mixtures?

Problems with community detection

The network is almost a complete graph!
This is already not too good.

Furthermore: highly mixed communities can be due to too much “glue” (common concepts) or to concepts that are shared by different physics domains…

We need to sparsify the network

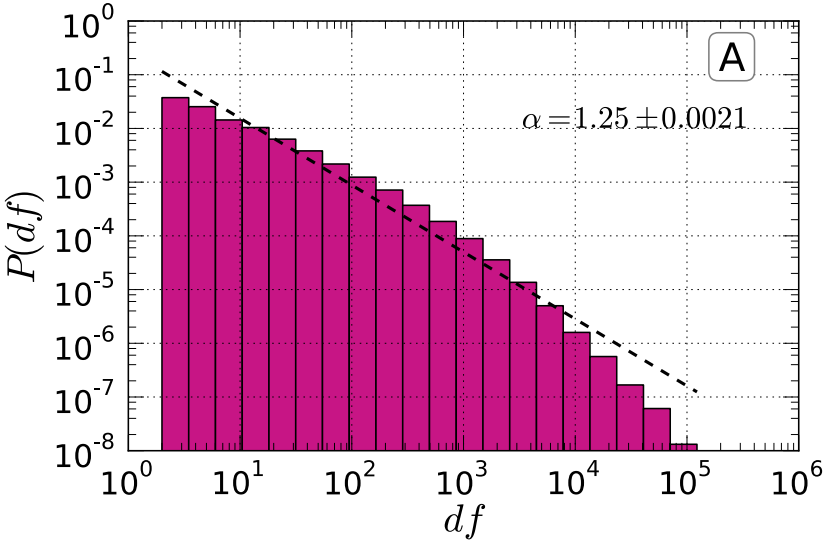
Various approaches:

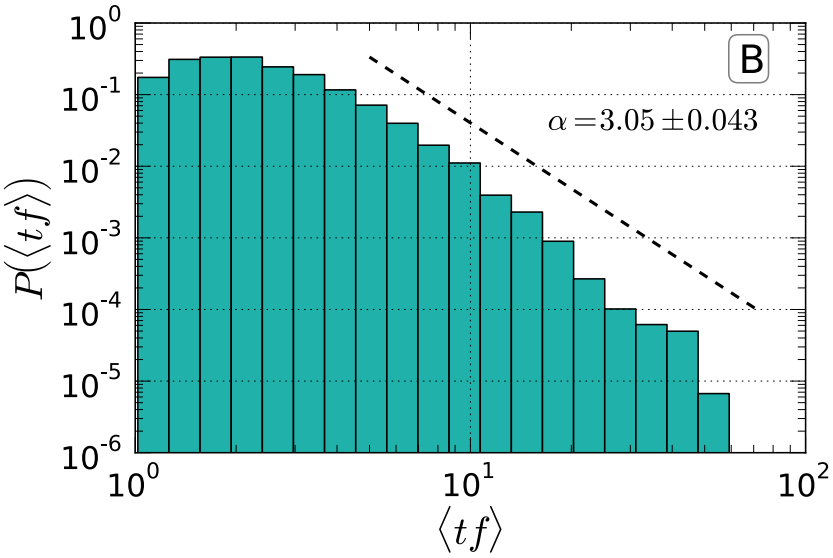
Filter with a threshold on edge weights

Filter with a threshold on Term Frequency

Filter with a threshold on Document Frequency

How do TF and DF look over our set?





Not surprisingly: broad distribution

Consequently… thresholds rather meaningless

Filtering out common concepts

Idea: why not looking for concepts that
are very common in the document set?

These concepts might be good to characterize the set as a whole, but not to tell sub-communities one from the other…

How to automatically detect them?

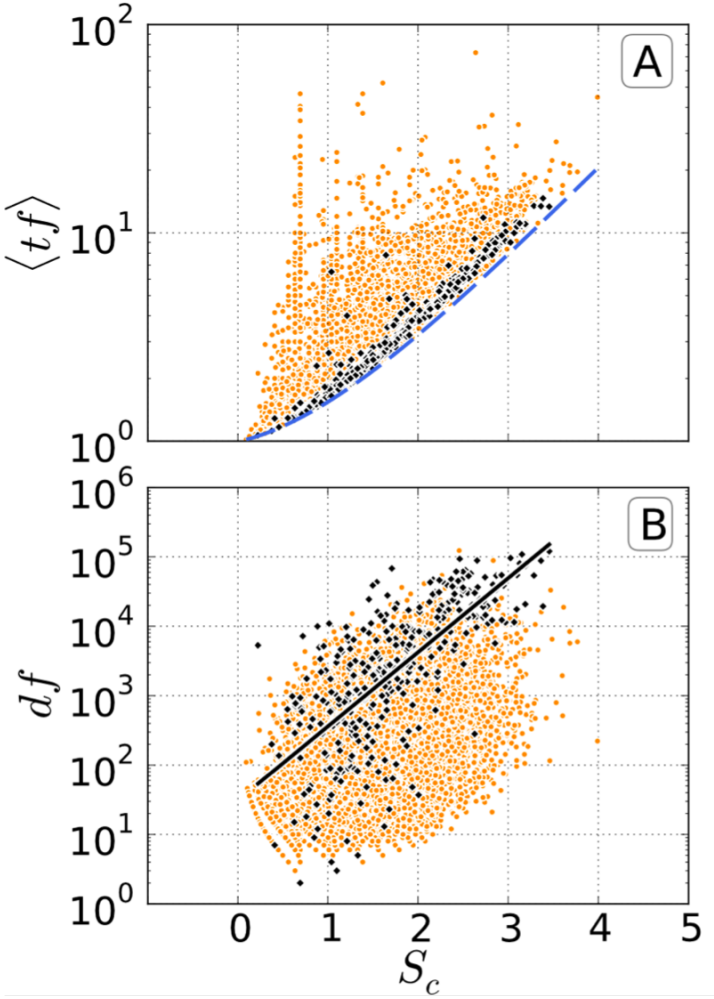
Looking at entropies…

We define the entropy of a concept as follows:

$$S_C = \sum p_k \ln p_k$$

where p_k is the fraction of documents where concept C appears k times (TF = k).

Looking at entropies…

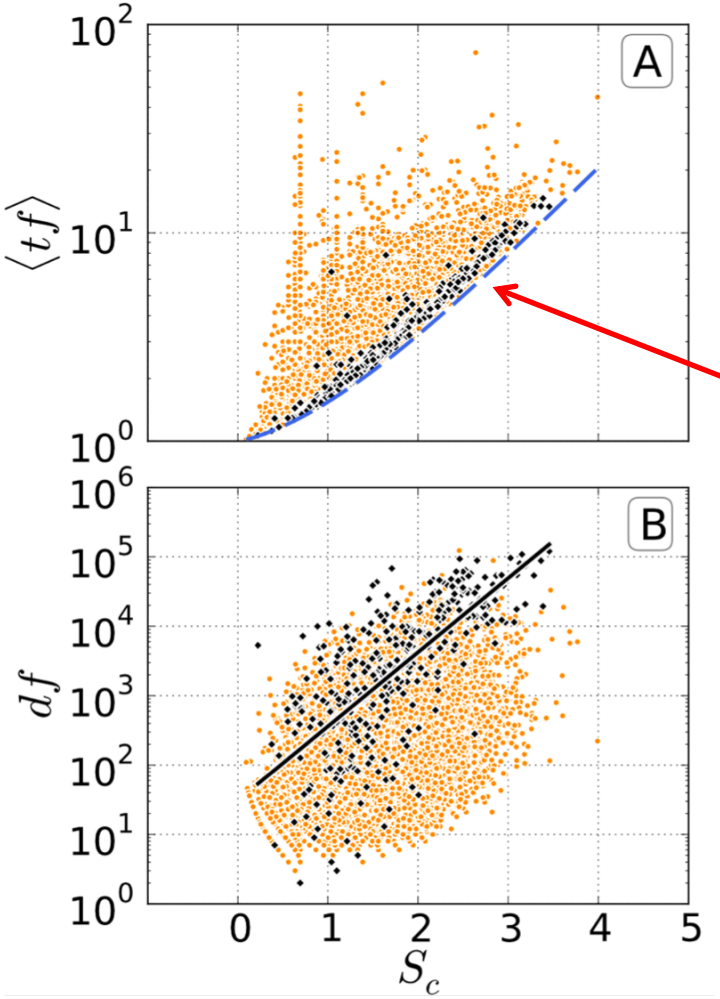


Black circles: supervised common concepts

Common concepts correlate with the “hull” of the scatter plot of the entropy *vs*. the average TF

Remarkably, common concepts do not trivially correlate with their document frequency.

Looking at entropies…



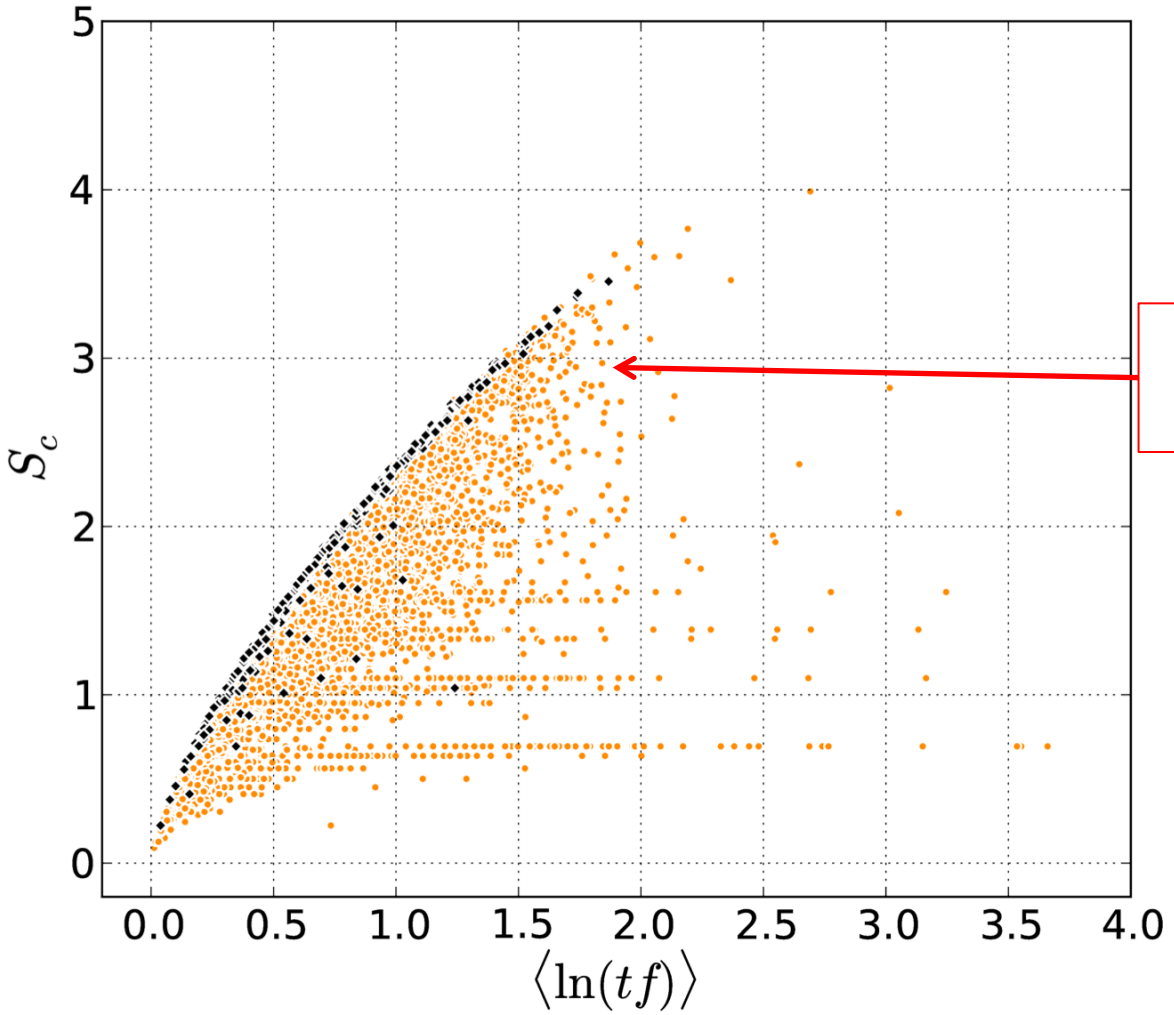
Black circles: supervised common concepts

Common concepts correlate with the “hull” of the scatter plot of the entropy *vs*. the average TF

Looks like there is a maximum possible entropy!

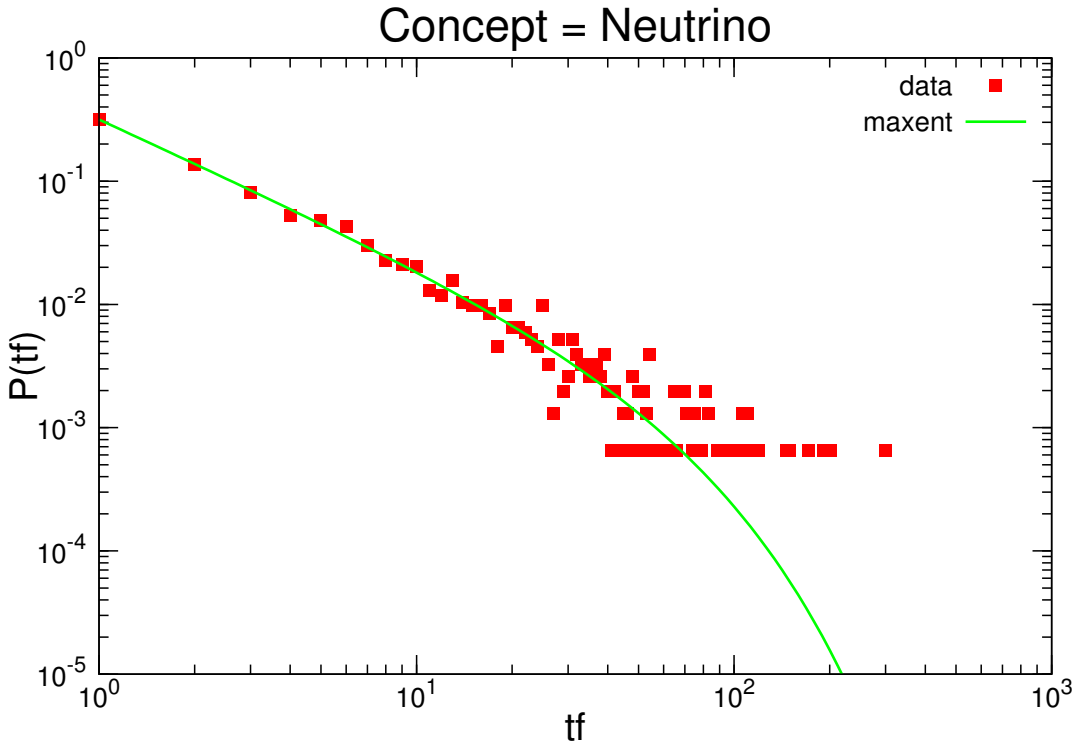
Remarkably, common concepts do not trivially correlate with their document frequency.

Looking at entropies…



Looks like there is a maximum possible entropy!

Shape of TF distributions



By inspection, the distribution looks like a power-law with cutoff

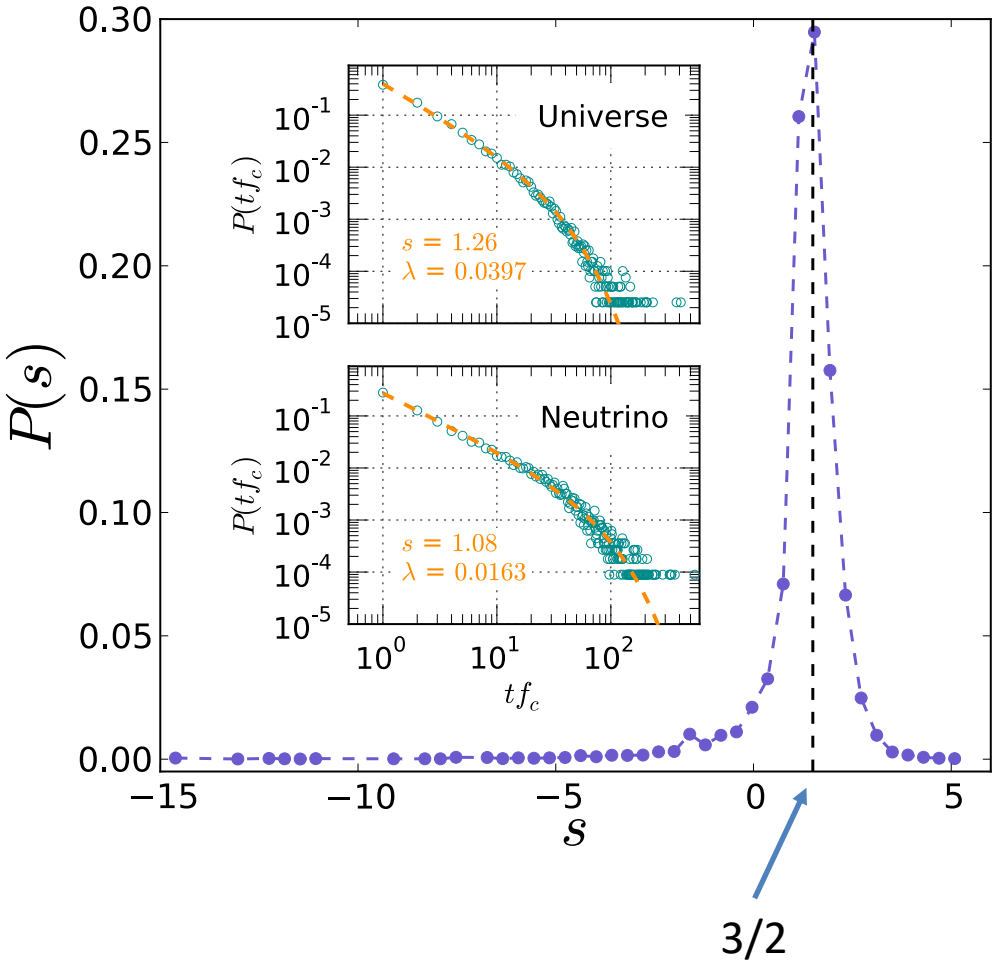
So… maximum entropy principle with two constraints:

<TF> <ln(TF)>

Nicely provides a power-low with cutoff.

The green-curve is the max-ent distribution

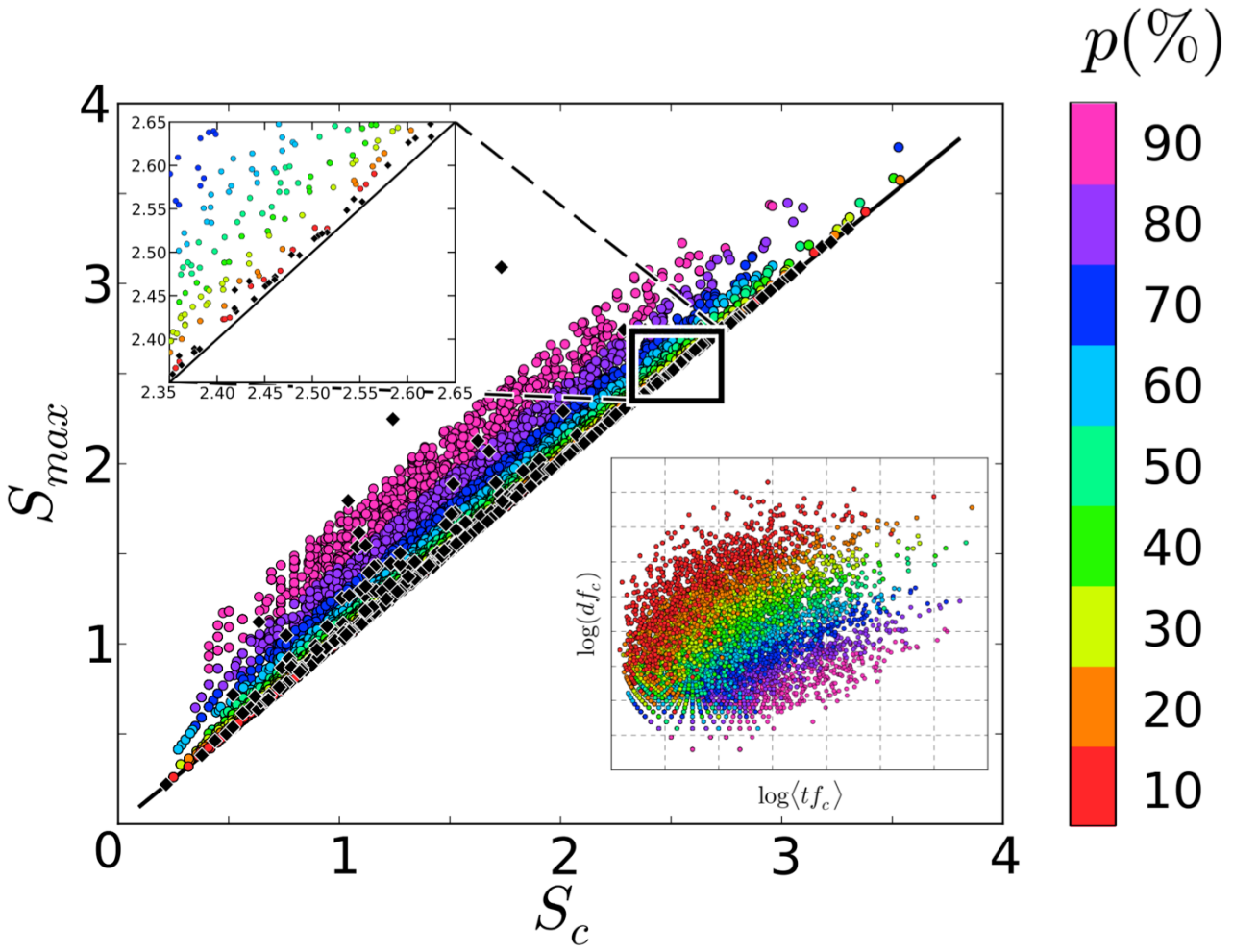
A curious interlude



$$p_k = k^{-s}\, e^{-\lambda k}$$

S=3/2 is reminiscent
of critical branching processes…

$S_{max\text{-}ent}$ vs. S_{data}

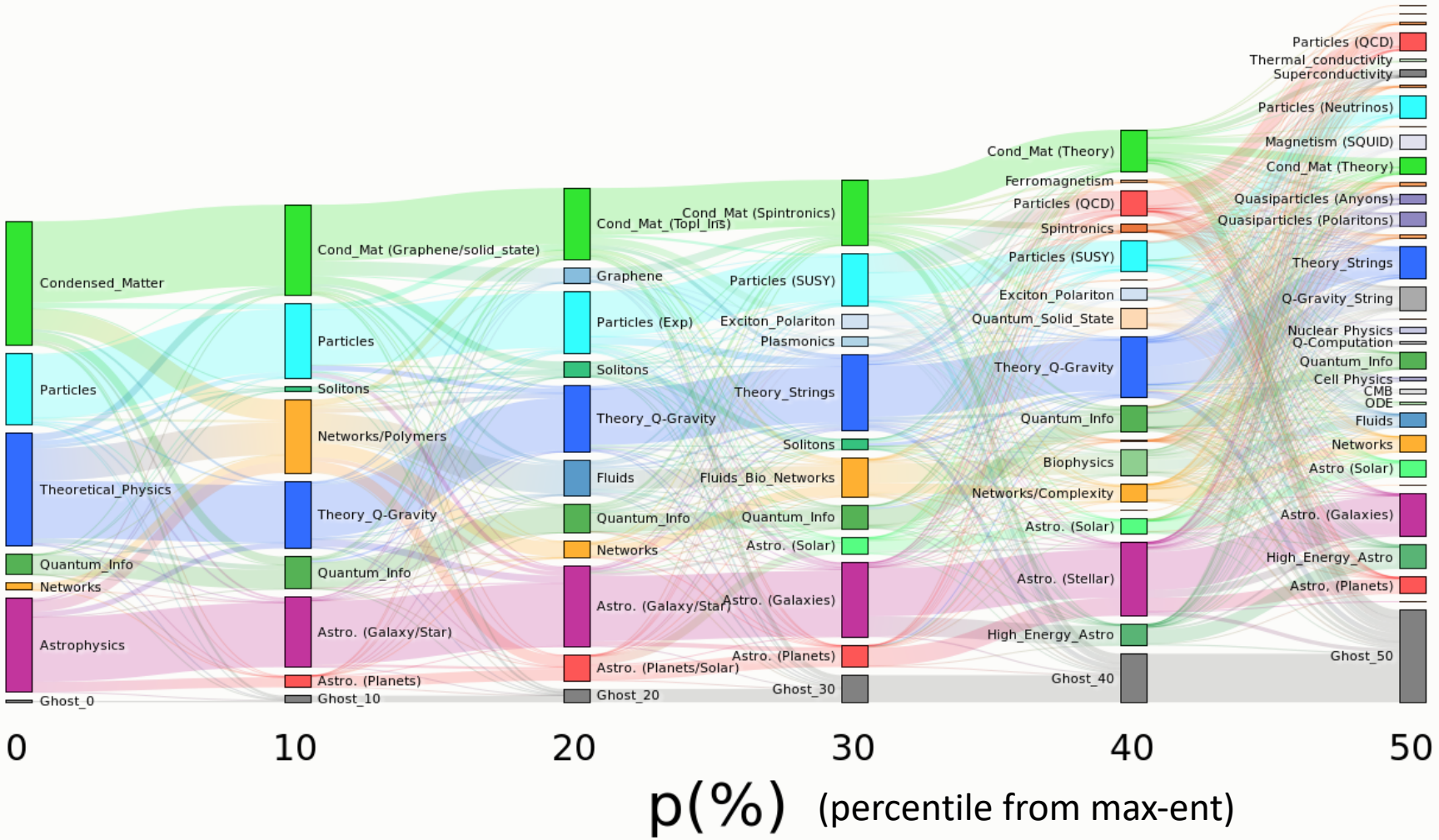


Most interestingly:
concepts hand-tagged as common (black dots) do show max-ent!!!

The next step is thus to filter-out concepts, progressively, given their proximity to their ideal $S_{max\text{-}ent}$

Progressive filtering

Data: Phys2013 $w_{min} = 0.01$



Conclusions

The great load of information available at a “fingertip” is a blessing, but comes with the “too much information” problem. How to get to the relevant one?

Here we propose to tackle the issue at the root: curing the cause of the problem (here, common concepts) rather than the effect.

Need for:

automatic detection methods for common concepts (entropy?)

fast and reliable classification techniques (community detection? Topic modeling?)

constant (but not complete) human validation?

Acknowledgments

EPFL
Karl Aberer
Alex Constantin

Leiden University
Alexei Boyarsky
Diego Garlaschelli

Fribourg University
Philippe Cudre-Mauroux

Niels Bohr Institute
Oleg Ruchayskiy

Funding:
Swiss National Science Foundation