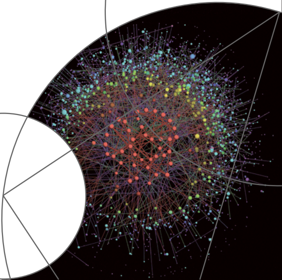
Entropic selection of concepts in networks of similarity between documents

Alessio Cardillo

Laboratory for Statistical Biophysics (LBS)

École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland

http://bifi.es/~cardillo/



ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Acknowledgements

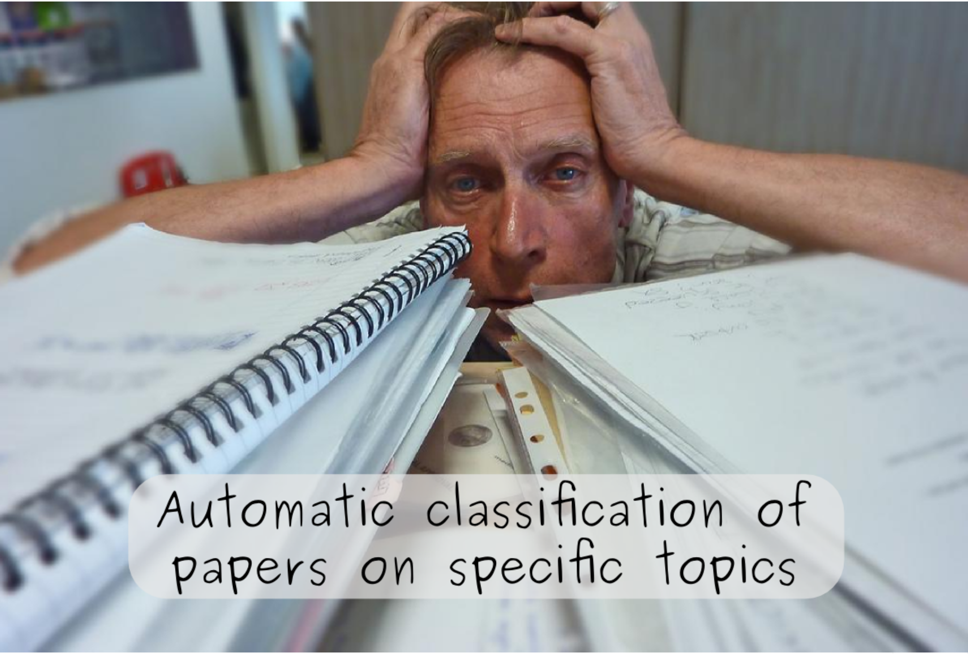
Main Collaborators

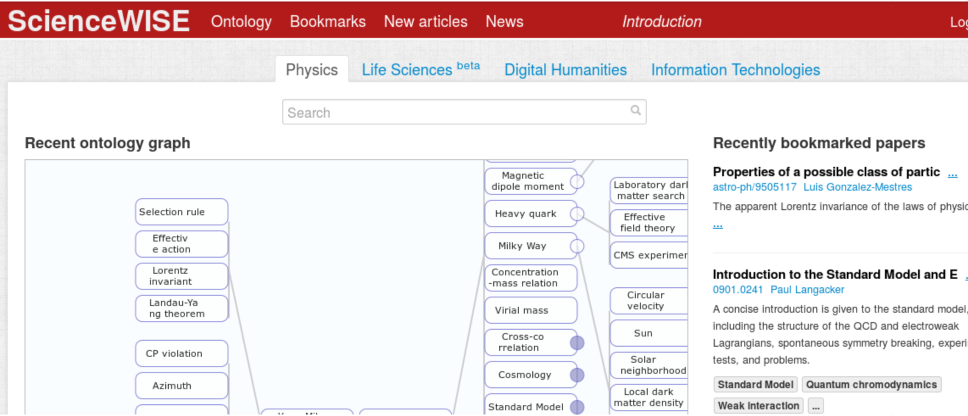
- Paolo De Los Rios (EPFL)
- Andrea Martini (EPFL)

Other Collaborators

- Vasyl Palchykov, Valerio Gemmetto, Diego Garlaschelli (Leiden - NL)
- ScienceWISE Team (Lausanne)

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE


Automatic classification of
papers on specific topics



http://sciencewise.info

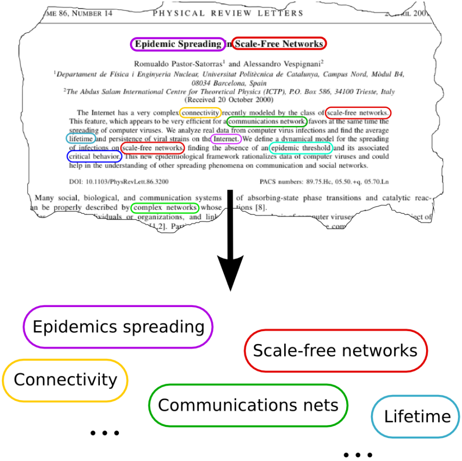
Outline

- Introduction on similarity networks.
- ★ Filtering of weighted networks.
- ★ Entropic filtering of concepts.
- ★ Results.
- Take home messages
- Questions

Section 1

Similarity networks

Networks of similarity between papers


ME 86, NUMBER 14 PHYSICAL REVIEW LETTERS

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

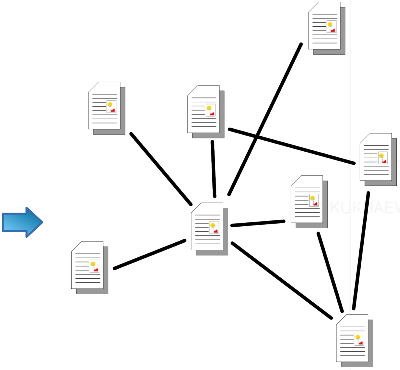
[2]The Abdus Salam International Centre for Theoretical Physics (ICTP), P.O. Box 586, 34100 Trieste, Italy

(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 ... of absorbing-state phase transitions and catalytic reactions [8].



Similarity network

TF-IDF and similarity

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

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

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

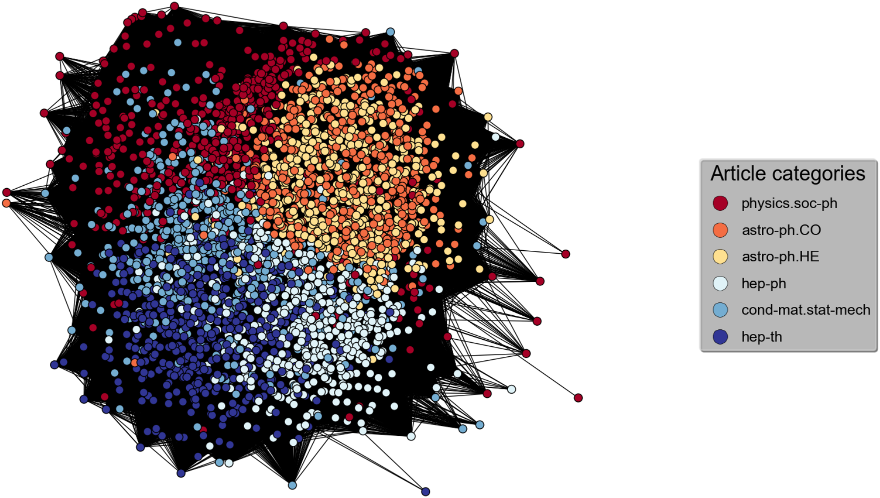
Edge weight

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

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

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

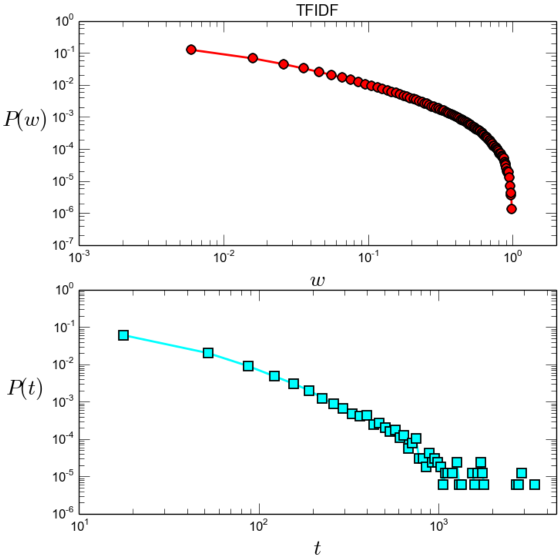
α	2	43	0	18	0	11	27
β	13	5	9	0	30	0	0
	C_1	C_2	C_3	C_4	C_5	C_6	C_7


Article categories
physics.soc-ph
astro-ph.CO
astro-ph.HE
hep-ph
cond-mat.stat-mech
hep-th

Section 2

Filtering

Available methods



Available 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

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

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Available methods


PNAS Celebrating 100 Years 1915-2015

Institution: EPFL

Proceedings of the National Academy of Sciences of the United States of America

CURRENT ISSUE // ARCHIVE // NEWS & MULTIMEDIA // FOR AUTHORS // ABOUT PNAS // COLLECTED ARTICLES / BROWSE BY TOPIC / EARLY EDITION

> Current Issue > vol. 106 no. 16 > M. Ángeles Serrano, 6483–6488, doi: 10.1073/pnas.0808904106

CrossMark click for updates

Extracting the multiscale backbone of complex weighted networks

M. Ángeles Serrano[a,1], Marián Boguñá[b] and Alessandro Vespignani[c,d]

Author Affiliations

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved March 2, 2009 (received for review September 9, 2008)

Abstract | Full Text | Authors & Info | Figures | SI | Metrics | Related Content | +SI

Abstract

This Issue

April 21, 2009
vol. 106 no. 16
Masthead (PDF)
Table of Contents

PREV ARTICLE | NEXT ARTICLE

View this article with LENS beta

Don't Miss

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

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Relevant concepts

Relevant concepts

Key features

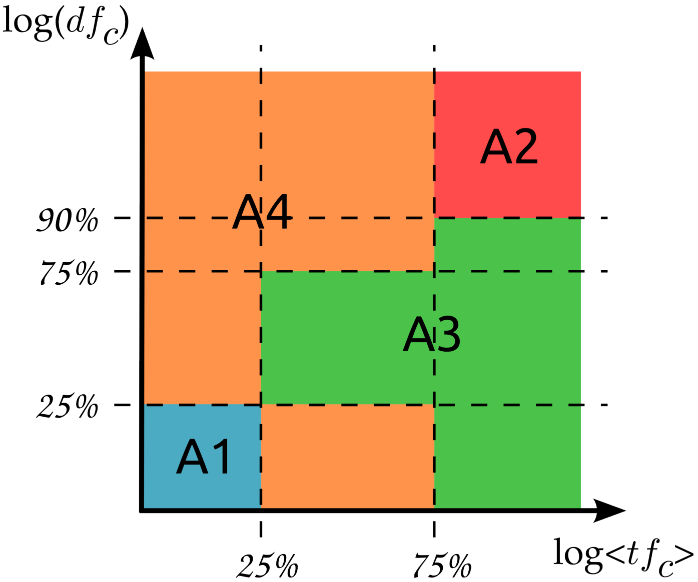
- # of papers a concept appears in

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

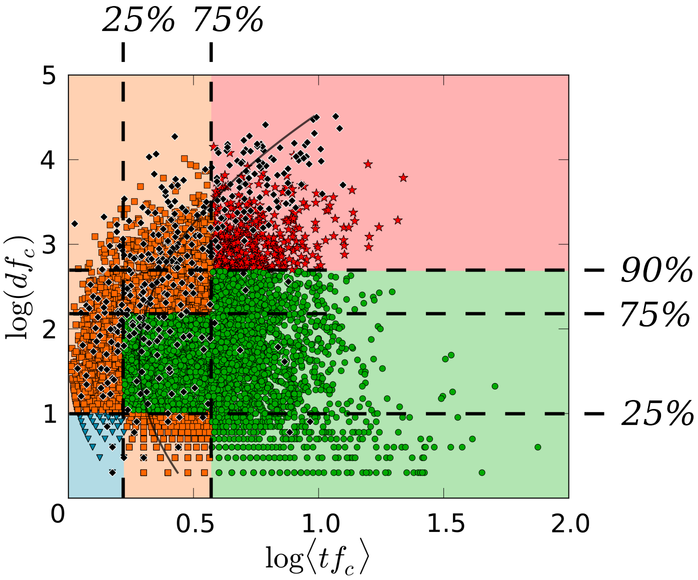
- average # of times a concept appears inside a paper

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

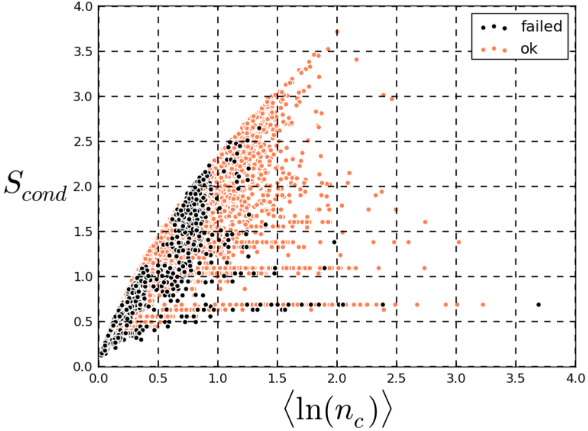
Bidimensional tessellation



Bidimensional tessellation



Maximum entropy



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

Maximum entropy

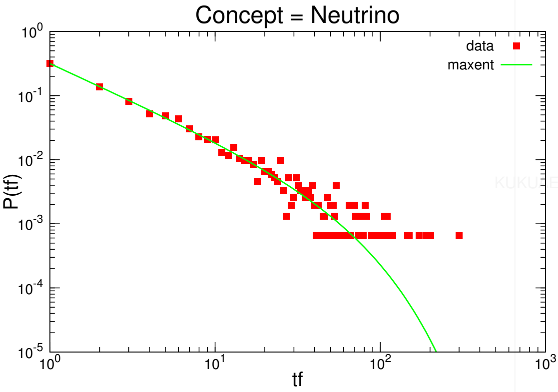
$$\sum_n p_n = 1$$

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

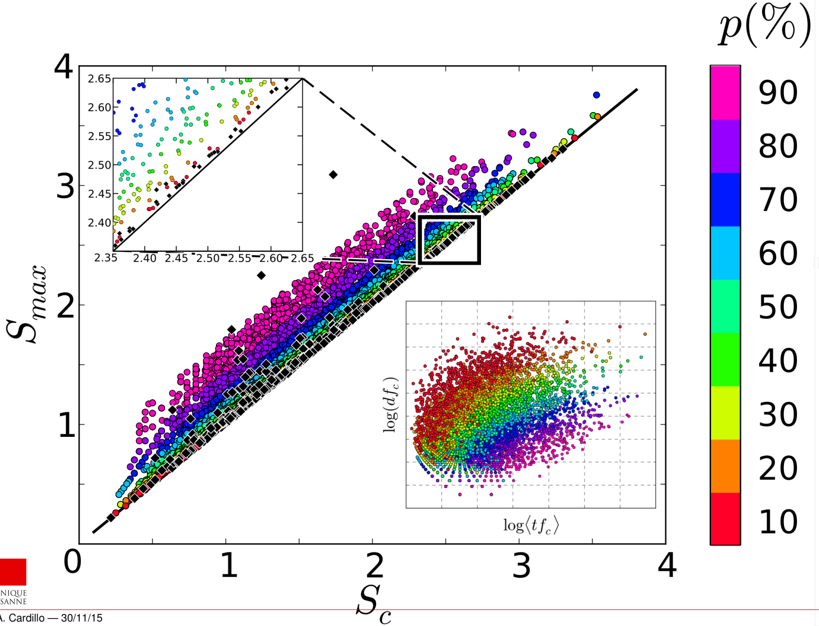
$$\sum_n p_n \ln n = \langle \ln n \rangle$$

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

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



Maximum entropy



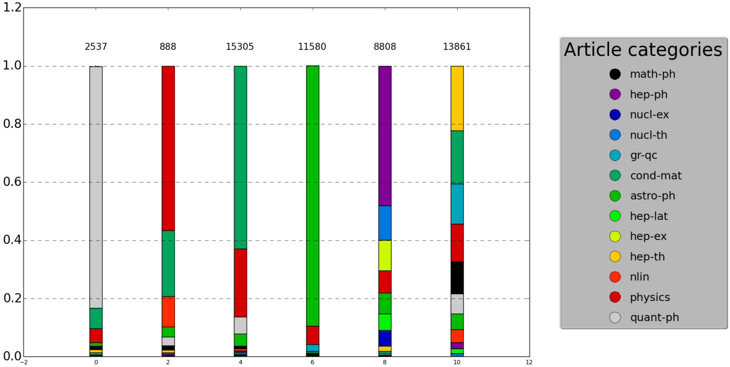
Section 3

Results

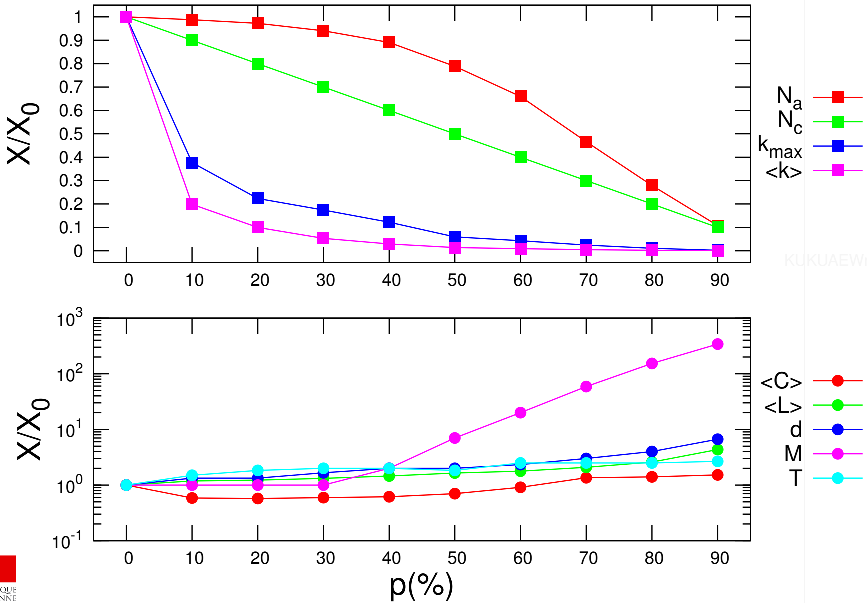
Topological properties

Original network

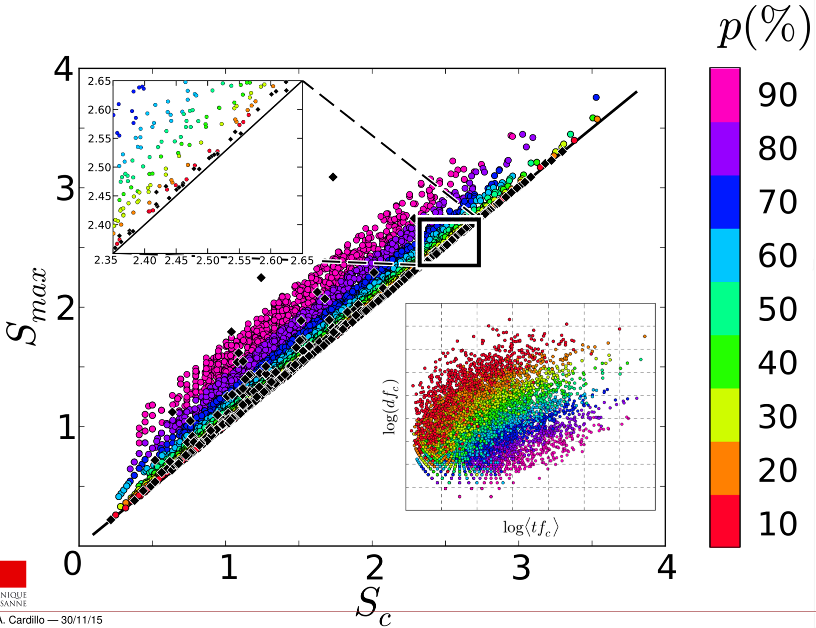
N_c	N_a	$\langle k \rangle$	k_{max}	$\langle C \rangle$	d	$\langle L \rangle$	M	T
10661	52979	19333.522	46504	0.557	3	1.635	1	6



Topological properties



Generic concepts



Community detection

arXiv.org > astro-ph > arXiv:1301.0319

Search or Article-id

Astrophysics > Solar and Stellar Astrophysics

Modules for Experiments in Stellar Astrophysics (MESA): Giant Planets, Oscillations, Rotation, and Massive Stars

Bill Paxton, Matteo Cantiello, Phil Arras, Lars Bildsten, Edward F. Brown, Aaron Dotter, Christopher Mankovich, M. H. Montgomery, Dennis Stello, F. X. Timmes, Richard Townsend

(Submitted on 2 Jan 2013 (v1), last revised 12 Jun 2013 (this version, v2))

We substantially update the capabilities of the open source software package Modules for Experiments in Stellar Astrophysics (MESA), and its one-dimensional stellar evolution module, MESA Star. Improvements in MESA Star's ability to model the evolution of giant planets now extends its applicability down to masses as low as one-tenth that of Jupiter. The dramatic improvement in asteroseismology enabled by the space-based Kepler and CoRoT missions motivates our full coupling of the ADIPLS adiabatic pulsation code with MESA Star. This also motivates a numerical recasting of the Ledoux criterion that is more easily implemented when many nuclei are present at non-negligible abundances. This impacts the way in which MESA Star calculates semi-convective and thermohaline mixing. We exhibit the evolution of 3-8 Msun stars through the end of core He burning, the onset of He thermal pulses, and arrival on the white dwarf cooling sequence. We implement diffusion of angular momentum and chemical abundances that enable calculations of rotating-star models, which we compare thoroughly with earlier work. We introduce a new treatment of radiation-dominated

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Community detection

arXiv.org > astro-ph > arXiv:1306.2314

Search or Article-id

Astrophysics > Cosmology and Nongalactic Astrophysics

Warm Dark Matter as a solution to the small scale crisis: new constraints from high redshift Lyman-alpha forest data

M. Viel, G.D. Becker, J.S. Bolton, M.G. Haehnelt

(Submitted on 10 Jun 2013 (v1), last revised 26 Aug 2013 (this version, v2))

We present updated constraints on the free-streaming of warm dark matter (WDM) particles derived from an analysis of the Lya flux power spectrum measured from high-resolution spectra of 25 z > 4 quasars obtained with the Keck High Resolution Echelle Spectrometer (HIRES) and the Magellan Inamori Kyocera Echelle (MIKE) spectrograph. We utilize a new suite of high-resolution hydrodynamical simulations that explore WDM masses of 1, 2 and 4 keV (assuming the WDM consists of thermal relics), along with different physically motivated thermal histories. We carefully address different sources of systematic error that may affect our final results and perform an analysis of the Lya flux power with conservative error estimates. By using a method that samples the multi-dimensional astrophysical and cosmological parameter space, we obtain a lower limit mwdm > 3.3 keV (2sigma) for warm dark matter particles in the form of early decoupled thermal relics. Adding the Sloan Digital Sky Survey (SDSS) Lya flux power spectrum does not improve this limit. Thermal relics of masses 1 keV, 2 keV and 2.5 keV are disfavoured by the data at about the 9sigma, 4sigma and 3sigma

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Community detection

p	**10**	**20**	**30**	**40**	**50**
Article					
	Galaxy	Planet	Planet	White dwarf	Andromeda galaxy
1301.0319	Accretion	Eccentricity	Eclipses	Initial mass function	Large Magellanic Cloud
	Stellar mass	Jupiter	Albedo	Globular cluster	Dark matter subhalo
	Extinction	Giant planet	Natural satellite	Eclipses	Maser
	Supernova	Magneto hydrodynamics	Terrestrial planet	Open cluster	Planetary nebula
	Galaxy	Supernova	Active Galactic Nuclei	White dwarf	Andromeda galaxy
1306.2314	Accretion	Spectral energy distribution	Quasar	Initial mass function	Large Magellanic Cloud
	Stellar mass	Neutron star	Pulsar	Globular cluster	Dark matter subhalo
	Extinction	Star formation rate	Point spread function	Eclipses	Maser
	Supernova	Active Galactic Nuclei	White dwarf	Open cluster	Planetary nebula

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Section 4

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.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

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 . . .

Reference

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

http://bifi.es/~cardillo/

alessio.cardillo@epfl.ch