TRACKING SCIENCE OVER TIME IN THE NEWS PRESS TROUGH TOPIC MODELING AND NETWORK ANALYSIS

Carlos G. Figuerola [**figue@usal.es**]
University of Salamanca - Spain

VIII STS Italia Conference, 2021

INTRO

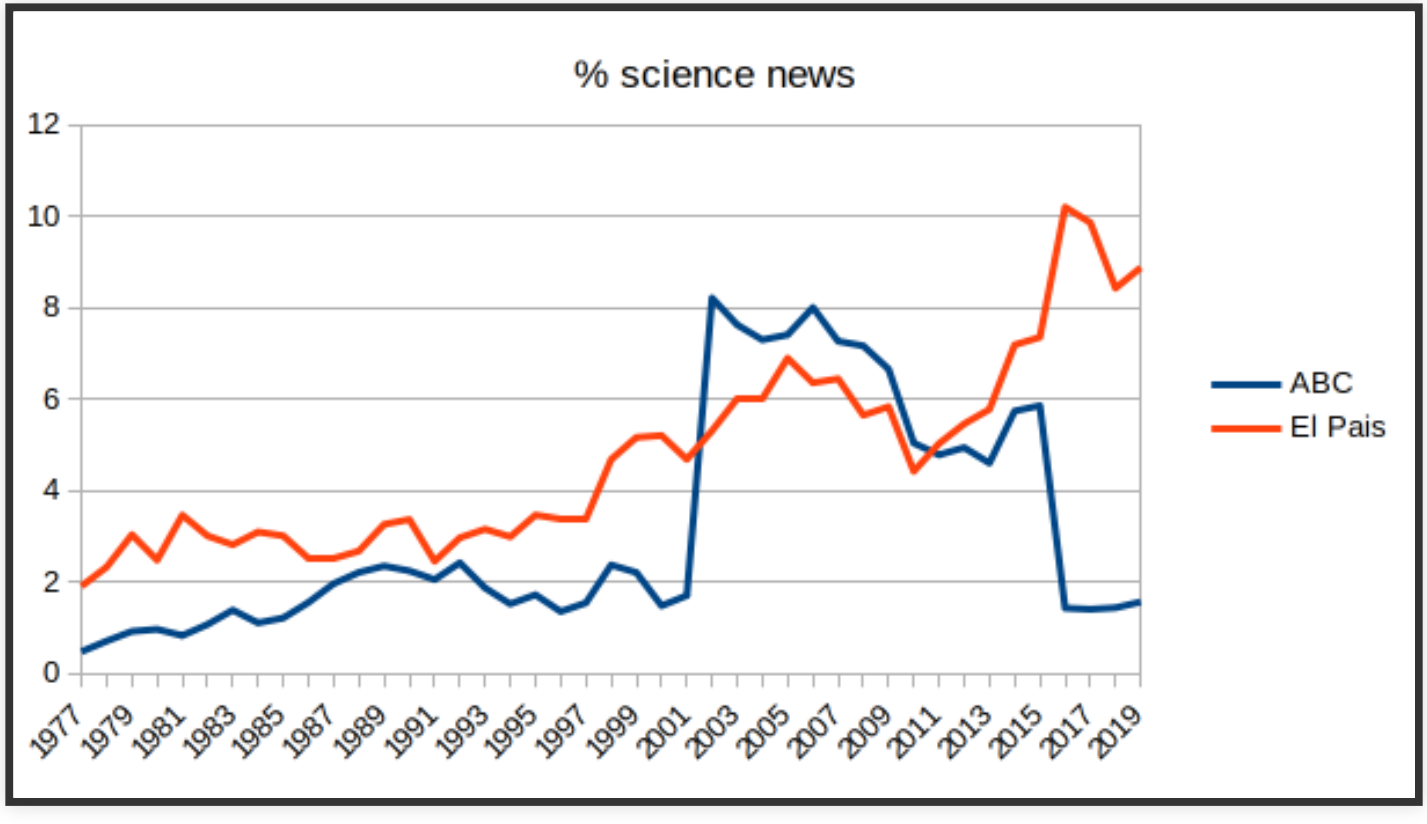
- digital newspaper: a lot of available information
- it is possible to download, clean, process and mining such information
- the aim of this work is to illustrate the use of some techniques for mining and analyzing news about science

DATA SOURCES

- 2 newspaper, **ABC** (conservative) and **El País** (progressive)
- from 1977 to 2019
- ABC: physical pages passed trough an scanner
  - dirty text, fuzzy and less acurate results

SCIENCE NEWS

- science news located and selected by an automatic classifier
  - requires supervised training



TOPIC MODELING

- Topic Modeling is a set of techniques that aim to detect, characterize and cuantify topics inside a collection of documents
- several techniques
- LDA the most used
  - performs well
  - tools available, some of them of easy use

LDA

- assumes a set of topics inside a collection of docs (or news about science)
- assumes that every doc has an specific amount of each topic (including 0 %)
- requires from the user the number of desired topics to detect

THE NUMBER OF TOPICS

- this can be done trough:
  - expertise of the user
  - automatic test optimizing specific metrics to determine optimum number
  - manual test, error check and retry
- in our case, automatic test sugest 51 topics

WHAT WE GET

a list of words intending to debscribe each topic

0	gen adn genoma genetico biologia proteinas celula proteina ratones molecular celular geneticas mutaciones nature geneticos descubrimiento mutacion mecanismos mecanismo cromosoma / **gene dna genetic genome biology proteins cell protein mice molecular cellular genetic mutations nature genetic discovery mutation mechanisms mechanism chromosome**
1	vino aceite carne sabor comida pan vinos pescado verduras leche ingredientes frutas platos plato sal fruta gramos arroz huevos patatas / **wine oil meat flavor food bread wines fish vegetables milk ingredients fruit dishes plate salt fruit grams rice eggs potatoes**
2	alumnos estudiantes profesores curso campus matematicas universitaria universitarios escuela universitario ensenanza cursos facultad politecnica clases autonoma carrera clase rector complutense / **students students professors course campus mathematics university university university school teaching courses polytechnic faculty autonomous classes career rector class Complutense**
3	hielo calentamiento clima antartida artico calor oceano carbono capa invernadero gases terremotos terrestre fenomeno glaciares fenomenos oceanos polo modelos erupcion / **ice warming climate antarctica arctic heat ocean carbon layer greenhouse gases earthquakes terrestrial phenomenon glaciers phenomena oceans pole models eruption**

TOPICS AND WORDS

topics can be grouped in meta-topics, in view of easy management

0	gen adn genoma genetico biologia proteinas celula proteina ratones molecular celular geneticas mutaciones nature geneticos descubrimiento mutacion mecanismos mecanismo cromosoma / **gene dna genetic genome biology proteins cell protein mice molecular cellular genetic mutations nature genetic discovery mutation mechanisms mechanism chromosome**
5	dieta obesidad diabetes alcohol grasa grasas nutricion colesterol leche comer ejercicio calorias azucar habitos vitamina bebidas comida sobrepeso proteinas ingesta / **diet obesity diabetes alcohol fat fats nutrition cholesterol milk eat exercise calories sugar habits vitamin drinks food overweight protein intake**
41	dolor sueno depresion cerebral sintomas estres mental trastornos alzheimer memoria ansiedad sindrome neuronas cerebrales trastorno perdida nervioso mentales afecta sufren / **pain sleep brain depression symptoms mental stress alzheimer's disorders memory anxiety syndrome brain neurons disorder nerve loss mental affects suffer**

WHAT WE GET

- the amount of every topic inside each of the docs
  - this is given in a normalized scale (0 to 1, usually)

TOPICS AND DOCS

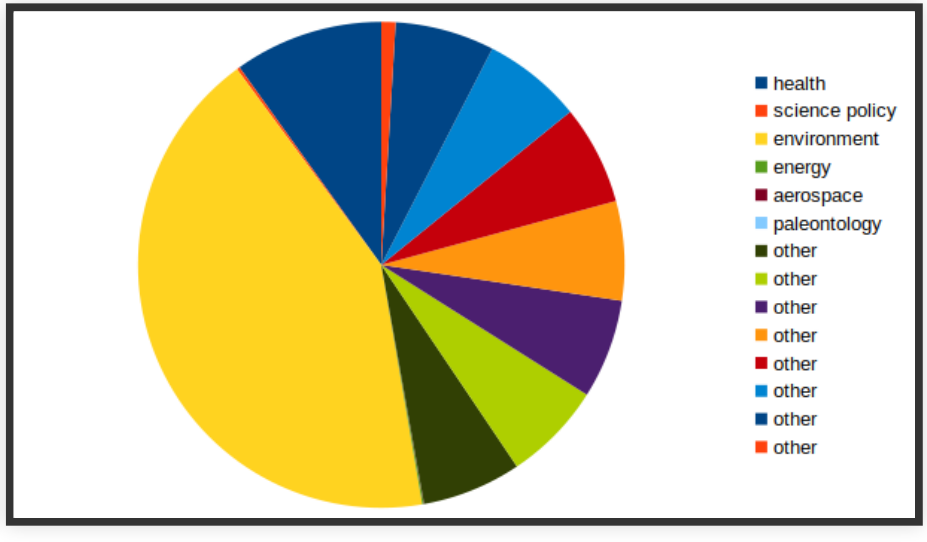
El barco hundido en Ibiza vuelve a perder fuel tras anunciarse el sellado de las fugas

Salvamento acelera al máximo el vaciado de combustible - Tres playas siguen cerradas

ANDREU MANRESA

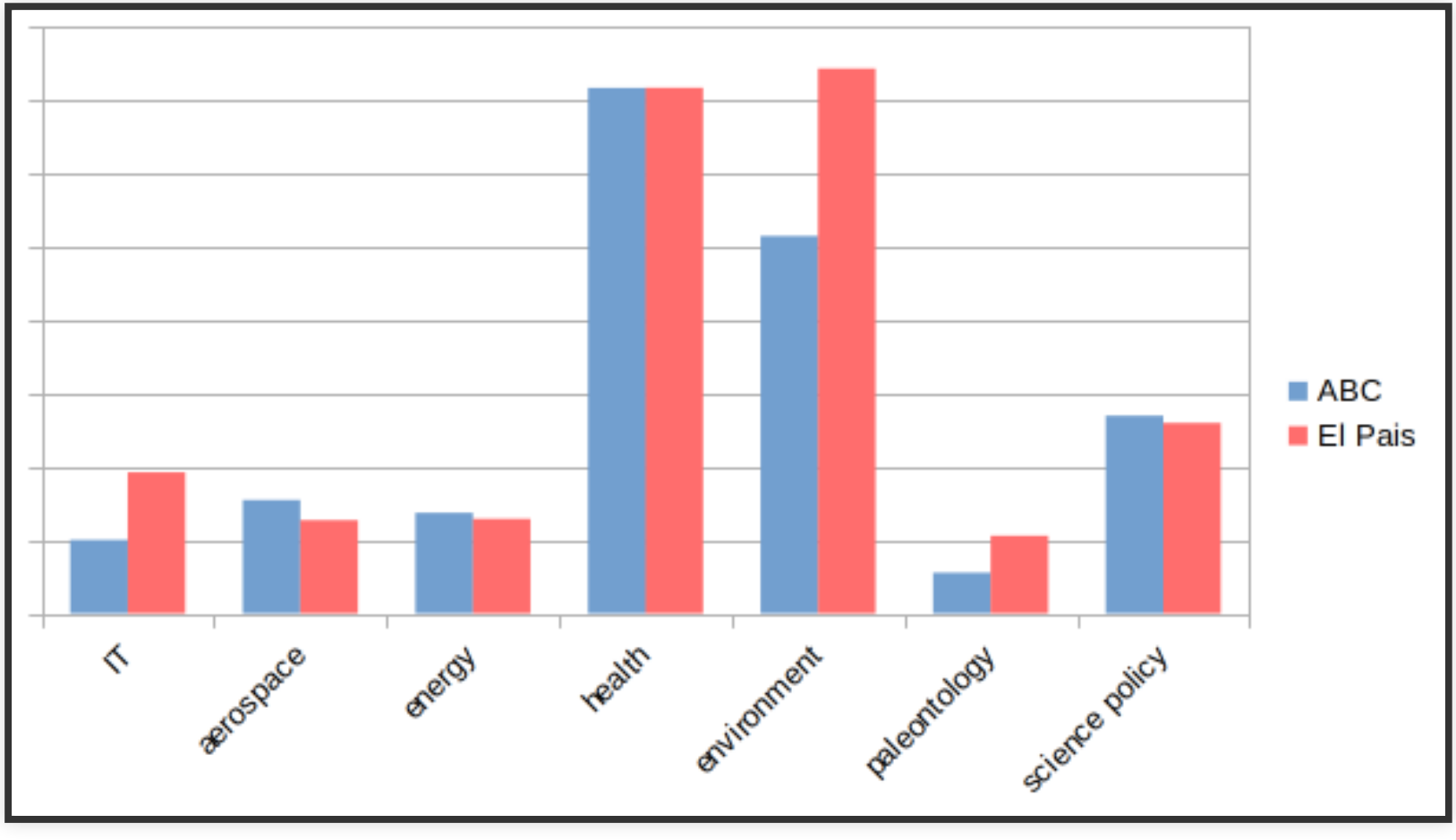
Palma De Mallorca - 15 JUL 2007 - 00:00 CEST

Tres nuevas fugas contaminantes, dos de combustible y una de aceite, emergieron ayer del casco hundido del mercante *Don Pedro* y reactivaron la inquietud provocada por la crisis medioambiental y económica en Ibiza. El revés por las nuevas fugas -que no se pudieron cerrar ayer tarde- llegó horas después de los "mensajes de tranquilidad" de la ministra de Fomento, Magdalena Álvarez, que el viernes afirmó que se habían sellado todos los escapes del pecio.



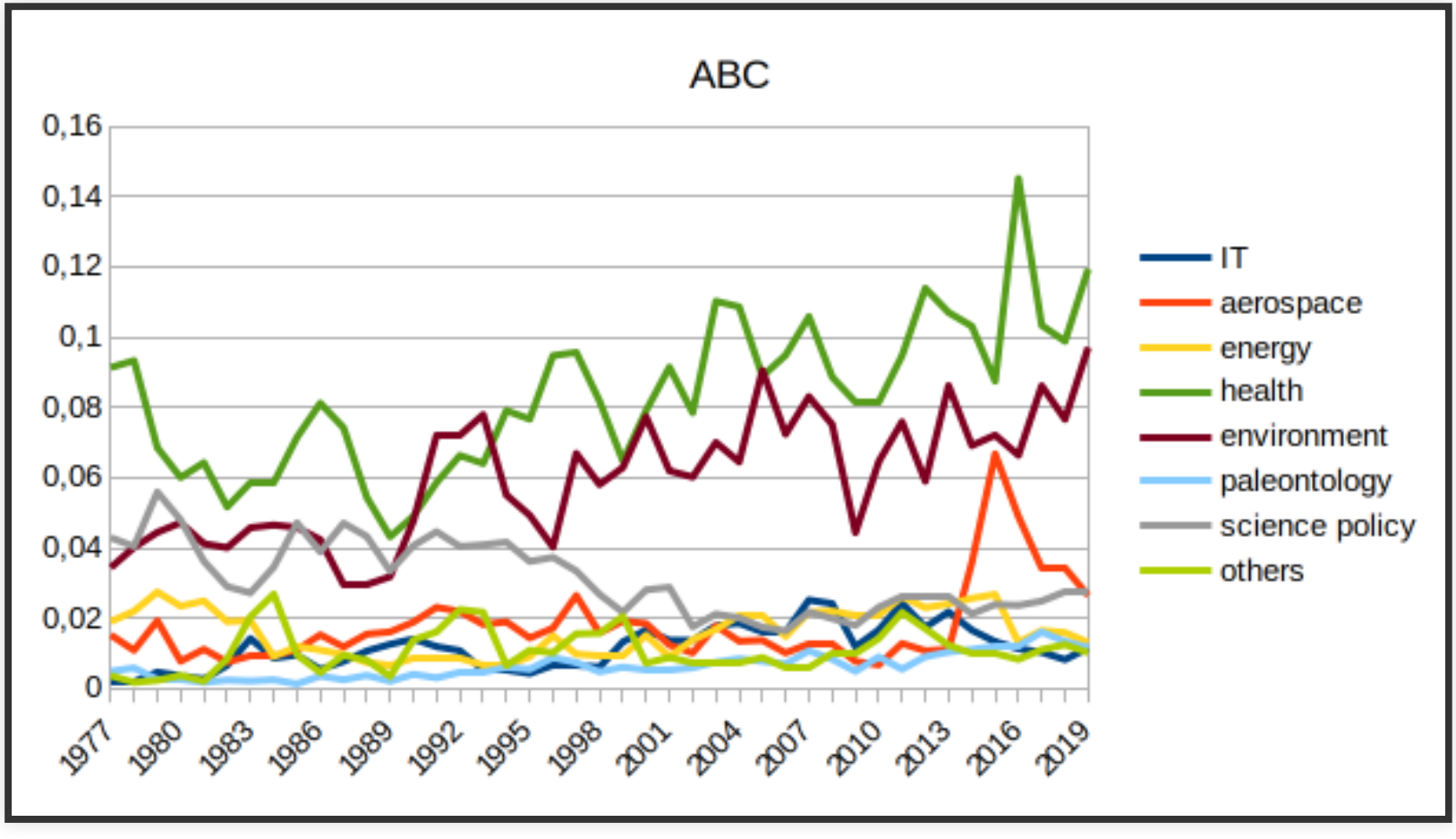
RESULTS

more or less the same for both newspapers



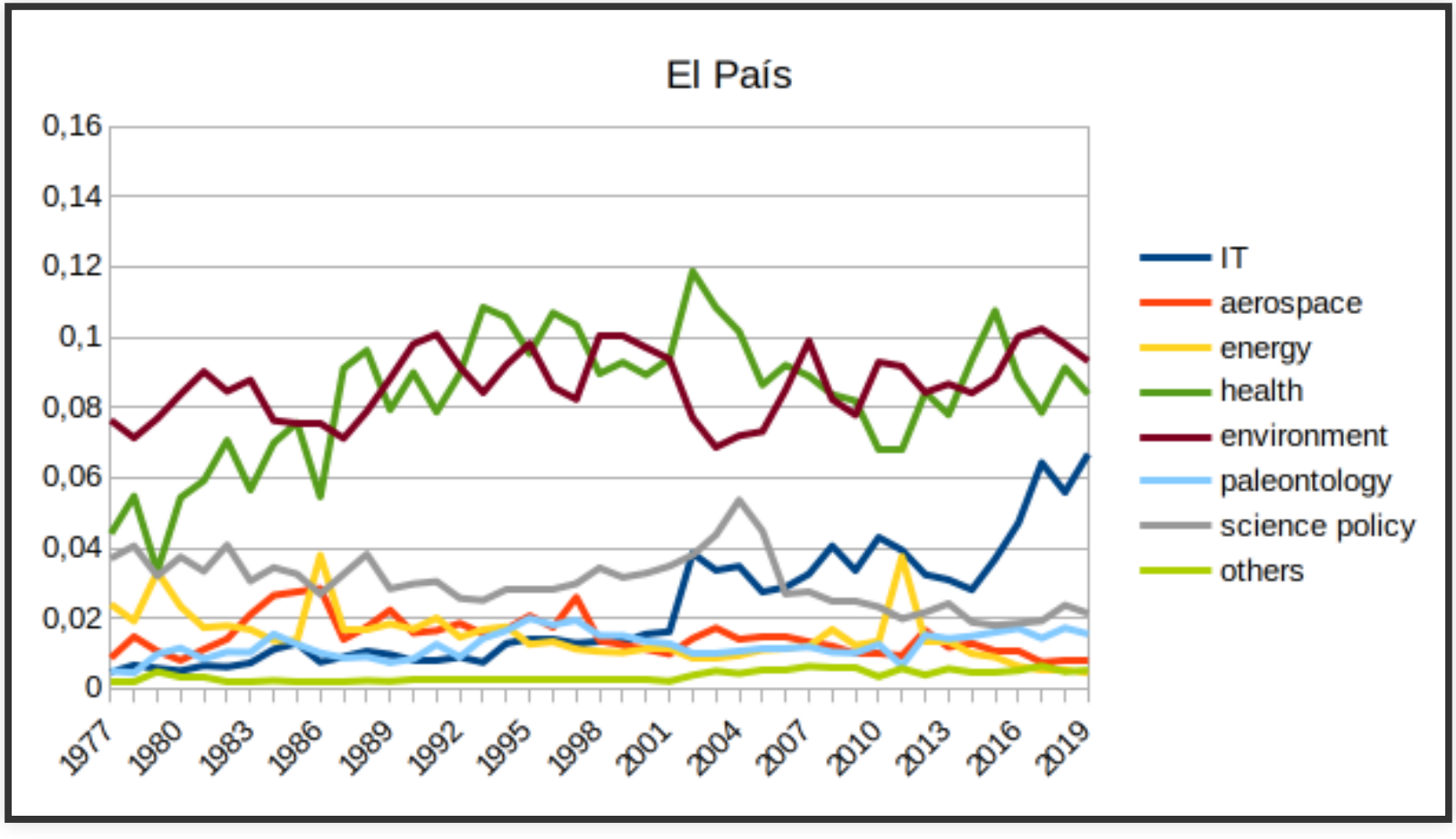
RESULTS

not exactly the same, but very similar curves for both newspapers



RESULTS

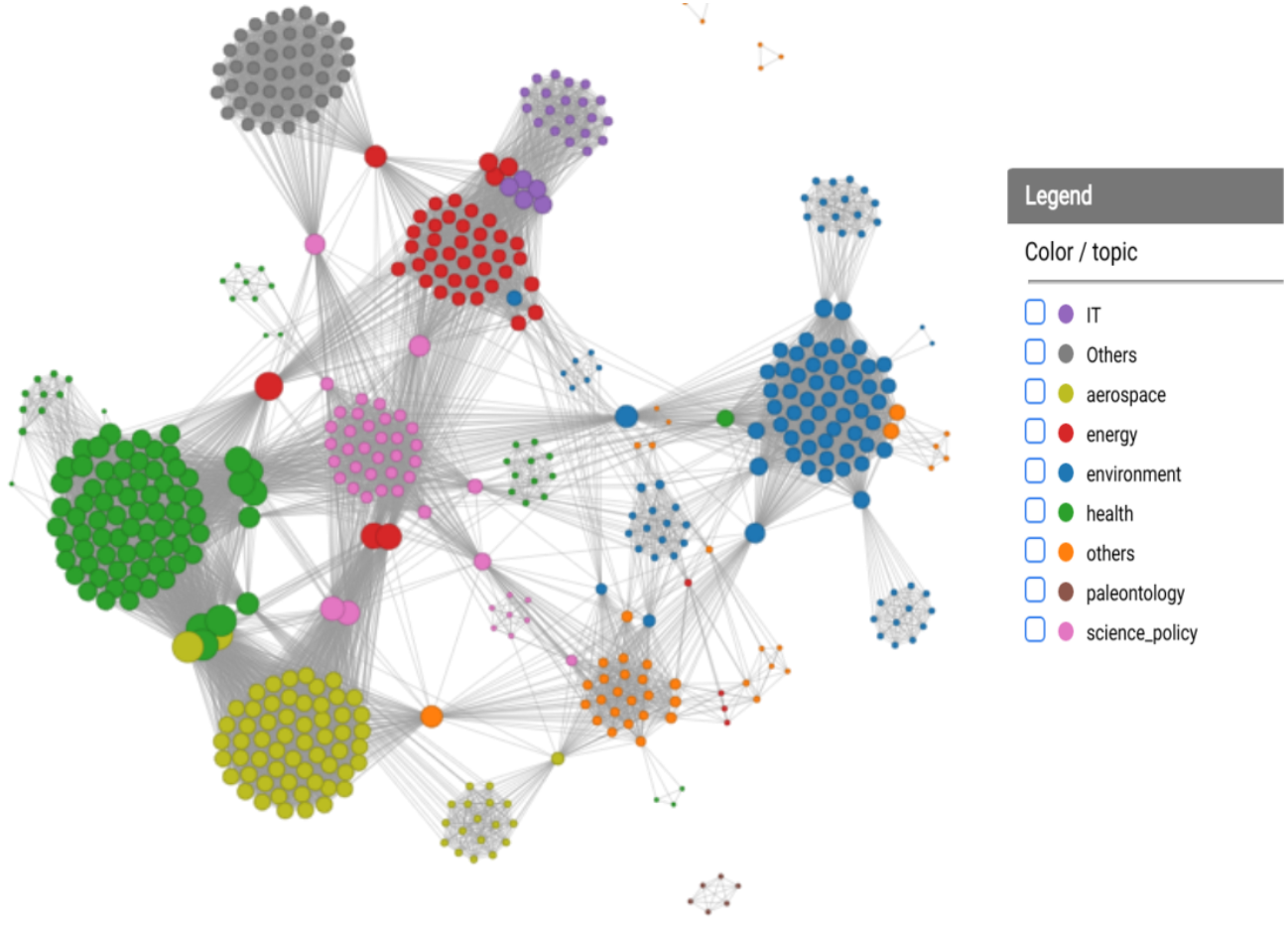
not exactly the same, but very similar curves for both newspapers



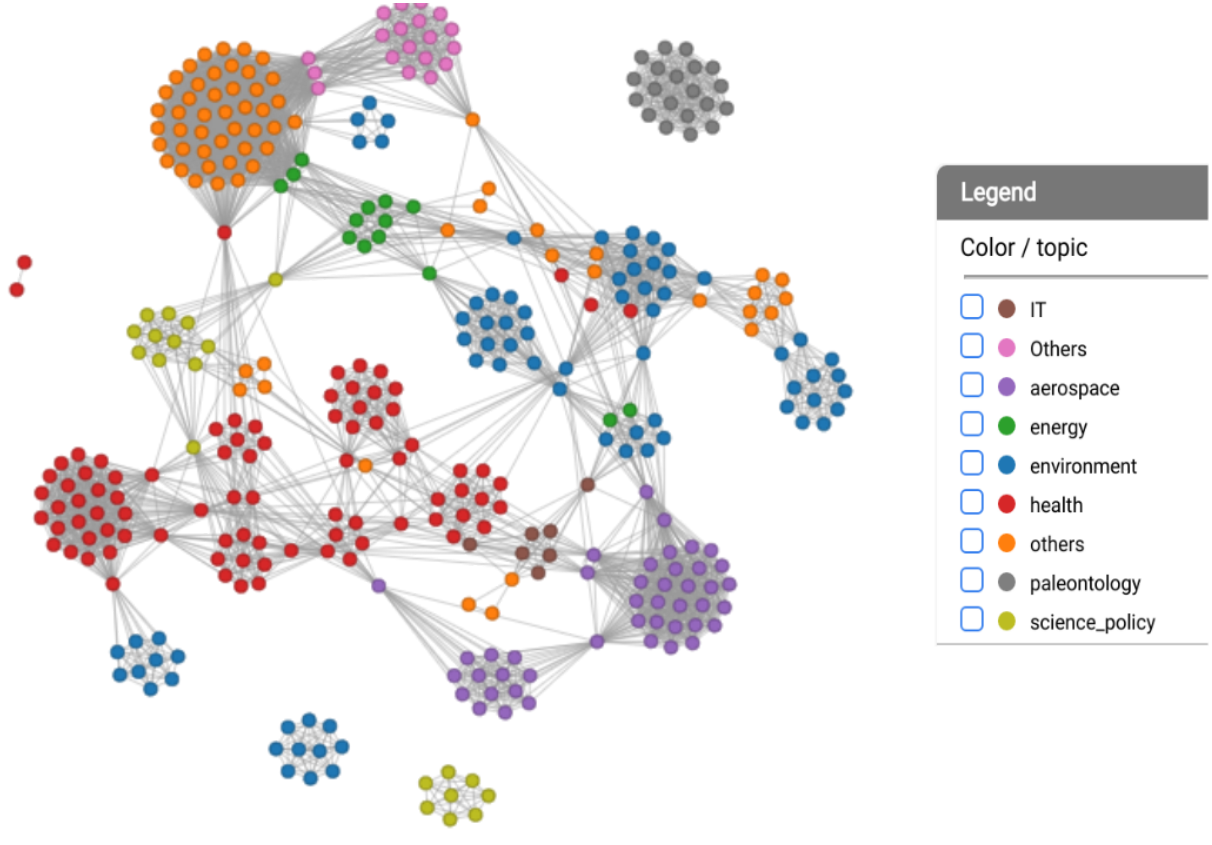
NETWORKS OF DOCS

- we can think in news as nodes if a network, which can be linked if they share a topic
- the amount of the shared topic can be seen as the intensity or weight of such a link
- we can apply this to the news inside a time period (say, for example, a year)

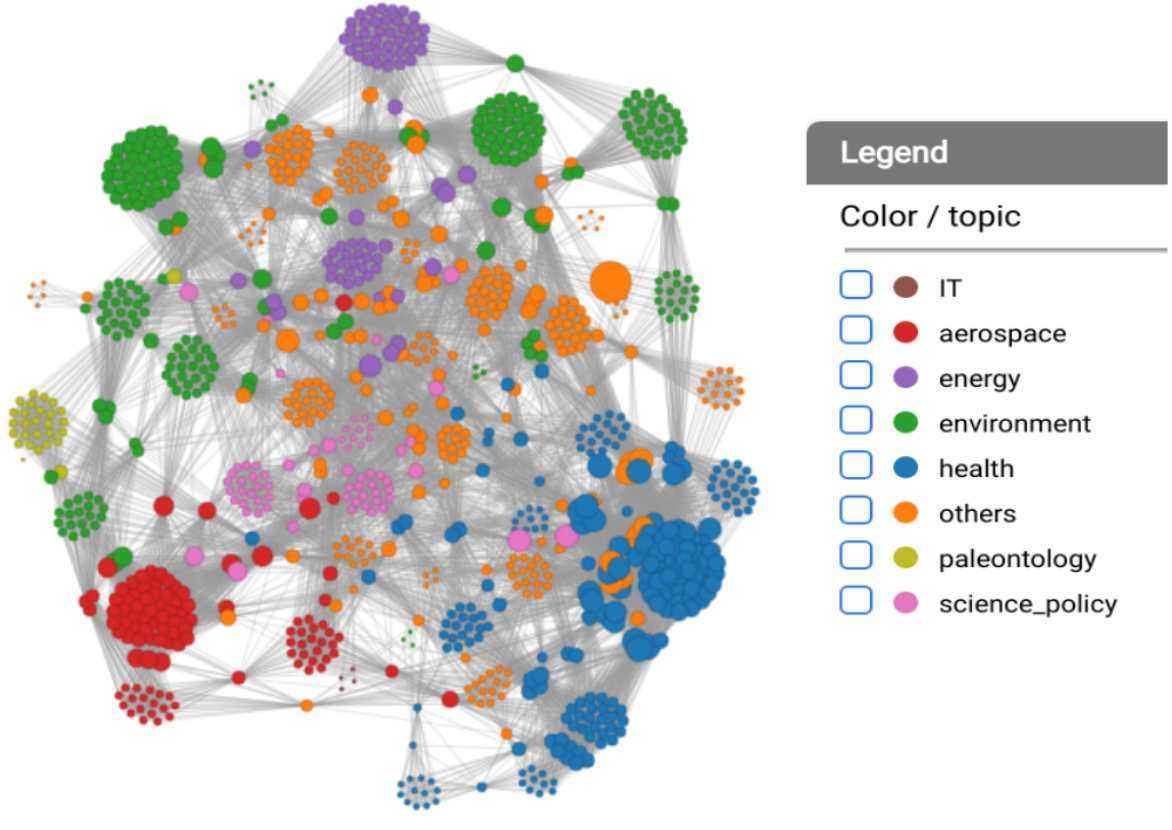
ABC 1991


Legend
Color / topic
IT
Others
aerospace
energy
environment
health
others
paleontology
science_policy

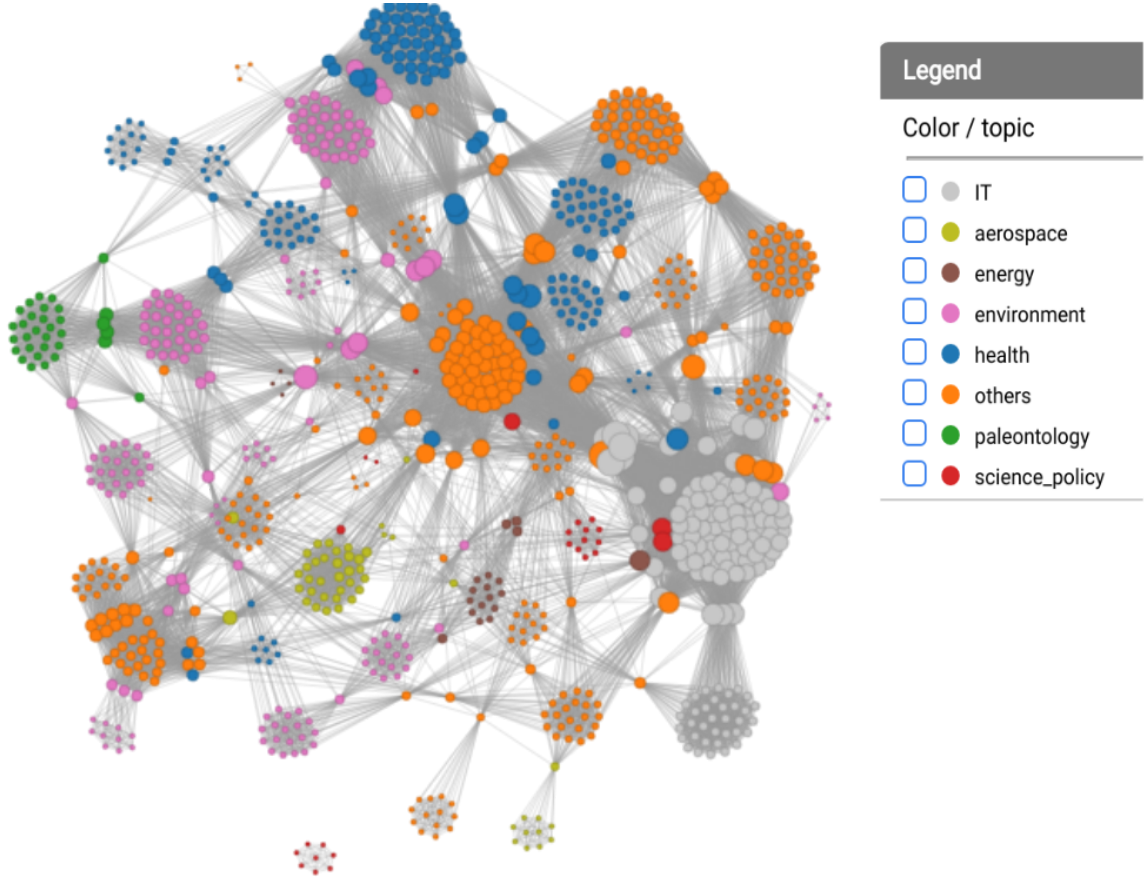
ABC 2017


Legend
Color / topic
IT
Others
aerospace
energy
environment
health
others
paleontology
science_policy

EL PAIS 1991


Legend
Color / topic
IT
aerospace
energy
environment
health
others
paleontology
science_policy

EL PAIS 2017


Legend
Color / topic
IT
aerospace
energy
environment
health
others
paleontology
science_policy

CONCLUSIONS

- Topic Modeling and Network Analysis can help us to analyze a hughe amount of texts
- They can show a wide view as well as a narrower picture of the thematic structure of a collection of documents
- As the date of the documents is also available, we can track the changes over time of such structure

THANKS YOU !!

figue@usal.es