The implications of Wikipedia for contemporary science education: Using Social Network Analysis Techniques for Automatic Organisation of Knowledge

Carlos G. Figuerola, Tamar Groves, Miguel Angel Quintanilla
ECyT Institute
University of Salamanca
TEEM 2015, Porto - 7-9/10/2015

Wikipedia

- knwoledge inside
- big impact, milions of users every day
- built in a collaborative way

The Reliability of Wikipedia

- often critized
- there is not personal responsability
- as anyone can edit articles, it can have non trusted content

but ...

- high amount of scientific papers about the reliability of *WKP*
- most of them conclude that the content of *Wikipedia* is highly reliable
- the strenght of *WKP* is its social nature
- errors, mistakes, even sabotages are fixed inmediatly by users

Our aim in this work

- this paper is not about *Wikipedia* quality
- our aim is to analyze the vision of Science in Wikipedia
- Wikipedia articles are used in teaching and education
- **still work in progress !!**

Wikipedia Content

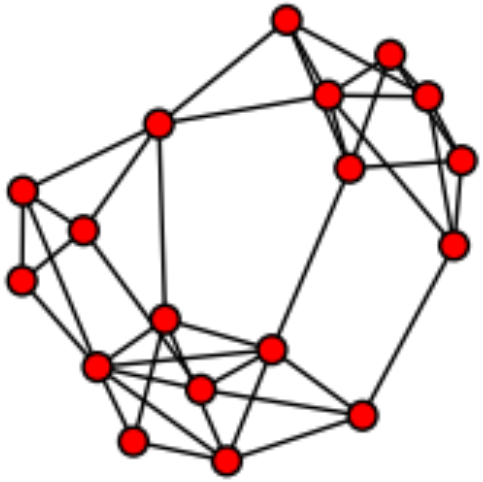
- first challenge: locating the articles on Science topics inside Wikipedia
- Wikipedia (spanish version, nov. 2013) has:
  - 1,027,168 articles
  - 30,007,372 links
- too big to explore manually

Wikipedia Categories

- *WKP* articles are classified in *categories*
- Wikipedia categories are like free keywords:
  - any editor can assign an article to any category (even the same article to several categories)
  - any editor can create new categories
- there are around 60,000 categories (after cleaning administrative only categories and the less frequent ones)
- still too big to explore by hand, we need Automatic Knowledge Organization methods

Wikipedia as a Network

- we can represent Wikipedia as a network
- every article is a node in such a network
- hyperlinks between articles are edges in this network
- edges are directed, as hyperlinks have also direction



Links between Wikipedia articles

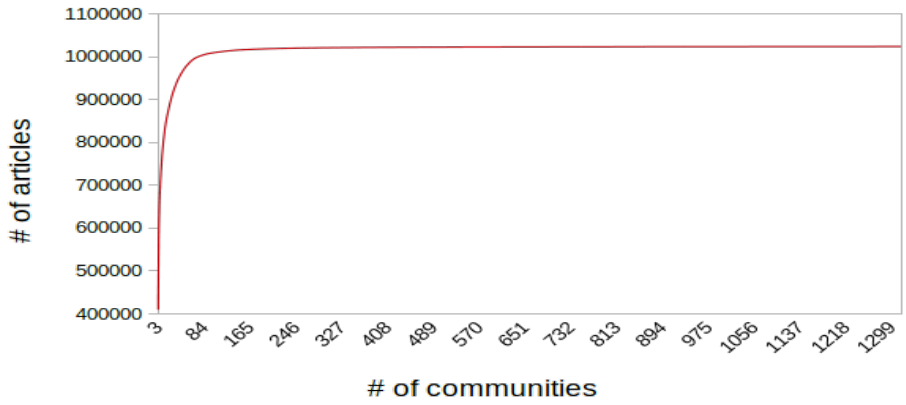
Communities in networks

- a network's community is a bunch of nodes that:
  - they link strongly among themselves
  - they link weakly with other nodes outside the bunch
- when visually represent a network as a force directed graph
  - nodes strongly linked are placed closer

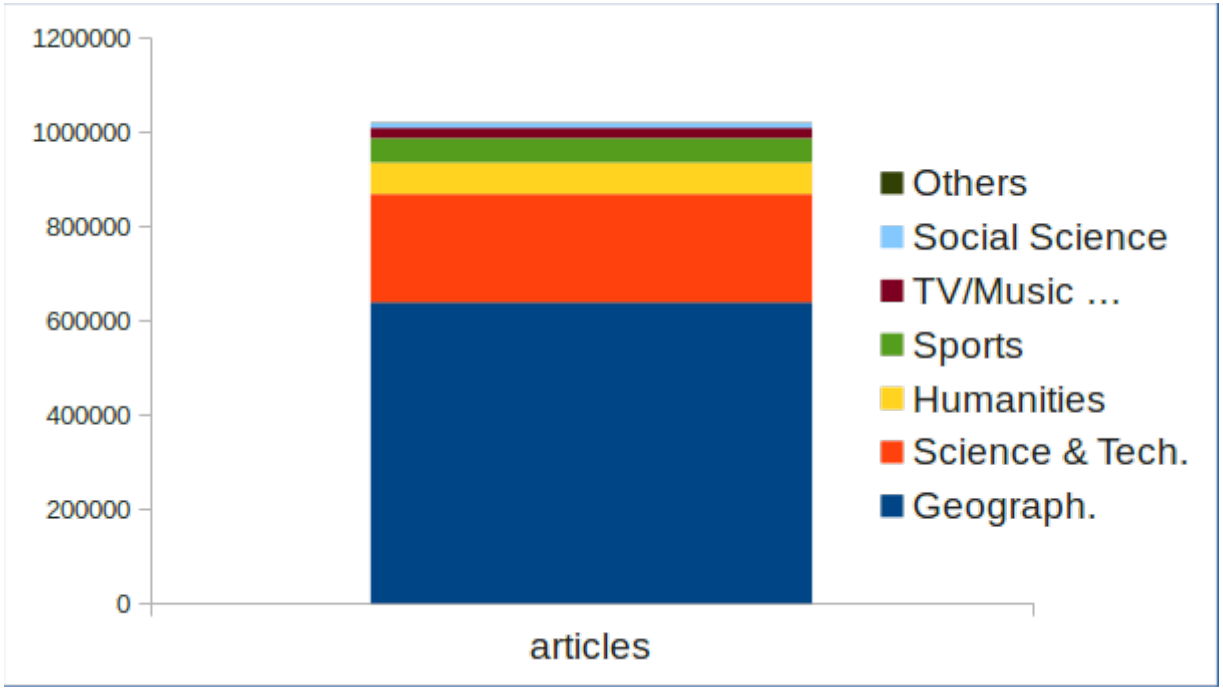
Communities in networks

Communities finding

- several algoritms
- we used *Infomap* (good with big networks)
- 1300 communities
- only 255 communities have more than 20 members
  - this means 98 % of all articles of wikipedia

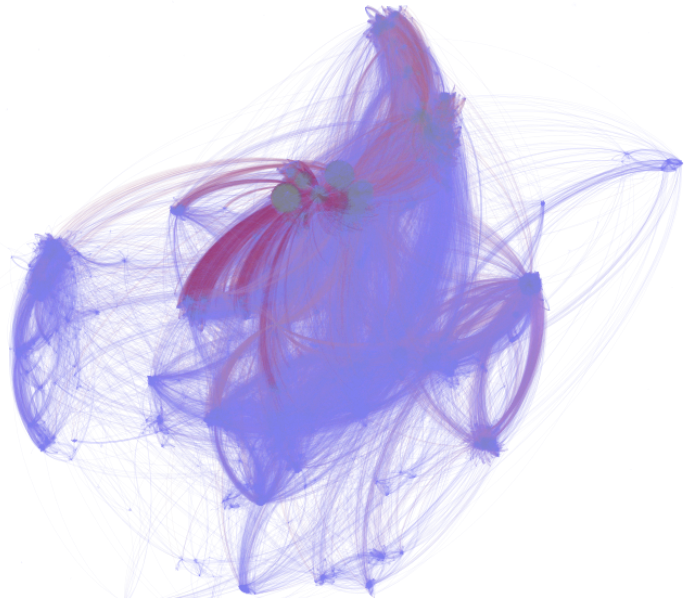


Broad topics in Wikipedia



Articles of Science

- after manual revision of 255 communities:
  - 177 K articles with Science content = 17.30 % of the whole Wikipedia



The taxonomy issue

- articles about elements belonging to taxonomic trees: animals, plants, asteroids ...
- many of them have no edits, only name and place in taxonomy
  - these are very common: about 57 K
- we can clean the landscape:
  - putting apart such only taxonomic articles
  - focusing on full articles with elaborated content

there are also taxonomic articles with plenty of information, we keep the focus also on these.

- We work with 119,797 articles

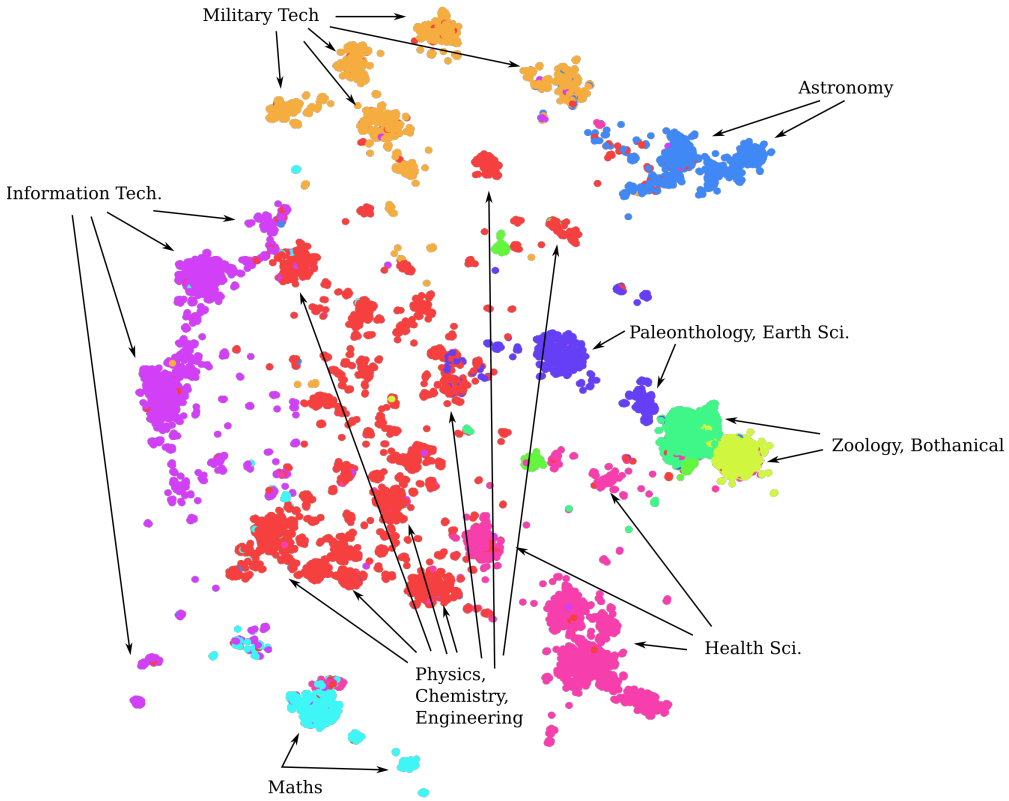
Science vs. non-Science articles

- Science articles link very little with no-Science
- except transversal articles: countries, places, ages, dates
- science articles tend to link with articles belonging to the same community
  - remarkable homophily index: all communities have negative values
  - homophily index for the whole wikipedia: 0.89

Subcommunities

Scientific field	Number of articles	Homophily
Zoology	35144	-0.33
Botanics	27554	-0.34
Health Sci	13260	-0.32
Chemistry, Physics, Engineer.	12629	-0.29
Inf. Tech.	9794	-0.71
Astronomy	7109	-0.68
Paleonthology, Earth Sci.	6080	-0.17
Military Tech.	5117	-0.56
Maths	3039	-0.37

Subcommunities of Science



Subcommunities on Science

- they don't fit a conventional classification of Science
- internal cohesion and spatial location
  - maths looks like an isolated topic
  - zoology and botanics are the biggest, but they have little connection with another science topics
  - physics, chemistry & engineering, the most related with the other science communities
  - health sci
    - a small part well linked with physics, chemistry & enginnering
    - another small part related to zoology and botanics
    - a big part with low connection with another science areas

Subcommunities on Science

- relevance of military technologies
- IT has a compact core, but also articles connecting with Physics, Chemistry & Engi
- Astronomy, some relationship with Physics, etc.
- Paleonthology & Earth Sci some related with Pysics, Chemistry & Engineering, some with Zoology & Botanics

Conclusions (1)

- applying Network Analysis techniques is useful to discovery main topics in Wikipedia
- Science is about 17.3 % of articles in spanish wikipedia
  - taxonomic articles are more than 5 % of wikipedia articles
- Science articles link little with non-science articles
  - except with transversal ones
  - they have a more internal cohesion (homophily) that non-science articles

Conclusions (2)

- Science main topics don't fit classical classification of scientific disciplines
    - Military technology has remarkable presence in Wikipedia
- Some scientific areas appear as unconnected of the rest of Science (maths, Paleonthology, ...)
- There are scientific areas with relevant internal cohesion
- There are areas more dispersed and well connected with the others (Physic, Chemistry & Engineering, part of Health Sci)
- In the future:
    - to analyze connections in fields in order to discern structures of scientific culture on the Web

Thank You!

Contact Information:

e-mail figue@usal.es

www ecyt.usal.es