Hierarchical Organisation of Scientific Journals

Gergely Tibély, Enys Mones, Péter Pollner, Tamás Vicsek,
Gergely Palla

Department of Biological Physics
Eötvös University, Budapest

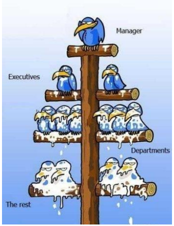
outline

1. About types of hierarchies
2. Dataset
3. Tools & results
4. Discussion
5. Hierarchy Construction Method

hierarchies

A very widespread form of organisation

- river basins
- animal flocks
- biological classification
- postal addresses
    - country
    - state
    - city
    - street
    - house
    - floor
    - door
- social organisation (universities, military, firms, ...)



hierarchi**es**

there are multiple type of hierarchies:

- a simple ordering (e.g., natural numbers)
- inclusion hierarchies (departments - faculties - institutes)
- flow hierarchies

multiple hierarchies may occur in the same system

- e.g., pidgeons have different ones for navigation and eating
- formal and informal channels of information spreading in organizations

investigated system: Web of Science database

Aim

- determine the hierarchy of scientific information sources
- from different perspectives,
  1. flow of information
  2. generality
- compare

Data

- 35M scientific papers
- 1975-2011
- 13k journals
- 11M journal-to-journal citation links

flow hierarchy – method

Trying to order journals according to influence in information flow

Let's use the reaching centrality ($C_R^{(m)}$):

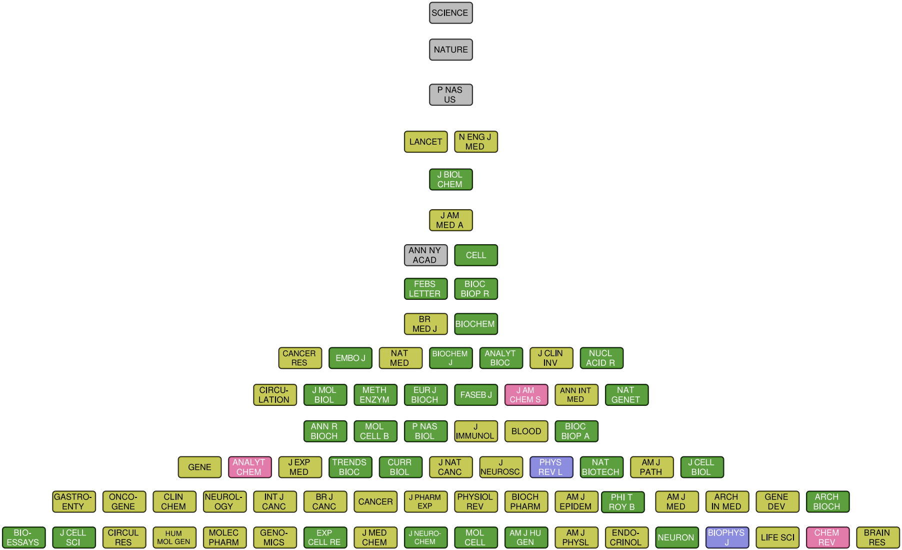
- for each node in the network, calculate the fraction of nodes reachable in m steps
- assign nodes to hierarchical levels by introducing ΔC intervals

No ancestor-descendants relationships, just hierarchical levels.

For the Web of Science data: calculate everything for individual articles, then take the union of each journal's paper's m-reachable sets:

$$C_R^{(m)}(\text{journal}) = \frac{|\{p | d_{\text{out}}(q,p) \leq m, q \in \text{journal} \wedge p \notin \text{journal}\}|}{N_{\text{papers}}}$$

flow hierarchy – results



SCIENCE

NATURE

P NAS US

LANCET · N ENG J MED

J BIOL CHEM

J AM MED A

ANN NY ACAD · CELL

FEBS LETTER · BIOC BIOP R

BR MED J · BIOCHEM

CANCER RES · EMBO J · NAT MED · BIOCHEM J · ANALYT BIOC · J CLIN INV · NUCL ACID R

CIRCULATION · J MOL BIOL · METH ENZYM · EUR J BIOCH · FASEB J · J AM CHEM S · ANN INT MED · NAT GENET

ANN R BIOCH · MOL CELL B · P NAS BIOL · J IMMUNOL · BLOOD · BIOC BIOP A

GENE · ANALYT CHEM · J EXP MED · TRENDS BIOC · CURR BIOL · J NAT CANC · J NEUROSC · PHYS REV L · NAT BIOTECH · AM J PATH · J CELL BIOL

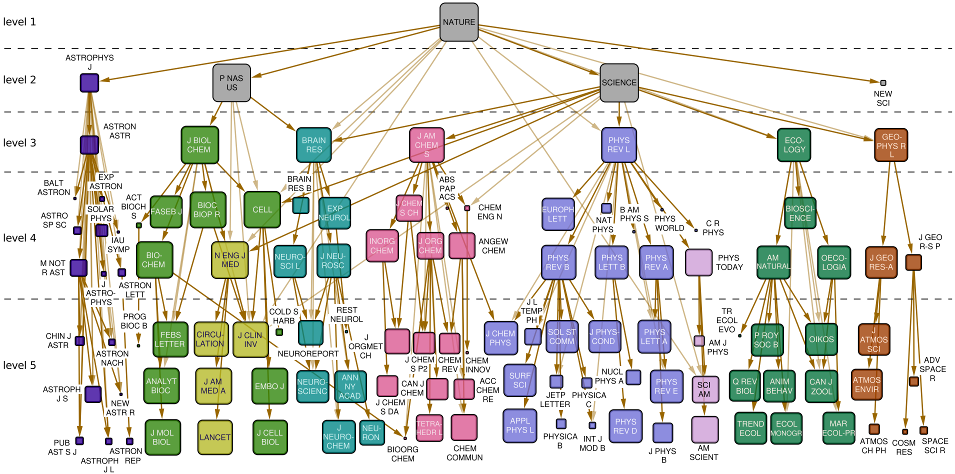
GASTROENTY · ONCOGENE · CLIN CHEM · NEUROLOGY · INT J CANC · BR J CANC · CANCER · J PHARM EXP · PHYSIOL REV · BIOCH PHARM · AM J EPIDEM · PHI T ROY B · AM J MED · ARCH IN MED · GENE DEV · ARCH BIOCH

BIOESSAYS · J CELL SCI · CIRCUL RES · HUM MOL GEN · MOLEC PHARM · GENOMICS · EXP CELL RE · J MED CHEM · J NEUROCHEM · MOL CELL · AM J HU GEN · AM J PHYSL · ENDOCRINOL · NEURON · BIOPHYS J · LIFE SCI · CHEM REV · BRAIN RES

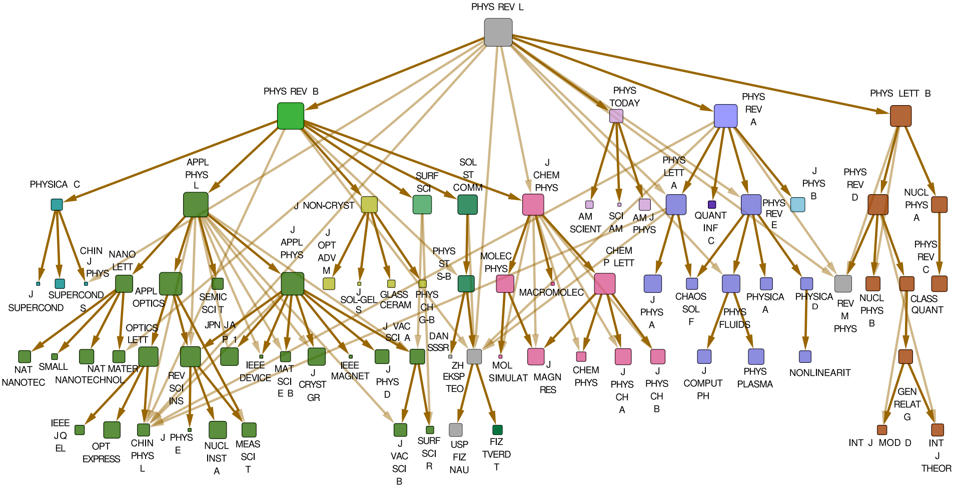
inclusion hierarchy – method

- trying to get more general-less general relationships between journals
- using the journal-to-journal citations
- we have a hierarchy reconstruction method for tagged datasets (e.g. photo tags on flickr)
- citations can be considered technically quite similar: journal names (citing & cited) appear on papers as tags
- however, it is directed (citing → cited)!

inclusion hierarchy – results (top)



inclusion hierarchy – results (physics branch)



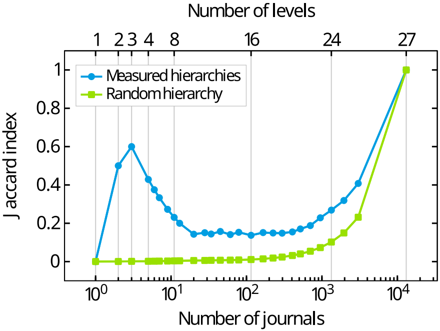
comparison of the top levels

Consider the set of journals from the top l levels, without any structure.

Calculate the Jaccard similarity for the sets of the 2 hierarchies:

$$J(l) = \frac{|S_{\text{flow}}(l) \cap S_{\text{incl}}(l)|}{|S_{\text{flow}}(l) \cup S_{\text{incl}}(l)|}$$

Do the same for random hierarchies (shuffle journals).



comparison by counting inverted relations

Kendall's tau distance (for 2 total orders):

$$\tau = \frac{\#\text{inversions}}{\#\text{max possible inversions}}, \qquad \tau \in [0, 1]$$

Here we have

- a bucket order (ties are possible), and
- a partial order (2 journals may not be directly comparable, i.e., in different branches).

Kendall's tau can be generalized for that case (by checking relations only of the partial order).

Result is 0.16
Random case: 0.80 ± 0.02

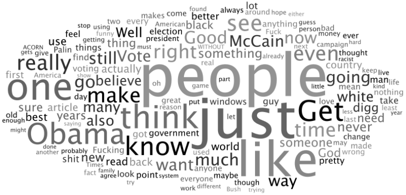
results

- there are multiple types of hierarchies
- multiple hierarchies (same type or not) may exist for the same system
- for the scientific journals, current data & methodology gives quite large, although not perfect similarity
- scientific fields correspond well to branches in the constructed hierarchy

hierarchy construction method: motivation

Traditional way of categorisation: hierarchies (directiories, phylogenetic tree of life, postal address, . . .)
top-down

Recent innovation: *tags*
bottom-up


lot
found always around hope either
makes come better
two every American black anything guess
stop using election see person bad
feel funny Well president Good McCain now money ever
ACORN use getting thing must WITHOUT next campaign hard
gets give Palin things right something already even thought
really find still Vote racist
voting actually real country
first America show people keep live
one go believe oh part going man
day make game little mean kind life
sure article many great put windows guy white nothing
old best years also reason think just Get love digg take
enough saying let less least
might Obama got government time last need year
know someone never
done probably Fucking world change may
another shit new used made
Times read back want much God wrong
fact family agree look point system anyone like pretty
everyone maybe
try work different though way
Bush trying

Handy for large datasets:

- easy search
- more tags on one object
- categorisation can be done by several, non-expert users

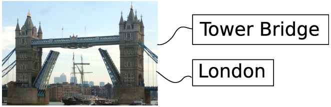
examples

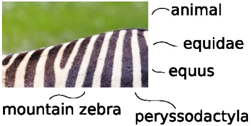
A few examples:

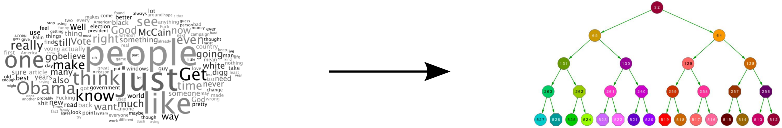
- photo tags on flickr
- Wikipedia categories
- book & film genres
- Gmail labels
- biological functions of proteins, genes, . . .
- recipes
- blog entries
- news portal entries

a few observations & reflections

- tags are usually created equal, no relation (e.g. ordering) is defined between them
- however, things are organised in the head of the user (hopefully), frequently as a hierarchy
- that should be reflected by the tags somehow







the question

How to to extract the underlying hierarchical organisation just from the coappearances of the tags?

Motivations:

- Nice challenge
- Appearing large datasets suggest possible applications
- or understanding the terminology of small, very specialised fields (maybe using common words in different meanings)

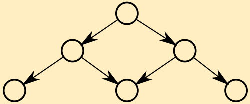
So, let's try!

(a bit more) precise definition of the problem

We are given a set of objects, each object having a set of tags.

We are looking for a directed acyclic graph of the tags, describing the tag-tag coappearances such that we like it the most.

a directed acyclic graph is a hierarchy, in which more parents are allowed



hierarchy construction

Algorithm:

- which tag is higher in the hierarchy: tags high in the hierarchy should connect strongly to other high ranking ones. Leads to PageRank on the weighted tag-tag coappearance graph.
- choose parent: we build the hierarchy bottom-up. Look for the most significant coappearing partner of the current tag & its descendants. (*aggregated z-score compared to the hypergeometric distribution*). Much more information than in any pairwise similarity measure.
- 2nd, 3rd, ... parents: take the z-score of the 1st parent, connect any other possible parents having at least as high z-score.

More roots, more components are possible.
Typical time complexity: $\mathcal{O}(N \log N)$

benchmark system

But how good are the results?

- look at them – only for small & familiar systems

- run on data having a known directed acyclic graph – very limited collection

- build synthetic data with a known directed acyclic graph, and compare the result – does not exists

benchmark – building synthetic datasets

Method:

- take a directed acyclic graph of tags, e.g. a binary tree
- generate sets of tags
  - choose tags randomly, according to some probability distribution (popularity),
  - or by a random walk starting from the last tag

benchmark – similarity of directed acyclic graphs

How similar are the original & reconstructed directed acyclic graphs?

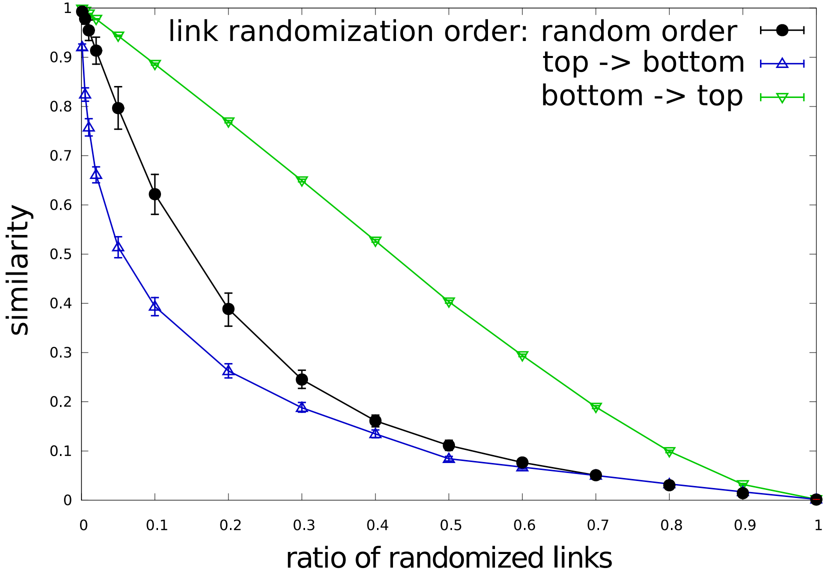
1. compare the set of descendants of each tag.
2. correct for the expected value of a random case

$$I_{o,r} = -\frac{\sum_{i=1}^{N} p_{o,r}(i) \ln\left(\frac{p_{o,r}(i)}{p_o(i)p_r(i)}\right)}{\frac{1}{2}\left(\sum_{i=1}^{N} p_o(i)\ln p_o(i) + \sum_{i=1}^{N} p_r(i)\ln p_r(i)\right)}$$

$$= -\frac{\sum_{i=1}^{N} |D_o(i)\cap D_r(i)| \ln\left(\frac{|D_o(i)\cap D_r(i)|\cdot(N-1)}{|D_o(i)|\cdot|D_r(i)|}\right)}{\frac{1}{2}\left(\sum_{i=1}^{N} |D_o(i)|\ln\frac{|D_o(i)|}{N-1} + \sum_{i=1}^{N} |D_r(i)|\ln\frac{|D_r(i)|}{N-1}\right)}$$

1 only for identical objects, 0 for independent ones.

benchmark – similarity of directed acyclic graphs

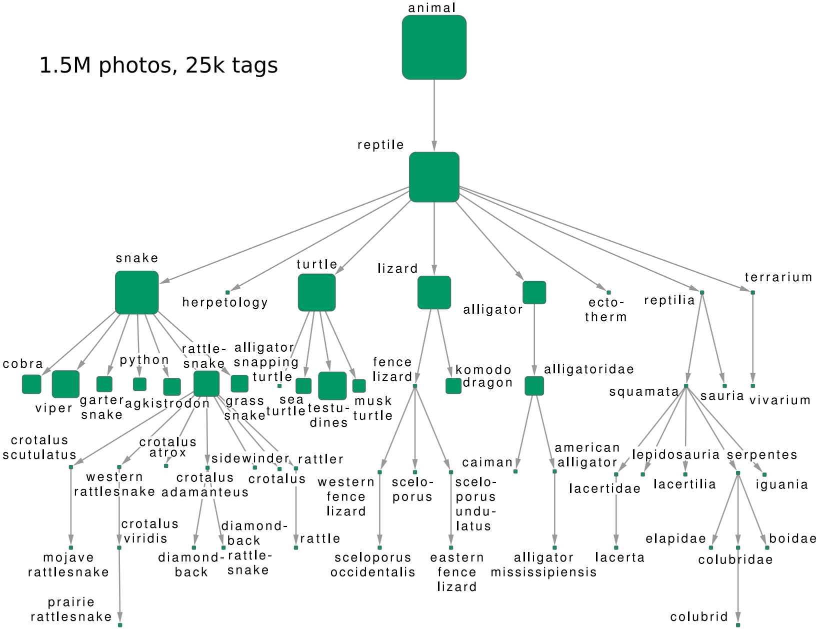


benchmark – results

method	#exact links	#OK links	$I_{\text{linearised}}$
our algorithm	89%	91%	97%
Heymann & Garcia-Molina	48%	54%	76%
Schmitz	1%	2%	5%

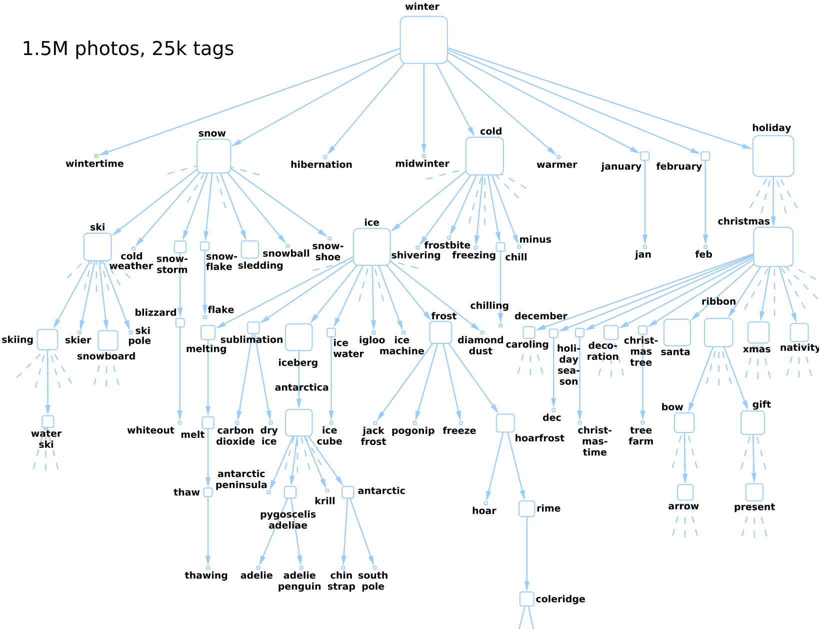
real-world examples: flickr

1.5M photos, 25k tags



real-world examples: flickr

1.5M photos, 25k tags



summary

We got...

- an automated method for reconstructing tag hierarchies
  - the most precise currently
  - fast – $\mathcal{O}(\#\mathrm{tags} \cdot \ln(\#\mathrm{tags}))$ running time
  - multiple parents are possible

- a directed acyclic graph similarity measure

- a benchmark system

Thank you!