

# **Predictive Modeling of the Emergence and Development of Scientific Fields**

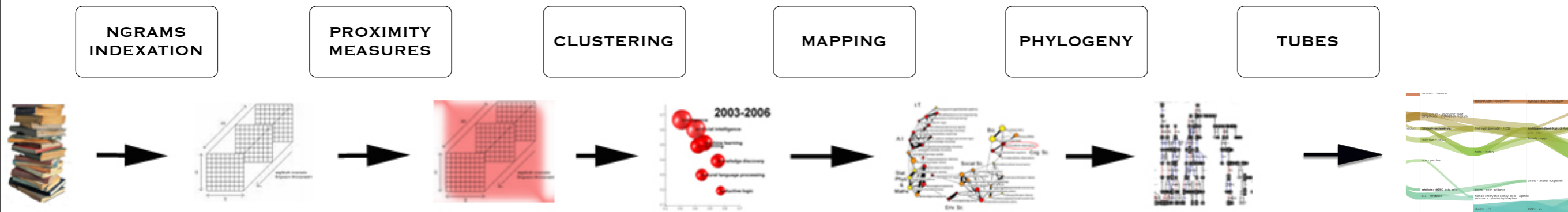
*MIT, 25-26 May 20*

---

## **From Textual Corpora to Lexical Networks**

*Jean-Philippe Cointet - IFRIS, INRA-SenS, ISC-PIF*

# Knowledge dynamics reconstruction



- Lexical networks analysis is a way to investigate **knowledge communities dynamics** based on the structure of the use of terms or concepts,
- Historically, keywords have been privileged as the basic unit of analysis for co-word analysis, but...
  - some datasets may not have keywords entries
  - **indexer bias** can be criticized

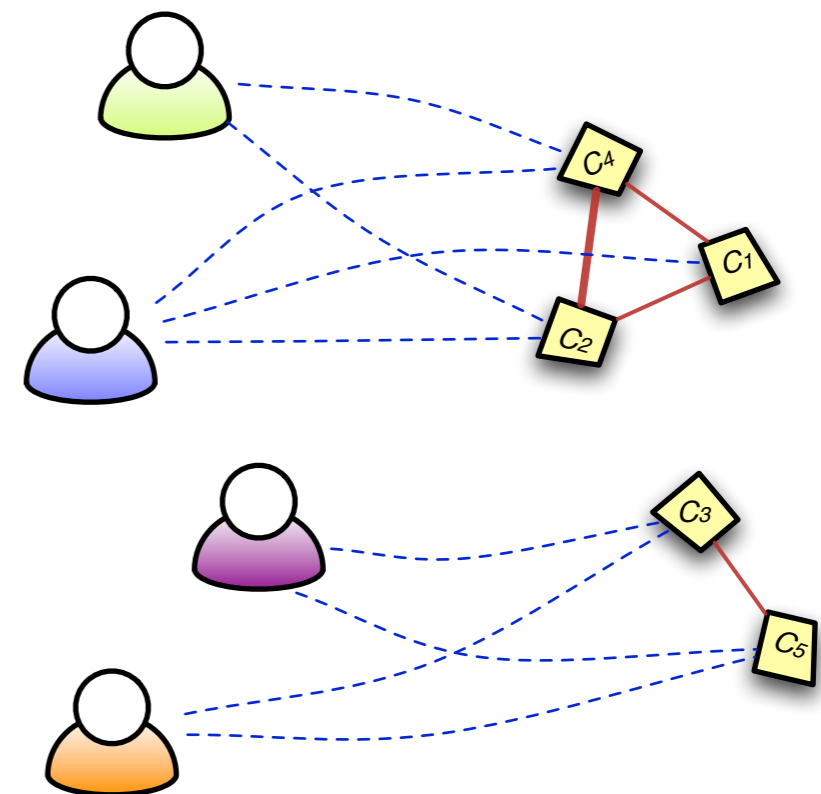
# What it is about a text that is interesting ?

---

«Indexing is an intervention between the text and the co-word analysis, and the validity of the map will depend, to a certain extent, on the nature of the indexing. Yet since indexers try to capture what it is about a text that is interesting, they partially reproduce the readings that the texts are given within the field itself'. Thus, despite the fact that indexing is not entirely reliable, validity is never totally absent.»

Callon, M.; Law,,J.; & Rip, A. (Eds.). (1986a). *Mapping the dynamics of science and technology: Sociology of Science in the real world*. London: The Macmillan Press 1,td.

- **grammatical** criterion, candidate terms are usually limited noun phrases,
- **unithood**, phrases should represent a proper semantic unit,
- **termhood**, terms should be domain specific to carry substantial information



# Linguistic approach

---

The phylogenetic position of the elephant shark (*Callorhynchus milii*) is particularly relevant to study the evolution of genes and gene regulation in vertebrates.

# Linguistic approach

---

## i.Part-Of-Speech Tagging

The phylogenetic position of the elephant shark (*Callorhynchus milii*) is particularly

*DT JJ NN IN DT NN NN ( NNS NN )VBZ RB*

relevant to study the evolution of genes and gene regulation in vertebrates.

*JJ TO VB DT NN IN NNS CC NN NN IN NNS*

# Linguistic approach

---

## i. Part-Of-Speech Tagging

## ii. Tag Chunking - Noun Phrases extraction

ex: Regexp={((Adj|Noun)+|((Adj|Noun)\*NounPrep?)(Adj|Noun)\*Noun}

The phylogenetic position of the elephant shark (*Callorhynchus milii*) is particularly

*DT JJ NN IN DT NN NN ( NNS NN )VBZ RB*

relevant to study the evolution of genes and gene regulation in vertebrates.

*JJ TO VB DT NN IN NNS CC NN NN IN NNS*

# Linguistic approach

---

## i. Part-Of-Speech Tagging

## ii. Tag Chunking - Noun Phrases extraction

ex: Regexp={((Adj|Noun)+|((Adj|Noun)\*NounPrep?)(Adj|Noun)\*Noun}

## iii. Stemming and filtering of empty words

gene regulation in vertebrate -> {gene regul vertebr}

phylogenetic position of the elephant shark : {eleph phylogenet posit shark}

phylogenetic position -> {phylogenet posit}

# Linguistic approach

---

## i. Part-Of-Speech Tagging

## ii. Tag Chunking - Noun Phrases extraction

ex: Regexp={((Adj|Noun)+|((Adj|Noun)\*NounPrep?)(Adj|Noun)\*)Noun}

## iii. Stemming and filtering of empty words

## iv. Output: classes of candidate multi-terms:

- cellular isoform prion protein = {isoform of cellular prion protein ; cellular isoform of the prion protein ; cellular prion protein isoform ; isoform of the cellular prion protein ; cellular isoform of prion protein}
- conform: {conformers ; conformational ; conformation ; conformer ; conformations}
- resist scrapi: {resistance against scrapie ; scrapie resistance ; scrapie resistant ; Scrapie resistance}
- associ genotyp prp = {association of PrP genotype ; associations between PrP genotypes ; association between PrP genotype ; associations of the PrP genotype ; associations between PrP genotypes}

# Unithood: extracting semantic units with C-value

---

- Simple frequency-based approach : «Real» Terms tend to appear more frequently than non-terms
- C-value approach (Frantzi K. & Ananiadou S., 2000):
  - Longer phrases are more likely to be relevant,
  - Nested terms may induce false positive, ex: self organizing maps.

$$C\text{-value}(a) = \log_2|a|(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b))$$

# Termhood

---

- Candidate terms should be thematically specific ; terms not specific to a specific thematic subfield have neutral meaning given the whole domain and should be excluded
- On the contrary, terms which distribution is biased toward certain topics are more likely to have interesting meaning.
- Co-occurrences between existing candidate terms are extracted to compute the **Khi2 score** of specificity of each term compared to other terms (Matsuo Y. & Ishizuka M., 2004).

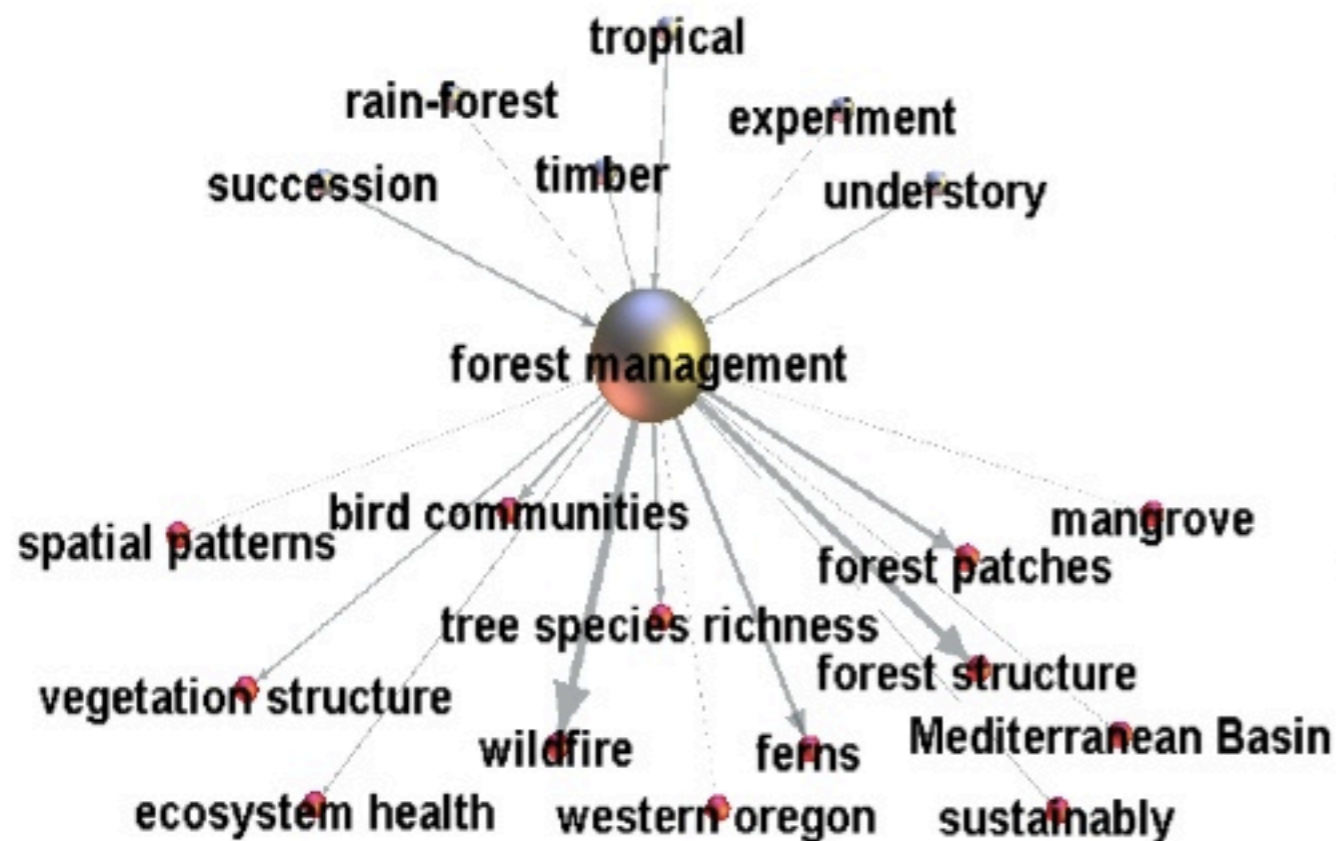
$$\chi^2(w) = \sum_{g \in G} \frac{(\text{freq}(w, g) - n_w p_g)^2}{n_w p_g}$$

Final output example:

stem	main form	forms	occurrences	specificity	C-value
brassica-campestri	BRASSICA-CAMPESTRIS	BRASSICA-CAMPESTRIS	10,0	686,4	7,0
oilse rape	OILSEED RAPE	OILSEED RAPE	7,0	778,1	9,5
cdna	cDNAs	cDNAsI&IcDNAI&ICDNA	16,0	468,3	7,0
brassica rapa	Brassica rapa	Brassica rapa	33,0	1144,8	44,4
alloplasm line	alloplasmic lines	alloplasmic lineI&Ialloplasmic lines	5,0	404,7	7,9
indian mustard	Indian mustard	Indian mustardI&IINDIAN MUSTARDI&Iindian mustard	18,0	2027,6	23,8
crop	crops	cropsI&ICropl&Icrop	58,0	708,8	35,0
hybrid intergener	intergeneric hybrids	INTERGENERIC HYBRIDSI&Iintergeneric hybridsI&Iintergeneric hybridization	16,0	2208,2	25,4
cm line	CMS line	cms lineI&ICMS lineI&ICMS lines	13,0	278,5	15,8
anther	anthers	anthersI&IAntherI&IANTHERI&Ianther	62,0	911,5	30,5
high level	high level	high levelsI&Ihigh level	5,0	252,5	7,9
express gene	gene expression	expression of genesI&IGENE EXPRESSIONI&Igene expressionI&Igenes in the	22,0	397,1	8,7
gene	genes	genesI&Igene	175,0	296,4	57,4
canola	canola	canolaI&ICANOLAI&ICanola	27,0	457,3	23,0
male-steril	male-sterility	MALE-STERILITYI&Imale-sterility	68,0	2606,9	8,3
radish	radish	RADISHI&Iradish	35,0	808,1	20,0
cybrid	cybrids	CYBRIDSI&IcybridI&ICYBRIDI&Icybrids	16,0	463,5	14,0
marker	markers	markerI&Imarkers	60,0	455,2	10,0
genom mitochondri	mitochondrial genome	mitochondrial genomel&Imitochondrial genomes	21,0	423,0	38,0
brassicacea	Brassicaceae	BRASSICACEAEI&IBrassicaceae	20,0	872,5	18,0
flow gene	gene flow	gene flow	15,0	919,6	22,2
fertil restor	fertility restoration	restoration of fertilityI&Irestorer of fertilityI&Ifertility restorationI&Ifertility restorerI	39,0	440,6	31,7
bud flower	flower buds	flower buds	6,0	311,0	7,9
brassica oleracea	Brassica oleracea	BRASSICA OLERACEAI&IBrassica oleracea	51,0	1399,2	42,8

# What next ?

- Reconstruction of the cognitive dynamics in science through the analysis of the lexical network built upon the temporal matrix of co-occurrences within our term list (asymmetric measure of proximity between terms).



Distributional approach :

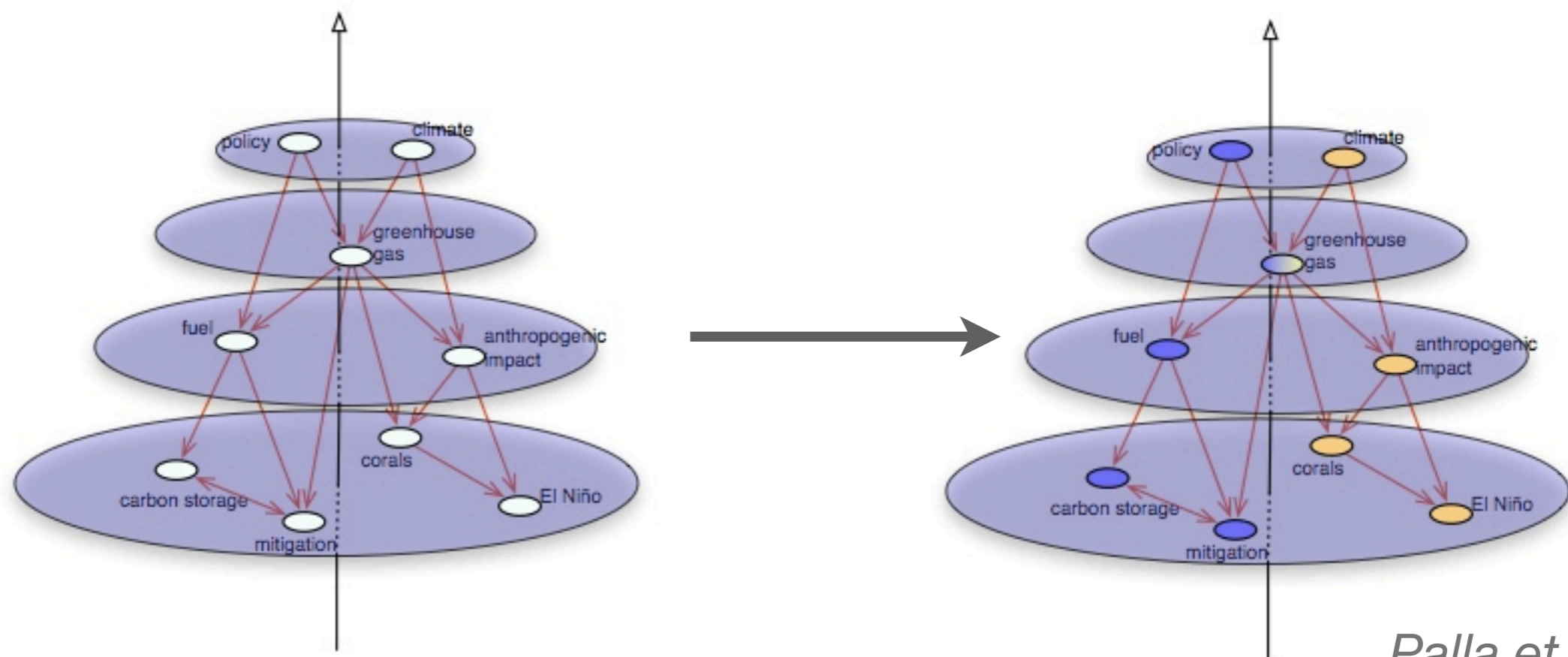
$$S(w_1, w_2) = \frac{\sum_{\{c, I(c, w_1) > 0, I(c, w_2) > 0\}} I(c, w_1)}{\sum_{\{c, I(c, w_1) > 0\}} I(c, w_1)}$$

$$I(c, w_1) = \log \frac{p(c, w_1)}{p(c)p(w_1)}$$

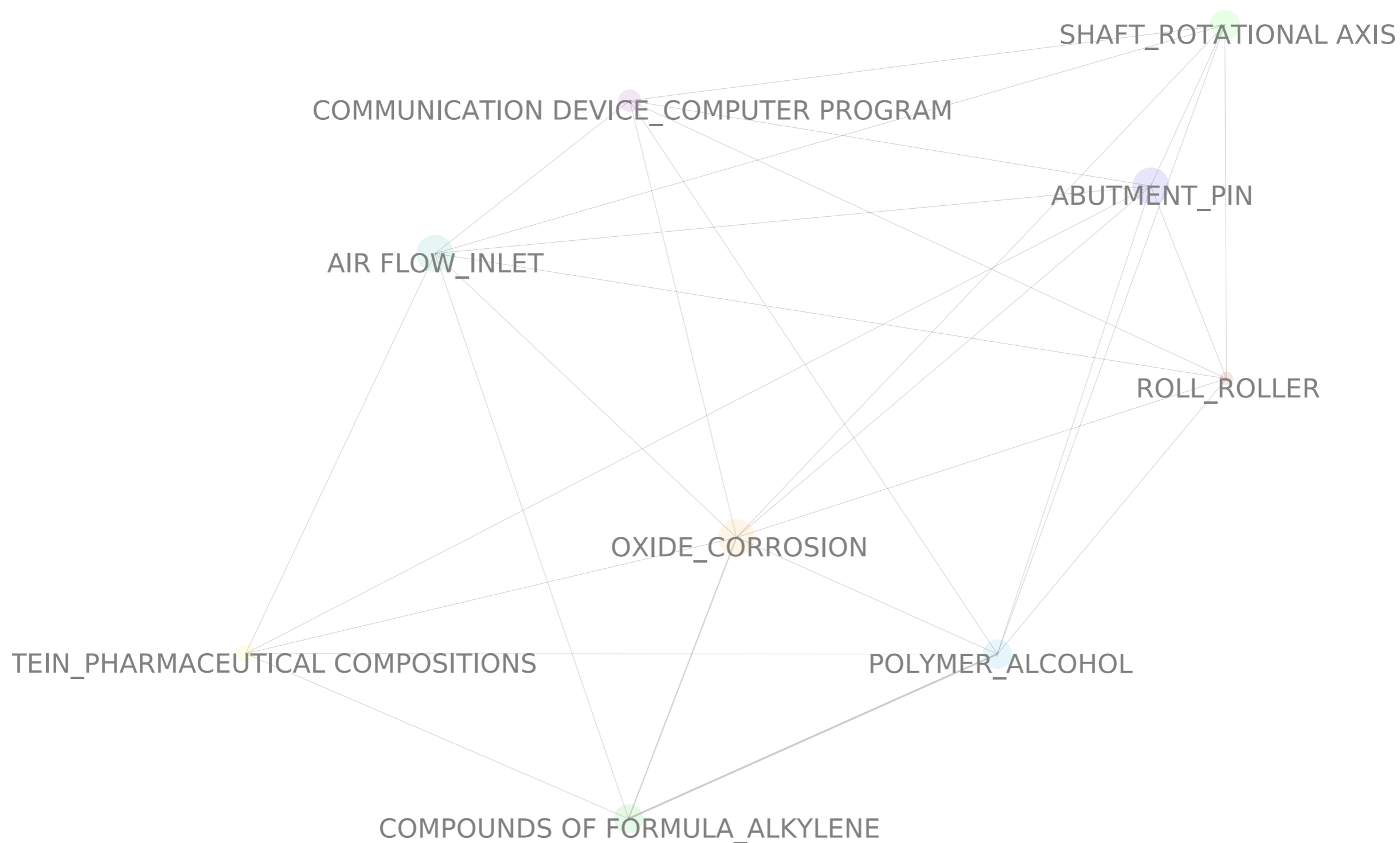
Weeds & Weir, 2005

# What next ?

- Reconstruction of the cognitive dynamics in science through the analysis of the **lexical network** built upon the temporal matrix of co-occurrences within our term list (asymmetric measure of proximity between terms).
- Overlapping clusters detection

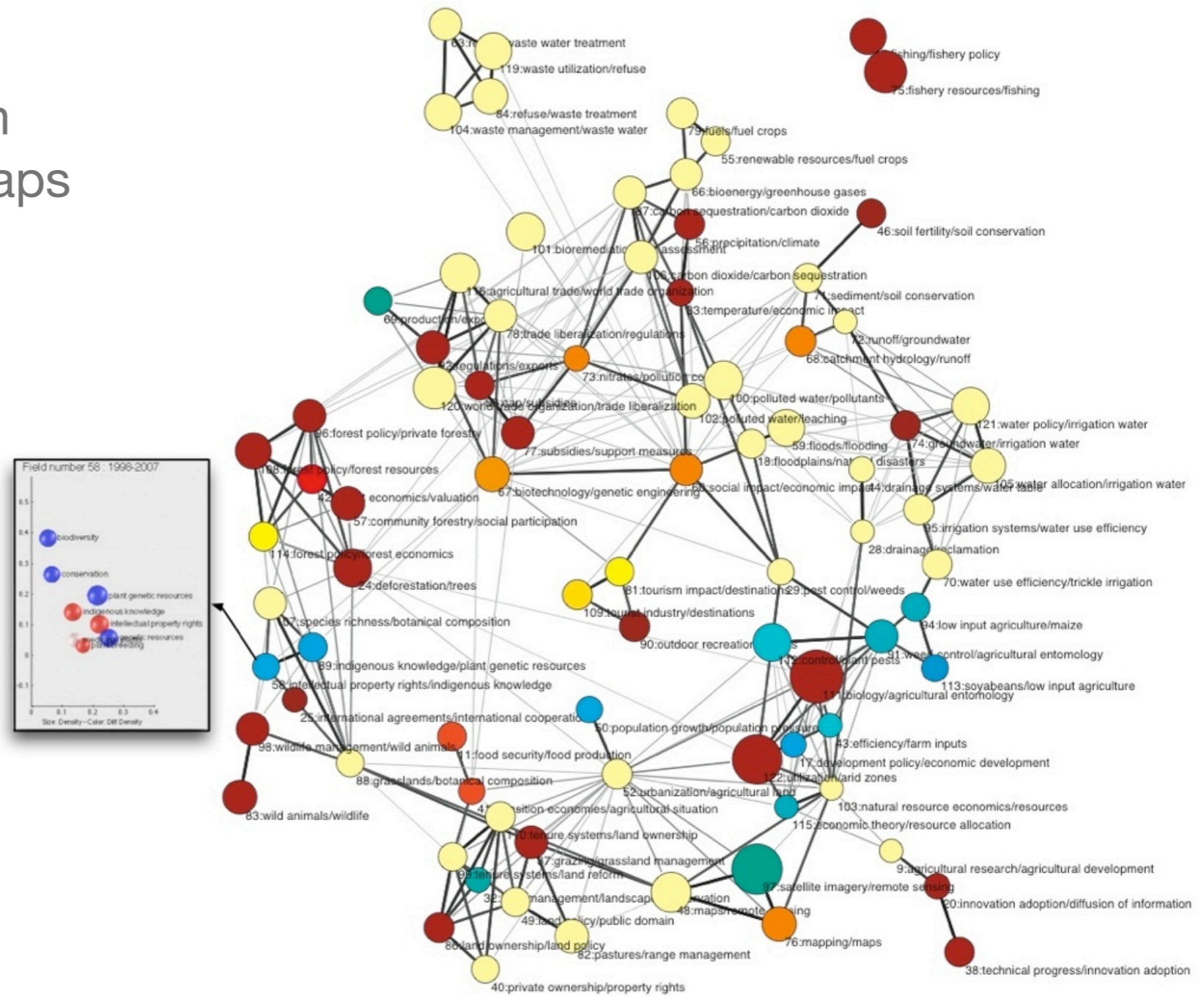






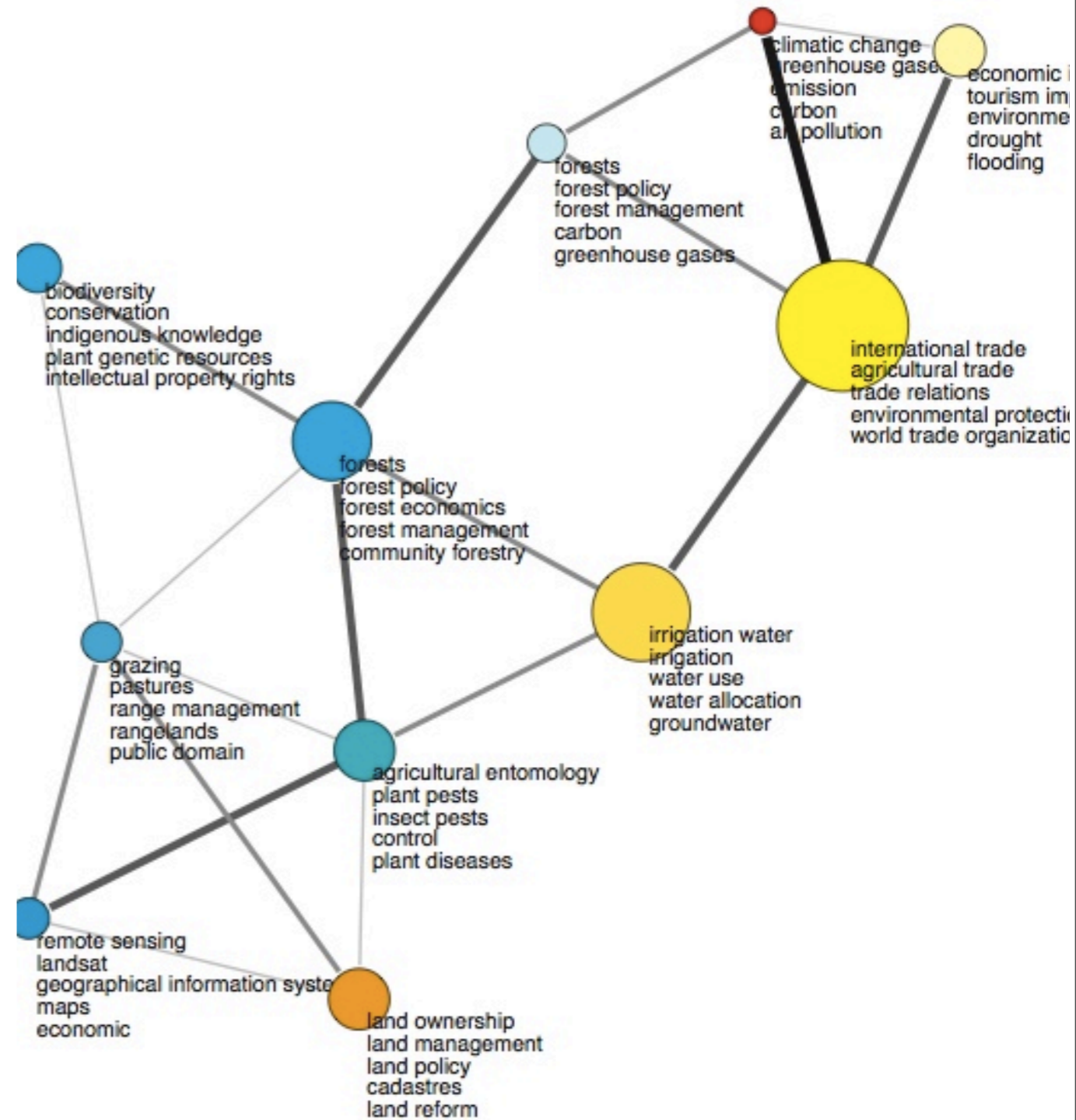
European Patents semantic cartography (High level)

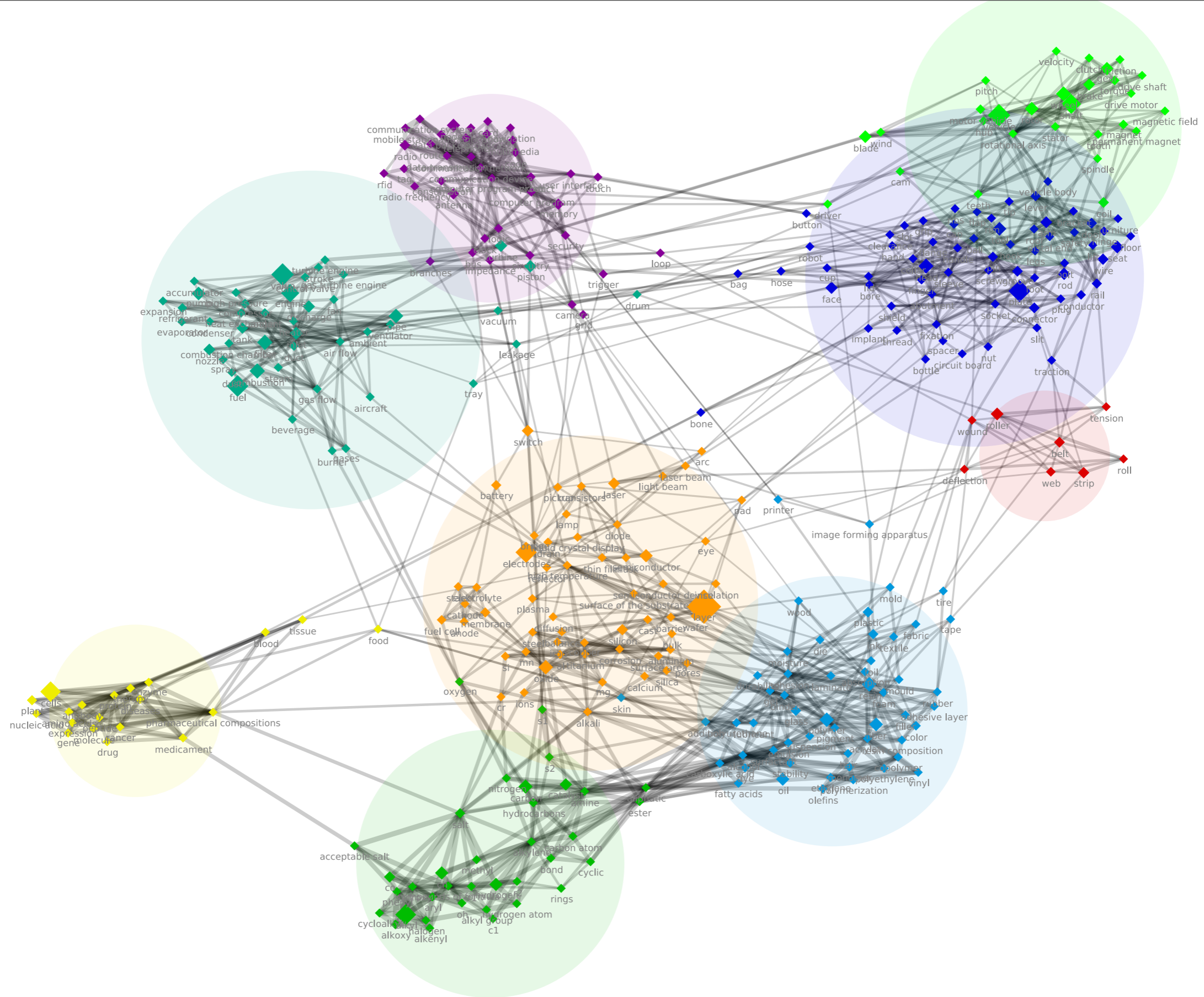
- Semantic distance between clusters build multi-level maps

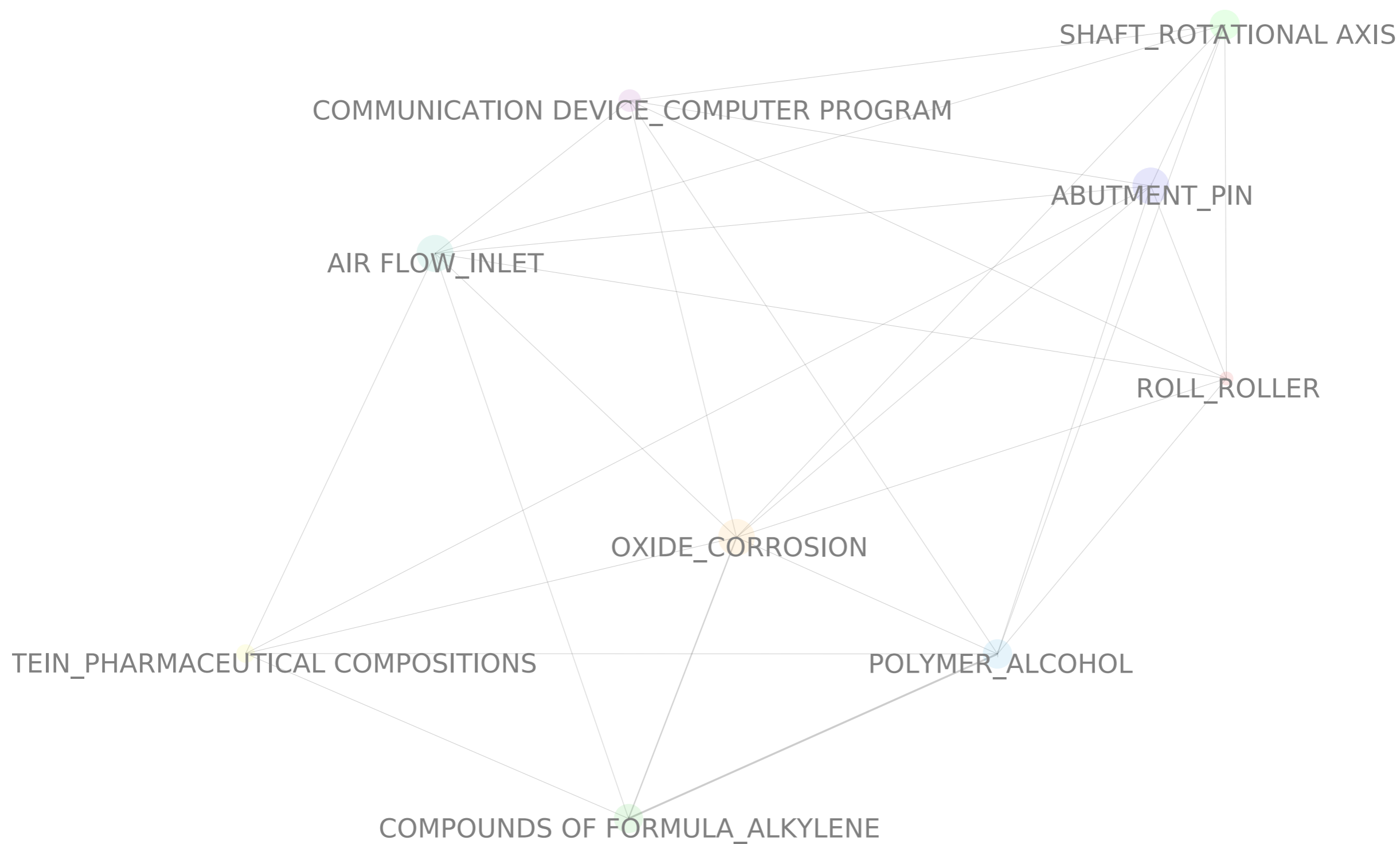


# From Clusters to tubes

- Semantic distance between clusters build multi-level maps

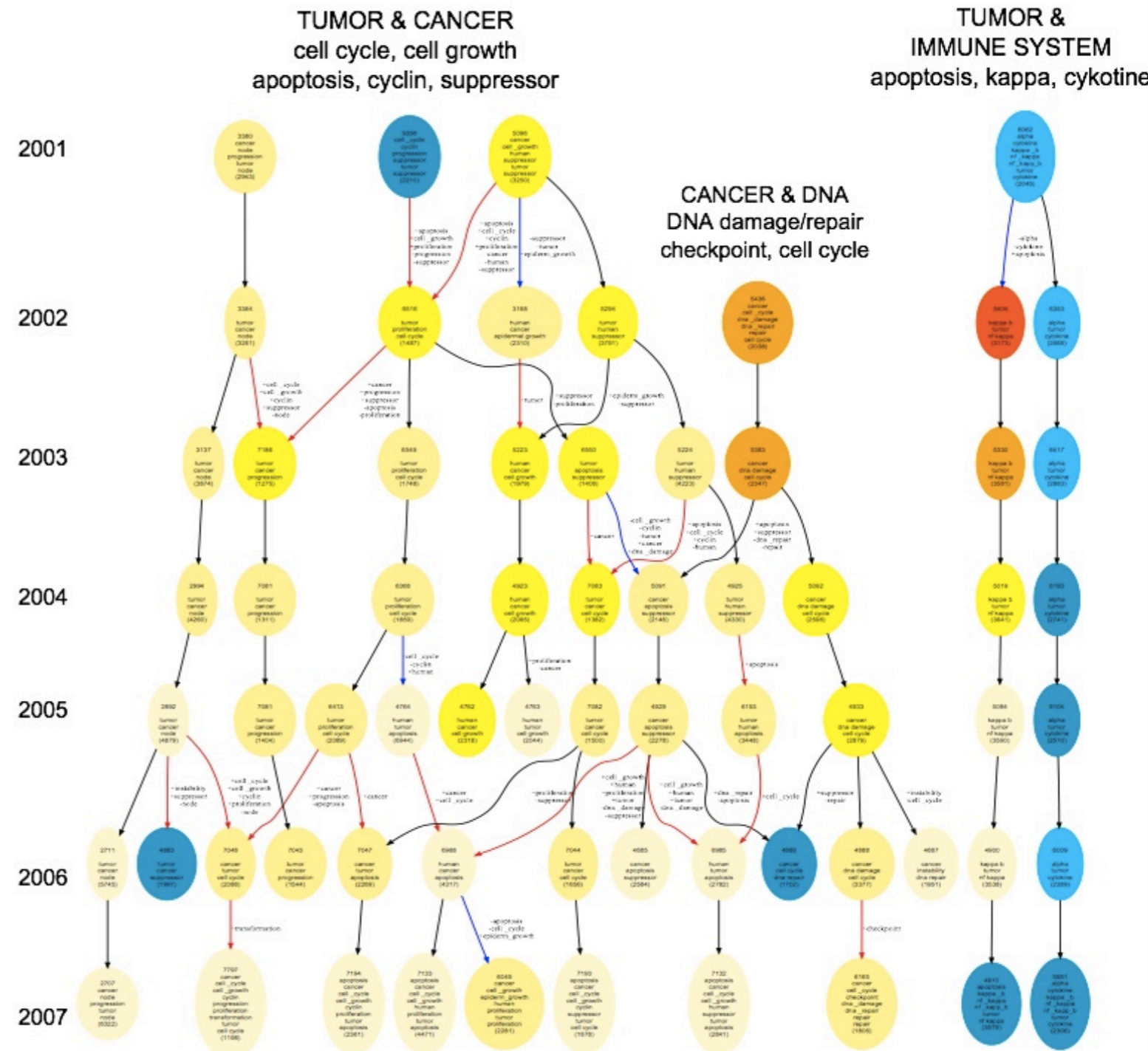






# From Clusters to tubes

- Semantic distance between clusters build multi-level maps
- A semantic phylogenetic network is built by matching thematic fields inter-temporally



# From Clusters to tubes

- Semantic distance between clusters build multi-level maps

- A semantic phylogenetic network is built by matching thematic fields inter-temporally

- This structure can be enriched by synchronic proximities to build knowledge tubes

