Capturing Oncology Dynamics from Textual Content of Conference Abstracts

---

Word Embedding and Stochastic Block Models

Ale & J-P

Sous le haut-patronage de Pascale Bourret & Alberto Cambrosio

IXXI - Lyon - 23/11/2017

Metaknowledge Project

Ludwig Fleck thought collectives

The force which maintains the collective and unites its members is derived from the community of the collective mood. This mood produces the readiness for an identically directed perception, evaluation and use of what is perceived, i.e. a common thought-style,

"*The Problem of Epistemology*", 1935

Thinking is a collective activity (…). Its product is a certain picture, which is visible only to anybody who takes part in this social activity, or a thought which is also clear to the members of the collective only. What we do think and how we do see depends on the thought-collective to which we belong,

Scientific Observation and Perception in General, 1936

Observations from the Field

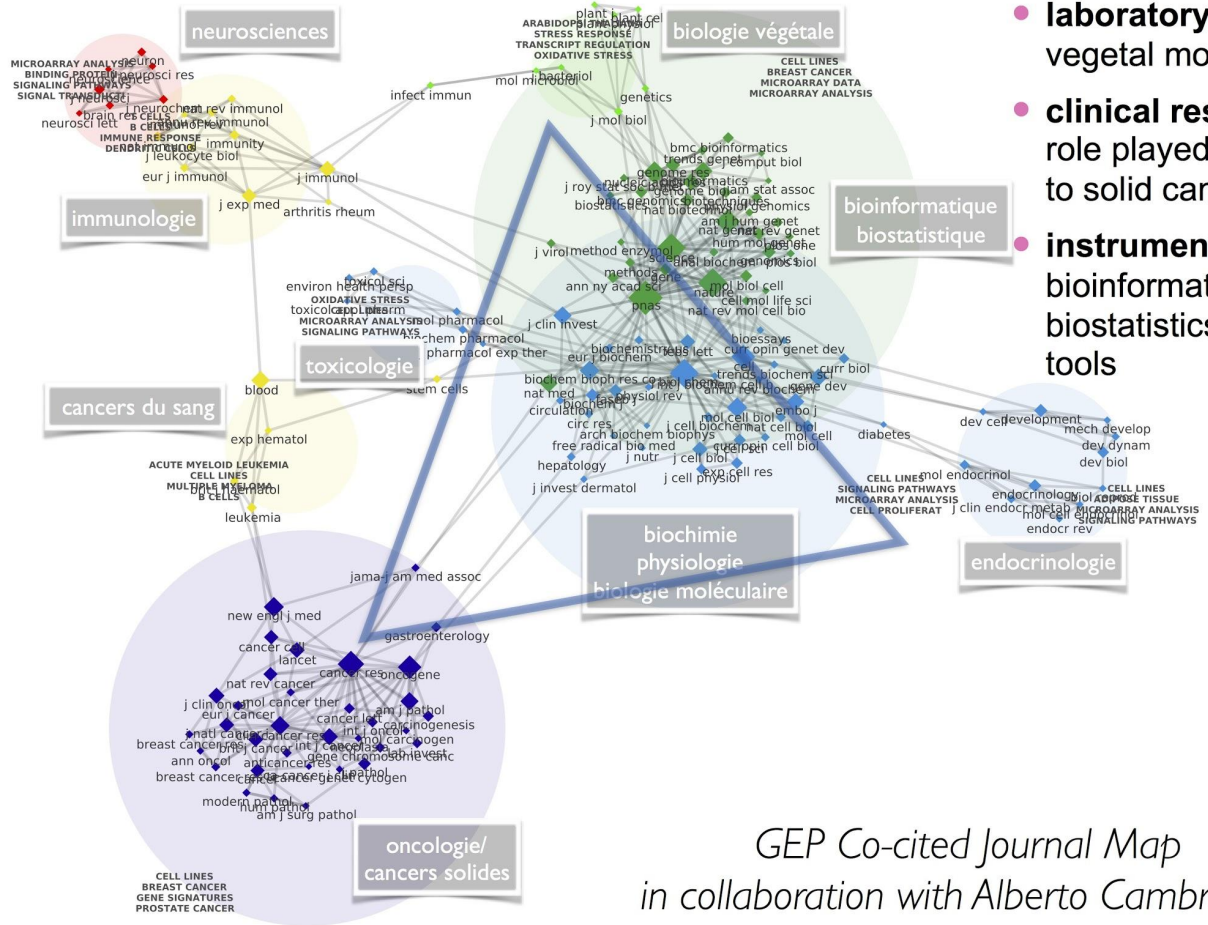
- Oncology has shifted from a small, marginal domain within biomedicine to one of the largest, most central and successful pioneering the most innovative approaches to translational research

- A relatively small number of oncologists—the oncology “core set”—appear to define the international research and treatment agenda.

- Oncology’s core-set is closely involved in onco-policy and politics—regulatory, organizational, and health policy implications of oncology.

- The core set is not homogeneous and stable but characterized by the presence of controversy and shifting fault-lines

Translational Research Triangle



- **laboratory research** - vegetal model organism
- **clinical research** - central role played by application to solid cancer
- **instrumental research** - bioinformatics and biostatistics as necessary tools

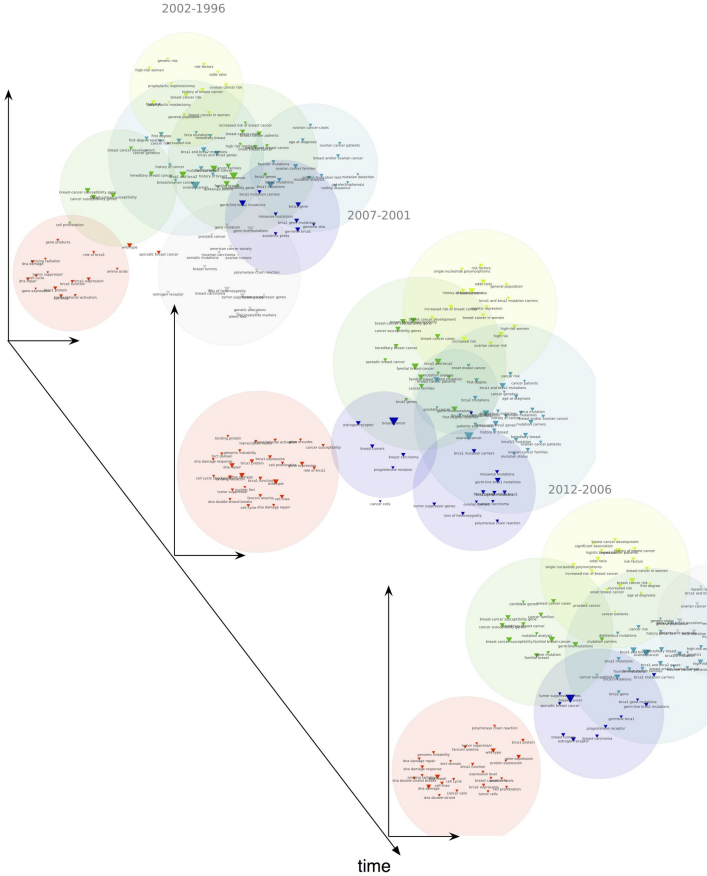
GEP Co-cited Journal Map
in collaboration with Alberto Cambrosio

Limits of traditional network analysis

- Network are unfit to model assemblages or “agencements”
- The collectives we are looking for are heterogeneous

,

- The transformation of the entities making up a heterogeneous collective as the cause, rather than consequence, of the dynamics of these collectives
- Attempts to account for dynamical processes often rely on the structural comparison of the ‘same’ network at different times, pointing to the elements that are held responsible for the observed changes…



Outline

- **0 - Data** - oncology related abstracts

- **1 - Word Embedding** - capturing influence of abstracts over time

- **2 - SBM** - characterizing temporal patterns of oncology subfields

Data

ASCO conference abstracts dataset:

- **What is the role of conferences in oncology research ?**
- **How is it different from publications ?**
- **Educated opinion is: innovative research, not routine papers**

Pubmed abstracts datasets, two fields in contrast:

- **Breast is an active field with several innovations and investment cycles**
- **Lung was an arid field until recent advances unlocked innovations**

“ A word is characterized by the company it keeps ”

Firth, John R. "A synopsis of linguistic theory, 1930-1955."

Word-embedding

Compress the (word, context) co-occurrence matrix M into a (word, s in S) matrix such that dot_S(word, sum(context)) approximates log(M).

If co-occurrence approximates meaning, distances in S are a proxy for semantic similarity between words:

Nearest(Sao Paulo) = [São Paulo, Rio de Janeiro]

If relative co-occurrence approximates semantic relationship (distributional hypothesis), arithmetic operations in S are semantically meaningful:

Paris / France ~ Rio de Janeiro / Brazil (tourism > politics)

Lyon - France + Brasil = ?

Word-embedding

Word-embedding

Studying temporal evolution of corpora

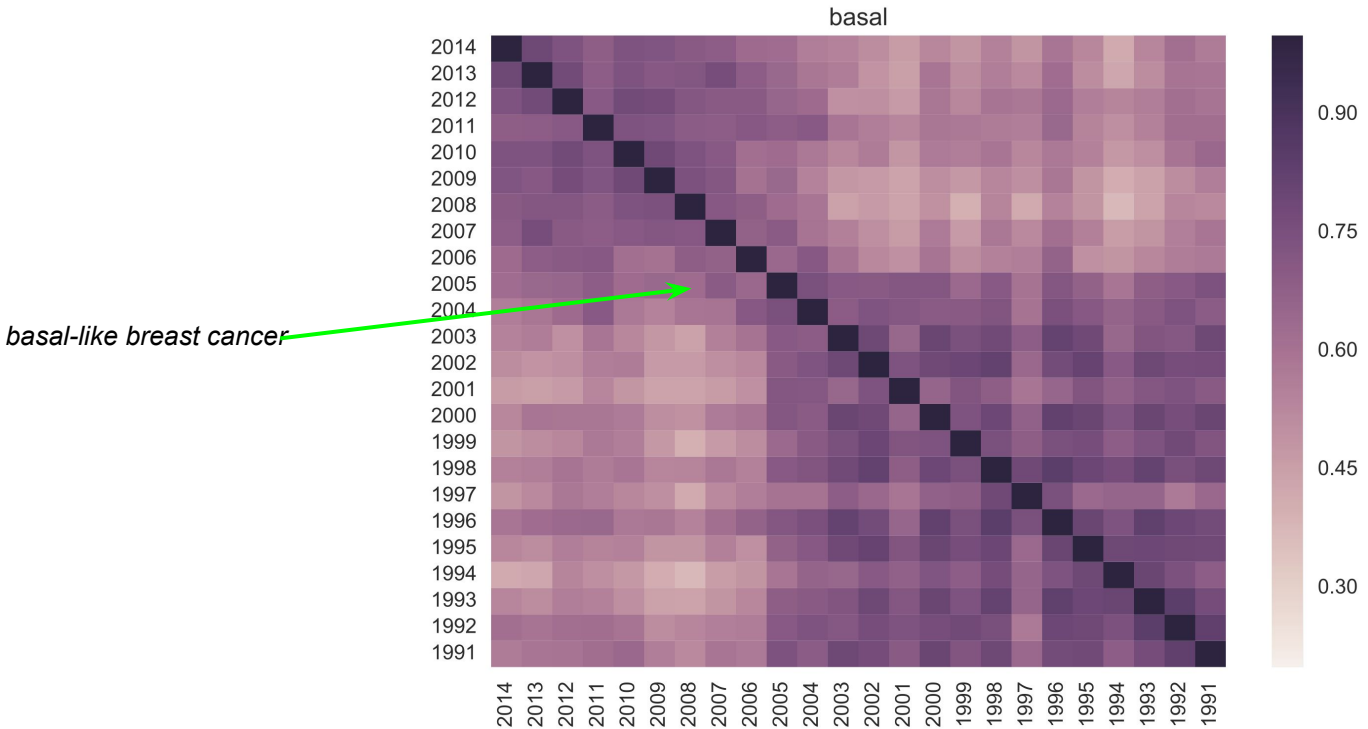
Train a different model for every year and:

- **Geometric semantic space (Mikolov)**
  - **Compare the bulk geometric structure between years**

- **Likelihood score (Taddy)**
  - **Compare the likelikhood of a document in different years**

Word-embedding

geometric semantic space



Capturing word most dramatic shifts of meaning

Word-embedding

likelihood scores

Goals:

- **Provide a contextualized reading experience**
- **Detect interesting abstracts**
- **Treat influence at different levels: paper, author, institution etc**

Word-embedding

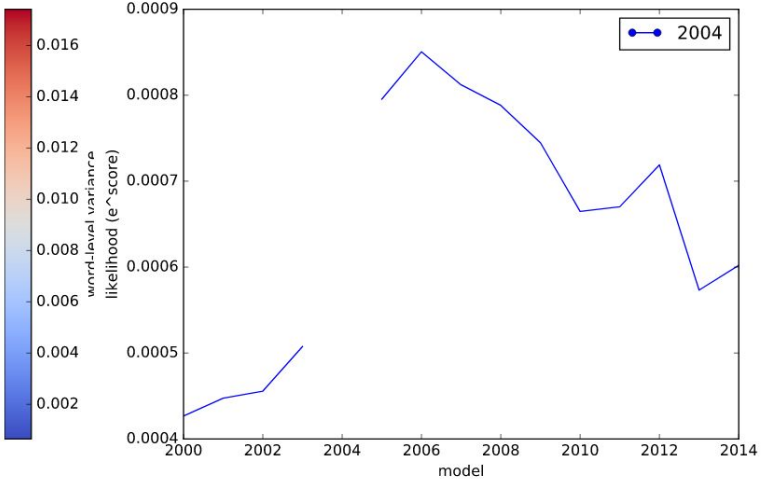
likelihood scores

15329413

EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib.
William Pao, Vincent Miller, Maureen Zakowski, Jennifer Doherty, Katerina Politi, Inderpal Sarkaria, Bhuvanesh Singh, Robert Heelan, Valerie Rusch, Lucinda Fulton, Elaine Mardis, Doris Kupfer, Richard ...
(Program in Cancer Biology and Genetics and Department of Medicine, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021, USA. paow@mskcc.org)
Proc. Natl. Acad. Sci. U.S.A.

somatic_mutations in the tyrosine_kinase tk_domain of the epidermal_growth factor_receptor egfr gene are reportedly associated with sensitivity of lung cancers to gefitinib_iressa kinase_inhibitor in-frame_deletions occur in exon whereas point_mutations occur_frequently in codon exon we found from sequencing the egfr_tk domain that of gefitinib-sensitive tumors had simila… …ations no mutations were found in eight gefitinib-ref… …rs we of seven tumors sensitive to erlotinib_tarceva relate… …inhibitor for which the clinically_relevant target is undocumented had anal… …tic_mutations as_opposed to none of erlotinib-refractory tumors because … …tion-positive tumors were adenocarcinomas from patients … …d cigarettes in lifetime_never smokers we screened egfr_exons 2-28 in adenocarcinomas resected from untreated never_smokers seven tum… …mutations in contrast to of non-small_cell lung cancers resected from untreated former or current_smokers immunoblotting o… …om cells transiently_transfected with various egfr constructs demonstrated that compared to wild-type protein an exon_deletion mutant i… …diminished levels of phosphotyrosine whereas the phosphorylation at tyrosine of an exon point mutant was inhibited at 10-fold_lower concer… …of drug collectively_these data show_that adenocarcinomas from never_smokers comprise distinct_subset of lung cancers frequently containing mutations within the tk_domain of egfr that are associated with gefitinib and erlotinib sensitivity

```
Var: 0.004282125418188857
Mean: 0.968740263042652 9
2000    0.970795
2001    0.970358
2002    0.967437
2003    0.959620
2004         NaN
2005    0.972311
2006    0.974393
2007    0.963792
2008    0.973453
2009    0.965205
2010    0.971221
2011    0.967645
2012    0.967476
2013    0.965260
2014    0.973398
Name: 34, dtype: float64
```



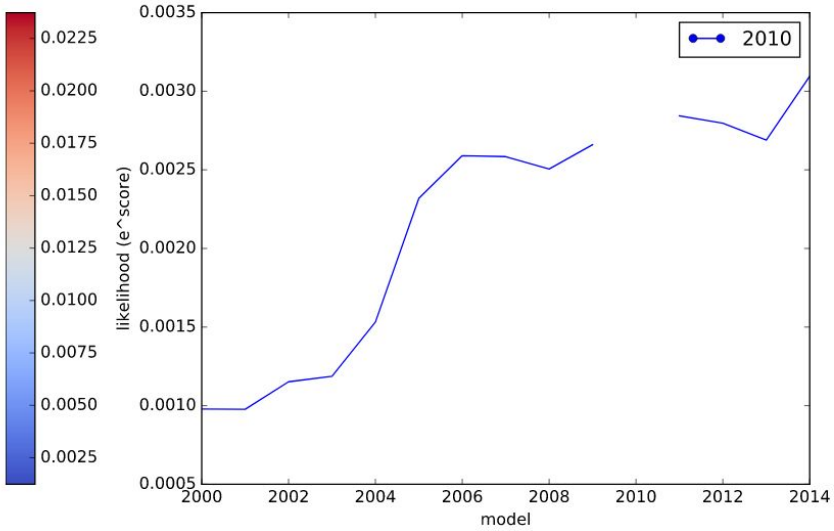
Word-embedding

likelihood scores

20186026

De novo resistance to epidermal growth factor receptor-tyrosine kinase inhibitors in EGFR mutation-positive patients with non-small cell lung cancer.
Masayuki Takeda, Isamu Okamoto, Yoshihiko Fujita, Tokuzo Arao, Hiroyuki Ito, Masahiro Fukuoka, Kazuto Nishio, Kazuhiko Nakagawa
(Department of Medical Oncology, Kinki University School of Medicine, Osaka-Sayama, Osaka, Japan.)
J Thorac Oncol

somatic_mutations in the epidermal_growth factor_receptor egfr gene are predictor of response to treatment with egfr_tyrosine kinase_inhibitors tkis in patients with non-small_cell lung cancer nsclc however mechanisms of de_novo resistance to these drugs in patients harboring_egfr mutations have remained_unclear we_examined whether the mutational_status of kras might_be associated with primary resistance to egfr-tkis in egfr_mutation-positive patients with nsclc forty patients with nsclc with egfr_mutations who were treated with gefitinib or erlotinib and had archival_tissue specimens available were enrolled in the study kras_mutations were analyzed by direct_sequencing three of the patients had progressive_disease and two of these three individuals had both kras and egfr_mutations our_results suggest_that kras_mutation is negative predictor of response to egfr-tkis in egfr_mutation-positive patients with nsclc



Under closer inspection, some words associate with the main jump on 2005, others with the small jump on 2014

Stochastic Block Models

- **Generate networks by partitioning nodes into blocks with similar connectivity patterns towards other blocks**

- **Patterns are generic, not assortative or dissortative: independent probabilities for links from nodes in one block to nodes in another**

$$p(b_i, b_j)$$

- **Can take into account degree sequences within groups**
- **Can be hierarchical: a block model is also a graph**

Stochastic Block Models

- **SBM model selection: given a graph, find the model (partition and probabilities) whose generated graphs *best resemble* the original**
- ***Resemblance* according to some criterion**
- **Provides an abstraction of the network based on connectivity patterns**

- **Minimum description length (MDL) is proposed as criterion (Peixoto)**
- **Finds the partition that minimizes the sum of model information and the information needed to recover the exact graph from the model**
- **Is non-parametric**

Stochastic Block Models

Modeling the oncology literature with hierarchical SBMs

Modeling a bipartite network of (documents, contents) lets us establish a simultaneous partition of publications and words

- **Publication partitions lets us connect the partition to exogenous variables such as authors, institutions, time and existing categories**
- **Word partitions lets us understand the semantic connections present in the corpus, as well as their relationships to those exogenous variables**
- **A reduced vocabulary makes networks manageable (|V|~200K, |E|~2M)**
- **Partition temporal sub-corpora to understand stability of total partition**

Stochastic Block Models

Analysis by our high patrons P.B. & A.C.

Document count across time: sanity checking, with some insights

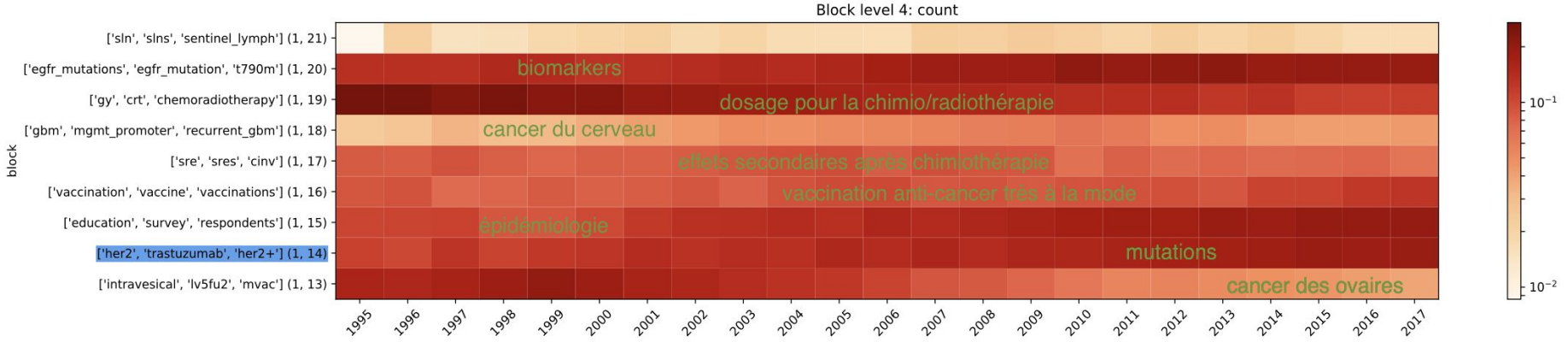
(pause to show the crazy matrices)

Stochastic Block Models

Analysis by our high patrons P.B. & A.C.

On the overall partitioning

- **It doesn't seem to follow professional specialities (surgery, radiotherapy etc) nor anatomic divisions (different cancers appear together), but instead “research fronts”**



Stochastic Block Models

On the research fronts

At the highest partition level

- **Growing importance of biomarkers and targeted therapies**
  - **(1, 20) egfr and egfr mutations; major receptor, oncogene, pathway component; biomarker for targeted therapies**
  - **(1, 14) her2 and trastuzumab: biomarker for a subtype of breast cancer; targeted therapy for her2**
- **Diminishing presence of traditional chemotherapy**
  - **(1, 13) mvac, llv5fu2, intravesicular; chemotherapies for different organs**
  - **(1, 17) cinv, sre/sres; side-effects of chemotherapy**

Stochastic Block Models

On the research fronts

At the highest partition level

- **Vaccines: less strong signal, but interpretable digging deeper**
  - **(1, 16) attempts to develop anti-cancer vaccines, with rise of more recent efforts**
- **Chemoradiotherapy: needs further investigation**
  - **(1, 19) combination of radiotherapy with chemotherapy, still present but way less**

Stochastic Block Models

On the research fronts

Down the ~~rabbit hole~~ hierarchy

- **(1, 20), (1, 14) biomarkers and targeted therapies**
  - **(1, 20, 77) the one rising: treatment mutations: egfr, kras**
  - **(1, 20, 91) rising not as much: immunotherapy, ipilimumab is the pioneer**
  - **(1, 20, 85); (1, 20, 88) descending, related to traditional therapies; stable, blood cancers**
- **(1, 13), (1, 17) traditional chemotherapy**
  - **(1, 13, 74) breast and ovarian cancer**
  - **(1, 13, 64) colorectal cancer**
- **(1, 16) vaccination**
  - **Actually some subgroups on vaccination, plus other assorted subgroups**
  - **Somewhat mysterious, but with coherent subgroups, deserving of further study**

Stochastic Block Models

On the research fronts

Free fall into wonderland!

- **(1, 20, 85, 306) rise of new endpoints and parameters for clinical trials**
- **(1, 20, 85, 301) decline of phase-based clinical trials**

- **(1, 13, 74, 279, {695, 670}) tamoxifen declines (metastatic breast cancer)**
- **(1, 13, 74, 250, 629) trabectedin declines (soft-tissue-sarcomas)**

Stochastic Block Models

Further Directions

- **We haven't baked the cake yet**
- **Passed the sanity check: work with other exogen variables**
  - **Authors, institutions, location**
- **Interactive, dynamic visualisation of the hierarchy**
- **Obvious improvements within our framework, at a computational cost:**
  - **Average over several fitted SBMs**
  - **Overlapping SBMs**
  - **Tripartite network: context layer**
  - **Synonymity and disambiguation**

Merci bien! Muchas gracias! Vai curíntia! Grazie mille! Thank you! شكرا لكم

Ale & J-P

abdo@member.fsf.org

jeanphilippe.cointet@sciencespo.fr

https://en.wikiversity.org/wiki/The_dynamics_and_social_organization_of_innovation_in_the_field_of_oncology