Topic models

Experiments’ initial results

Hierarchical Dirichlet Process

Pros:

- No need to specify number of topics

Cons:

- Difficult to tune
- Poor topics

To do:

- Tune on cluster to identify hyperparameters

Online Hierarchical Dirichlet Process

Pros:

- Adds documents in an online approach
- As HDP does not require number of topics (only maximum)

Cons:

- As HDP poor topic quality
- Topics and their distribution over documents change with streaming

To Do:

- Improve pre-processing

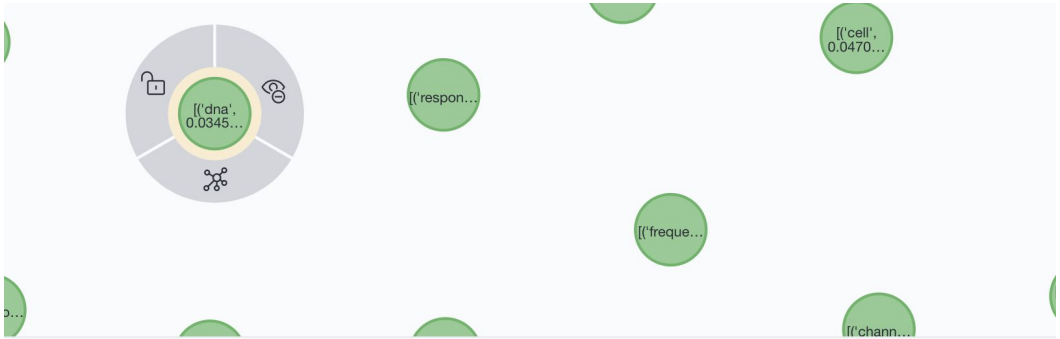
New Avenues to Explore

- Dynamic Topic Models - TM-LDA
  - Developed for online topic modelling transition in Social Media

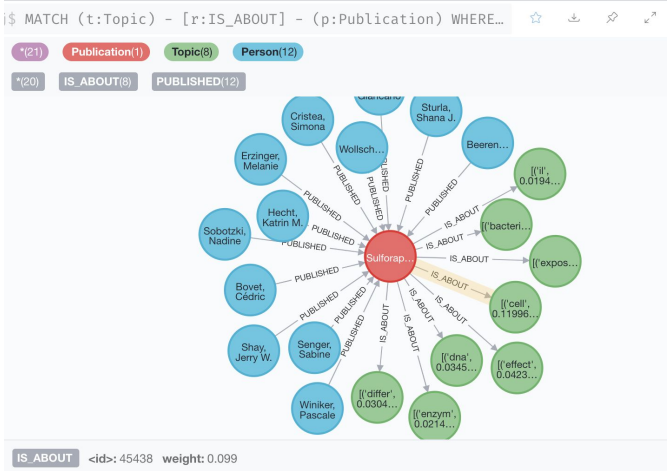
- (Deviate from classical generative approaches $\longrightarrow$ embedding + similarity + clustering)

Updated Graph Structure

- 440k nodes of publications + authors + topics + depts (15) + …
- Each topic has a weighed list of words




```
MATCH (t:Topic) - [r:IS_ABOUT] - (p:Publication)
WHERE p.title = "Sulforaphane Preconditioning Sensitizes Human Colon
Cancer Cells towards the Bioreductive Anticancer Prodrug PR-104A"
RETURN p
```



Leitzahl & Author Disambiguation

- Current info by Leitzahl limited (i.e. maps to single Dept/Org)
- Ideally Leitzahl per person file - more data but still issues with Disambiguation
- Leitzahl hierarchy
- Leitzahl name matching coverage (# complete author+dept pairs):
  - With research areas data: 49086
  - With plain cost center data: 62998
  - With pre-processed cost center data: 68999

Graph growth

Upper bound on growth with **O**(T x A $\mathbf{n^2}$) where n is the number of publications

Current graph takes up ~ 500 MB for 170k publications (12% having abstracts)