Inductive Learning of Concept Representations from Library-Scale Bibliographic Corpora

Lukas Galke[1,2] & Tetyana Melnychuk[2] & Eva Seidlmayer[3] & Steffen Trog[1] & Konrad U. Förstner[3] & Carsten Schultz[2] & Klaus Tochtermann[1]

[1] ZBW, [2] Kiel University, [3] ZB MED

INFORMATIK 2019, Kassel, 26.09.2019

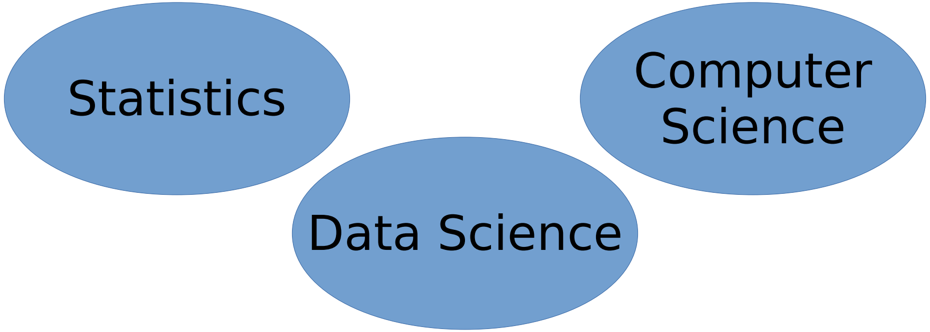
Outline

1) Motivation: **analyses of research dynamics**
2) Problem statement: **learning concept similarity from graph data**
3) Approach: **unsupervised training objective for graph neural nets**
4) Quantitative and (small-scale) qualitative evaluation

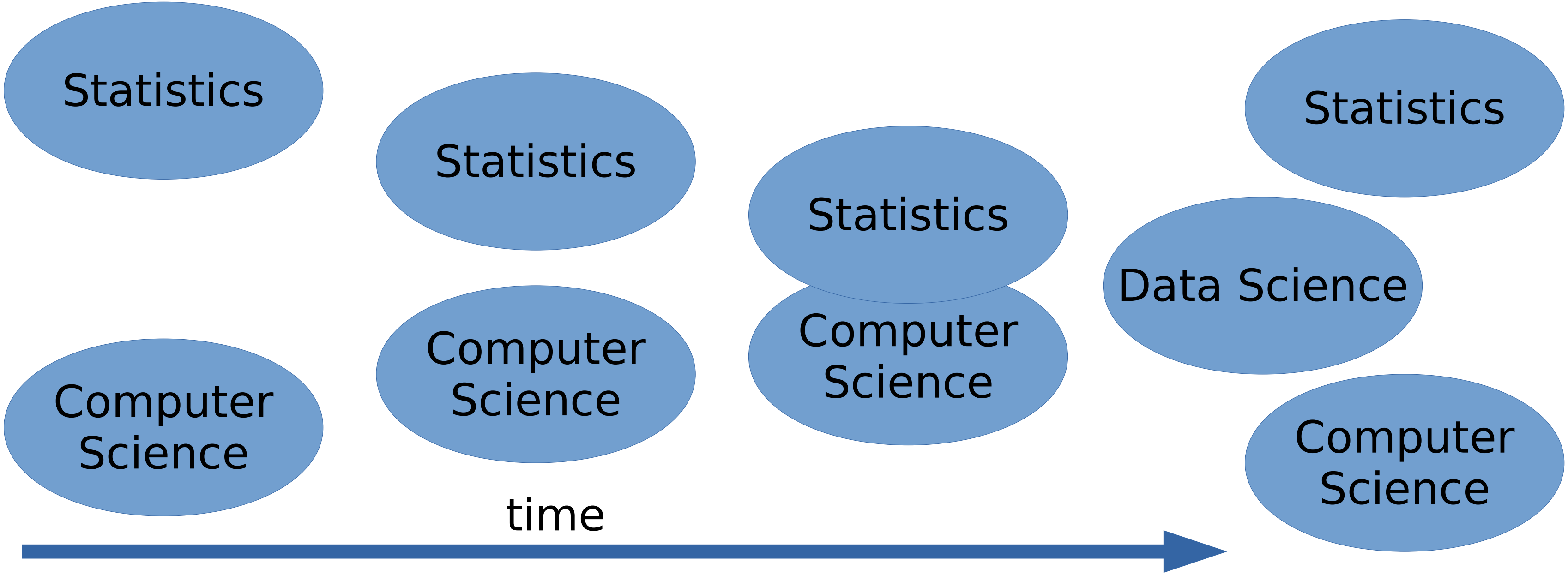
Motivation

- Digital libraries accumulate a large amount of bibliographic data
- Include valuable annotations with controlled vocabularies (concept hierarchies)
- Used for multi-label classification
- Used for information retrieval
- Used for recommender systems
- What about analyses of research dynamics?

Motivation



Motivation: Research Dynamics



Statistics

Computer Science

Statistics

Computer Science

Statistics

Computer Science

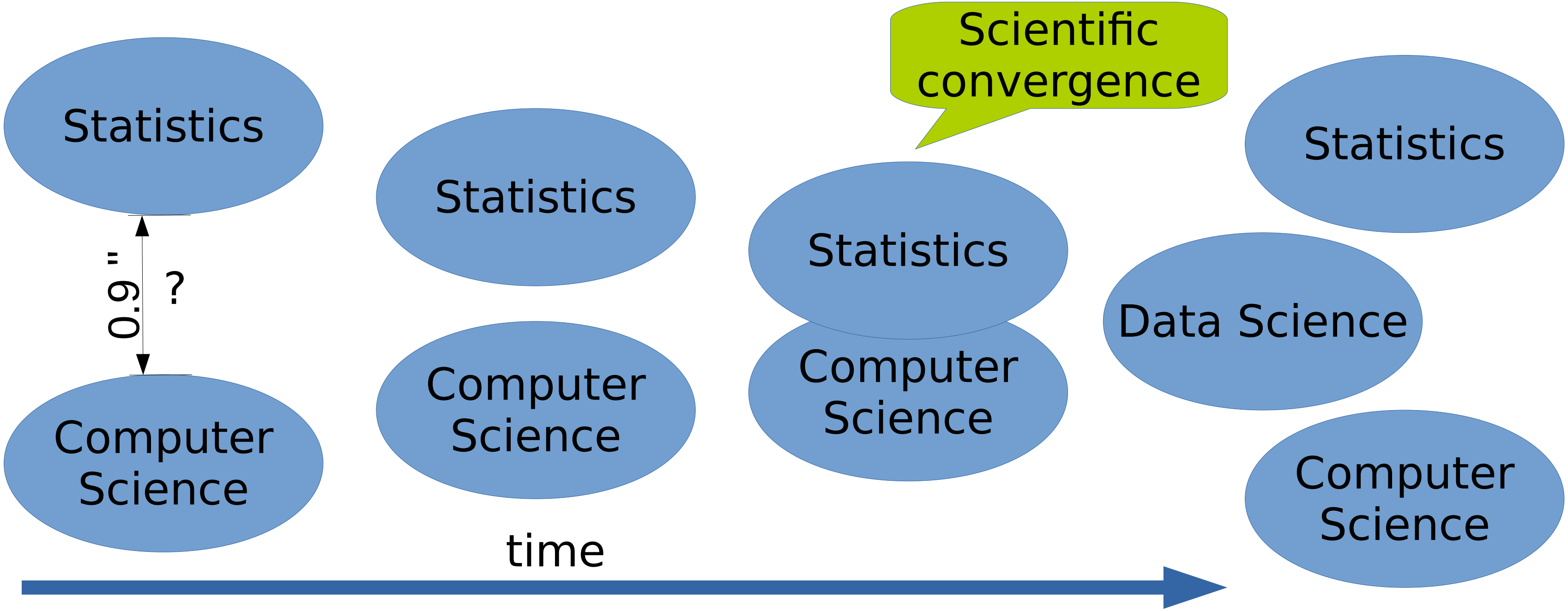
Statistics

Data Science

Computer Science

time

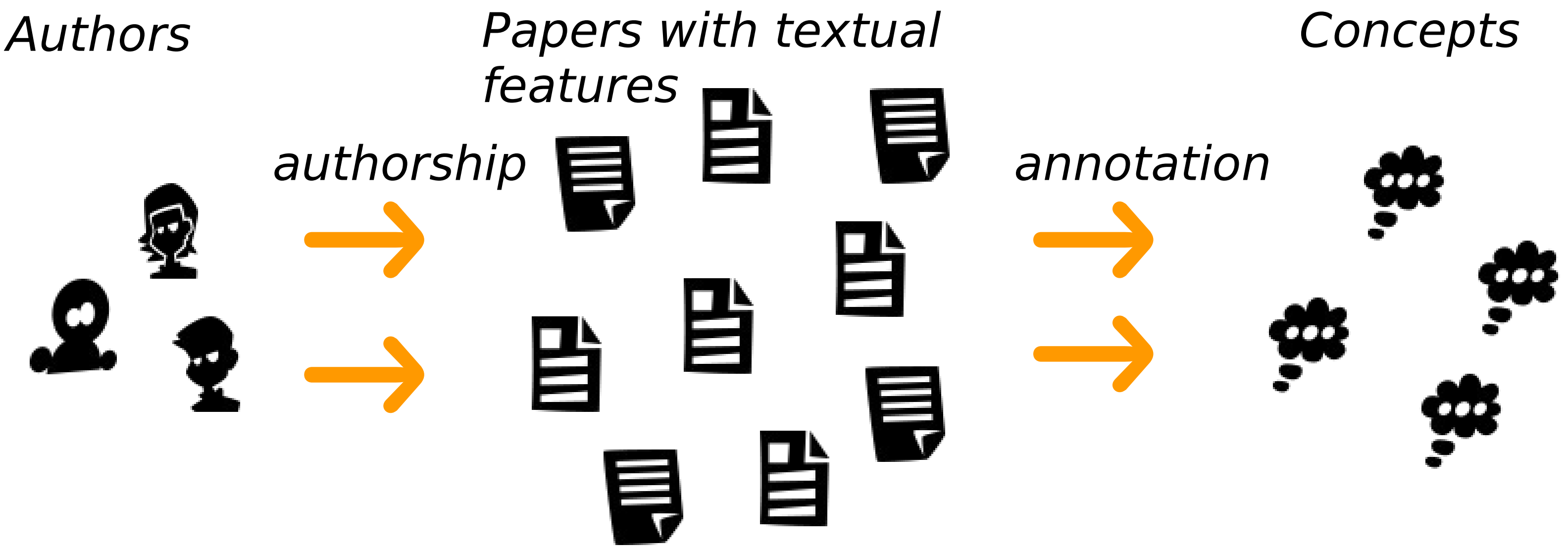
Motivation: Research Dynamics



How to identify scientific convergence?

- Decreasing distance over time
- $\frac{d_{t+1}(a,b)}{d_t(a,b)} < 1$ for t in [t_{start},…, t_{end}], distance metric d, arbitrary concepts a, b
- Metric learning in pairwise concept space
- Replace d(a, b) by $d^{euclidean}$(f(a), f(b)), where f maps concept to vectors
- **This paper:** *We can learn function* f *from data*

Bibliographic Data



Authors

authorship

Papers with textual features

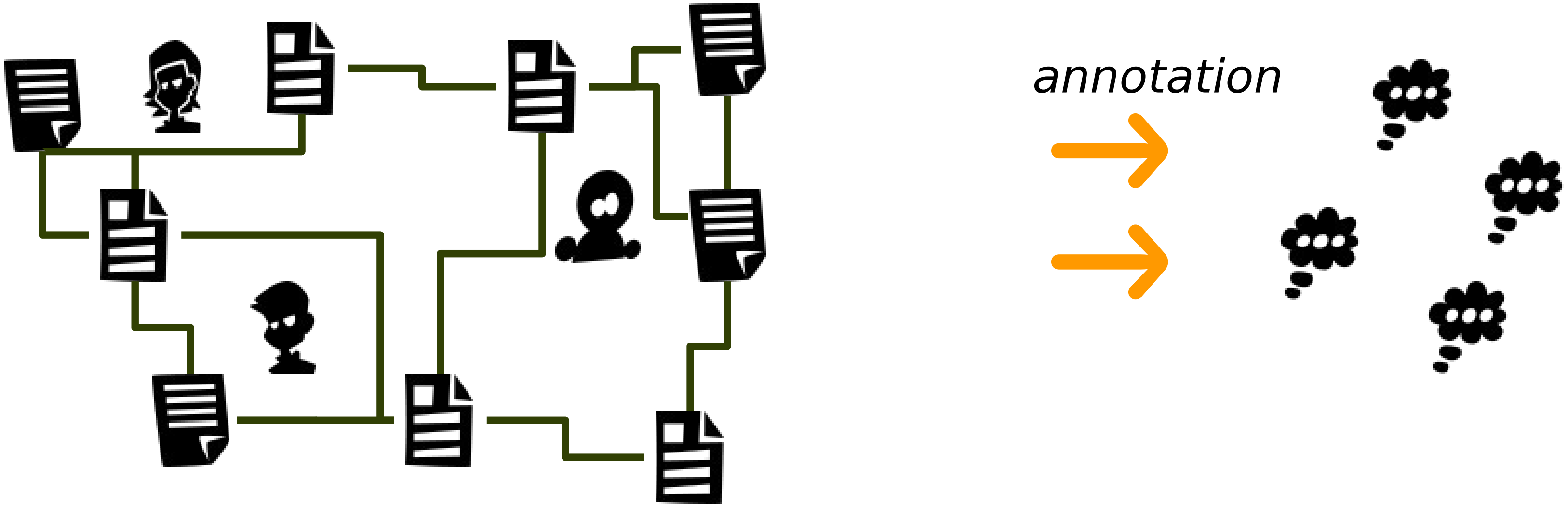
annotation

Concepts

Bibliographic Data

Coauthorship edges between Papers

Concepts



annotation

Problem Statement

Learning Concept Representations

Given:

- a graph (V, E)
- V consists of paper nodes P and concept nodes C
- E consists of co-authorship and annotation edges
- Papers P have textual features (e.g. their title)
- Concepts C have no features

Desired output: *Meaningful* and *useful* vector representations for concepts C

Problem Statement

Learning Concept Representations

Given:

- a graph (V, E)
- V consists of paper nodes P and concept nodes C
- E consists of co-authorship and annotation edges
- Papers P have textual features (e.g. their title)
- Concepts C have no features

Desired output: *Meaningful* and *useful* vector representations for concepts C

Induced similarity corresponds to human judgments

useful for downstream tasks

Transductive vs Inductive Learning

Transductive Learning

- Look-up table for node representations
- Need further training whenever new nodes/edges appear
- Approaches: DeepWalk (Perozzi et al., KDD 2014), node2vec (Grover & Lescovec, SIGKDD 2016), TransE, (Bordes et al., NeurIPS 2013), … (many more)

Inductive Learning

Valuable for dynamic graphs

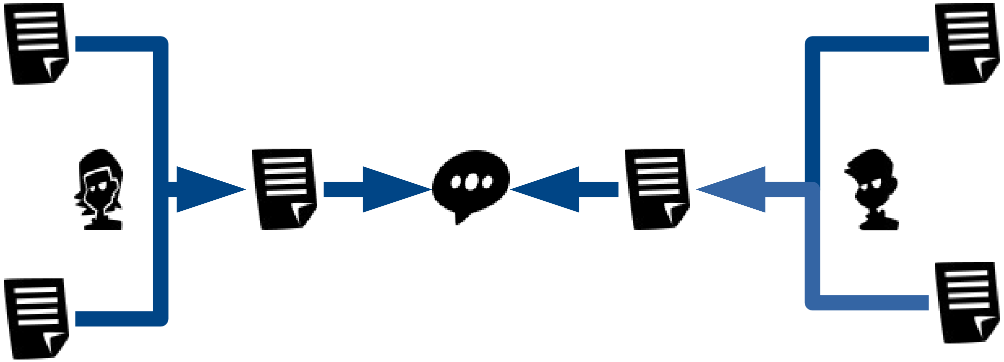
- Node representations solely induced by node features
- *Capable of dealing with unseen nodes/edges without retraining*
- Approaches: GCN (Kipf & Welling, ICLR 2017), GraphSAGE (Hamilton et al., NeurIPS 2017), … (many more)

Graph Convolution (Kipf & Welling 2017)

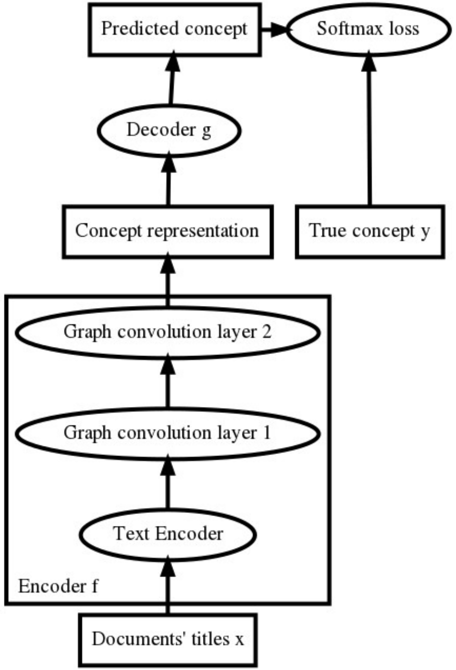
$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{c_{ij}} W^{(l)} h_j^{(l)} + b^{(l)}\right)$$

From layer l to l+1

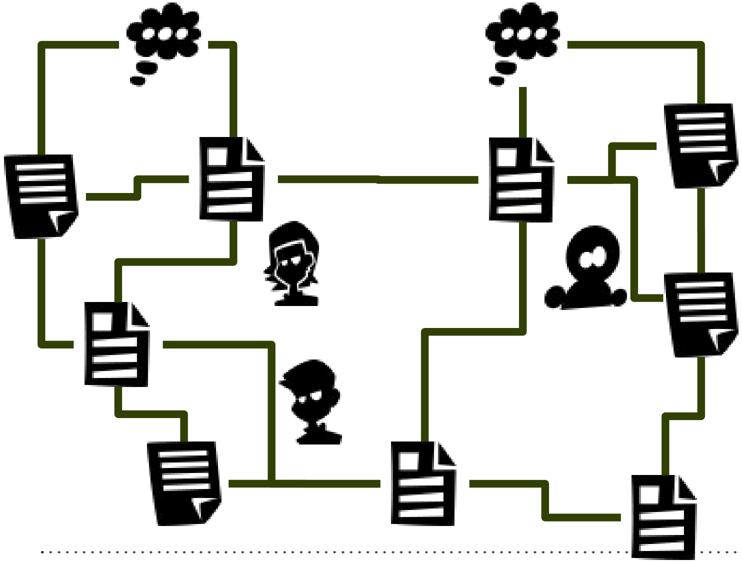
1) Transform via parameters W, b
2) Aggregate neighbor representations
3) Nonlinear activation

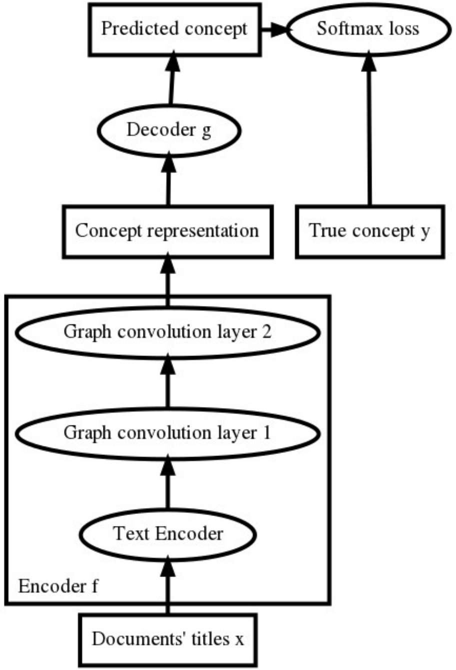


Reconstruction-based Training

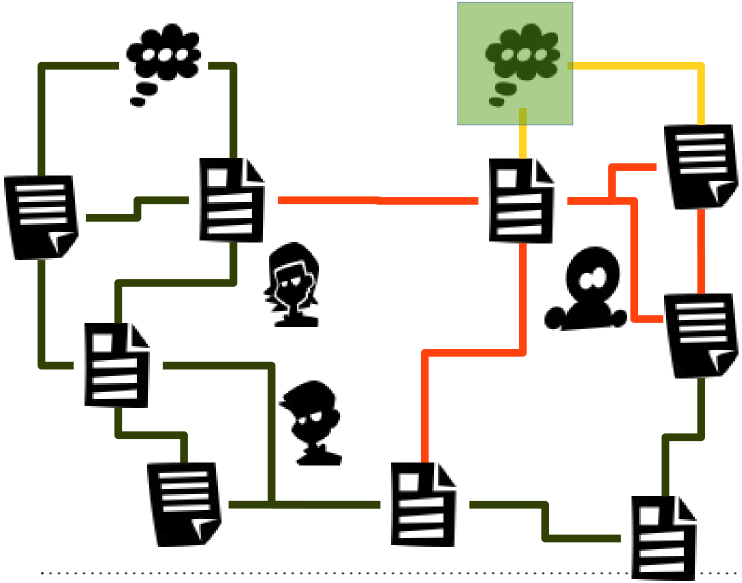


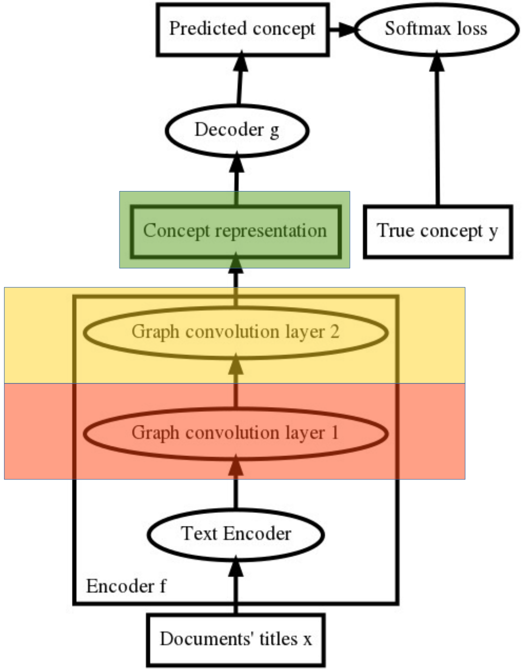
Reconstruction-based Training



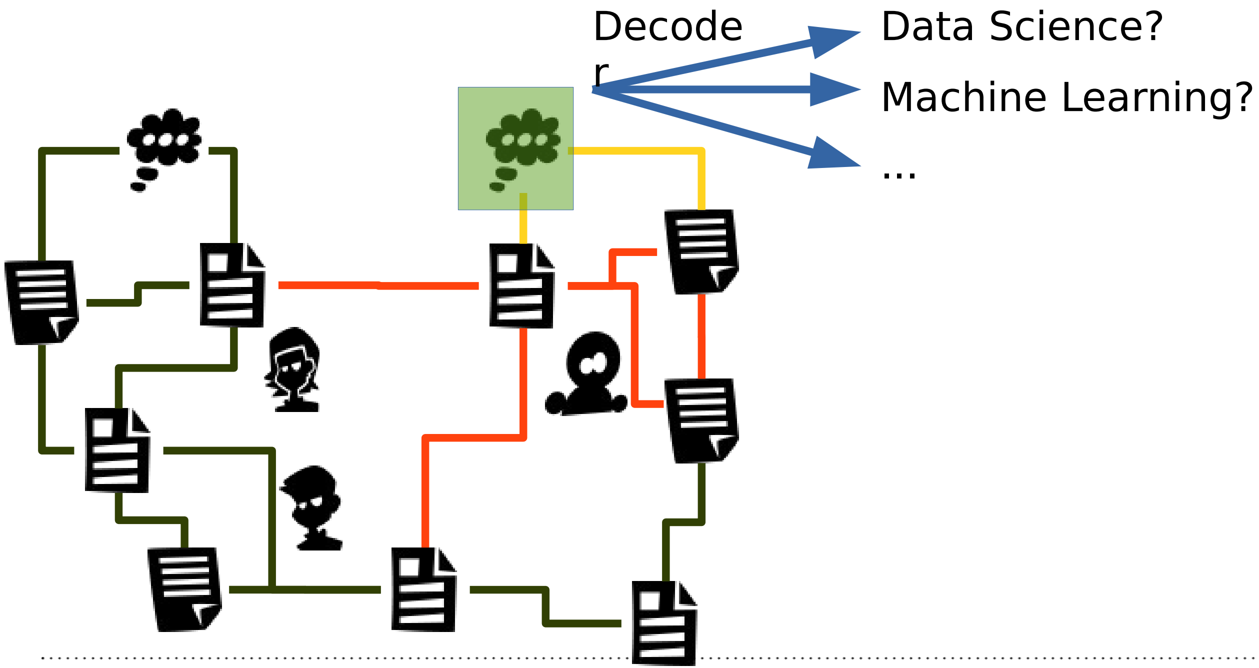


Reconstruction-based Training





Reconstruction-based Training

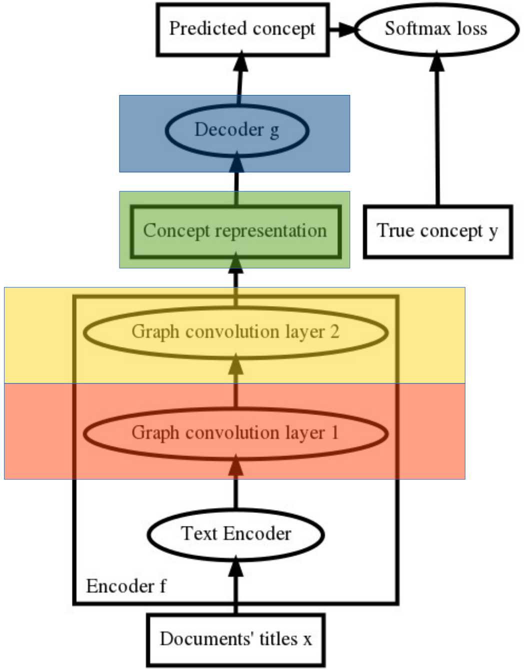


Decode
r

Data Science?

Machine Learning?

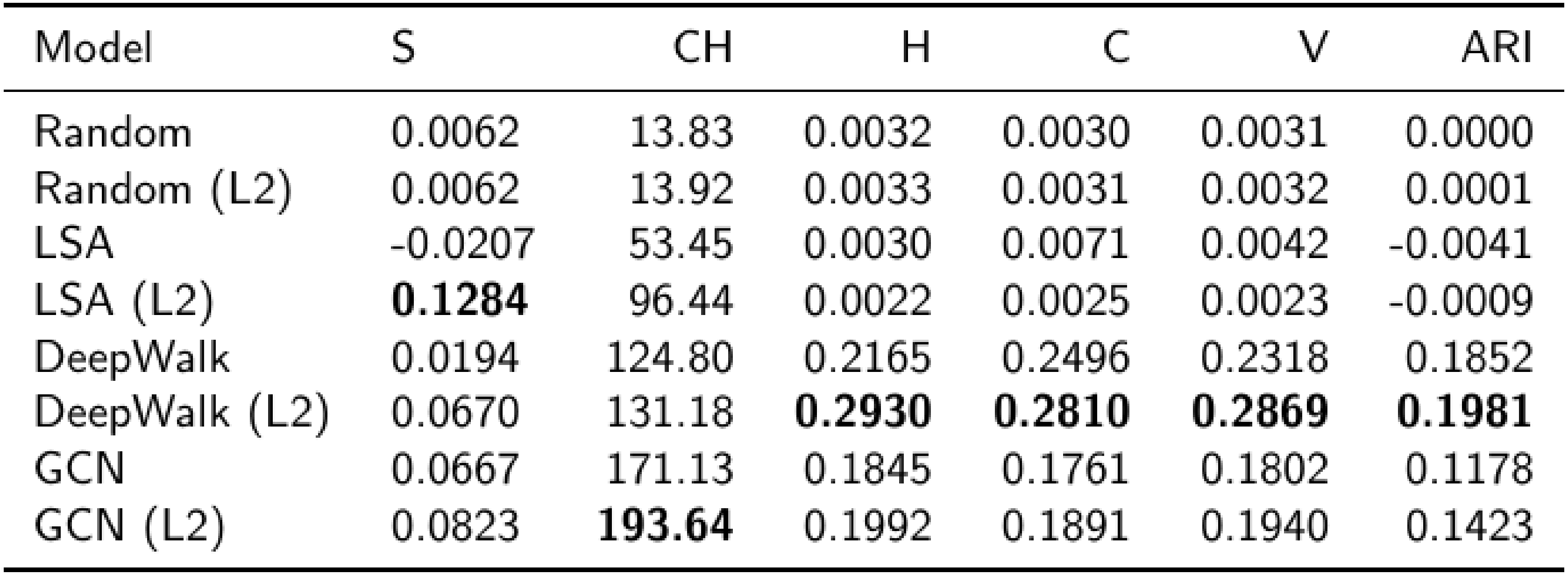
...



Experiments

- 2.1M English research papers from economics and business economics domain
- 5,688 concepts from Standardthesaurus Wirtschaft (http://zbw.eu/stw)
- Quantitative Evaluation
  - Subset of 3,113 concepts, which belong to only one of 7 subthesauri
  - Downstream tasks: clustering and classification
- Qualitative Evaluation:
  - Nearest concept queries
  - Linear operations in latent space

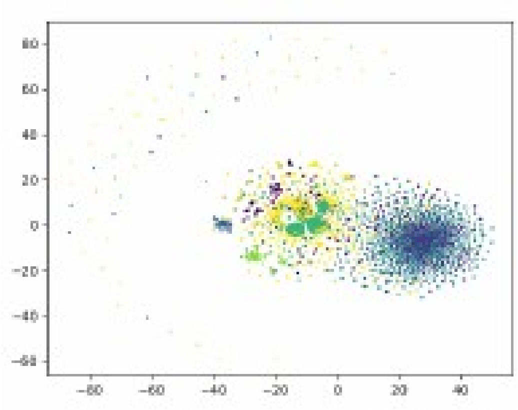
Results: Clustering with k-Means



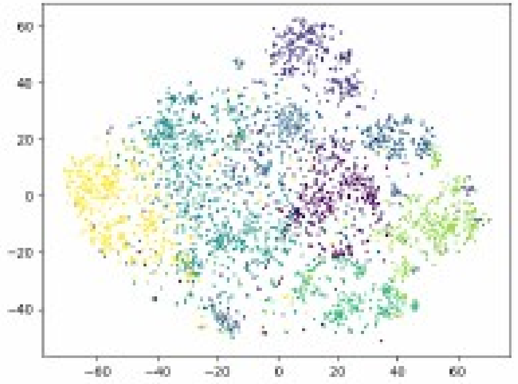
Model	S	CH	H	C	V	ARI
Random	0.0062	13.83	0.0032	0.0030	0.0031	0.0000
Random (L2)	0.0062	13.92	0.0033	0.0031	0.0032	0.0001
LSA	-0.0207	53.45	0.0030	0.0071	0.0042	-0.0041
LSA (L2)	**0.1284**	96.44	0.0022	0.0025	0.0023	-0.0009
DeepWalk	0.0194	124.80	0.2165	0.2496	0.2318	0.1852
DeepWalk (L2)	0.0670	131.18	**0.2930**	**0.2810**	**0.2869**	**0.1981**
GCN	0.0667	171.13	0.1845	0.1761	0.1802	0.1178
GCN (L2)	0.0823	**193.64**	0.1992	0.1891	0.1940	0.1423

Avg. of 100 k-Means runs

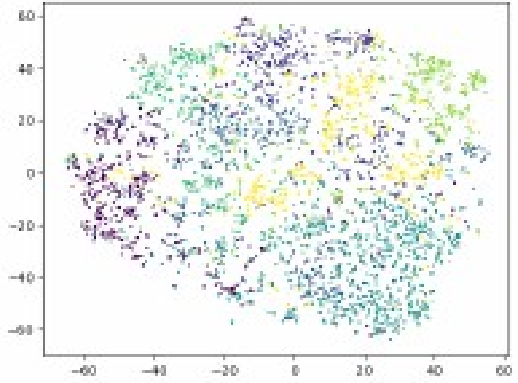
t-SNE Visualization



LSA

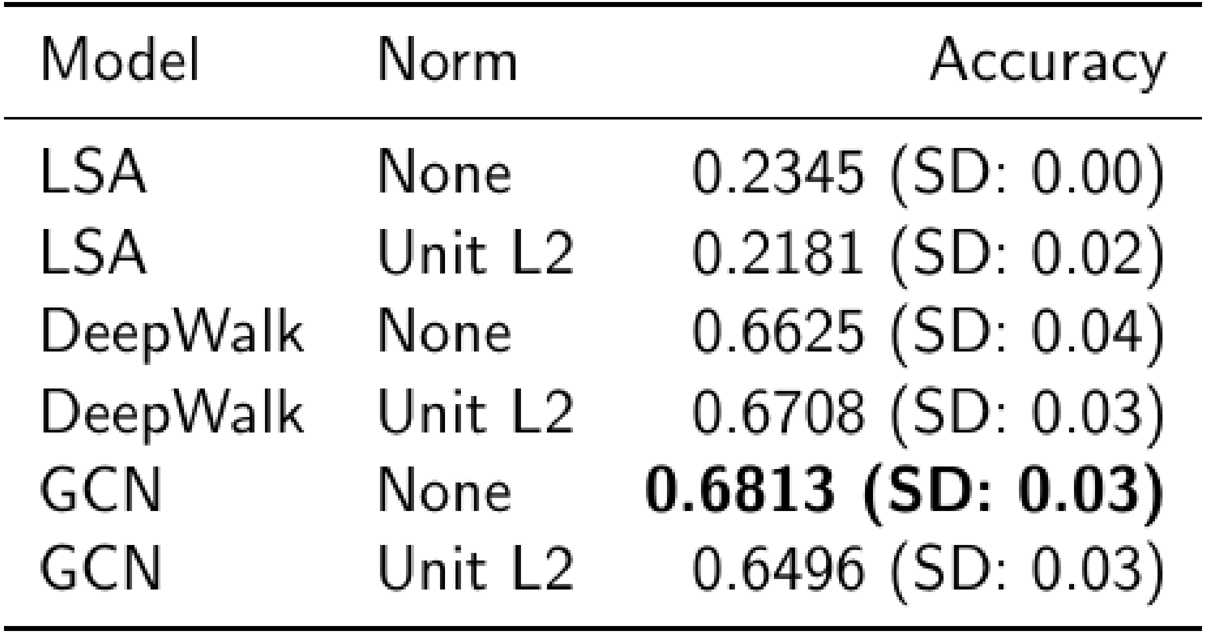


DeepWalk



GCN

Results: Classification with linear SVMs

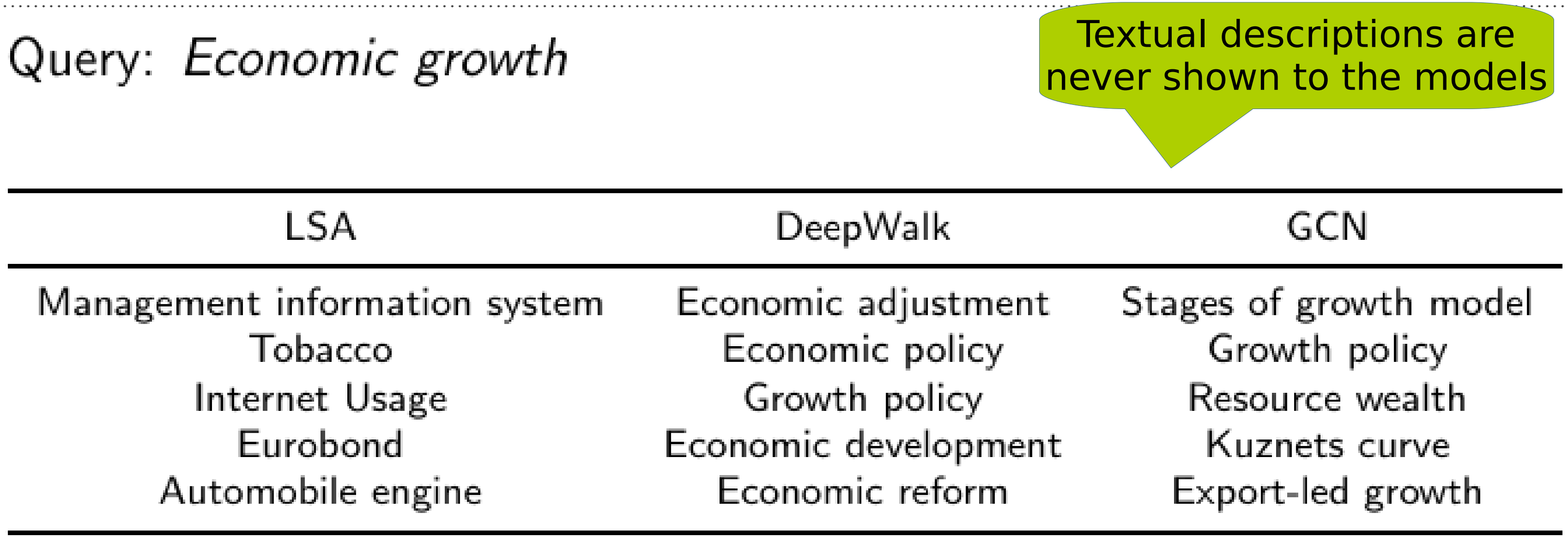


Model	Norm	Accuracy
LSA	None	0.2345 (SD: 0.00)
LSA	Unit L2	0.2181 (SD: 0.02)
DeepWalk	None	0.6625 (SD: 0.04)
DeepWalk	Unit L2	0.6708 (SD: 0.03)
GCN	None	**0.6813 (SD: 0.03)**
GCN	Unit L2	0.6496 (SD: 0.03)

Avg. and SD of 10-fold CV

Nearest Concept Queries 1/4

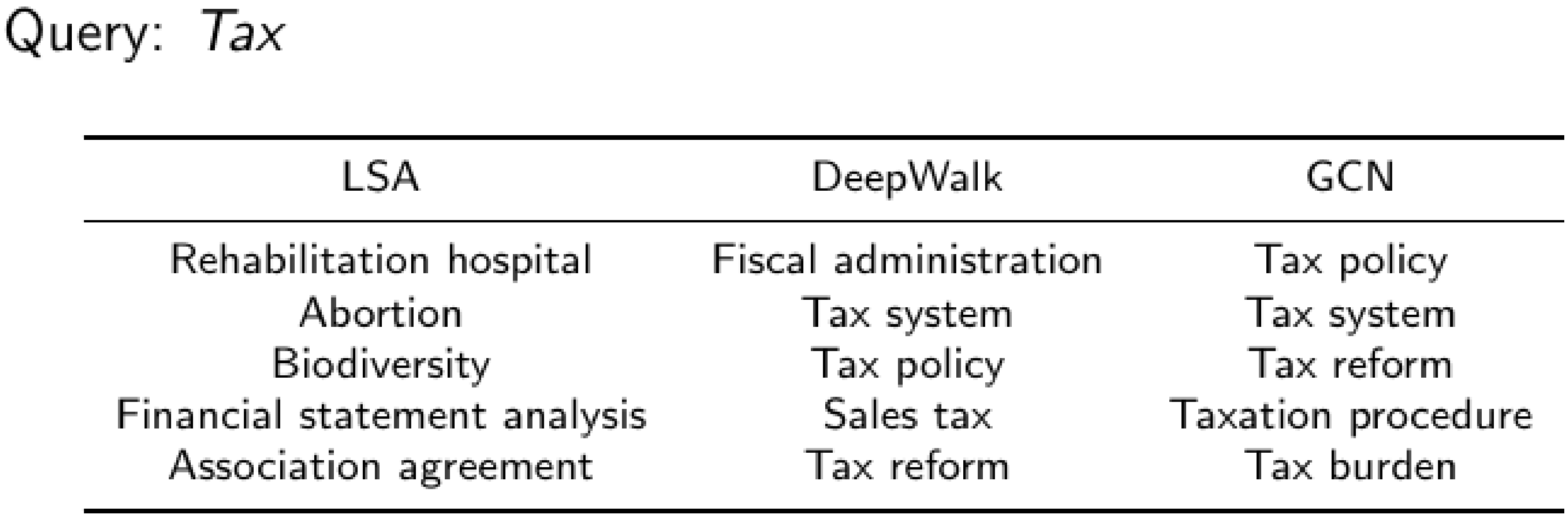
Textual descriptions are never shown to the models



Query: *Economic growth*

LSA	DeepWalk	GCN
Management information system	Economic adjustment	Stages of growth model
Tobacco	Economic policy	Growth policy
Internet Usage	Growth policy	Resource wealth
Eurobond	Economic development	Kuznets curve
Automobile engine	Economic reform	Export-led growth

Nearest Concept Queries 2/4



Query: *Tax*

LSA	DeepWalk	GCN
Rehabilitation hospital	Fiscal administration	Tax policy
Abortion	Tax system	Tax system
Biodiversity	Tax policy	Tax reform
Financial statement analysis	Sales tax	Taxation procedure
Association agreement	Tax reform	Tax burden

Nearest Concept Queries 3/4

Query: *Germany*

LSA	DeepWalk	GCN
Debt crisis	Italy	East Germany
Mesoeconomics	France	Austria
Population policy	Comparison	West Germany
Complaint management	Netherlands	Lower Saxony
Unemployment theory	Austria	Western Europe

Nearest Concept Queries 4/4

Query: *Vehicle*

LSA	DeepWalk	GCN
Pigouvian tax	Transport research	Sustainable mobility
Cargo shipping	Transport economics	Passenger transport
Cyclical unemployment	Waste treatment	Freight transport
Wage subsidy	Battery	Major electrical appliances
Financial Statement analysis	Microsystems	Traffic

Linear relationship queries 1/2

Query: *Tax* + *Theory*

LSA	DeepWalk	GCN
Tax	Tax	Theory of taxation
Theory	Theory of taxation	Theory
Financial statement analysis	Tax system	Second best
Nursing profession	Capital income	Optimal taxation
Rehabilitation hospital	Public economics	Welfare economics

Linear relationship queries 2/2

Query: *Economic growth* + *Theory*

LSA	DeepWalk	GCN
Economic growth	Economic growth	Growth theory
Banking services	Growth theory	Neoclassical growth model
Producer cooperative	Economic model	Unbalanced growth
Licence	Theory	Balanced growth
Laboratory	Endogenous growth model	Functional income distribution

Conclusion & Limitations

- DeepWalk works well despite using only structural features (confirms orig. paper)
- GCNs can be used for **inductive** representation learning
- Learned GCN representations are comparably meaningful & useful as DeepWalk‘s
- **Limitations:**
    - No ground truth for pairwise concept similarity
    - Only one dataset, but large-scale!

Next steps

- Add more structure (concept hierarchy, journals, institutions, ...)
- Use publication years for truly dynamic research analyses
- Create a ground truth for pairwise concept similarity

Github: lgalke/INFORMATIK2019-concept-representation-learning

Twitter: _lpag

E-mail: l.galke@zbw.eu

Acknowledgment

This work was supported by BMBF within the programme “Quantitative Wissenschaftsforschung” under grant numbers 01PU17013A, 01PU17013B, 01PU17013C.