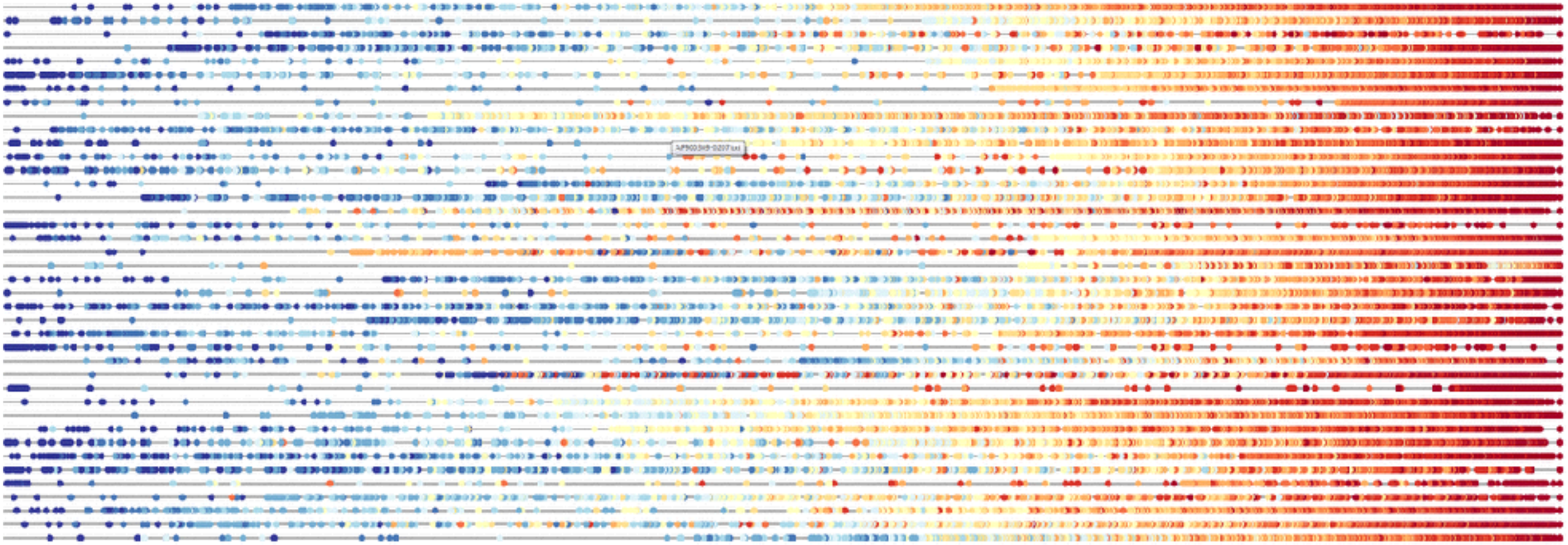


Semantic Networks and Topic Modeling

A Comparison Using Small and Medium-Sized Corpora

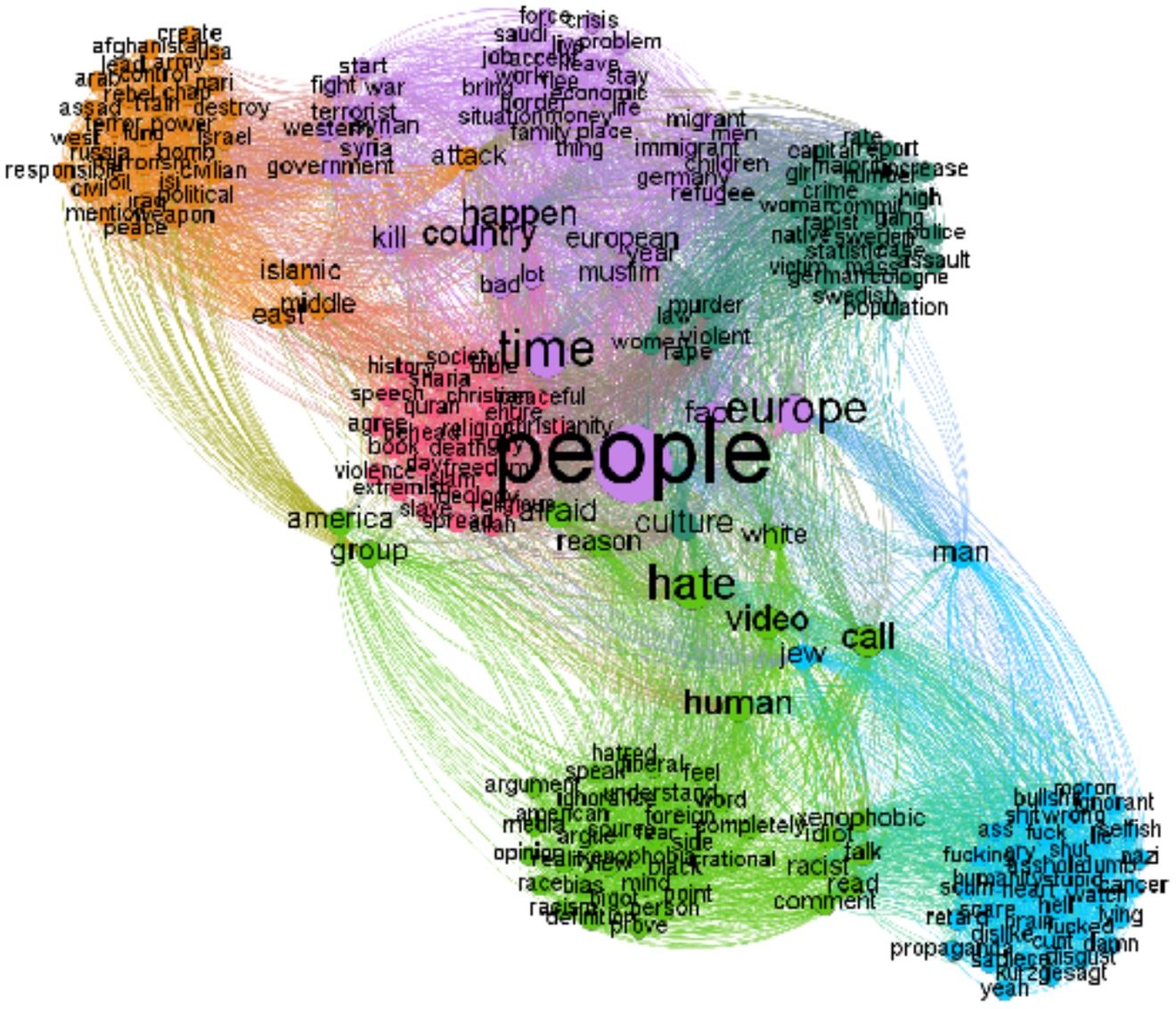
Loet Leydesdorff & Adina Nerghes

KNAW Humanities Cluster DIGITAL HUMANITIES LAB

Networks of words

Semantic Networks

Networks of concepts

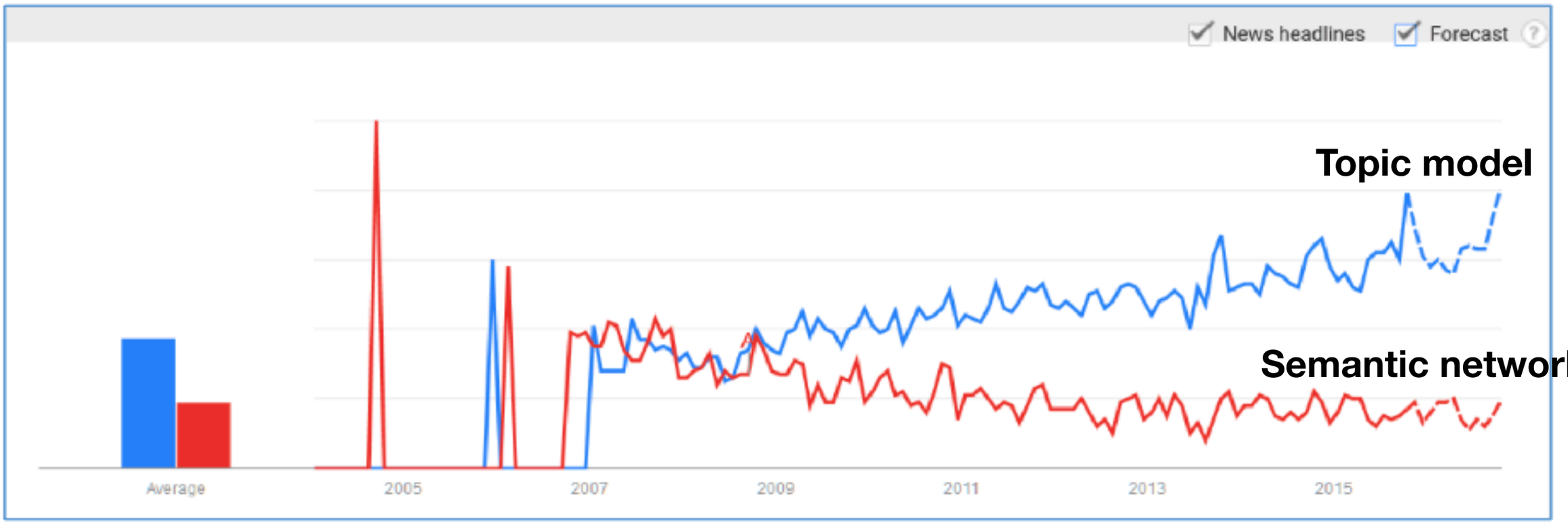


Content networks

Co-word maps

Maps

Semantic networks and Topic Models



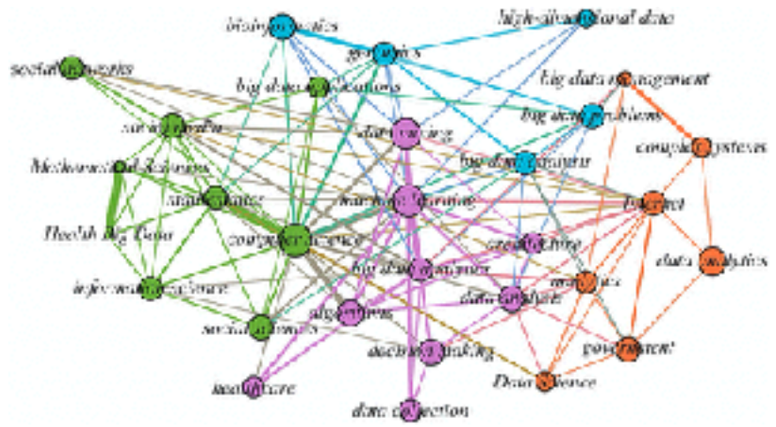
Google Trends for “topic model” (blue) and “semantic network” (red) on November 1, 2015.

Semantic networks

- Defined as: ``representational format [that would] permit the `meanings' of words to be stored, so that humanlike use of these meanings is possible'' (Quillian, 1968, p. 216)

- The meaning of a word could be represented by the set of its verbal associations

- Basic assumption: language (is) can be modeled as networks of words and the (lack of) relations among words



What makes semantic networks interesting?

- Correspond to a natural way of **organizing information** and the way humans think
- Semantic networks allow to **model** semantic relationships (Sowa, 1991)
- Investigate the meaning of texts by detecting the **relationships** between and among words and themes (Alexa, 1997; Carley, 1997a)
- Allow the analysis of words in their **context** (Honkela, Pulkki, & Kohonen, 1995)
- Expose semantic **structures** in document collections (Chen, Schuffels, & Orwig, 1996)
- Very **flexible** way of organizing data: you can easily **extend** the structure of semantic networks if needed
- You can easily convert **almost any** other data structure into semantic networks
- To represent **knowledge** or to support automated systems for reasoning about knowledge.

Semantic networks and the philosophy of science

- Hesse (1980)—following Quine (1960) argued that networks of co-occurrences and co-absences of words are shaped at the **epistemic** level and can thus reveal the **evolution** of the sciences in considerable detail (Kuhn, 1984)

- The latent structures in the networks can be considered as the **organizing principles** or the **codes of the communication**

  (Luhmann, 1990; Rasch, 2002)

- This “linguistic turn in the philosophy of science” makes the sciences amenable to **measurement** and sociological **analysis**

  (Leydesdorff, 2007, Rorty, 1992)

Software for semantic network generation and analysis

- **Callon** was the first to introduce semantic networks (co-word maps) on the research agenda of science and technology studies (**STS**) (Callon et al., 1983)

ti.exe

- However, the development of software for the mapping remained **slow** during the 1980s (Leydesdorff, 1989)

fulltext.exe

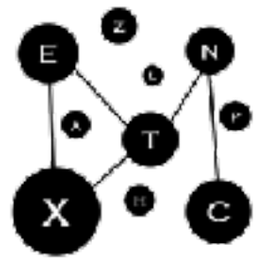
- From the second half of the **1990s**, many software packages became freely available

AutoMap

- Similar purpose —visualization of the latent structures in textual data (Lazarsfeld & Henry, 1968) — different results

VOSviewer

- Two highly relevant parameter choices:
  - similarity criteria
  - clustering algorithms



Wordjj.exe

Topic models

- A type of statistical model for discovering the **abstract** "topics" that occur in a collection of documents

- Frequently used text-mining tool for discovery of **hidden** semantic structures in a text body

- The "topics" produced by topic modeling techniques are **clusters** of similar words

Why topic models?

- To help to **organize** and offer insights for us to **understand** large collections of unstructured text bodies

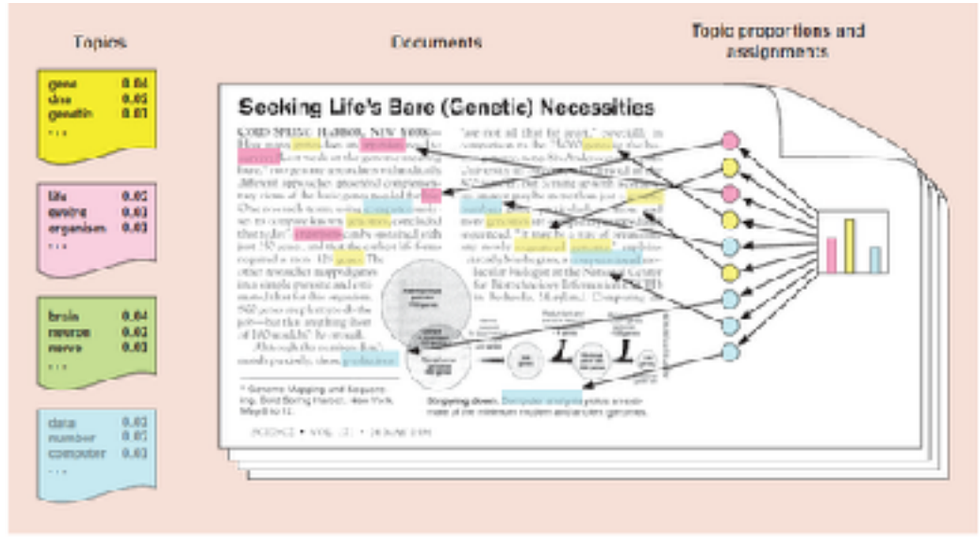
- Used to detect **instructive** structures in data such as genetic information, images, and networks

- **Annotating** documents according to these topics

- Using these annotations to **organize**, **search** and **summarize** texts

- Applications in other fields such as bioinformatics

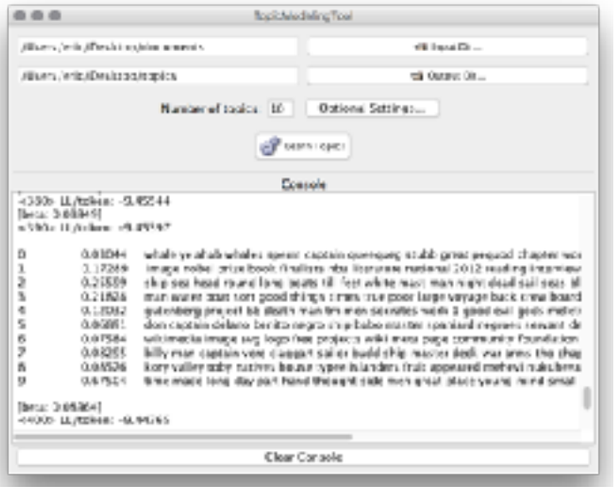
Latent Dirichlet allocation (LDA)



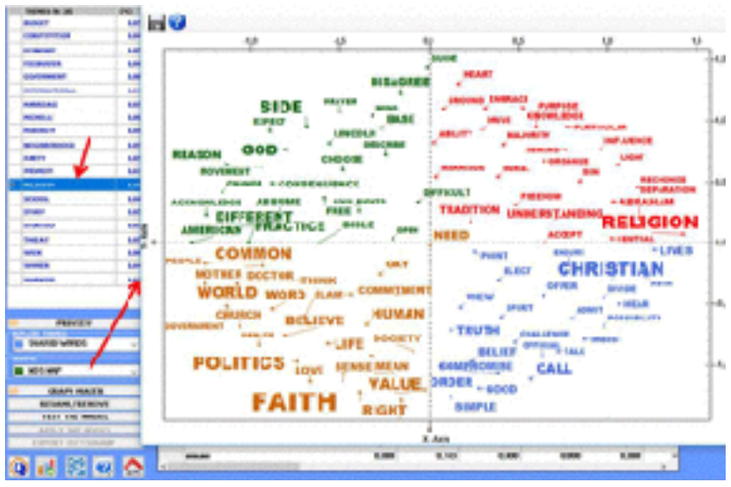
- "LDA is a statistical model of language."
- The most **common** topic model currently in use
- A generalization of **probabilistic latent semantic** analysis (PLSA)
- Developed by David Blei, Andrew Ng, and Michael I. Jordan in 2002
- Introduces sparse **Dirichlet** prior distributions over document-topic and topic-word distributions
- Assumption: documents cover a **small number of topics** and that topics often use a small number of words
- Other topic models are often **extensions** on LDA
- Currently **more popular** than semantic maps for the purpose of summarizing corpora of texts

Tools for topic modeling

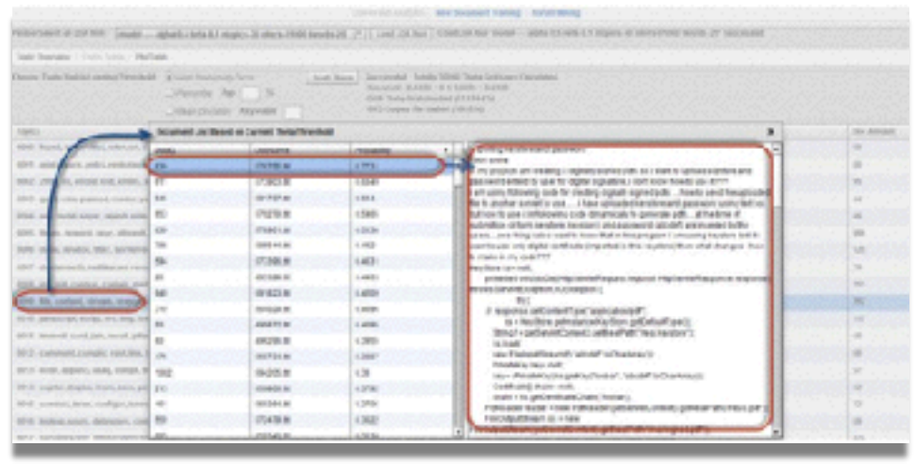
Mallet



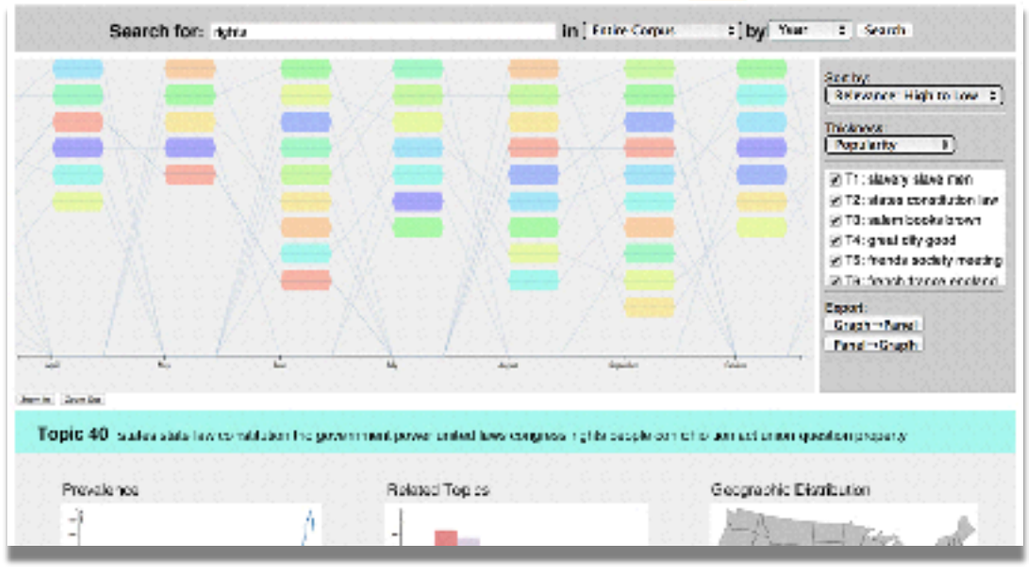
T-LAB PLUS



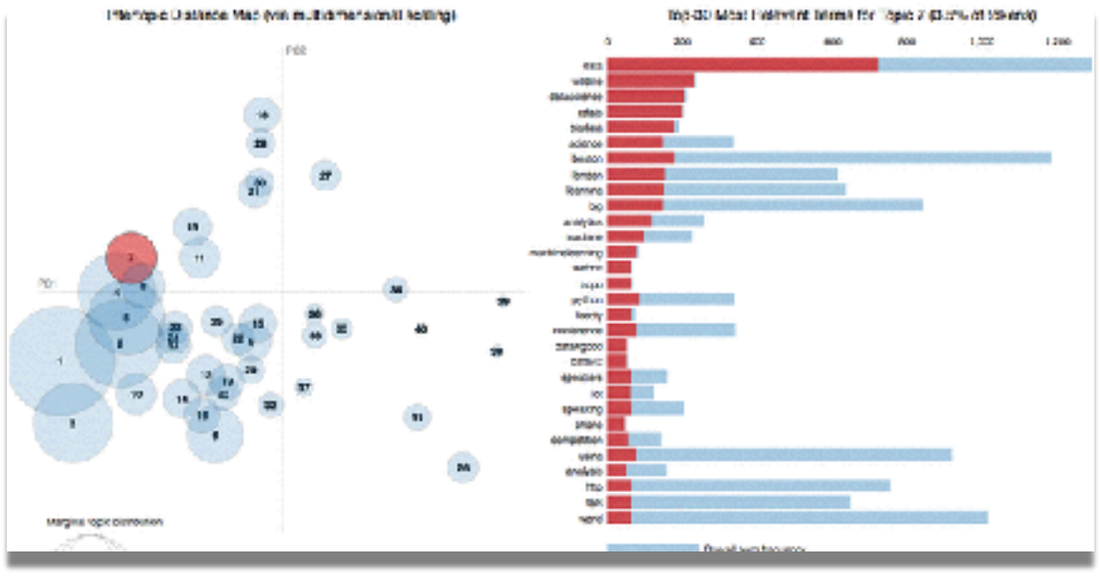
LDA Analyzer



TOME



LDAvis



A bottom-up perspective

- Large text corpora are **beyond the human capacity** to read and comprehend

- **Validity** of the results with large text corpora remains a problem

- One can almost always provide an **interpretation** of groups of words *ex post*

Aims:

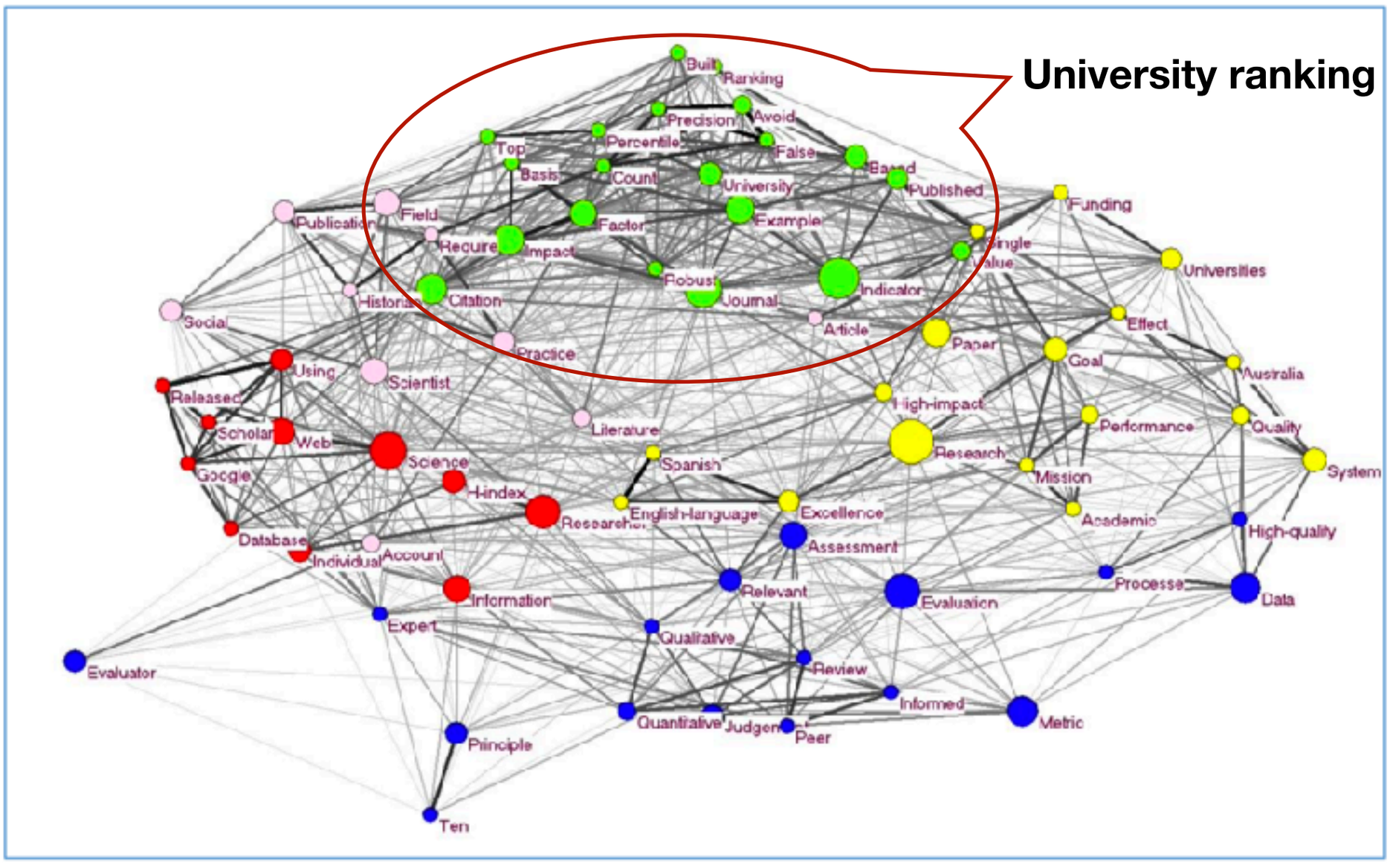
- Taking a **bottom-up** perspective, we compare semantic networks and topic models step-by-step

- Does topic modeling provide an **alternative** for semantic networks in research practices using moderately sized document collections?

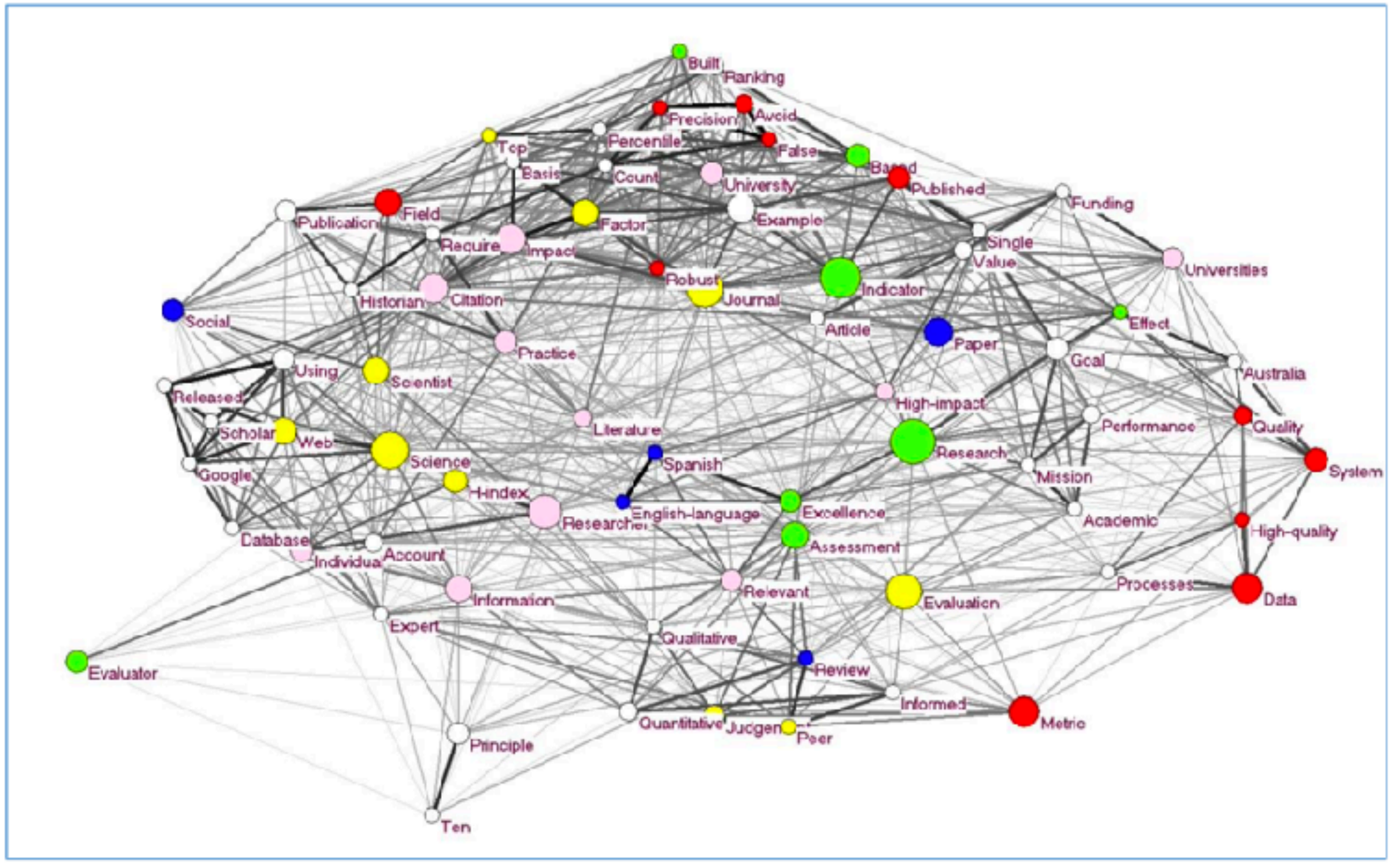
Data

- **The “Leiden Manifesto”** (Hicks et al., 2015)
    - Nature on April 23, 2015
    - Guidelines for the use of metrics in research evaluation
    - Translated into nine languages
    - Units of analysis: 26 substantive paragraphs
- **Leiden Rankings** (Waltman et al., 2012, at p. 2420)
    - Google Scholar: "Leiden ranking" OR "Leiden rankings"
    - Units of analysis: 687 documents retrieved

- **The “Leiden Manifesto”**
    - 429 stop words list
    - 550 unique words
    - 75 occur more than twice
    - Normalized word vectors by cosine
    - Treshold cosine > 0.2
- **Leiden Rankings**
    - 429 stop words list
    - noise words in languages other than English
    - 56 words occur > 10 times



Five clusters of 75 words in a cosine-normalized map (cosine > 0.2) distinguished by the algorithm of Blondel et al. (2008); Modularity Q = 0.27. Kamada & Kawai (1989) used for the layout.



Nodes are colored according to the LDA model.
(Words not covered by the LDA output are colored white.)

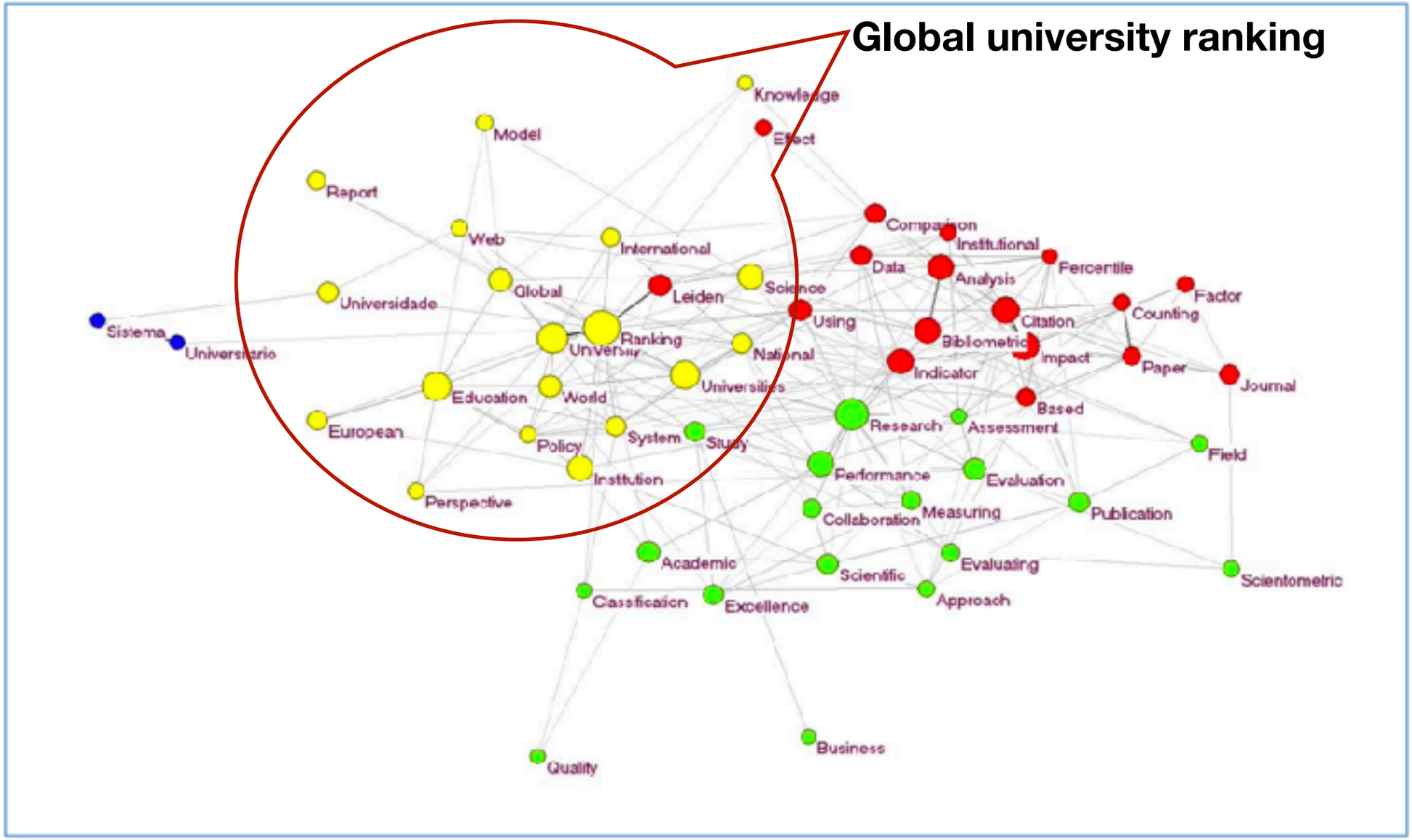
Cramér's $V = .311$ ($p = .359$)

“The Leiden Manifesto”: Semantic networks vs. LDA

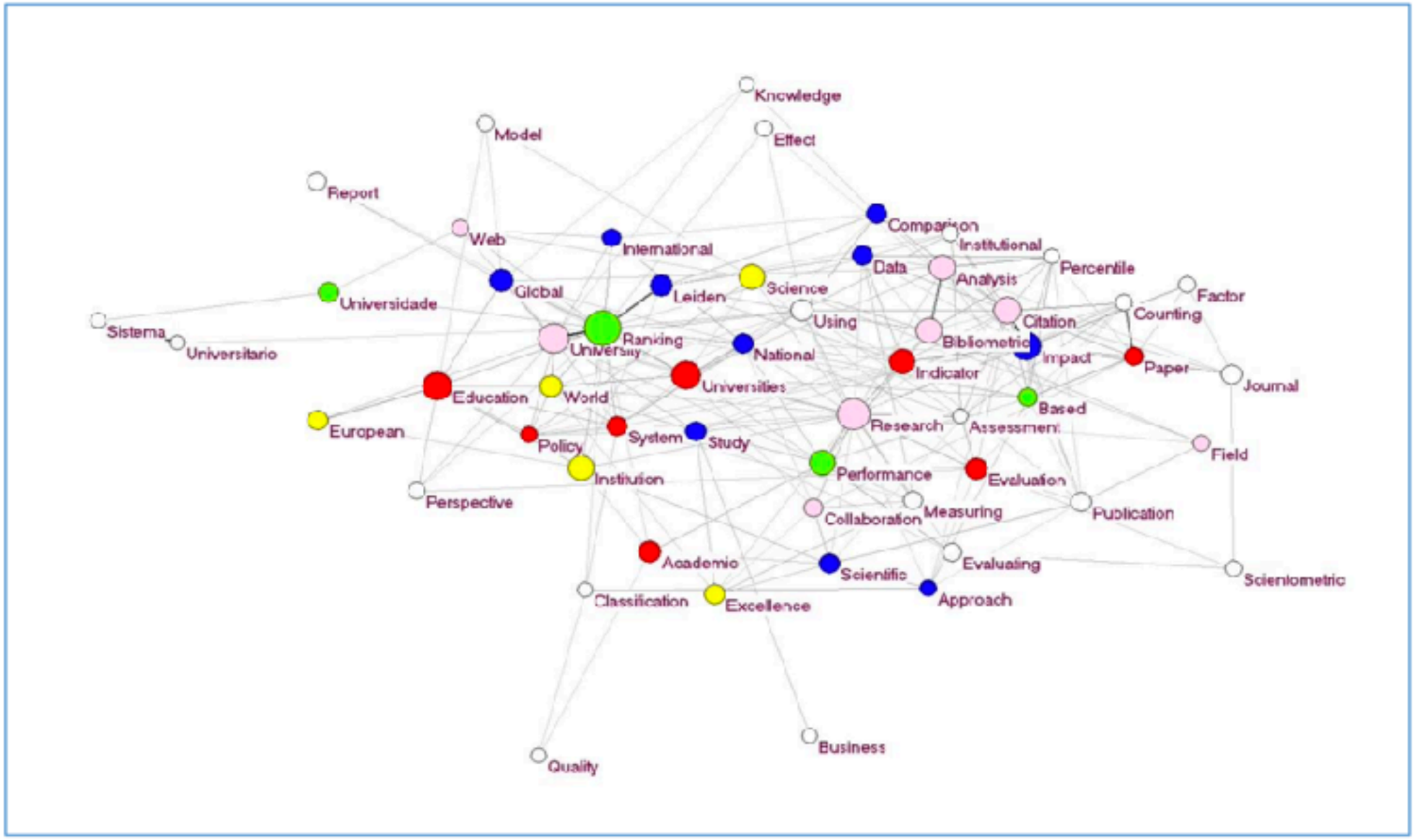
- The topic model is **significantly different** in all respects from the maps based on co-occurrences of words

- The results are **incompatible** with those of the co-word map

- The results of the topic model were significantly non-correlated and **not easy to interpret**



Four clusters of 56 words in a cosine-normalized map (cosine > 0.1) distinguished by the algorithm of Blondel et al. (2008); modularity Q = 0.36. Kamada & Kawai (1989) used for the layout.



Nodes are colored according to the LDA model.
(Words not covered by the LDA output are colored white.)

Cramér's $V = .240$; $p = .811$

The Leiden Rankings: Semantic networks vs. LDA

- The two representations are **significantly different**.
- Even when using a larger set, the topic model still distinguished topics on the basis of considerations other than semantics (e.g., statistical or linguistic characteristics).

Conclusion

- Topic modeling have become **user-friendly** and very **popular** in some disciplines, as well as in policy arenas
- We were **not able** to produce a topic model that **outperformed** the co-word maps
- The differences between the co-word maps and the topic models were **statistically significant**
- As topic models are further developed in order to handle “big data,” **validation** becomes increasingly difficult
- However, the computer algorithm may find **nuances and differences** that are not obviously meaningful to a human interpreter (Chang et al., 2010; Jacobi et al., 2015, at p. 6).
- The robustness of LDA topic model results is unaffected by the lack of semantic and syntactic information (Mohr & Bogdanov, 2013), our results suggest differently in the case of small and medium-sized samples
- Further steps: Hecking, T., & Leydesdorff, L. (2019). ***Can topic models be used in research evaluations?*** Reproducibility, validity, and reliability when compared with semantic maps. Research Evaluation, 28(3), 263-272.

IDEAS WITH IMPACT:
How connectivity shapes idea diffusion

Dirk Deichmann, Julie M. Birkholz, Adina Nerghes, Christine Moser, Peter Groenewegen, Shenghui Wang

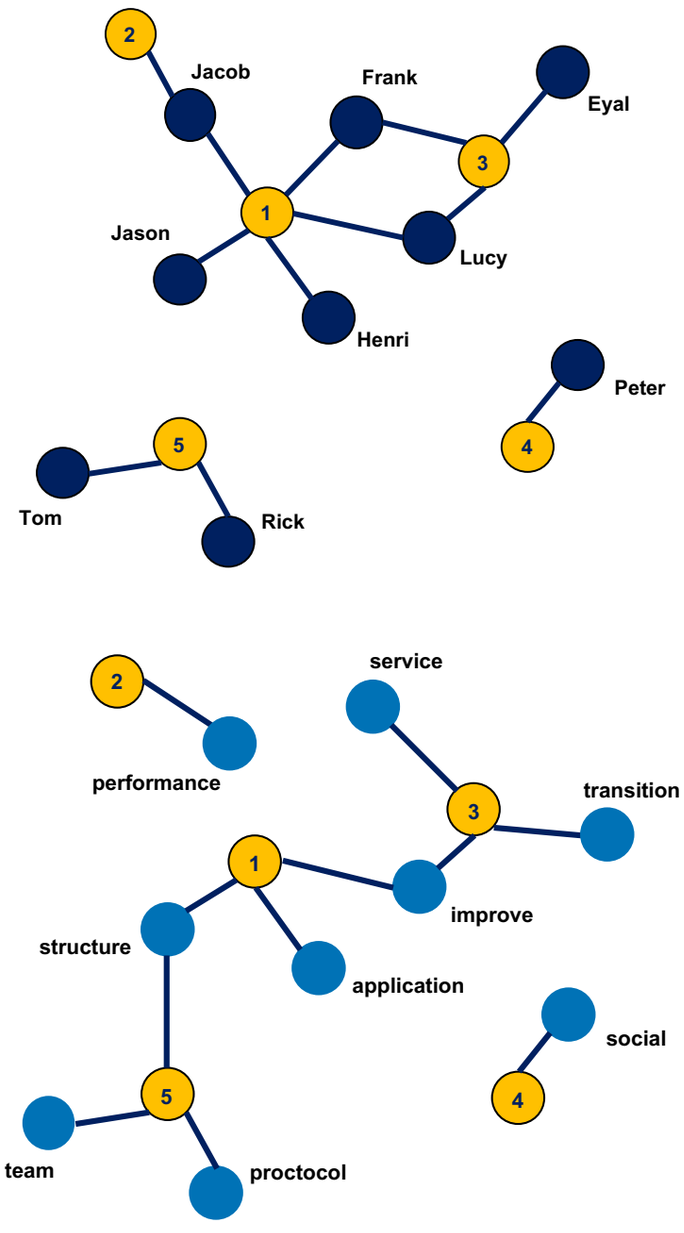
Context of science

- Goal of science: **Produce (new) knowledge**
- Increasingly done in co-authorship **teams**
- **Disseminated** through journal articles, conference proceedings, workshop presentations, demos, etc.
- These “dissemination events” are **documented events** of both a **team** of co-authors and **idea** content
- Recognition of ideas through **citations**

How to semantic and social networks relate to successful idea diffusion?

- MOTIVATION:
  - Better understand the **idea diffusion process**
  - Not only focus on the social network position of the team of inventors of an idea, but **shed light on the characteristics of the idea itself**
  - **Disentangle** the effects of a team's position in the social network from effects that are driven by the idea's position in the content network

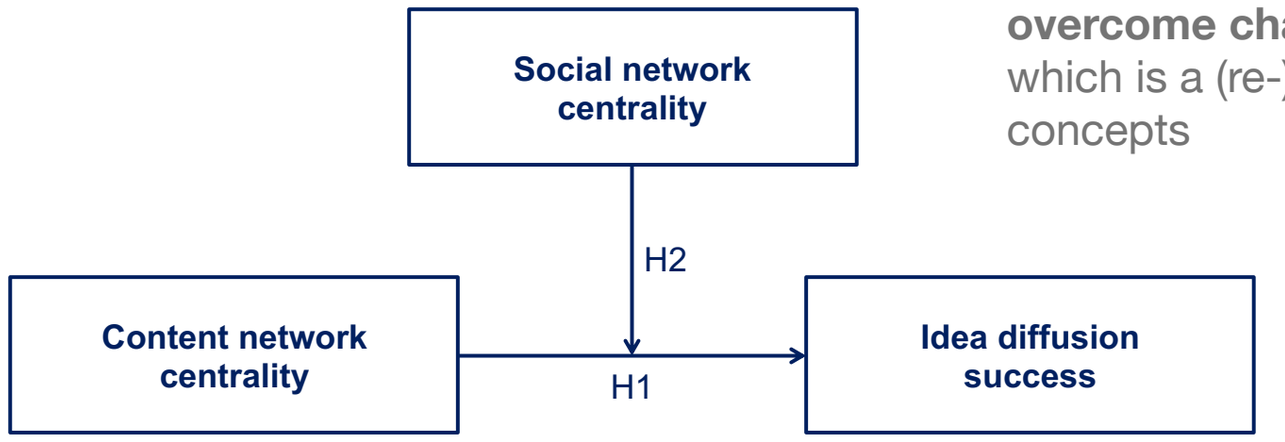
- SOCIAL VS. CONTENT NETWORK CENTRALITY:



Hypoteses

- CONTENT NETWORKS
  - Content network centrality
    - (Re-)**combination** of different concepts
  - A central content network position is argued to fuel the idea diffusion process:
    - **Overlap** – easier for others to identify the focal idea as relevant
    - **Popularity** – get more attention from the community

- SOCIAL NETWORKS
  - Social network centrality
    - **Status** and **access** to expertise
  - Social network centrality is argued to moderate the effect of content network centrality on idea diffusion
    - A highly central team working on a highly central idea reaches the **outskirts of the network**
    - Status of a central team helps to **overcome challenges of an idea** which is a (re-)combination of different concepts



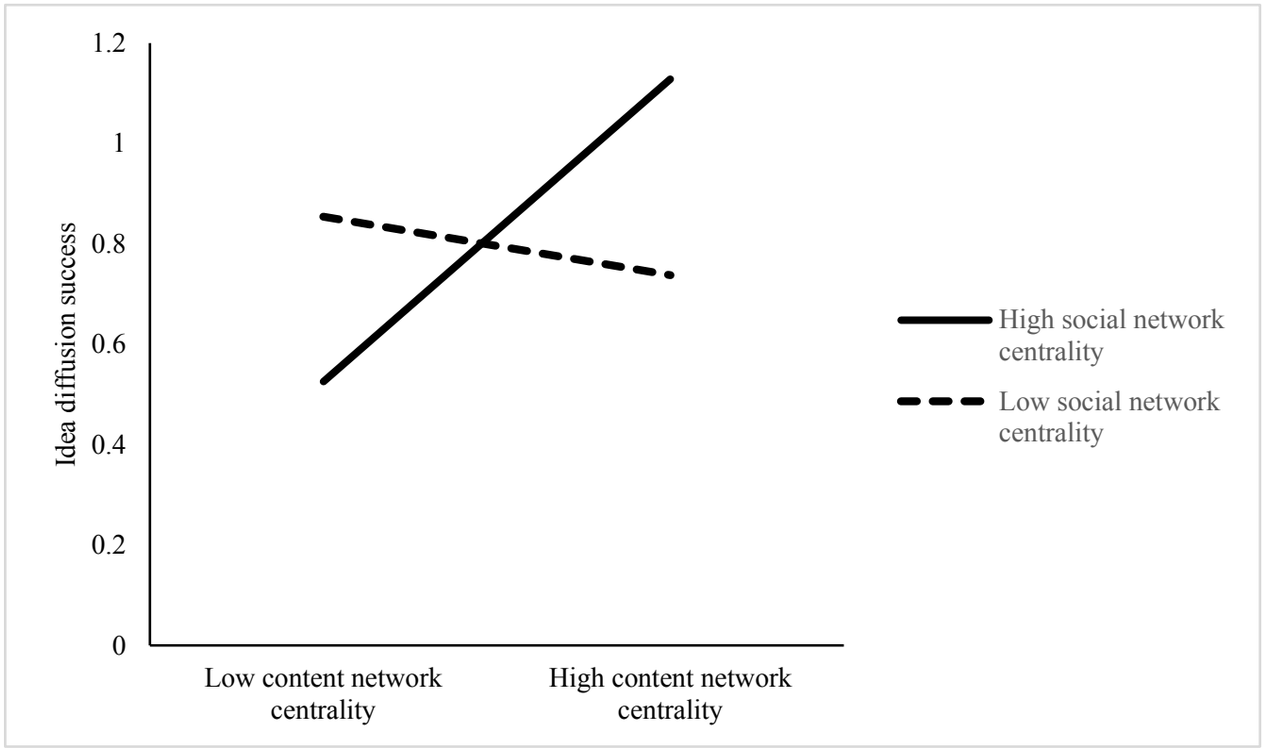
Data & Method

- Conference publication data
    - Source: Semantic Web - subfield of Computer Science
    - 31 conferences from 2006 - 2012
    - 2,492 conference items (proceedings, posters, demos)
    - 5,456 unique co-authors
- Dependent variable: **Idea diffusion success**
    - Citation score after two years
- Independent variable: **Content network centrality**
    - Two-mode betweenness centrality (the number of times a node acts as a bridge along the shortest path between all other nodes)
    - Embeddedness in other ideas
- Moderating variable: **Social network centrality**
    - Two-mode betweenness centrality (the number of times a node acts as a bridge along the shortest path between all other nodes)
    - Embeddedness in other co-authorship teams
- **Controls:**
    - Number of title words
    - Number of authors
    - Scientific age (average)
    - Prior citations (average) / prior publications (average)
    - Conferences attended (average)

	Idea Diffusion Success				
Variables	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Constant	-0.18	-0.07	-0.18	-0.08	-0.08
	(0.18)	(0.18)	(0.18)	(0.18)	(0.18)
Number of title words	-0.01	-0.03+	-0.01	-0.03+	-0.02+
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of authors	0.17***	0.17***	0.17***	0.17***	0.17***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Scientific age (average)	-0.02	-0.02	-0.02	-0.02	-0.02
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Prior citations (average)	0.20***	0.20***	0.20***	0.20***	0.20***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Conferences attended (average)	-0.03	-0.02	-0.03	-0.02	-0.02
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Content network centrality		0.13***		0.13***	0.12***
		(0.03)		(0.03)	(0.03)
Social network centrality			-0.00	-0.00	0.02
			(0.03)	(0.03)	(0.03)
Content network centrality x Social network centrality					0.18**
					(0.06)
Variance of constant	0.37	0.37	0.37	0.37	0.37
Variance of residual	1.58	1.57	1.58	1.57	1.56
Log likelihood	-3479.07	-3469.05	-3479.06	-3469.04	-3464.37
Publications	2,096	2,096	2,096	2,096	2,096
Conferences	26	26	26	26	26

Results

- Ideas which are **highly connected in the content network** perform better and receive more citations
- A **positive interaction** between content and social network connectivity
- The highest diffusion success can be attributed to publications with **high content connectivity** and **high social connectivity**
- Ideas which **bridge** different knowledge domains in the content network will amass even **more citations** when they are developed by teams that are **highly connected** in the social network of co-authorship teams



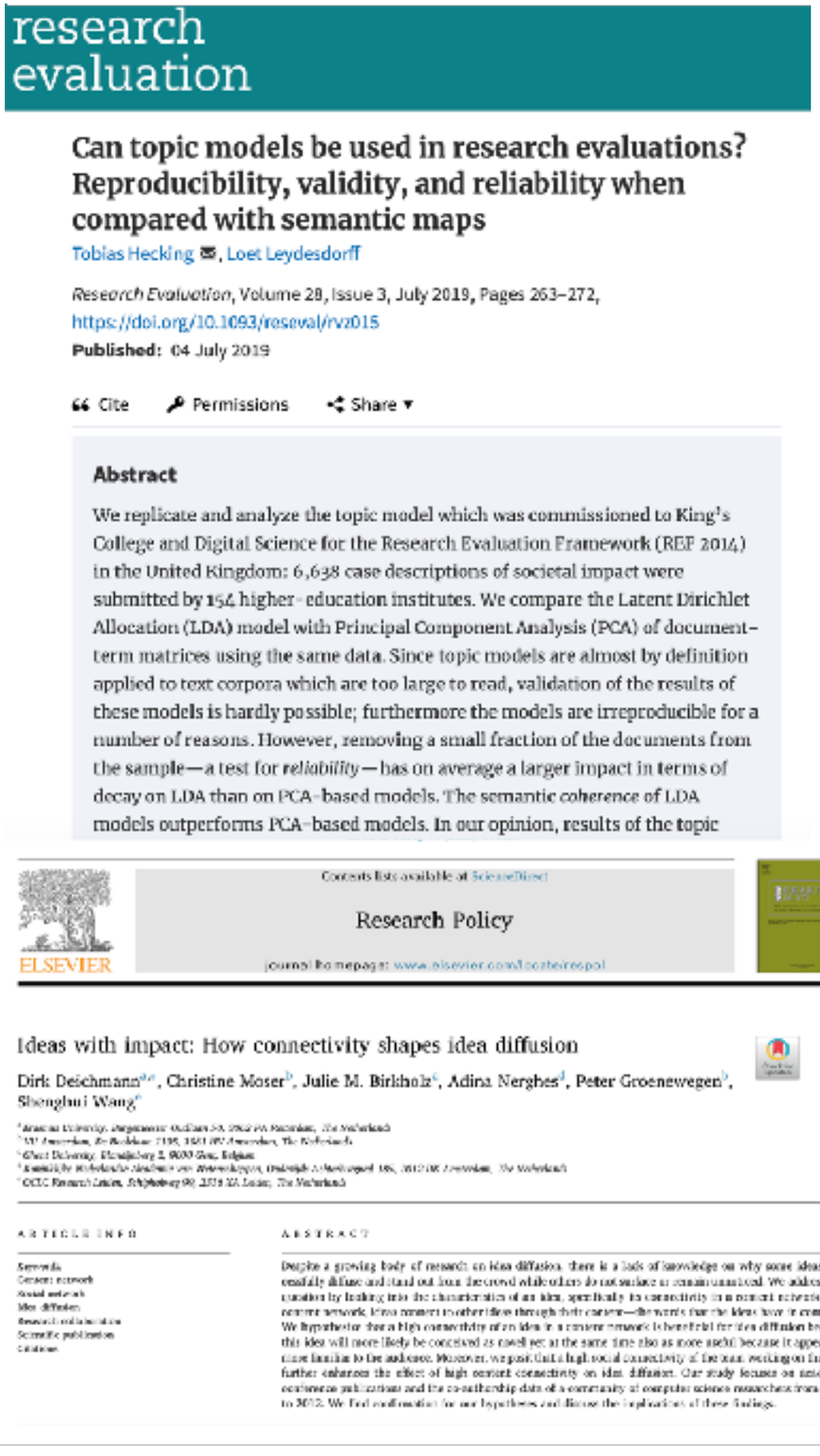
Resources

- Leydesdorff, L. and Nerghes, A. (2017), **Co-word maps and topic modeling: A comparison using small and medium-sized corpora (N<1,000)**. *Journal of the Association for Information Science and Technology*, 68: 1024-1035. doi:10.1002/asi.23740

- Ti.exe: http://www.leydesdorff.net/software/ti

- Fulltext.exe: http://www.leydesdorff.net/software/fulltext

- Pajek: http://vlado.fmf.uni-lj.si/pub/networks/pajek/



research evaluation

Can topic models be used in research evaluations? Reproducibility, validity, and reliability when compared with semantic maps

Tobias Hecking ✉, Loet Leydesdorff

Research Evaluation, Volume 28, Issue 3, July 2019, Pages 263–272, https://doi.org/10.1093/reseval/rvz015

Published: 04 July 2019

❝ Cite Permissions Share ▾

Abstract

We replicate and analyze the topic model which was commissioned to King's College and Digital Science for the Research Evaluation Framework (REF 2014) in the United Kingdom: 6,638 case descriptions of societal impact were submitted by 154 higher-education institutes. We compare the Latent Dirichlet Allocation (LDA) model with Principal Component Analysis (PCA) of document-term matrices using the same data. Since topic models are almost by definition applied to text corpora which are too large to read, validation of the results of these models is hardly possible; furthermore the models are irreproducible for a number of reasons. However, removing a small fraction of the documents from the sample—a test for *reliability*—has on average a larger impact in terms of decay on LDA than on PCA-based models. The semantic coherence of LDA models outperforms PCA-based models. In our opinion, results of the topic

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

ELSEVIER

Ideas with impact: How connectivity shapes idea diffusion

Dirk Deichmann[a,*], Christine Moser[b], Julie M. Birkholz[c], Adina Nerghes[d], Peter Groenewegen[b], Shenghui Wang[e]

ARTICLE INFO

ABSTRACT

Despite a growing body of research on idea diffusion, there is a lack of knowledge on why some ideas successfully diffuse and stand out from the crowd while others do not surface in communication. We address this question by looking into the characteristics of an idea, specifically its connectivity in a content network. In a content network, ideas connect to other ideas through their content—the words that the ideas have in common. We hypothesize that a high connectivity of an idea in a content network is beneficial for idea diffusion because this idea will more likely be conceived as novel yet at the same time also as more useful because it appears as more familiar to the audience. Moreover, we posit that a high social connectivity of the team working on the idea further enhances the effect of high content connectivity on idea diffusion. Our study focuses on academic conference publications and the co-authorship data of a community of computer science researchers from 2006 to 2012. We find confirmation for our hypotheses and discuss the implications of these findings.

Contact

ADINA.NERGHES@DH.HUC.KNAW.NL

@ADINANERGHES

@DHLABHUC

HTTP://WWW.DHLAB.NL

KNAW Humanities Cluster