	00-084-0-0-00-00101010101010000000000000	and a second	
			מיניים הערכו בארגעיים במספר הימה בבניה הבני מינה בניים המנייב היכור היה מזור ביילי בי המבני בינור ו- הבי בי המ
• • • • • • • • • • • • • • • • • • •		80 81 338 1-0-0-038 8-30004-038-0-000 - 838 - 0	
	-818 3329 3-837 - 372 - 199 - 44 23 - 38 - 68 32 - 7.4 3	- 3 - 312000.300-033-30.00333.0-3.003.302-4	1 (1) (1) (1) (1) (1) (1) (1) (1) (1) (1
		Order 1 (1990) - 10 (1990) - 10 (1990) - 10 (1990) Order 1 (1990) - 10 (1990) - 10 (1990) - 10 (1990) Order 1 (1990) - 10	
	· · · · · · · · · · · · · · · · · · ·		ייין אין אין אין אין אין אין אין אין אין
			1. (1. (1. (1. (1. (1. (1. (1. (1. (1. (
		** ******	
31-0 B -00 -0 -0 -0 -0 -0 -1 33 3 333		בביו הסבוסו את המכבה מכבירות. את כי היאתה את היו להיכוב ביותר זבה את המכביר היו היא את א	<b>אין אין המבערות אין אין אין אין אין אין אין אין אין אין</b>
B08010-0-0000-04 +0000-04 (30000-000000000-00-00-00-000-000-0000000	0 \$2)-\$191-310(310 \$1)-01-8-39(0.303) \$3-01 \$338 \$33	-01030-30.3 -01010-11-330300-000.5 3333000010-5 338 20-3	2003-100000000-0030-0032-0032-0032-0032-
	AP9003x8-0207xx1	10 11-6 2 - 2 10 10 10 10 10 10 10 10 10 10 10 10 10	מאר היה איר היה אל היה לי שלה היה היה המתוכח ההתוכחה הלה המאר להליחת שלה להית שלה להלול ללובר הלי לי לה השת לה ל
• • • • • • • • • • • • • • • • • • •	12-10-10-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1		
CONTENTS			• מיות את המאמים הבלאות כבו מהמות הלי הלותה והלוה מהביצה הלי ב", ביי הלי מני הלי ", הי ל - ב' - הי ל - ל' - ל'
	• • • • • • • • • • • • • • • • • • •	249-03-124200-21-035-32-025-00-030-0910	
	1999-199-199 200 1992 20 20 20 20 20 20 20 1999 - 199 - 199 20 20 20 20 20 20 20 20 20 20 20 20 20	10 - 10 - 10 - 10 - 10 - 10 - 10 - 10 -	אין איז
	- 3) 3 34 3 30 <b>- 4</b> - 5 30 - 30 - 30 - 30 - 30 30 13 33 30 13 30 30 30 30 30 30 30 30 30 30 30 30 30		an a
	••••••••••••••••••••••••••••••••••••••	······································	0
			- 32 TO B O B O B O B O B O B O B O B O B O
	0403-0-0-030-030-0403-0403-0330-0-0330-03030-03030-0-0400-0-0-0-	39 80 - 3 \$ 130 330 1312-0015 \$0000 (313) 3 130 20130 (39 8	
	• • • • • • • • • • • • • • • • • • • •	(0.13) to (0.110 (0.111 - 0.111 (0.111 (0.110 (0.1111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.1111 (0.111 (0.1111 (0.1111 (0.1111 (0.1111 (0.1111 (0.1111 (0.1111 (0.1111 (0.1111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111 (0.111))))))))))))))))))))))))))))))))))	איז
- C - CONTA - CO- 3 - CONTA - CONTA - CONTA - CO- C - C - C - C - C - C - C - C - C	HE 42-1303-1-01-0HB 18-04 30313/3HD 3-0-0 315-03 88	4.3.30 - 3 - 3 2303800:00 303 - 3380 0.00 - 4 - 10 2320 - 4 2 \$4083	
• • • • • • • • • • • • • • • • • • •	30300-0333 (0-0.00-0.303-0.300-1.300.303) 303-0.11 33-030-330 (0.00-	31-0 3 1 332 2- 38 939 31 310 3 31-0 30 - 0 3 3 9 30	ס 🛉 אומרמסופר פומופורמטונים פארכום נמינוסטור ופינו פנו פניור פווי נעייי נעייי בעום פיובני ב- פר ב-2 פנוע אביר ב
- • • • • • • • • • • • • • • • • • • •		· · · · · · · · · · · · · · · · · · ·	(010310-131018)-(0040-34)-4030-0438-3030-1-60304-040304-0413800000000000000000000000000000000000
			2 - 1 - 4 31 0 12 - 9 - 1 - 0 - 0 - 0 0 0 0 0 0 0 0 0 0 0 0 0
		000000000000000000000000000000000000000	• • • • • • • • • • • • • • • • • • •
		x- 3049-33808309-34-34-6-433-32324-3-4-9486-3486-343	
		(1) 2000 (1) 10 2 (10) (20) (1) (0) (20) (1) (0) (20) (20) (20) (20) (20) (20) (20)	
B B B B B B B B B B B B B B B B B B B	40 ) 0	0 - 19 19 19 19 19 19 19 19 19 19 19 19 19	ייין אין איז
	Internet (1990) - Internet	2014 0 3 110 13 11 0 0 13 3 10 0 0 13 0 10 0 0 0	1000 C C C C C C C C C C C C C C C C C C
<b>00 + 0 + 00 + 00 + 00 + 00 - 00 + 0 + 00</b>	0-	• • • • • • • • • • • • • • • • • • •	איז
	-04) 3-4)-5 40 5 330 8 30 3-0 -0 333 3-0 30 3-0 -0 -0 -	0/300-e331e3-0-e301003-01e-3e-303003-031e0e-	- 30 3 300 30303010 3 40840 @050030300 30030 310 320 300 000 40300000 000000 0000000000
	2010-0120-0-000-00-00-00-00-00-00-000-00	0.000-00000000000000000000000000000000	ייייים אוריינגע איז אינגע איז
	and and and pressesses and a second s	- · · · · · · · · · · · · · · · · · · ·	4 · · · · · · · · · · · · · · · · · · ·
• • • • • • • • • • • • • • • • • • •	3/3 0 - 40 - 4 - 0 - 0 0 30 1 40 10 40 40 40 40 10 10 10 10 40 40 10 10 10 10 10 10 10 10 10 10 10 10 10	8 3 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4	
		-310 (0 -1 - 133 ) 3 0 3 0 3 0 3 0 3 3 3 3 30 (3430) 30	10.144 (10.101)   0.1
• • • • • • • • • • • • • • • • • • •		400 - 2010 2010 2010 2010 2010 2010 2020 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010	A REAL PROPERTY AND A REAL
	4-0-0(8)(0-0(0)-4-00-	-0031313003300330333003300310030030030303031333333	איז

### Semantic Networks and Topic Modeling

#### A Comparison Using Small and Medium-Sized Corpora

Loet Leydesdorff & Adina Nerghes





**Networks of words** 

#### Semantic Networks

**Networks of concepts** 



## Semantic networks and Topic Models



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

KNAW Humanities DIGITAL HUMANITIES LAB

## 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 cooccurrences 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)
- However, the development of software for the mapping remained slow during the 1980s (Leydesdorff, 1989)
- From the second half of the **1990s**, many software packages became freely available
- Similar purpose —visualization of the latent structures in textual data (Lazarsfeld & Henry, 1968) — different results
- Two highly relevant parameter choices:
  - similarity criteria
  - clustering algorithms





fulltext.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

DIGITAL HUMANITIES LAB

Currently more popular than semantic maps for the purpose of summarizing corpora of texts



## Tools for topic modeling

#### Mallet



#### **T-LAB PLUS**



#### LDA Analyzer

Showed a 29 NR (make.	option is a first of the life.	states 31 step-1000 tes	4428.2711.1441.00.ho	CONTRACTOR AND ADDRESS OF ADDRESS	
Inter Destation 1 Protes Salities, 1984	Table -				
Secone Tradic Statistic contractive de	A Provinsion "Venete "Oberphik	hip han the second seco	e han South Res Off Solution With Coper for	de Sillell Tune Lettour Familiei 1 F. Kulto - Sandi Mandel (FEREN) ander (Milling	
New York Comments	Numeri Joille	ed a Canal Brighten	·	Contraction of the second s	2
and have a president of the	and a		and the second s	· · · · · · · · · · · · · · · · · · ·	-
the side and put in the second	19.	(N78 N	1 773		(1) (B) (B) (C)
1912 (TTAL 10, 1920) 108, 1981, 3	11	DOM: N	104	pacies/services use to oper spatial ion con treat on en-	1000
int an own parent carter pa	1.0	0707m	1818	art um Riterig och for rinding ogsåt opretode Noets sent heudrocke	2 41
the support they want also	80	P129.8	1586	sufficients are contralioners and dimansials to generate pdf at heating of	COLUMN AND
OR. Don. American. Manuf.	-	matia	+200	autorities offers andrea to start and appendig all diff and used by the	
ter an every ser service	788	(1993 + 1.44)	1.402	partness on approximation provides the system that the gas that	1000
	54	17.16.8	1431	a mark of out the	1000
and the second sec	an .	400 M	1.4801	presiei musulephysenefepes report hyserefepera source	in the second
_	-141	015218	1400	anne pave of an order of	
the set opened where we are	100	torigan m.	1.0091	2 spots attaction and an art	and the second second
the property state of the lot	**	4947191	1.4785	to - Northise petrolarise (a disor prOxfuel (part))	A
10.2 month and pro-most play	-	05291.0	4,280	Stor - point cred attended to recently 1	1. 1. 1. 1.
10.2 - Calmin Complex Fail, Im. 1	100		1.000	au Telestinanti andif sifeatearit	-
17.7 NOR. 82491, MAL 19921, S	192	9-05.8	1.38	Television and the second seco	1.18
Will sugar stephe from here per	10	streets as	+3780	Cost and Age with	
Well screens, here, configur, have	-	-months	1349	PUBLICATION FOR THE AREA CONTRACTOR AND A DESCRIPTION OF T	· · · · ·
IT I HARD AND ADDRESS, COR	-	01438	4362	Fecologistment is + tax	1000
the second s	-	00568	476.00	· · · · · · · · · · · · · · · · · · ·	and a second

#### 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

KNAW Humanities DIGITAL HUMANITIES LAB



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 noncorrelated 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

DIGITAL HUMANITIES LAB

KNAW

Humanities Cluster

- 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					0.18**	
Social network centrality					(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

KNAW

Humanities Cluster

- 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 coauthorship teams

DIGITAL HUMANITIES LAB



## Resources

- Leydesdorff, L. and Nerghes, A. (2017), Co-word maps and topic modeling: A comparison using small and medium-sized corpora (N<1,000).</li>
  Journal of the Association for Information Science and Technology, 68: 1024-1035. doi:10.1002/ asi.23740
- Ti.exe: <u>http://www.leydesdorff.net/software/ti</u>
- Fulltext.exe: <u>http://www.leydesdorff.net/software/</u> <u>fulltext</u>
- Pajek: <u>http://vlado.fmf.uni-lj.si/pub/networks/</u> <u>pajek/</u>

### Contact

ADINA.NERGHES@DH.HUC.KNAW.NL

#### @ADINANERGHES

#### @DHLABHUC



#### HTTP://WWW.DHLAB.NL

#### 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



#### Ideas with impact: How connectivity shapes idea diffusion

Dirk Deichmann<sup>a</sup>", Christine Moser<sup>b</sup>, Julie M. Birkholz<sup>6</sup>, Adina Nerghes<sup>4</sup>, Peter Groenewegen<sup>b</sup>, Sbenghui Wang<sup>6</sup>

<sup>4</sup> Manuna University, Dergementer Ostfann 55, 2002 PA Returniser, The Madwelands <sup>1</sup>/10 Amerikan, Ro-Raddauer (1985, 1981) IW Amerikan, The Norbeitands <sup>4</sup> Sharib Delvaring, Electriqueberg 2, 9000 Gent, England <sup>1</sup> Manini De Verbeinarde and Sentersite (2007), Orabeitado Achteriturguel 1985, 2012 DE Amerikan, The Weberlands <sup>1</sup> OCCC Research Indian, Schlieberg 98, 2015 XA Lation, The Madwelands

ARTICLE INFO

#### ABSTRACT

Serverik Gener: estwork Social anti-tak Mer difficien Merschich collaboration Seimstike publication Citation Despite a growing body of meanch on idea diffusion, there is a lack of isocolledge on why some ideas successfully diffuse and stand out from the crowd while others do not surface or centain unmitted. We address this quoties by italing into the characteristics of an idea, specifically its connectivity in a context reducers. In a context network, letter context to other idea through their context—the work fract the likes have in contexts. We hypothesise that a high connectivity of an idea in a context network is beneficial for idea diffusion because this idea will more likely be conceived as movel yet at the same time also as more useful because it appears as there finalities to the addresse. Measure, we posit that a high total connectivity of the term working on the idea further enhances while each event is constitutely on idea diffusion. Car study focuses on another a subcline content of a context of a contextual of a contextual of consectivity of the same tenses from a subcline to contextual the either of high contactivity on idea diffusion. Car study focuses on another is a 2012. We find confirmation for our hypotheses and discuss the implications of these findings.