Research Evaluation, 28(3), 2019, 263–272
doi: 10.1093/reseval/rvz015
Advance Access Publication Date: 4 July 2019
Article

OXFORD

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

Tobias Hecking [1,*] **and Loet Leydesdorff** [2]

[1]Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, Lotharstraße 63, 47057 Duisburg, Germany and [2]Amsterdam School of Communication Research (ASCoR), University of Amsterdam, PO Box 15793, 1001 NG Amsterdam, The Netherlands

*Corresponding author. Email: hecking@collide.info.

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 models are statistical and should not be used for grant selections and micro decision-making about research without follow-up using domain-specific semantic maps.

Key words: topic models; LDA; co-word models; validation; decay; reliability.

1. Introduction

Topic modeling is as a technique originally developed to structure large text corpora into latent topics with the objective of application in document retrieval. Since its launch Latent Dirichlet Allocation (LDA; Blei, Ng, and Jordan 2003) has become the gold standard for the field (Lancichinetti et al. 2015: 011007-2); LDA inspired the development of a number of similar methods (Girolami and Kabán 2003). LDA belongs to a family of probabilistic models in which a document is modeled as a probability distribution of topics, while each topic is considered as a distribution of words.

Apart from the original application of structuring document collections, LDA has increasingly been adopted to support decision-making in other domains such as digital humanities, journalism, and policy-making. However, the potential impact of decisions based on probabilistic models raises questions about the reliability and validity of such probabilistic approaches in comparison to alternative methods. More specifically, we investigate in this study to what extent science-policy decisions can legitimately be based on these models.

The issue is urgent because of the application of topic models in research evaluation exercises and meta-evaluations. In the USA, for example, both the National Institute of Health (Talley et al. 2011) and the National Science Foundation (NSF; Nichols 2014) experimented with 'topic models' for organizing their research-grants portfolios. The Research Evaluation Framework (REF 2014) in the UK commissioned a topic model for organizing the social impact statements of the research output under evaluation (Grant and Hinrichs 2015). In this study, we try to replicate this latter model in order to study its reliability, validity, and semantic coherence. We compare the results with semantic (co-word) mapping.

2. LDA and PCA

Since its introduction in 2003, LDA-based topic models have increasingly been used for structural representations of large bodies of texts. A major problem with analyzing large corpora of texts that are (almost by definition) beyond the human capacity to comprehend by reading, however, remains the validity of the results

(Grimmer and Stewart 2013). Words are so flexible that one can almost always provide an interpretation to a group of words *ex post*. When specific words are not organized within topics, but spread across categories, for example, one can consider these words as a special group of 'methodological' words which are used among substantive categories (e.g., Draux and Szomszor 2017: 12). However, such an explanation remains *ex post*. *Ex ante*, it would be difficult to specify which words can be classified as 'methodological'.

In other words, it is not possible to specify an expectation of the outcome of topic models. In this regard, topic modeling can be considered as an example of a 'big data' methodology: using computer routines one can analyze large text corpora inductively as bottom-up processes without the need of theoretical justification (Anderson 2008; cf. Graham 2012). The discursive nature of knowledge and the need of justification of the results can thus be denied in principle with the arguments of pragmatic utility and the availability of computer power.

Is this a problem of bad research practices or is the problem more systemic and perhaps methodological? Lancichinetti et al. (2015: 9) conclude on the basis of an experimental comparison between LDA with community detection algorithms applied on word co-occurrences in semantic networks, as follows:

> 'Ten years since its introduction, there has been surprisingly little research on the validity of LDA optimization algorithms for inferring topic models [35]. Our systematic analysis clearly demonstrates that current implementations of LDA have low validity'.

In a similar vein, Leydesdorff and Nerghes (2017) compared co-word analysis with LDA and found statistically significant differences using the same input data and the same list of stop words in both cases. The results of the topic model were significantly non-correlated with the co-word maps and not easy to interpret. A limitation to this study was the use of small and medium-sized document sets. Goldstone and Underwood (2012), however, argued that the results of separate runs on the *same* data represent only representations, but are compatible by definition. The different interpretations can be considered as opportunities for raising further research questions.

Professional practitioners know that a very large number of topic models can fit the same data almost equally well (Lancichinetti et al. 2015: 2). The competition among models poses a serious challenge for algorithmic stability. Surprisingly, even though it is well established that the problem of fitting topic models is computationally hard, little is known about how the vastness and roughness of the likelihood landscape impact algorithmic performance in practice. These authors conclude that standard techniques for likelihood optimization are significantly hindered by the roughness of the likelihood-function landscape, even for very simple cases (at p. 1).

Topic models were developed for machine learning and natural language processing using computers, whereas co-word models have been developed in the information sciences with the objective of a meaningful interpretation of the results (Callon and Courtial 1989; Rip 1997). Is it possible to bridge this divide? As topic models were further developed in order to handle large datasets, validation became increasingly difficult, since no human interpreter can capture such large volumes of text. However, the algorithm may find nuances and differences that are not obviously meaningful to a human interpreter (Chang et al. 2009; Jacobi, van Atteveldt, and Welbers 2016: 6).

In the NSF model, for example, thousand topics were constructed on the basis of 170,000 awards granted between 2000 and 2012 (Gretarsson et al. 2012). In an evaluation of this model, Nichols (2014: 747) concluded that 89% of the awards granted by the directorate of the Social and Behavioral Sciences were classified as 'interdisciplinary research'. However, one can question whether this 'interdisciplinarity' were perhaps a consequence of the mixing of disciplinary terminologies by the topic model (Leydesdorff and Nerghes 2017: 1034).

The vision behind topic modeling echoes the older programs of 'semantic maps' in artificial intelligence by Landauer et al. (1998; cf.van Atteveldt 2008) and 'co-word mapping' in Science and Technology Studies by Callon et al. (1983; cf. Law and Lodge 1984; Latour 1986). While semantic maps and co-word analysis have roots in the social sciences and the humanities, topic modeling appeals to computer scientists because of its relation to 'big data' and computational power. The incentive is not the understanding, but the upscaling. For example, in their study entitled 'Quantitative analysis of large amounts of journalistic texts using topic modelling', Jacobi, van Atteveldt, and Welbers (2016: 89) formulate the challenge of topic modeling as follows:

> 'The huge collections of news content which have become available through digital technologies both enable and warrant scientific inquiry, challenging journalism scholars to analyse unprecedented amounts of texts. We propose Latent Dirichlet Allocation (LDA) topic modelling as a tool to face this challenge. LDA is a cutting-edge technique for content analysis, designed to automatically organize large archives of documents based on latent topics, measured as patterns of word (co-)occurrence'.

Whereas a co-word map is meant to be illustrative to an argument and thus to open the discussion, topic models tend to be black-boxed for users and authoritative. In the meantime, however, user-friendly routines for LDA are available at the internet and consequently the rate of adoption of topic models is much higher than semantic maps. Figure 1 shows the numbers of search queries for both approaches on the basis of an analysis using Google Trends.

More specifically, LDA is available as a user-friendly app at https://code.google.com/p/topic-modeling-tool/ since 2013. LDA is also included in several software packages (e.g., the Stanford Topic Modeling Toolbox at https://nlp.stanford.edu/software/tmt/tmt-0.4/; Ramage et al. 2009, MALLET at http://mallet.cs.umass.edu/, or Diesner's (2014) program ConText for semantic mapping). Thus, the instrument of topic modeling has become intensively used in both scholarly and political contexts, making the validity and reliability of the results of LDA models urgent topics.

2.1 LDA and the research excellence framework (REF 2014)

In the context of the 2014 Research Excellence Framework (REF 2014), data was collected and analyzed in order to demonstrate the societal impact of research performed at 154 participating UK higher-education institutes. The universities submitted 6,975 case studies showing the impact of scientific works on society using a standardized template. This data collection was widely acclaimed in the literature and used in several evaluation studies (Derrick, Meijer, and Van Wijk 2014; Briggle, Frodeman, and Holbrook 2015; Samuel and Derrick 2015; Van Noorden 2015; Hicks and Holbrook 2017).

As part of the REF (2014) analysis of the data a project using text mining techniques was commissioned to a combined group of researchers at King's College in London and an expert group at Digital Science—an offspring of the Nature Publishing Group/Macmillan.

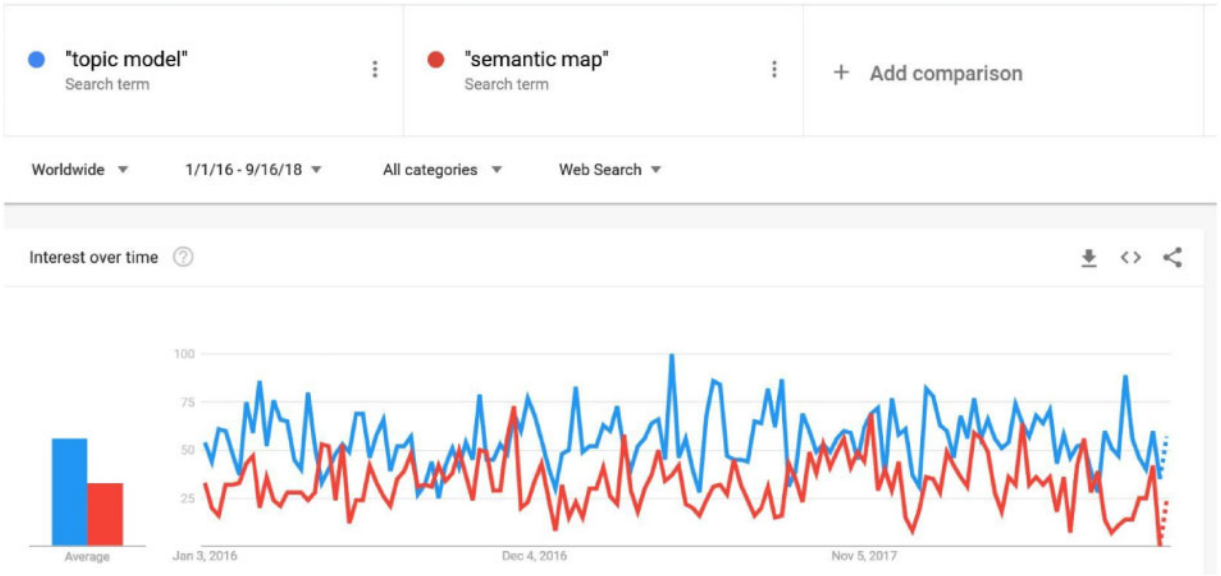

Figure 1. Comparison of the frequency of search queries on topic models and semantic maps.[1]

Both the data and the report (Grant 2015) are available online. LDA was used to investigate the collected impact studies. The authors of the report programmatically stated (at p. 85) that, in comparison to other techniques, LDA would have the following advantages:

> One of the most important aspects of topic modelling as implemented in LDA is that rather than simply basing topics on word features occurring in documents together, it uses contextual information of word occurrences in documents, and so can capture words with similar meanings that are used interchangeably within similar contexts.
> [. . .]
> LDA is the accepted state-of-the-art in topic modelling and is implemented in many standard toolboxes for machine learning.

The investigations reported in this study build on the topic model of 6,638 case descriptions of societal impact of research for which the data is complete.

Our first step was to contact the authors of the REF study at Digital Science with a request for access to the data, output, and routines, and an invitation to collaborate and possibly co-author such a replication. We obtained the following answer (Martin Szomszor, pers. comm., October 4, 2017):

> Generating the same topic model that King's produced would be practically impossible. Even if we were able to obtain the original source data, the code for preprocessing, and the topic modelling parameters, the output could still differ depending on the software libraries used and their versions. As you may also be aware, most topic modelling implementations will rely on a random seed that may not be known.
>
> To complicate matters further, the researcher responsible for the analysis [. . .] has now left King's. For this analysis, King's did not use the text that is now made available on the case studies website. Those cleaned versions were not available at the time so they made use of the text that could be automatically extracted from the original PDFs.
>
> Our current messaging around topic modelling is that no single topic model is more correct than another, but one may be more suited to answering a particular question than another. The target number of topics is the main consideration here, and usually needs to be made relevant to the likely use-case, with small numbers giving very broad generalisations and higher numbers giving more detail. If granularity is pushed too high, the topics start to degrade into incoherent nonsense. We tune these parameters on a per-dataset basis depending on diversity and volume of text.

In other words, Szomszor articulates that the results of a topic model are irreproducible because of technical factors, namely random seeding of Gibbs sampling and ongoing updates of the hard and software. Moreover, the customer may have a considerable say in the results because parameters have to be tuned to the use-case and its objectives. From this perspective, the development of the model is a practice which is not precisely reproducible.

In our opinion, one can use Gibbs sampling with a fixed seed for solving the problem of the random feed. However, there may be also problems more inherent to the nature of these kinds of studies, in particular, biases introduced by document sampling and the interpretability and verifiability of the results. Given our concerns about the validity of the resulting topics and these reservations about the reproducibility of a model in different runs, we approach the issue of reliability by studying the stability of the results in a space of *possible* solutions. We use the same case materials and, as much as possible, similar or comparable techniques for thus addressing the reliability issue.

2.2 Comparison of methods

Following Leydesdorff and Nerghes (2017), we use factor analysis (principal component analysis (PCA) with Varimax rotation) of the word/document matrix for the comparison. Both LDA and PCA can be used to attribute values (weights) to documents as cases and words as variables, and both can be used for grouping words or documents. In the case of PCA, one can use factor loadings of the words and factor scores of the documents, respectively. In LDA a document is considered a probability distribution of topics and a topic a distribution over the words. The probability of the participation of words and documents in each topic can thus be estimated. Different from PCA, LDA

is a generative model: probability distributions over topics and words are learned such that the resulting model most likely generates the observed corpus when the data is sampled from these distributions.

The results of PCA and LDA can be compared in several respects despite the two very different approaches. First, one can compare the differences between the resulting classifications of words into topics (LDA) or clusters (PCA) using Cramèr's *V* which offers a summary statistics between zero and one based on the chi-square. Despite this comparability of LDA and PCA in terms of the results, the two techniques are very different. LDA is based on a probabilistic model, whereas PCA is based on matrix algebra. PCA can in principle be done analytically (with pencil and paper). However, the number of documents and memory requirements are limiting factors in PCA, while LDA can be used for analyzing very large sets.

In the case of PCA, the number of factors to be extracted requires as much a decision as the number of topics in LDA. In the case of PCA, several statistical tools are available such as scree plots and the percentage of variance explained to guide this decision, while in the case of LDA the parameter choices are made in the practical context of the application.

In summary, we can compare LDA and PCA in terms of the following respects:

1. *Stability* of the model. Topic modeling is well-suited for structuring large document corpora, for example, for building document retrieval systems, especially when reading is beyond the human capacity. In these cases the corpus is considered to be a closed collection of documents that can be structured by grouping similar documents according to latent topics. However, when it comes to topic modeling in sampled and open corpora, stability issues become more salient.
   Since training a LDA model requires sampling from probability distributions, models of the same corpus can be expected to differ as seeds of the random number generator used for sampling vary. Notwithstanding the inherently non-deterministic nature of LDA, the sensitivity of topic models to relatively small corpus changes can be used to estimate the model's reliability and reproducibility. In other words, one objective is to investigate to what extent small changes in the text corpus have an effect on the outcome when drawing a sample. If a model shows a high sensitivity to minor variations in the samples, the corpus size, and the sampling procedure, non-systemic contingencies can have an impact on the results and thus the model outcomes can be expected to lead to erroneous conclusions and unwarranted interpretations.
2. *Validity* and *Interpretability* of the assignment. In the social sciences, one often does not have an external ground-truth of the data to validate the topics in terms of their meanings. Furthermore, assessment of validity is domain-specific, and thus, one needs domain expertise (Mimno et al. 2011). However, if an expert would be able to specify the topics in a corpus of documents manually, automatic topic modeling would no longer be needed.
   Validation and interpretation of the discovered topics is closely coupled with the validity of conclusions based on these models. Thus, possibly incorrect or hardly interpretable models can have an unwanted impact when the results are used for decision support.As in PCA, topic models are based on representations of words and documents in latent spaces. It is commonly assumed that the latent space is semantically meaningful but this assumption has to be supported by quantitative evaluations of the coherence and meaningfulness of topics (Chang et al. 2009).

2.3 Data and design

We use the 6,638 'impact case studies' of REF-2014 which are available for download at http://impact.ref.ac.uk/CaseStudies/Results.aspx? val=Show%20All. The texts in the column of the spreadsheet headed 'Summary of the impact' were transformed into lower case and pre-processed as in the original study (Grant and Hinrichs 2015) by using stemming (Porter 1980), stop-word removal, and the removal of punctuations. This leads to 30,934 words which occur 517,211 times in the set. Factor analysis further requires removal of noisy and infrequent words. Of normalized corpus words, 898 occur more than hundred times. These words occur 352,205 times (86.1% of the 517,211 occurrences) in total. In addition, eight words were removed for technical reasons.

Our basic word/document matrix (Salton and McGill 1983) thus contains 890 words as column variables attributed to 6,638 documents as rows. This matrix is factor-analyzed using SPSS (v. 22). We also derive from this matrix a cosine-normalized co-occurrence matrix among the 890 words which is analyzed and visualized using the implementation of the so-called Louvain-algorithm for community-finding in Pajek (Blondel et al. 2008) and VOSViewer for the visualization (van Eck and Waltman 2010). (The cosine is similar to the Pearson correlation underlying the factor analysis, but without the normalization to the mean). Cosine-normalization scales the numbers of word occurrences between zero and one; the resulting visualization is more focused on structural components than without this normalization (see for more details Vlieger and Leydesdorff 2011).

In order to avoid variation in the topics among runs induced by the non-determinism (i.e., the initial seed) of LDA, we fix the random seed of the random number generator used for Gibbs sampling so that multiple runs on the same corpus yield the same results. Using the 6,638 texts as input, we perform LDA with the following parameters: (1) 40 burn-in iterations; (2) 1,500 iterations; (3) alpha = 50/# texts; (4) thinning = 50. These values are akin (if not similar) to the ones used by Grant and Hinrichs 2015 for generating the original topic model.

3. Results

3.1 First observations

Not surprisingly given the sparsity of the matrix, the scree-plot of the factor analysis is very flat: 361 eigenvalues are larger than 1.0 (the default cut-off in SPSS). This can hardly be considered as a reduction of the complexity. However, this is a well-known problem when considering texts as bags of words; words are used flexibly. Citations, for example, are more specific than words by an order of magnitude (Leydesdorff 1989; Braam, Moed, and van Raan 1991). Decomposition of the cosine-matrix using the Louvain algorithm suggests six to eight distinct communities with modularity $Q = 0.10$. Visual inspection of the scree-plot makes an eight-factor solution also plausible. However, eight factors explain only 3.05% of the variance in the matrix.

For the orientation of the reader Figure 2 provides a visualization of the eight components using the two-mode matrix of 890 words versus eight clusters of words based on PCA (Vlieger and Leydesdorff 2011). The factor designation is ours (Table 1). The eight factors (PCA) are compared in Table 1 with an eight-topic solution of LDA in Table 2. Six topics can be unambiguously mapped to topics suggested by factor analysis (Industrial, Medical, Education, Policy,

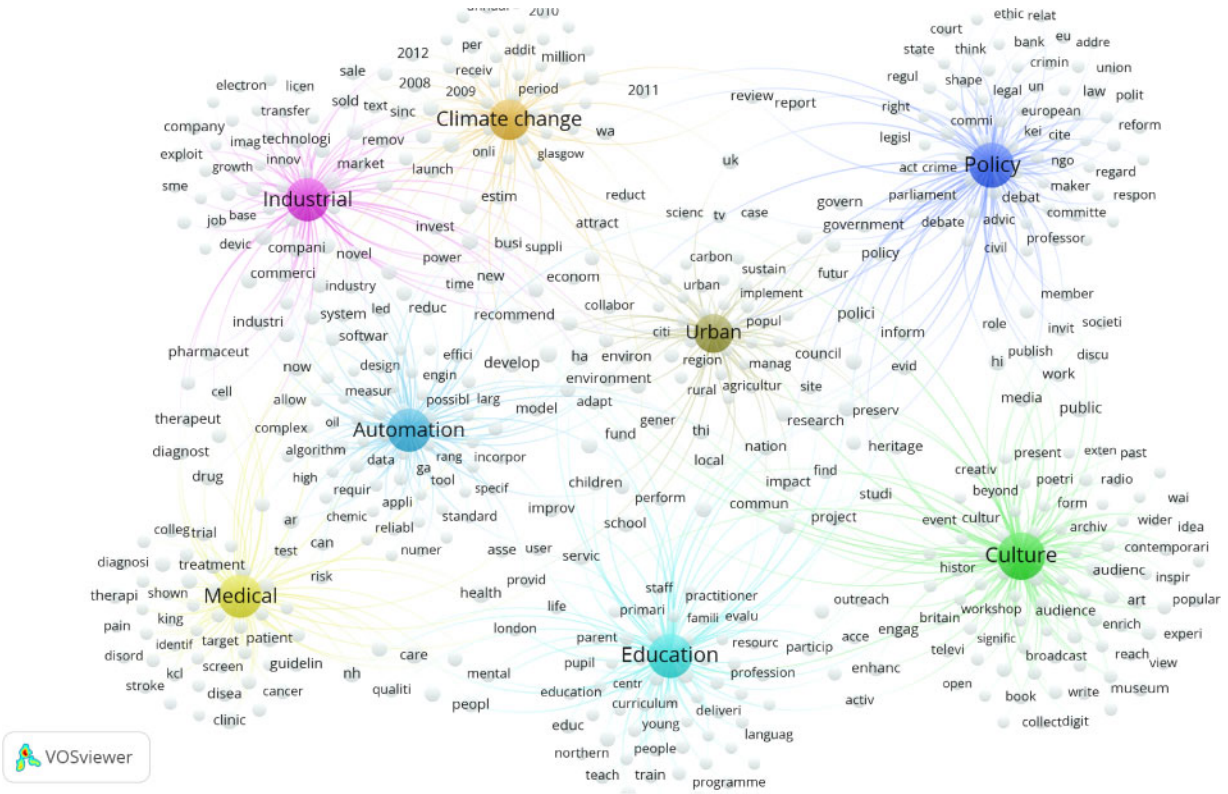

Figure 2. Eight topics based on principal component analysis of the word-document matrix.

Table 1. Eight topics identified by Factor Analysis.

Cultural	Policy advice	Medical	Industrial	Education	Automation	Climate	Urban
Public	Polici	Clinic	Company	Educ	Softwar	Wa	Manag
Culture	Debat	Patient	Spinout	School	Method	Million	Local
Audienc	Law	Treatment	Technologi	Teacher	Model	Year	Plan
Engag	Govern	Trial	Ltd	Learn	Data	2008	Environment
Exhibit	Right	Disea	Company	Train	Ar	2011	Climat
Museum	Reform	Therapi	Product	Teach	System	Per	Region
Art	Committee	Guidelin	Commerciali	Children	Tool	2013	Urban
Histori	Influenc	Drug	Commerci	Profession	Industri	Estim	Citi
Audience	Research	Care	Patent	Practic	Comput	2012	Sustain
Artist	Legisl	Nh	Market	Servic	Algorithm	2010	Environ

Table 2. Eight topics identified by LDA.

'Textual'	Industrial	Medical	Education	Policy	Climate	Culture	Economy
Thi	Product	Health	Educ	Polici	Univers	Public	Manag
Work	New	Patient	Servic	Nation	World	Cultur	Model
Studi	Company	Clinic	Practic	Inform	Sinc	Understand	Improv
Intern	Industry	Treatment	Programm	Govern	Thi	Work	Provid
Case	Design	Care	Support	Influenc	Led	Media	Assess
Signific	Technologi	Improv	Commun	Intern	2008	Engag	Chang
Practice	System	Now	Local	European	Major	Project	Environment
Within	Commerci	Led	Profession	Chang	Base	New	Data
Contribut	Market	Result	School	Debat	8212	Art	Tool
Relat	Process	Drug	Social	Contribut	Million	Histori	New

Culture, and Economy). The remaining two topics cannot or only partially be interpreted. The first topic suggests a topic that is distributed as a layer of 'methodological' terms distributed across the texts (Draux and Szomszor 2017: 12). However, PCA focuses by definition on specific densities.

3.2 Stability of topics

As noted above, a problem arises when topics are modeled on the basis of empirically sampled text collections as it is done in research impact studies. In these cases observing the whole set of relevant documents is practically impossible, and thus, models based on subsequent samples can be considered as an approximation of the actual research topics. We can study how sensitive are PCA- and LDA-based models to incomplete samples of a text corpus by using random drawings of different sizes from the original REF corpus successively.

First a set of 1,000 (C_{rem}) documents from the entire 6,638 documents in the REF corpus (C_{orig}) was randomly selected. Using this sample, 20 new text corpora in steps of fifty: C_{-50}, C_{-100}, C_{-150}, ..., C_{-1000} were created by removing the first 50, 100, 150, ... 1,000 documents of C_{rem} from the original corpus (C_{orig}). Topics were then modeled for each of these 20 new corpora and compared to the topics derived from the original text collection C_{orig} in order to investigate the impact of these variations in the text sampling on the outcome.

In both LDA and PCA models based on the word-document matrix, a weight is attributed to each word indicating the association with a topic or factor. In LDA this weight corresponds to the probability of that the word is used given a topic and in case of PCA the weight is equal to a factor loading. It is common practice to characterize each topic by its top-10 words.

Since words can have high weights for multiple topics, it can happen that one word is a representative of multiple topics. In this study, the reference model is the top-10 words representation of topics derived from the entire text corpus C_{orig}, and the models to test are based on the twenty reduced samples C_{-50}, C_{-100}, C_{-150}, ..., C_{-1000}.

Figure 3 shows the similarities of the topic representations for the 20 reduced text corpora and the original corpus for different numbers of topics generated by LDA (left-hand panel) and PCA of the word-document matrix (right-hand panel), respectively. The details of the similarity calculations are outlined in Appendix A.

As expected, the similarity between the models tends to decrease the more documents are removed from the original corpus, especially for larger numbers of topics. This tendency is much more salient for PCA-based models than for LDA. However, the deviation of the LDA models based on samples from the reference model is already large when only a small fraction of documents is removed from the original corpus. In the case of LDA, removing only 50 documents can lead to topics very different from the original model, whereas the PCA-based model is more robust and less sensitive for this first intervention.

In another experiment, surprisingly the steep drop of topic similarity in LDA models could already be observed if only 5 documents were removed. This leads to the conclusion that any variation of the text corpus can cause unpredictable variation, which makes this problem in empirically sampled topics even worse.

Furthermore, the more topics are declared the larger the deviation of the topics found in the samples compared with the reference. This result can be expected for the following reason: when more topics are extracted, more degrees of freedom are introduced and therefore topics can be more differentiated. In sum, the relationship between the number of topics and the sensitivity to corpus changes raised questions when applying topic modeling to the REF study. Therefore, we turn in the next section to the relationship between the instability of topics after data sampling and the number of topics as input parameters in more detail.

Figure 4 depicts the aggregated similarity of the topic models derived from the different reduced corpora: C_{-50}, C_{-100}, C_{-150}, ..., C_{-1000} and the original text corpus C_{orig} over varying number of topics for LDA and PCA, respectively. In both cases the decline in the similarity with increasing numbers of topics provides further evidence that the sensitivity to corpus variations is correlated with the number of topics. However, the comparison of LDA and PCA models shows that the *minimum* similarity of a PCA-based model is in most cases larger than the corresponding *maximum* value for LDA. In other words: the largest distortion by sampling in the case of PCA is smaller than the smallest distortion in the case of LDA. PCA thus outperforms LDA in this respect.

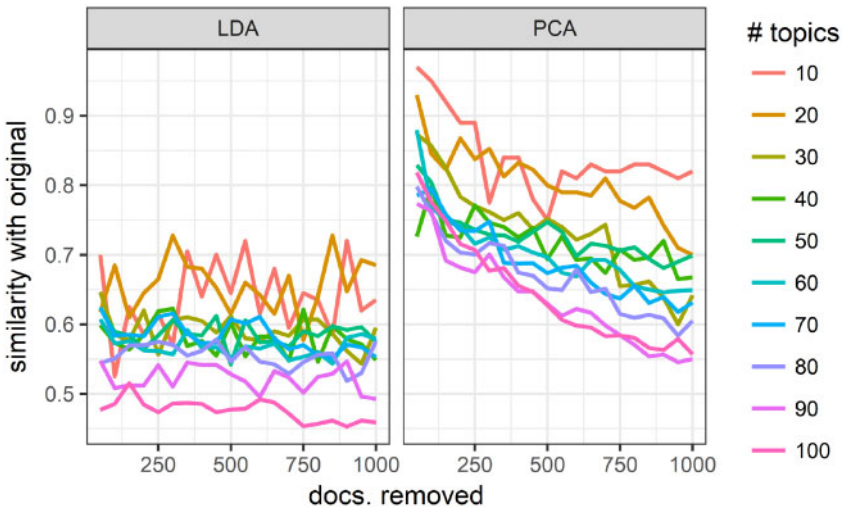

Figure 3. Similarity of topic and factor models derived from samples (C_{-50}, C_{-100}, C_{-150}, ..., C_{-1000}) compared with the original collection (C_{orig}.).

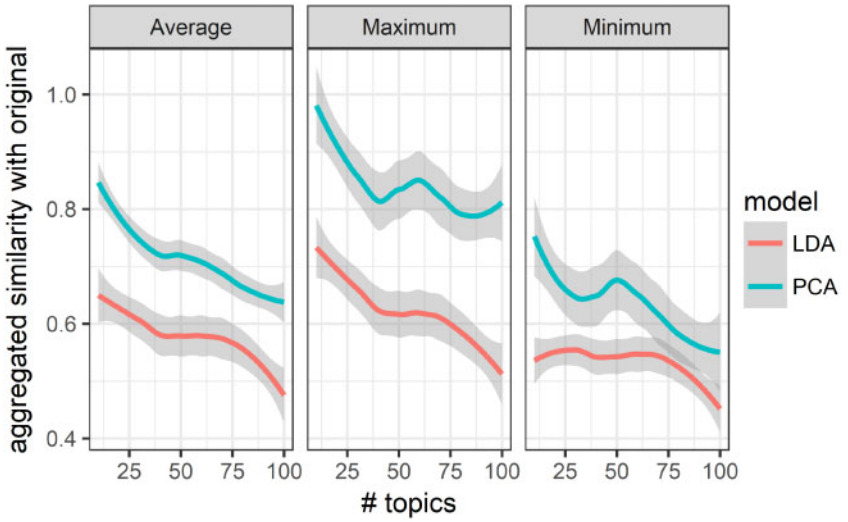

Figure 4. Maximum, mean, and minimum similarity of topics discovered from the reduced text samples and the topics discovered in the original corpus for different number of topics based on LDA and PCA, respectively.

The practical implication of this finding for probabilistic topic models, is that one has to be very careful when applying topic modeling on empirical data sampled as a subset of the corpus in a domain. A small bias in the sampling or incidentally missed documents can have a large impact on the inference of topics and thus on the conclusions and decisions based on the models. These models may not be sufficiently robust for serving as legitimation of decisions.

As already described above, another source of instability of probabilistic topic models is the initialization of the Gibbs sampling during the topic inference process. Since stochastic sampling introducing some randomness, the results are not deterministic. In practice, this can lead to reproducibility problems of topics if an analyst is not aware of this limitation and may require methodological adjustments as pointed out by Agrawal, Fu, and Menzies (2018). In order to ensure reproducible result the seed of the applied random number generator used for sampling can be fixed as done for the results reported above. To assess the instability of LDA topics if no preparations for stabilization of the results are applied, we compare the variations in the resulting topics if the stochastic sampling process is initialized with different seeds in each of 10 different runs. The average pairwise topic similarity of the 10 resulting models is depicted in Figure 5 for different number of topics.

This shows very clearly that the more topics are derived the higher is the impact of the topic variation resulting from the non-determinism of LDA, if it is naively applied.

3.3. Interpretability and semantic coherence

The background problem of working with empirically sampled data is the absence of ground-truth data that allows for assessing different models according to external validity criteria. Thus, internal validity criteria are needed that allow an analyst to obtain insights into how adequate the actual topics of a domain are covered by a topic model. One of these measures is the semantic coherence of topics as introduced by Mimno et al. (2011)—see Appendix B—which has also been applied in other studies on assessing the outcome of topic modeling (cf. Stevens et al. 2012).

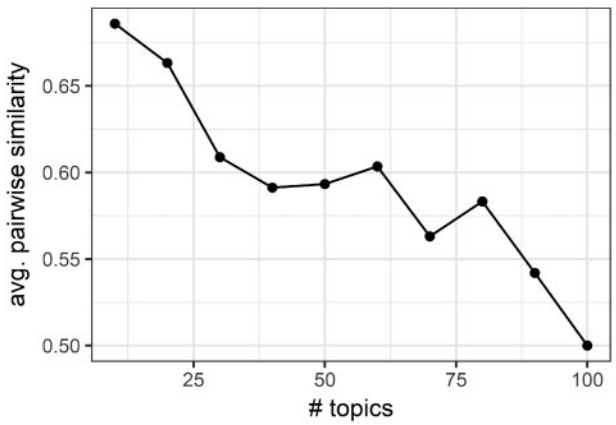

Figure 5. Pairwise topic similarity of 10 models using different seeding of the Gibbs sampler for different number of topics.

The comparison of the average internal coherence of the topic models produced by LDA and PCA for a different number of topics (on the *x*-axis) is shown in Figure 6. The coherence is higher in the case of LDA-based topic models when compared with PCA-based models. This result accords with Stevens et al. (2012) who also found a high IC for LDA-based models. However, the internal coherence (IC) of LDA-based models decreases with increase in the number of topics, while in PCA-based models the number of factors does not affect the internal coherence.

The decline in the curve for LDA suggests that there is a trade-off between coherently capturing the topics and the sensitivity to the number of topics. However, when we ran a number of tests with eight topics as in Table 2, the topics varied in terms of the top-10 words, but the classification was always significantly the same (with Cramèr's *V* by the order of 0.7, $P < .001$).

4. Discussion and conclusions

The enthusiasm for topic modeling as a technique to summarize large amounts of documents in the format of a limited number of

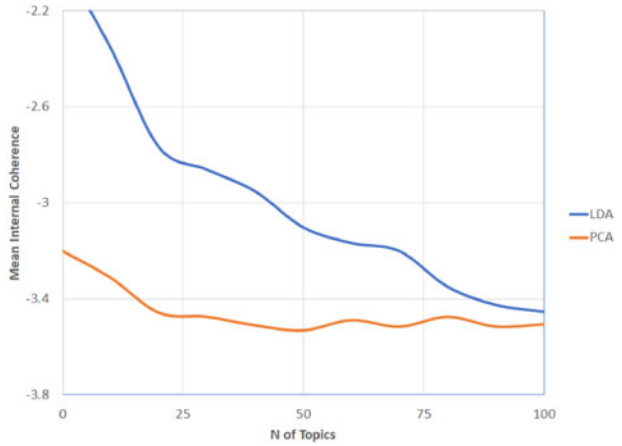

Figure 6. Average internal coherence of topics using LDA and PCA, respectively.

words cannot be justified because of the validity problems which are inherent to this methodology. The probabilistic character of the results may easily lead to misunderstandings outside the context of the production of these models. Producers sometimes go along with clients to give the results an interpretation and to 'train' the model for choosing parameters leading to results which are plausible and lead to useful results. From the perspective of careful decision-making, however, these models can be considered as quicksand on which one should not build. They provide an equivalent to the practice of catalogue systems but seem to be more objective because the generation is computer-assisted.

We raised this question initially in relation to co-word modeling of samples of a size which allowed us to have substantive understanding of the results given our own background. As noted, Leydesdorff and Nerghes (2017) found sometimes very counter-intuitive the results of LDA on the basis of medium-sized sets ($n \approx 1,000$). Proponents of topic modeling assured us that these problems were generated because of the relatively small sizes of the samples. However, sample sizes should be large (much larger than 1,000 documents) so that human validation is impossible in practice. Added to this is the irreproducibility of topic models and the dependency of a host of parameters which are usually not under control such as new versions of computer programs.

In this study, we controlled for the reproducibility issue. We found that the topic structure can reliably be reproduced from run to run if the parameter settings are fixed. This has increased our trust in the reliability of the technique: the system seems not to get stuck in sub-optima. The validity, however, remains a serious matter of concern. We showed that LDA-based topic models are more sensitive than PCA-based models when relatively small changes are made in the corpus or the number of topics to be extracted. The most drastic distortion of the PCA model had less effect on the results than the most modest distortion of the LDA model. LDA, however, scored much better than PCA on internal cohesion. This means that LDA based models reflect a certain thematic structure of the corpus resulting in a proper internal coherence but the emphasis of words representing the topics, and thus, the interpretation might differ if the input corpus is modified. Within the scope of this article, only the thematic structure of corpora, namely, associations between words and topics was studied. It would be also possible to assess the associations between documents and topics. These associations might be more robust against small corpus changes as the results on internal cohesion suggest. However, since the word lists are essential for topic representations, interpretation of by human analysts from the word lists alone is still difficult. Therefore, considering, both, word lists and a sample of documents that is strongly affiliated to a topic may be one possible approach to come to a reliable labeling of latent topics from LDA as well as PCA-baesd models.

In summary, LDA-based models provide reliable statistical results about the corpus under study, which is appropriate and desirable for structuring 'closed' (static) collections of documents in information retrieval. However, if applied to 'open-ended' (changing) corpora where only samples can be obtained as in the REF study, LDA-based topic models are difficult to validate; they may not be valid because of the sensitivity to small variations in the document corpus (Section 3.2 above). One can make the argument that the problem with topic models is not the method itself—the statistics—but the way they are used in decision-making—the semantics. In our opinion, topic models should not be used as the basis for decision making or intellectual delineations of domains in scholarly works, but for statistical purposes. Models based on co-occurrences of words in word/document matrices are a preferable alternative in situations where the content itself and not only the statistics count such as in micro-decision making and in scholarly work. As a general recommendation one can say that if probabilistic topics models based on LDA are applied, an analyst has to be aware of these limitations, and possibly perform different runs of the algorithm with varying parameter settings and different subsets of the input corpus in order to come to a consensus model, which allows much more reliable conclusions.

Since this work has shown different strengths and weaknesses of the two modeling approaches, a further question can be how to balance the advantages and disadvantages of probabilistic and co-word based methods. One solution may be, for example, the triangulation of various methods as proposed by T-Lab (Cortini and Tria 2014). Triangulation of models may be an approach to come to more reliable decisions based on unknown or dynamic corpora.

4.1 Limitations of the study

The results presented in this study provide some general insights into the basic properties of topic modeling methods in open corpora when it comes to the question of reliability and interpretability of the results. Stability issues of topic models have been addressed in relation to the number of topics (Greene, O'Callaghan, and Cunningham 2014), and variation in the input vocabulary (De Waal and Barnard 2008). Agrawal, Fu, and Menzies (2018) lists several studies that mention different aspects of topic instability. This study extended these works by investigating the effect of corpora variation on the extraction of topics. However, to come to a general conclusion and guidelines for how to support policy-making by topic models, more extensive experiments with different corpora in more areas of application are required.

In terms of interpretability and semantic coherence only internal validation methods could be used in the absence of ground-truth topics. It is very difficult to produce such ground-truth from large corpora and a lot of expert knowledge is required. However, the reason for applying topic modeling in science-policy making is that the thematic organization of the impact of science over several years is not known. If a human being would be able to specify the ground-truth topics, automated analysis of science impact studies would no longer be needed. A way out of this dilemma could be to combine

internal and external validation, where initially detected topics—using topic modeling in an explorative manner—are further empirically validated in follow-up studies using semantic maps. However, one risks to 'train' programs not only in terms of their utility, but also in terms of the suitability of results.

Note

1. The method of data collection of Google Trends was improved on January 1, 2016.

References

Agrawal, A., Fu, W., and Menzies, T. (2018) 'What is Wrong with Topic Modeling? And How to Fix It Using Search-Based Software Engineering', *Information and Software Technology*, 98: 74–88.

Anderson, C. (2008) 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete', *Wired Magazine*, 16/7: <https://www.wired.com/2008/06/pb-theory/> accessed 11 June 2019.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003) 'Latent Dirichlet Allocation', *Journal of Machine Learning Research*, 3: 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008) 'Fast Unfolding of Communities in Large Networks', *Journal of Statistical Mechanics: Theory and Experiment*, P10008.

Braam, R. R., Moed, H. F., and van Raan, A. F. J. (1991) 'Mapping of Science by Combined Co-Citation and Word Analysis. I. Structural Aspects', *Journal of the American Society for Information Science*, 42/4: 233–51.

Briggle, A., Frodeman, R., and Holbrook, B. (2015) *The Impact of Philosophy and the Philosophy of Impact: A Guide to Charting More Diffuse Influences across Time*. Impact of Social Sciences Blog <https://blogs.lse.ac.uk/impactofsocialsciences/2015/05/26/the-impact-of-philosophy-and-the-philosophy-of-impact/> accessed 11 June 2019.

Callon, M. (1986) 'The Sociology of an Actor Network: The Case of the Electric Vehicle', in M. Callon, J. Law, and A. Rip (eds) *Mapping the Dynamics of Science and Technology*, pp. 19–34. London: Macmillan.

Callon, M., and Courtial, J.-P. (1989) *Co-Word Analysis: A Tool for the Evaluation of Public Research Policy*. Paris: Ecole Nationale Sup,rieure des Mines.

Callon, M. et al. (1983) 'From Translations to Problematic Networks: An Introduction to Co-Word Analysis', *Social Science Information*, 22/2: 191–235.

Chang, J. et al. (2009) 'Reading Tea Leaves: How Humans Interpret Topic Models', in Bengio Y., Schuurmans D., Lafferty J. D., Williams C. K. I., and Culotta A. (eds) *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS'09), Vancover, BC, Canada*, Curran Associates Inc.

Cortini, M., and Tria, S. (2014) 'Triangulating Qualitative and Quantitative Approaches for the Analysis of Textual Materials: An Introduction to T-Lab', *Social Science Computer Review*, 32/4: 561–8.

De Waal, A., and Barnard, E. (2008) 'Evaluating Topic Models with Stability'. In*: 19th Annual Symposium of the Pattern Recognition Association of South Africa* <http://hdl.handle.net/10204/3016> accessed 11 June 2019.

Derrick, G., Meijer, I., and Van Wijk, E. (2014) 'Unwrapping "Impact" for Evaluation: A Co-word Analysis of the UK REF2014 Policy Documents Using VOSviewer', in Noyons E. (ed.) *Proceedings of the Science and Technology Indicators Conference 2014 Leiden: Context Counts: Pathways to Master Big and Little Data. Universiteit Leiden, Leiden*, pp. 145–54.

Diesner, J. (2014) 'ConText: software for the integrated analysis of text data and network data'. Paper presented at the Social and Semantic Networks in Communication Research Preconference at International Communication Association (ICA), Seattle, WA, USA.

Draux, H., and Szomszor, M. (2017) *Topic Modelling of Research in the Arts and Humanities: An Analysis of AHRC Grant Proposals*. London: Digital Research Reports <https://figshare.com/articles/Topic_Modelling_of_Research_in_the_Arts_and_Humanities/5621260/5621261>.

Eco, U. (1976) *A Theory of Semiotics (Transl. W. Weaver)*. Bloomington: Indiana University Press.

Girolami, M., and Kabán, A. (2003) 'On an equivalence between PLSI and LDA'. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03), ACM, New York, NY, USA, pp. 433–4.

Goldstone, A., and Underwood, T. (2012) 'What Can Topic Models of PMLA Teach us about the History of Literary Scholarship', *Journal of Digital Humanities*, 2/1: 39–48.

Graham, M. (2012) 'Big Data and the End of Theory', *The Guardian*, 9.

Grant, J. (2015) *The Nature, Scale and Beneficiaries of Research Impact: An Initial Analysis of Research Excellence Framework (REF) 2014 Impact Case Studies*. London: King's College and Digital Science.

Grant, J., and Hinrichs, S. (2015) *The nature, scale and beneficiaries of research impact: An initial analysis of the Research Excellence Framework (REF) 2014 impact case studies*. HEFCE - Higher Education Funding Council for England.

Greene, D., O'Callaghan, D., and Cunningham, P. (2014) 'How Many Topics? Stability Analysis for Topic Models', in T. Calders, F. Esposito, E. Hüllermeier, and R. Meo (eds) *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2014*. Lecture Notes in Computer Science, Vol 8724. Berlin, Heidelberg: Springer.

Greimas, A. J. (1983) *Du Sens II: Essais Sémiotiques*. Paris: Éditions du Seuil.

Gretarsson, B. et al. (2012) 'Topicnets: Visual Analysis of Large Text Corpora with Topic Modeling', *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3/2: 23.

Grimmer, J., and Stewart, B. (2013) 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts', *Political Analysis*, 21/3, 267–97.

Hicks, D., and Holbrook, J. B. (2017) *The impact of philosophy: Evidence from the UK research excellence framework*. <http://works.bepress.com/diana_hicks/48/> accessed 11 June 2019.

Jacobi, C., van Atteveldt, W., and Welbers, K. (2016) 'Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling', *Digital Journalism*, 4/1: 89–106.

Lancichinetti, A. et al. (2015) 'High-Reproducibility and High-Accuracy Method for Automated Topic Classification', *Physical Review X*, 5/1: 011007.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998) 'An Introduction to Latent Semantic Analysis', *Discourse Processes*, 25/2: 259–84.

Latour, B. (1986) 'Visualisation and Cognition: Drawing Things Together', *Knowledge and Society: Studies in the Sociology of Culture past and Present*, 6: 1–40.

Latour, B. (1996) 'On Interobjectivity', *Mind, Culture and Activity*, 3/4: 228–45.

Law, J., and Lodge, P. (1984) *Science for Social Scientists*. London: Macmillan.

Leydesdorff, L. (1989) 'Words and Co-Words as Indicators of Intellectual Organization', *Research Policy*, 18/4: 209–23.

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*, 68/4: 1024–35.

Mimno, D. et al. (2011) 'Optimizing semantic coherence in topic models' in *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA*, pp. 262–72. Association for Computational Linguistics.

Nichols, L. G. (2014) 'A Topic Model Approach to Measuring Interdisciplinarity at the National Science Foundation', *Scientometrics*, 100/3: 741–54.

Porter, M. F. (1980) 'An Algorithm for Suffix Stripping', *Program*, 14/3: 130–7.

Ramage, D. et al. (2009) *Topic Modeling for the Social Sciences*. Proceedings of the NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond. Whistler, Canada.

Samuel, G. N., and Derrick, G. E. (2015) 'Societal Impact Evaluation: Exploring Evaluator Perceptions of the Characterization of Impact under the REF2014', *Research Evaluation*, 24/3: 229–41.

Salton, G., and McGill, M. J. (1983) *Introduction to Modern Information Retrieval*. Auckland McGraw-Hill.

Stevens, K. et al. (2012) 'Exploring topic coherence over many models and many topics' in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea*, pp. 952–961. Association for Computational Linguistics.

Talley, E. M. et al. (2011) 'Database of NIH Grants Using Machine-Learned Categories and Graphical Clustering', *Nature Methods*, 8/6: 443–4.

van Atteveldt, W. H. (2008) *Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content*. Charleston, SC: BookSurge.

Van Eck, N. J., and Waltman, L. (2010) 'Software Survey: VOSviewer, A Computer Program for Bibliometric Mapping', *Scientometrics*, 84/2: 523–38.

Van Noorden, R. (2015) 'Seven Thousand Stories Capture Impact of Science', *Nature*, 518/7538: 150.

Vlieger, E., and Leydesdorff, L. (2011) 'Content Analysis and the Measurement of Meaning: The Visualization of Frames in Collections of Messages', *The Public Journal of Semiotics*, 3/1: 28–50.

Appendix A: Similarity calculation between topic models

Using LDA, each topic model m comprises T topics; each topic is represented by a word vector of length S, $t = (w1, w2, \ldots w_S)$. Using Equation 1, it is possible to formulate an adapted version of the so-called purity measure called *topSim* for comparing a topic model m_{test} to a reference model m_{ref} by calculating the overlap of the top S words of a topic in m_{test} and the best matching topic in m_{ref}. The values of *topSim* range between 0 and 1.

$$topSim(m_{test}, m_{ref}) = \frac{1}{TS} \sum_{t_i \in m_{test}} \underset{t_j \in m_{ref}}{\operatorname{argmax}} |t_i \cap t_j| \quad (1)$$

Note that *topSim* is not a strict similarity function since it is not symmetric, i.e., $topSim(m_{test}, m_{ref}) \neq topSim(m_{ref}, m_{test})$. However, it provides an indication of how much a given model deviates from the reference model.

Appendix B: Internal cohesion of topics

According to Mimno et al. (2011) semantic coherence of a given topic t can be measured using Equation 2: the parameter V_t denotes the set of words representing the topic (here the top 10 most associated words); $D(w_i, w_j)$ is the number of documents containing the words w_i and w_j; and $D(w_j)$ is the number of documents containing w_j. Stevens et al. (2012) provide further evidence that this measure is adequate to compare the outcome of different topic modeling approaches.

$$C(t; V_t) = \sum_{w_i, w_j \in V_t} \log\left(\frac{D(w_i, w_j) + 1}{D(w_j)}\right) \quad (2)$$