Special Issue Title: Topic Models and the Cultural Sciences

John W. Mohr and Petko Bogdanov

Special issue of *Poetics* (Vol. 41, no. 6), December 2013

Topic models: What they are and why they matter

John W. Mohr [a*], Petko Bogdanov [b]

[a] *Department of Sociology, 3103 Social Sciences & Media Studies, University of California, Santa Barbara, Santa Barbara, CA 93106-9430, USA*

[b] *Department of Computer Science, University of California Santa Barbara, CA 93106-5110, USA*

Abstract

We provide a brief, non-technical introduction to the text mining methodology known as topic modeling. We summarize the theory and background of the method and discuss just what kinds of things are found by topic models. Using a text corpus comprised of the eight articles from the special issue of *Poetics* on the subject of topic models, we run a topic model on these papers both as a way to introduce the methodology and also to help summarize some of the ways in which social and cultural scientists are using topic models. We review some of the critiques and debates over the use of the method and finally, we link these developments back to some of the original innovations in the field of content analysis that were pioneered by Harold D. Lasswell and colleagues during and just after World War II.

Keywords: Automated Text Analysis; Topic Models; latent Dirichlet allocation (LDA); Hermeneutics and Interpretation; Harold D. Lasswell

[*] Corresponding author.

Email addresses: mohr@soc.ucsb.edu (J.W. Mohr); petko@cs.ucsb.edu (P. Bogdanov).

1. Introduction

> Content analysis is a technique which aims at describing, with optimum objectivity, precision, and generalizability, what is said on a given subject in a given place at a given time.
>
> Harold Lasswell, et. al. (1952, p. 34), *The Comparative Study of Symbols: An Introduction*

In this short essay we provide a brief, non-technical introduction to the text mining methodology known as topic modeling. We start with the basic question, what is a topic model? We summarize the theory behind the method and then focus on the question of what exactly is a topic? (Or, to put it the other way round, we ask just what does a topic model measure?) We address this issue by describing the work published here. For each article we pose three questions—what topics have these researchers found? How have they interpreted the meaning of their topics? And how have they used them as a component within a larger research project? We turn then to briefly discuss some of the demands, dilemmas and limitations of topic models and proceed to the second question telegraphed by our title—why do topic models matter? We answer this by describing some of the ways that we think these methods can change how scholars in the social and cultural sciences approach (and use) texts and textual analysis and we end by taking the long view of just how topic models represent a certain kind of closure on one chapter in the history of content analysis and the beginning of another.

2. What is a topic model?

Topic models are a promising new class of text analysis methods that are likely to be of interest to a wide range of scholars in the social sciences, humanities and beyond.[1] The most distinctive feature of

---

[1] New is a relative term here. The original article on latent Dirichlet allocation (LDA) by Blei, Ng and Jordan (2003) was published a decade ago. As that article quite usefully explains, there is also a long pre-history to the method—including the early work on Latent Semantic Indexing (LSI) by Deerwester et al (1990) and Hoffman's (1999) probabilistic Latent Semantic Indexing (pLSI) approach. There is also another tradition of topic modeling using Gibbs sampling techniques that dates back to work by Griffiths and Steyvers (2004) (see, also Griffiths, et. al,

topic models is that they provide an automated procedure for coding the content of a corpus of texts (including very large corpora) into a set of substantively meaningful coding categories called "topics". The algorithms can do this with a minimum of human intervention and this makes the method more inductive than traditional approaches to text analysis in the social and human sciences.[2] Instead of starting with pre-defined codes or categories of meaning (like those we generate when we start to hand-code a text), the researcher begins by specifying the number of topics for the algorithm to find. The program then identifies that number of topics and returns the probabilities of words being used in a topic as well as an accounting of the distribution of those topics across the corpus of texts. While not infallible, when used thoughtfully and applied carefully, the method seems to consistently yield very plausible readings of the texts, demonstrating what DiMaggio, Nag and Blei, (this issue) describe as high levels of "substantive interpretability."

2.1. The theory behind the method

So, how do topic models work? How does an automated procedure reliably find textual meanings that prove to be useful? A simple answer is that the method depends upon the

---

2005; Newman et. al, 2007). McNamara (2010) provides a broad view of thirteen classes of Latent Semantic Analysis (LSA) that she describes as representing different "statistical models of semantics" (of which topic modeling is one). McNamara also traces the field back to the original work of Osgood, Suci and Tannenbaum (1957). Nonetheless it is still largely a new class of methods for most social scientists and humanists. There are some exceptions. A few political scientists have been quick to pick up these methods and employ them in useful ways (Grimmer, 2010; Grimmer and King, 2011; Grimmer and Stewart, 2013). Also, the digital humanities community is way ahead on the use of topic models. In part thanks to workshops funded by the NEA such as the "Networks and Network Analysis for the Humanities" conference held at the Institute for Pure and Applied Mathematics at UCLA (organized by Tim Tangherlini, also an author of one of the papers published here). See, also the special issue on topic models in the *Journal of Digital Humanities*, edited by Elijah Meeks and Scott Weingart (Vol 2, no. 1) as well as new books by Matt Jockers (2013) and Franco Moretti (2013). Among sociologists, there is the early work by Moody and Light, (2006), and also some interesting new work coming out by Sophie Mützel (2012) and by Sarah Kaplan and Keyvan Vakili (2012) among others.

[2] From a machine learning perspective topic modeling is an unsupervised task, meaning that no prior human annotation, labeling or hand-coding is necessary to infer a model. But of course it is also important to point out that what is inductive for the analyst is deductive for the method, in the sense that LDA topic models presume a particular theory of the meaning of a text and this theory is expressed in way in which the LDA model is constructed. We say more about these assumptions below.

presumption that meanings are relational (Saussure, 1959). In this case, the meanings that define a coherent topic of conversation are constructed from a set of word clusters. Thus a topic might be thought of as the constellation of words that tend to come up in a discussion (and thus to co-occur more frequently then they otherwise would) whenever that (unobserved and latent) topic is being discussed. Note that topic models capture co-occurrences regardless of these words' embeddedness within other complexities of language such as syntax, semantics, or location within the text. Instead each document is treated as if it were a so-called "bag of words." The goal of a topic model analysis is then to analyze these various word bags, to identify word co-occurrence patterns across the corpus of bags and to then use these to produce a mapping of the distribution of words into the topics and of the topics into the bags.

Within the more general data science field, topic modeling is an instance of probabilistic modeling.[3] The simplest and most widely used model is latent Dirichlet allocation (LDA) introduced by Blei and colleagues (Blei 2003).[4] As DiMaggio, Nag and Blei explain (in the paper published here), "LDA is a statistical model of language." The generative process behind the model is a convenient way to introduce its intuition. Each document (text) within a corpus is viewed as a *bag-of-words* produced according to a mixture of themes that the author of the text intended to discuss. Each theme (or topic) is a distribution over all observed words in the corpus,

---

[3] Another computer science branch that deals with text is natural language processing (NLP). The latter differs from topic models in that many of the developed methods require human/expert training. It is important to note that these different model families are compatible and in fact could be combined to get closer to meanings in text. In this special issue the paper by McFarland (et. al.), Mohr (et. al.) and Jockers and Mimno all take this issue up explicitly. For an example of another kind of combination of topic models with other modalities of text analysis see Diesner and Carley (2010).

[4] Since the inception of early topic models like pLSI (Hofmann 1999) and LDA (Blei 2003), a family of approaches have been proposed to address and develop further some of the assumptions in the original models and make them more applicable to specific real world analysis tasks by uncovering more sophisticated structures within texts. Some extensions relax the bag-of-words assumption by modeling the word order (Griffiths et al., 2005, Wallach 2006). Other extensions deal with dependent documents in the corpus by modeling links (Chang and Blei 2010) and dynamic topic models incorporating a temporal order of documents within the corpus (Blei and Lafferty 2006). The assumption of apriori known number of latent topics is addressed by Teh and colleagues (Teh 2006). A more complete list of extensions and new topic models is available in (Blei 2011, Jelisavcic, et. al, 2012). McFarland et. al (published here) also provide a useful review.

such that words that are strongly associated with the document's dominant topics have a higher chance of being selected and placed in the document bag. Given the above distributions the author repeatedly picks a topic, then a word and places them in the bag until a document is complete. The objective of topic modeling is to find the parameters of the LDA process that has likely generated the corpus. This is also referred to as inference in the LDA literature and in essence it is the task of reverse-engineering the intents of the author(s) in producing the corpus.[5]

Among the outputs of the inference is a set of per-word topic distributions associating a probability with every topic-word pair and a similar set of per-topic document distributions describing the probability of choosing a particular topic for every specific corpus document. But note again, the obtained structure is latent, which means that the learned per-word topic distributions are not associated with an explicit topic label, but instead with a set of word probabilities which when ordered by decreasing probability often relate closely to what a human would call a topic or a theme.

2.2. What is a topic?

This then brings us back to the question of just what exactly is a topic? For an answer to this question we will focus on the way that the authors published in this special issue have used the method and we will look to see what types of topics they have found. Table 1 provides a summary of the papers allowing us to see at a glance the range and diversity of topic model applications published in this special issue.

---

[5] Formally, the LDA algorithms are founded on a Bayesian probabilistic model. The DiMaggio, Nag and Blei paper (this issue) does a nice job of offering a simple explanation of the formal logic behind the approach. Rhody (2012) also has a useful explanation of the probabilistic logic behind a topic model in which she uses the homespun analogy of trying to guess what the proportions of vegetables were being sold at the local farmer's market based on a post-hoc examination of one's neighbors' shopping bags. Brett (2012) provides a non-technical overview of the broader topic modeling methodology. Other papers by Blei (2012a, 2012b) also provide very accessible introductions.

Table 1

The nature and scope of topic model applications in articles published in *Poetics* Vol. 41, no. 6 (part–1)

Authors	Discipline	Source	Size of Corpus	# topics	A sampling of "topics" identified by analysis	Measured Object/ Use of Measure
Mohr & Bogdanov	7 1	Articles Published in this special issue of *Poetics* (Vol. 41, no. 6)	8 (Tot # articles) 92,260 (Tot # Words)	25	"public-terror-alerts" "forgotten-versions" "Topic model" "female-male-authors/data"…	Topics measure themes in research articles. TMs used to identify paper specific themes and common themes across papers and to illustrate method.
DiMaggio, Nag & Blei	7 7 1	Newspaper Articles, (if: "NEA", "Arts Agencies", "public funding of arts") (Houston Chron., NY Times, Seattle Times, WSJ & Wash. Post) (1986-1997)	7,958 (Tot # articles) 54,982 (Tot # terms) 3,381,574 (Tot # Words)	12	"NEA grant controversies" "Congressional deliberations" "1990s culture wars" "all kinds of musical performances & orgs" "museum exhibits & visual arts" "theater and dance" …	Topics measure media frames within a policy domain. TMs used as part of research design that focuses on the use of different frames by different types of newspapers and the more general questions about the drop in popularity of public funding for the arts.
McFarland, Ramage, Chuang, Heer, Manning & Jurafsky	7 1 1 1 4 1	Dissertation Abstracts from 240 U.S. Research Universities (Proquest) (1980-2010)	1 million+ (tot # abstracts) (here: a sub-corpus just Anthropology related abstracts)	40	"social structures" "physical anthropology" "archeology" "identity studies" "cultural anthropology"…	Topics measure group language conventions. Paper reviews series of uses of TMs to understand language differentiation in academic communities. Includes summary of different types of TMs.
Miller	2	Qing Dynasty Veritable Records (1723-1911)	—	50	"crime" "unrest" "sedition" "rebellion" "border rebellion" "major rebellion" …	Topics measure how the central state (during Qing dynasty) thought about and categorized mass violence. Used here to gain new insights into the crime rates & state record-keeping practices of 18th and 19th century China.
Bonilla & Grimmer	5 5	News stories on nightly newscasts by ABC, CBS and NBC & Newspapers from across the country (from Lexis-Nexis) (2002-2005)	51,766 (tot # news stories)	24	"Memorial" "Local small business" "Criminal prosecution" "Iraq/World" "Local philanthropy" "Law and order" "Personal interest stories" "2004 Presidential campaign" "Iraq war" …	Topics measure broad, thematic categories for newspaper stories. Used to show that Bush's *Terror Alerts* raise the public's perceived likelihood of a terror attack, but not opinions about President's job performance, foreign intervention, or willingness to restrict civil liberties.
PhD Discipline:	1. Computer Science, 2. East Asian Languages & Civilizations, 3. English, 4. Linguistics, 5. Political Science, 6. Scandinavian, 7. Sociology, 8. Swedish Literature.					

Table 1 (cont.)

The nature and scope of topic model applications in articles published in Poetics Vol. 41, no. 6 (part–2)

Authors	Discipline	Source	Size of Corpus	# topics	A sampling of "topics" identified for analysis	Measured Object/ Use of Measure
Mohr, Wagner-Pacifici, Breiger & Bogdanov	7 7 7 1	U.S. National Security Strategy reports (1990-2010)	11 (# NSS Documents) 6,102 (Tot # Agents) 572,358 (tot # Words)	15	"Terrorism" "Economic development" "Human rights" "Global security strategy" "Military operations" "Peace"…	Topics measure dramatistic "scenes." Incorporated into a model for graphing the Burkean "grammar of motives" of official United States National Security Strategy texts.
Marshall	7	*Population Studies* (if: "fertility") *Population* (if: "*fecondité*" "*natalité*") & select newspaper: *Times* & *Guardian* (1946-2005)	1,623 (tot # Articles in *Pop. Studies*) 1,835 (tot # Articles in *Population*)	75 75	British: "Africa and data" "Economics & transition" "married fertility" "nuptiality" … French: "*housing*" "*war & France*" "*abortion & contraception*" …	Topics measure content of professional discourse. TMs used as part of a cross-national comparison of research discourse (and its impact) in the (British & French) academic discipline of demography.
Tangherlini & Leonard	6 8	*Topic Model Data* 1. "The Origin of Species" & "The Descent of Man" 2. Modern Breakthrough authors: Jacobsen, Schandorf & Drachman 3. Folk legends collected by Tang Kristensen 1892-1901; 1928-1939 *Trawl data:* Google Books Danish corpora (1860-1920)	1. two books 2. selections from several books 3. ~34,000 legends from Kristensen's collection	100 50 100	1. "social instinct" "struggle for survival" … 2. "a woman's thoughts" "her self" "intelligence" "Men, little girls, god, black robes and shouting" … 3. "death and churchyards" "shooting and witches" "horses & wagons" "the minister" "serpents" …	Topics measure literary *feel*. Sub-corpus topic modeling (STM) is presented as a new tool for discovering meaningful passages in a larger corpora. 3 tests of STM trawls here: 1. Tracing the diffusion of Darwin's ideas. 2. Finding unknown authors of the Modern Break-through. 3. Finding the *feel* of Danish folklore in other Danish literature.
Jockers & Mimno	3 1	British, American & Irish works of fiction (from Chadwyck Healey collection, Project Gutenberg & the Internet archive) (1750-1899)	3,346 (Tot # books)	500	"female fashion" "enemies" "convents & abbeys" "religion" …	Topics are a measurable, data-driven proxy for literary themes. Used here to assess how meta-data (like date of pub, gender…) predict fluctuations in the use of themes and the individual word choices within themes. Tests whether this evidence is statistically significant.
PhD Discipline:	1. Computer Science, 2. East Asian Languages & Civilizations, 3. English, 4. Linguistics, 5. Political Science, 6. Scandinavian, 7. Sociology, 8. Swedish Literature.					

The data sources vary widely — both by type of data and by size of corpus. Ian Miller analyzes over a hundred years of the Qing Dynasty's "Veritable Records" containing comprehensive archives of "zouzhe," or messages of concern that were reported directly to the Chinese emperor. McFarland, Ramage, Chuang, Heer, Manning and Jurafsky draw on a corpus of over a million dissertation abstracts (for dissertations filed between 1980-2010) as a way to help map out the changing contours of academic fields. DiMaggio, Nag and Blei analyze a corpus of nearly 8,000 newspapers articles (published between 1986 and 1997) that were concerned with the National Endowment for the Arts (NEA) or with publicly funded art projects in general. Bonilla and Grimmer study over 51,000 news stories (taken from both newspapers and nightly news broadcasts) sampled after days in which the color coded terror alert level had been raised by the Bush Administration. Tangherlini and Leonard analyze (among other things) more than 34,000 Danish folk legends. Jockers and Mimno use a corpus of over 3,000 British, American and Irish 19th century novels, Marshall studies more than 3,000 post-war academic journal articles (written by British and French demographers) while Mohr, Wagner-Pacifici, Breiger and Bogdanov have a corpus that consists of eleven official National Security Strategy documents (containing about a half million words).

What do the topic modelers get from topic modeling all this data? Both as a way to introduce the papers and also to help us think more deeply about these methods, we will ask three questions of each article — what topics have they found? What are the meanings and understandings that the authors attribute to the topics? And how are the topic data deployed to help advance a specific research agenda?

The first two articles provide broad introductions to the method. DiMaggio, Nag and Blei investigate the controversies that erupted over U.S. federal funding of the arts during the 1980's and 1990's. They use an LDA algorithm to code 7,958 newspaper articles selected from five newspapers (culled for stories relevant to the subject published between 1986-1997). They suggest

that when applied to data of this type, topic models provide a useful way to measure what social scientists have generally called frames. DiMaggio, Nag and Blei define a frame as "a set of discursive cues (words, images, narrative) that suggests a particular interpretation of a person, event, organization, practice, condition, or situation" (this issue, p. xx). Media frames are important because they are powerful interpretive devices that "prime particular associations or interpretations of a phenomenon in a reader" (p. xx). Different media frames are promoted by different institutional actors as a way to try to influence the course of public discourse or the shape of political debate.

DiMaggio, Nag and Blei ask for twelve topics when they model their corpus. Looking at their results, we see that some of their topic-frames capture what appear to be generic news discussions of the arts, for example, one concerns "all kinds of musical performances and organizations," another describes "museum exhibits and visual arts" but other topics clearly reflect more politicized frames, such as the "NEA grant controversies" or "1990s culture wars" (for all these examples, see Table 1). By mapping out the distribution of these different topic-frames both across types of newspapers and across time, DiMaggio and colleagues are able to use topic model coding of their corpus as a tool to answer basic questions about the changing dynamics of policy debates for public support of the arts during this volatile decade. In the process they also provide what is probably one of the best introductions to the use of LDA topic modeling for social scientific research.

To help us better demonstrate these methods we ran an LDA model on the articles published in this special issue. Of course this is a much smaller corpus than the techniques were designed for, but we think that even at this scale it can be a useful exercise. After exploring some alternatives, we settled on a 25-topic model. Table 2 presents our results. The leftmost column lists

Table 2.
Topic distribution across articles published in *Poetics* Vol. 41, no. 6. (Articles listed by first author)

Topic and its Description	DiMaggio	McFarland	Miller	Bonilla	Mohr	Marshall	Tangherlini	Jockers
Doc Word Count (tot) 92,260 =	*18,440*	*8,394*	*12,013*	*8,285*	*12,756*	*12,174*	*12,345*	*7,853*
T-1 Engaging the canon	0.0008	0.0030	0.0001	0.0005	0.0002	0.0039	**0.0466**	0.0019
T-2 Predicting economic expectations	0.0168	0.0002	0.0001	**0.0441**	0.0013	0.0019	0.0001	0.0019
T-3 Archives & struggles	0.0001	0.0055	0.0001	0.0002	0.0002	0.0010	**0.0430**	0.0025
T-4 Identification & extraction of nouns	0.0003	0.0012	0.0008	0.0005	0.0035	0.0022	**0.0308**	0.0002
T-5 Computer models of language	0.0001	0.5225	0.0008	0.0002	0.0527	0.0001	0.0001	0.0002
T-6 Forgotten versions	0.0002	0.0023	0.0023	0.0002	0.0039	0.0121	**0.0244**	0.0005
T-7 Earlier efforts	0.0003	0.0005	0.0070	0.0008	0.0002	0.0072	**0.0181**	0.0002
T-8 Topic models	0.2720	0.3492	0.2546	0.1574	0.2201	**0.4186**	0.1976	0.1789
T-9 Bauditz largely deliberately missing	0.0001	0.0080	0.0027	0.0002	0.0002	0.0086	**0.0428**	0.0022
T-10 Anniversary result	0.0122	0.0002	0.0020	**0.0300**	0.0028	0.0001	0.0001	0.0029
T-11 themes, authors, gender	0.0010	0.0062	0.0003	0.0002	0.0125	0.0001	0.0019	**0.2307**
T-12 Original, arbitrary and begrudgingly famous	0.0003	0.0023	0.0008	0.0005	0.0024	0.0001	**0.0302**	0.0005
T-13 Author turn began	0.0005	0.0002	0.0033	0.0011	0.0006	0.0031	**0.0466**	0.0015
T-14 Frames for coverage of Art news	**0.3896**	0.0027	0.0001	0.0254	0.0106	0.0013	0.0001	0.0019
T-15 Great Britain, WWII & uninformative topics	0.0003	0.0002	0.0029	0.0002	0.0024	**0.0471**	0.0005	0.0005
T-16 Studying the media effects of terror alerts	0.2091	0.0002	0.0010	**0.7197**	0.0103	0.0001	0.0003	0.0008
T-17 Films & meanings	**0.0706**	0.0012	0.0001	0.0018	0.0088	0.0004	0.0001	0.0002
T-18 Novels as bags of character names	0.0010	0.0116	0.0057	0.0002	0.0140	0.0147	0.0157	**0.0328**
T-19 Standard, relational, predicted, occurrences	**0.0212**	0.0002	0.0001	0.0133	0.0144	0.0001	0.0001	0.0015
T-20 Authors' Gender	0.0003	0.0119	0.0038	0.0015	0.0088	0.0004	0.0037	**0.5275**
T-21 Crime, banditry, unrest & rebellion	0.0002	0.0002	**0.6503**	0.0005	0.0062	0.0004	0.0001	0.0002
T-22 Communities of authors: Research on literary passages & demography journals	0.0003	0.0350	0.0048	0.0002	0.0013	**0.4525**	0.4417	0.0002
T-23 Honor, position & conscience	0.0009	0.0005	0.0079	0.0008	0.0002	0.0025	**0.0288**	0.0002
T-24 Banditry as an ontological question	0.0001	0.0002	**0.0477**	0.0002	0.0002	0.0101	0.0230	0.0002
T-25 Texts, meaning & national security	0.0023	0.0347	0.0008	0.0005	**0.6224**	0.0109	0.0037	0.0103
Column Total	**1.0006**	**.9999**	**1.0001**	**1.0002**	**1.0002**	**.9995**	**1.0001**	**1.0004**
	x ≥ .25	.25 > x ≥.10	.10>x≥.020	**RowLargest**				

the topics, the other columns report the probability that a word in a given article will have been drawn from the topic in each row (note that the columns sum to 1).[6] Reading down the first column of data, we can see that just a handful of the topics had a very high probability of occurrence in the DiMaggio, Nag and Blei article. Only five of the twenty-five topics have a probability greater than .025. Topic 14 (which we have labeled "Frames for coverage of art-news") is the most important topic in this article (words have nearly .4 probability of being "on this topic"). As the label suggests, this is a word constellation that captures the main intellectual themes of the article; it is defined by terms like —topic, arts, assigned, Times, NEA, art, coverage, grants, York, frames, prevalence, culture, funding, newspapers, government, and controversial.[7] Notice that none of the other articles in the special issue discuss Topic 14 (Bonilla and Grimmer, the only other newspaper study, has the highest probability at just over .025). This illustrates an important (but not surprising) result of our analysis, most of the topics that we have identified are unique to a specific paper.

In fact just a few of the topics are shared across the articles and the only topic that is shared across *all* of the article is the subject of this special issue. Described by words such as: topic, topics, words, model, analysis, corpus, time, texts, terms, related, models, modeling, documents, social, results, word, number, document, Topic 8 captures the topic of topic modeling itself. Words chosen for the Marshall article (which devotes extra attention to the question of how to go about choosing the proper number of topics) have more than .4 probability of being generated (introduced into the paper) from Topic 8. At the other end of this scale is the article by Bonilla and Grimmer with a probability of .1574 for Topic 8, a reflection of the fact that much of that article is not concerned with topic models at all but rather with survey data (that was serendipitously collected during the

---

[6] To be precise, since the number of texts (eight) is small, we trained the model first by running it with each paragraph in the corpus as a separate document. Then we ran each of the eight complete documents against this existing LDA model, asking for the probability of each topic occurring in the eight whole documents.

[7] Here and elsewhere, likely capitalizations of the words have been added by us — the actual terms used in the analysis were not case specific.

same time periods and which Bonilla and Grimmer brilliantly use as a way to assess the effects of the media framing that they show—using topic models—is linked to the escalating terror alerts).

Words in the DiMaggio article have a probability of .2720 of being on the topic of topic modeling. They also have .2091 probability of being linked to Topic 16. This is more curious since it is labeled "Studying the media effects of terror alerts" suggesting that it is the topic that captures the thematic focus of the Bonilla and Grimmer article (and in fact, words in that article have a probability of .7197 of being on this topic). But if we dig a little deeper into the list we also see words like — percent, arts, attention, support (note: the top 8 words for each topic are listed in Appendix 1). Looking further down the list adds — media, Bush, increase, stories, articles, news, figure, effect, policy, percentage, terms, surveys. Having traced these words and the identified paragraphs back into the articles leads us to suggest that the hybridity of this topic (the mixing of paragraphs from the DiMaggio and the Bonilla articles) reflects the fact that, beyond its first few words concerning the public terror alerts, Topic 16 is also capturing a broader, shared discussion on "media effects research." In other words, the algorithm is finding common passages about the use of news stories, studied statistically, that are linked to the study of public policy. When seen from this perspective, the connection between the DiMaggio and the Bonilla articles makes sense.

The DiMaggio article is also linked to Topic 17 (.0706). The most important terms here include: film, solutions, produced, films, museum, Hollywood, meanings, core, solution, percent, observed, appendix, cases, today, university, independent. This is an interesting case because the topic is capturing an extended discussion in the article (that continues into an appendix) about how topic models respond to the difficult analytic problem of polysemy. DiMaggio, Nag and Blei explain that the word "film" is used in several of their topics but that the word has different meanings in the different thematic contexts, thereby helping to validate the power of the LDA method.[8] The last

[8] In each case, we have gone back to the actual texts (and textual contexts) and inspected the mapping of words and topics in order to facilitate our interpretation of these topic word lists. The reader can approximate this

topic of any note is Topic 19, (with a probability of just .02). The first four words—standard, relational, predicted, occurrences—suggests a kind of mantra for the style of formalization being discussed in this article (and elsewhere — note that both the Mohr and the Bonilla articles resonate to this theme).[9]

The McFarland, Ramage, Chuang, Heer, Manning and Jurafsky article reports on a stream of work that the group has published over the last few years employing various types of topic model as well as other text mining methodologies. Their paper provides another useful introduction to topic models by focusing on some alternative types and applications of the method. The main research described here analyzes 30 years of dissertation abstracts (1980-2010) drawn from the Proquest database (of 240 U.S. research universities). They use topic models to identify intellectual streams in this data. A topic identifies constellations of words that co-occur inside the discourse of an intellectual sub-field. For example, looking just at the data from anthropology, their model identifies one topic (they label it "Archeology") which is associated with these (stemmed) terms of art: site, archaeology, period, region, evid, pattern, late, popul, settlement, materi, suggest, valley, earli, prehistory, etc. A different topic/discourse frame (labeled "Identity studies") is defined by terms such as: ident, practice, discours, culture, nation, construct, global, etc. By observing the flow of these topics across the data, McFarland and colleagues are able to track the differentiation and blending of academic disciplines across time.

From Table 2 we can see that words in the McFarland article are likely to be distributed into just two main topics. More than a third of the words are sorted into the topic model topic (Topic 8). A bit more than half are sorted into Topic 5 which we have labeled "Computer models of language" to capture the main themes of their paper. This topic is defined by words such as: language, LDA,

---

experience themselves as they read along in the special issue by searching for the particular word constellations described here.

[9] Note that the top terms for this topic includes the word box which refers to George Box, the statistician, and also to a box in a figure containing topic model probabilities.

models, field, document, fields, labels, identified, knowledge, domains, label, Ramage, supervised, categories, work, identify, subject, labeled, applied, and anthropology. There are also low probabilities of words being sorted into Topic 22 "Communities of authors: Research on literary passages and demography journals" and Topic 25 "Texts, meaning & national security" (both of which we will discuss presently).

The next three articles focus on the use of topic models as a method of studying forms of state discourse. The paper by Miller uses topic models to analyze an archive of official notes sent to the Chinese emperor during the years of the Qing dynasty (1723-1911). At the time, as Miller explains, terminology for different categories of illegitimate public violence was fluid and the distinctions had great significance up and down the chain of command. Thus the official records on violent social episodes are ambiguous and difficult to interpret. Miller, in a very Foucault-like maneuver, uses topic models to sort through the corpus of 'zouzhe' to identify constellations of words (actually, Chinese characters) that capture coherent classifications of imperial concern, including socially constructed categories of violent offenders.

As Miller puts its, "Of the fifty topics in this model, six are clearly related to areas where the dynastic apparatus encountered illegitimate violence" (this issue, pp. ??). He labels these—crime, unrest, sedition, rebellion, border rebellion, and major rebellion. The topic of "crime" is defined by a constellation of terms that include: crime, case, punishment/sentence, board/ministry, try/interrogate, precedent/sub-statute and behead. Miller describes this as a topic that captures the regular pattern of court proceedings. It differs from the topic/social category of "sedition" which is defined by words like: capture, investigate, confession, case, and teach religion. Miller explains that the latter describes a collection of crimes such as heresy, the printing of banned books and the cutting off of queues. Research applications for these measures are primarily historical. Miller graphs the frequency of these topics across time and in doing so he is able to contribute to our understanding of (at least) three specific issues—crime rates (and their reporting) during 18th-

19th century China, the social processes affecting the cultural construction (and prosecution) of the crime of sedition, and the responses of the state to rebellions (especially during the later years of the Qing dynasty).

Looking at Table 2, only three of our topics apply to Miller's paper. The highest probability topic (.6503) is T 21, "Crime, banditry, unrest & rebellion" which well captures the main themes of Miller's article. The second most important is the topic model topic at .2546. Third is T-24, "Banditry as an ontological phenomenon" (.0477) which is an ambitious, engaging and clearly articulated sub-theme in Miller's article (that interestingly, also resonates with a similar theme running through Tangherlini and Leonard's paper).

Bonilla and Grimmer is the second article concerned with the study of state discourse practices. Their research concerns the Bush administration's color-coded terror alert system and its effects on news coverage of terrorism and on public opinion more broadly. They sample front-page news stories and nightly news broadcasts for the two days before, the day of and then two days after each of the terror alert escalation events. Like the DiMaggio paper, Bonilla and Grimmer primarily see topic models as a way to identify media frames in news-stories but they modify the standard LDA model by forcing every news-story to have just one topic. This allows them to more easily calculate when the terror alerts becomes a central focus. They search for twenty-four topics in their corpus which results in stories (or media frames) such as: "local small business," "law and order," "the Iraq war, " as well as one topic which is focused on terror alerts. This research design allows them to directly measure the effect of the terror alert announcements on the content of what is subsequently reported in the news. Then, by drawing on a series of surveys that happened to be conducted during the same time windows, they are able to also directly assess the impacts of the terror alerts on public opinion regarding matters such as economic expectations and support for Bush administration policy agendas.

In terms of our topic model, the Bonilla and Grimmer article is mostly focused on Topic 16 (.7197), "Studying the media effects of terror alerts" which captures the main thematic focus of their article and is defined by words such as — alert, terror, public, percent, arts, attention, support, media, Bush, increase, stories, etc. (Note, this is the same topic that we looked at earlier because it overlaps with the DiMaggio article). Next is the topic model topic (at .1574), followed by two topics that capture more specific sub-themes in the article, Topic 2 (.0441) "predicting economic expectations" and Topic 10 (.0300) "Anniversary result" (which looks at the effects of 9/11 anniversary commemoration events on news coverage and public opinion).

The last article in this section on the study of the state is by Mohr, Wagner-Pacifici, Breiger and Bogdanov. Data come from a series of publications by the U.S. Office of the President regarding the National Security Strategy of the United States (1990-2010). The goal of the article is to better understand the rhetoric that is used by the U.S. state for describing and characterizing the strategic situation of the world (and the U.S. posture there). To do this Mohr and colleagues draw upon the dramatistic theory of rhetoric developed by the literary theorist Kenneth Burke (a half century ago). Topic models are used here as part of a suite of text mining methods that are applied to measure the different elements within Burke's theory of motives. Specifically topic models are used to measure Burke's concept of a scene, which he defined as the setting in which a dramatistic act occurs. For the Mohr et. al. article then, topic models are used to capture the kinds of thematic scenes that tend to re-occur again and again in the U.S. strategy discussions that unfold around global security. It is within dramatistic scenes that the other elements of Burke's grammar of motives —the acts, actors, agencies and purposes—are combined and combusted. Terrorism is one such thematic scene that emerges from the corpus. Others topic/scenes include economic development, human rights, and military operations.

According to Table 2, most of the Mohr article is focused on just two topics—the topic model topic (.2201) and Topic 25, "Texts, meaning and national security" (.6224) which is defined

by the words—text, states, security, texts, united, scene, national, act, figure, meaning, motives, terms, semantic, documents, basic, terrorism, coding, strategy, acts, and automated. There is also some overlap (.0527) with the McFarland article over the use of the "Computer models of language" theme.

The last three papers in the special issue have in common their use of topic models as a strategy for measuring academic and literary fields. In a wonderful paper reminiscent of a classic style of work in the sociology of knowledge, Emily Marshall compares two academic communities—one French, one British, all demographers—by analyzing the intellectual ideas that they use for constructing theories about the world. Specifically, she explores the differential embrace by the French of the theory of the "demographic revolution" in contrast to the British demographers' commitment to the "demographic transition" theory. To test out the implications of this difference, Marshall collects all articles published between 1946-2005 on the subject of "fertility" ("fecondité" or "natalité" in French) in the British and French flagship demography journals. Then, using correlated topic models (CTM)[10] she identifies 75 topics (or intellectual frames) in both the French and in the British corpora. She hand codes these topics (by closely combing through the texts and passages identified with each topic) to discern which reflect high-fertility subjects (like family planning programs, a preoccupation of the British) or low-fertility subjects (like working mothers, a concern of the French). She then maps these categories of topics across time and context (supplemented with a second set of topic models of newspapers over the same period) to demonstrate the tangible persistence of intellectual frameworks (or logics) in academic communities and the way in which those frames endure even in the face of demographic (e.g., objective) trends that challenge them.

[10] In contrast to LDA models (which assume topics are not correlated across documents) CTM is a variation of topic models that assumes that topics are correlated across texts (Blei and Lafferty, 2007).

In our model, the Marshall paper is split into two themes. Reflecting her attention to the question of how to find the best number of topics, the words in Marshall's article have a .4186 probability of being in the topic model topic. And the most important (.4525) is T22 which we have labeled "Community of authors: Research on literary passages and demography journals." What is most intriguing about this topic is that it is shared almost equally with Tangherlini and Leonard (.4417). After reading through the two texts and relevant (identified) passages again, we weren't that surprised.

The Tangherlini and Leonard article is the second in this section focused on measuring academic and literary fields. Like the McFarland et. al. paper, it reports on several topic modeling projects, in this case, three demonstrations of a procedure Tangherlini and Leonard call "sub-corpus topic modeling" (STM). Their idea is to take a small, well-understood corpus of texts and to use them to provide a training logic that can then be applied to larger, less well understood corpora in order to identify examples of textual passages containing similar literary forms. They describe this is a kind of targeted fishing expedition. In the article, Tangherlini and Leonard offer three different STM experiments.

First , they train their algorithm on two of Charles Darwin's books (*The Origin of Species* and *The Descent of Man*) which they use as "bait" to trawl through the Google Danish books corpus (1860-1920) looking for matches. What they "catch" is a splendid array of fish in which Darwinian ideas are woven into unexpected literary passages. As Tangherlini and Leonard explain, these borrowings are in no sense innocent because Darwin's ideas played a critical role in a bitter intellectual dispute waged in late 19th century Denmark between a progressive looking Naturalism (which admired Darwin's works and held them up as an ideal) and a reactionary Romanticism that held to a theocentric scientism and a conservative political vision. The focused trawl of the STM enables Tangherlini and Leonard to pull entirely unknown works of literature up for display and

examination in a way that begins to fill in a much broader and non-canonical history of these intellectual movements.

In their second example Tangherlini and Leonard again trawl for unknown authors but this time they start out by training their algorithm on a collection of works by three canonical figures (Jacobsen, Schandorf and Drachman) in the Danish literary movement known as the Modern Breakthrough. The STM analysis enables Tangherlini and Leonard to locate a variety of (non-canonical) authors (often women) who were also early innovators in the literary form, all of whom had become lost to modern scholarship. And, finally, in what is perhaps their most intriguing experiment, Tangherlini and Leonard go one step beyond Propp (1958) by applying STM to search for the 'feel of the rural.' This time they trained their algorithm on 34,000 Danish folk tales and returned a series of topics such as "death and churchyards," "shooting and witches," "horses & wagons," "serpents," and "the minister" that turn up in all kinds of interesting ways across a range of other literary genres.

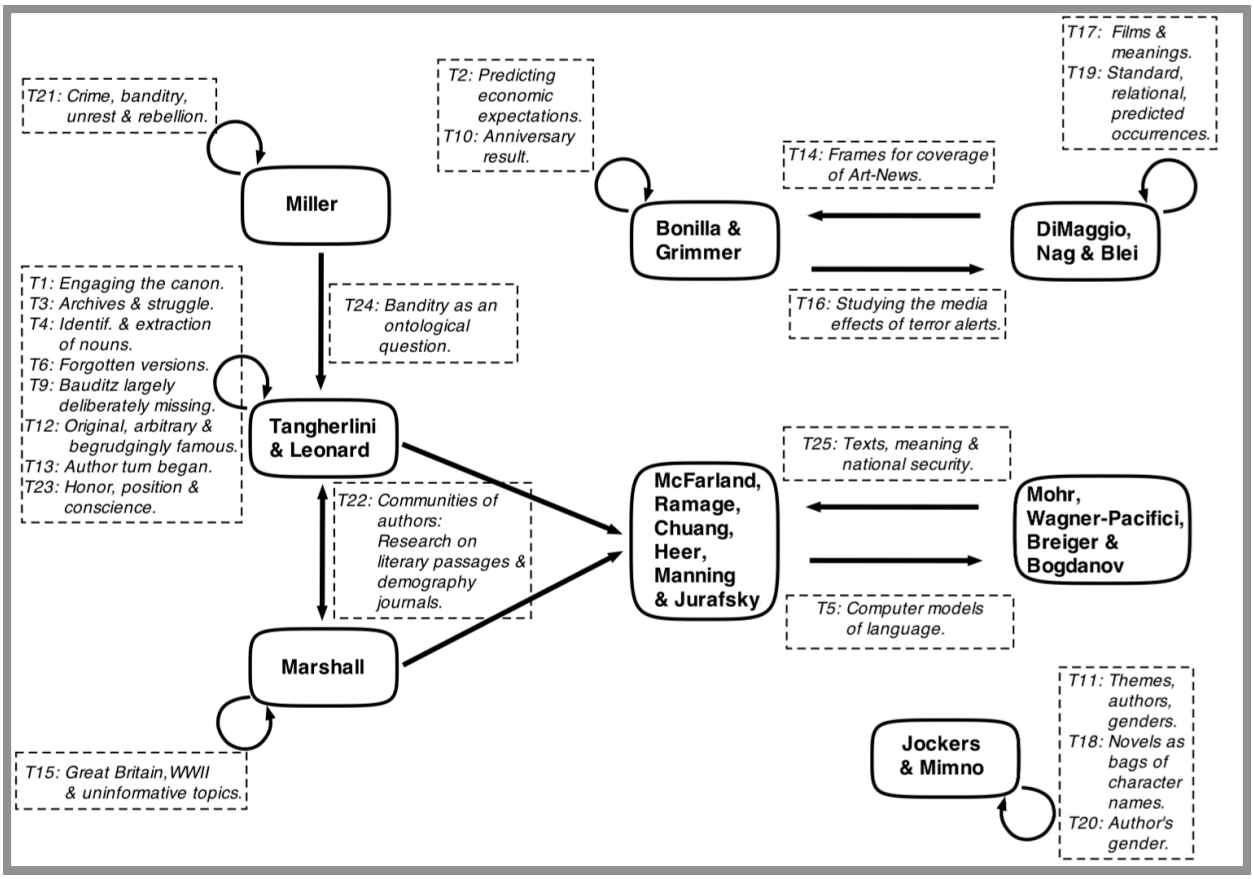
In terms of our analysis, Tangherlini and Leonard are unusual for being the source of so many unique topics (T1, T3, T4, T6, T9, T12, T13, and T23). Some of this reflects the fact that the article is broken into a series of three discrete experiments, each with its own scholarly context and set of intellectual problems (but we suspect we may also be encountering variations in the sensitivity of topic models to different intellectual and rhetorical styles). For Tangherlini and Leonard T 22 (which is the theme shared with Marshall) is also their most important topic (.4417). The hybridity between the two papers was a surprise at first though it makes sense when we recognize that both articles use topic models to identify communities of authors who share common ways of writing and common styles of thinking, both articles focus on national communities of authors, and both address matters concerning alternative styles of scientific thought. Beyond this, however, they are substantively far apart and this is apparent in the key words that are braided together here—fertility, literary, demographic, passages, British,

Breakthrough, journal, research, population, work, articles, Danish, modern, literature, French, works, Darwin, authors, language, academic, etc.

The last paper in this section (and the last in the special issue) is by Matthew Jockers and David Mimno. More than the other papers collected here, Jockers and Mimno are especially focused on calibrating the methodology. Using a clear and intellectually precise research design, they take a corpus of over 3,000 British, American and Irish novels (published between 1750-1899) and sort them into three groups, novels with male authors, novels with female authors and those by authors of unknown gender. Topic models identify coherent literary themes across the corpus (they ask for 500 topics) and then Jockers and Mimno explore the ways in which the gender of an author affects the selection of particular themes. Rather than just looking at simple distributions, however, Jockers and Mimno develop a series of formal assessments—a permutation test, a bootstrap test and a classification test—to assess the reliability of inferences from meta-data for all of these kinds of models. Table 2 shows that the most important themes for this paper are T20 (.5275) which we have labeled "author's gender" and T11, "Themes, authors, gender."

Figure 1 summarizes the information we have presented so far. We have constructed a graph of the articles with arcs drawn to represent shared topics (excluding T8 which was shared by all the articles). With the exception of T22, whenever one article has a higher proportion of a given topic we have represented this with an asymmetric arrow. At the dyad level it is interesting to see that in the both the Bonilla/DiMaggio and the McFarland/Mohr pairings there is a balance to the topic sharing (with each article sharing its main topic with the other article). It is also useful to see the overall mapping of the articles. The Tangherlini/Marshall/McFarland articles are tied together around the communities of authors topic. The Bonilla/DiMaggio pair shares a common focus on newspaper frames and media effects. The McFarland/Mohr dyad has a common focus on applying a range of text-mining tools to tackle problems in the social sciences.

Figure 1. A Mapping of Articles and Topics



2.3. Some demands, dilemmas and limitations of the method

> Content analysis should begin where traditional modes of research end. The man who wishes to use content analysis for a study of the propaganda of some political party, for example, should steep himself in that propaganda. Before he begins to count, he should read it to detect characteristic mechanisms and devices. He should study the vocabulary and format. He should know the party organization and personnel. From this knowledge he should organize his hypotheses and predictions. At this point, in a conventional study, he would start writing. At this point, in a content analysis, he is, instead, ready to set up his categories, to pretest them, and then to start counting.
>
> Harold Lasswell, et. al. (1952, p. 65), *The Comparative Study of Symbols: An Introduction*

The most common complaint that is heard about topic models is that they rely upon the "bag-of-words" assumption, disregarding the order of words within a text (Meeks and Weingart, 2012). To many, it seems hard to believe that one can discard all of that critical

information and not be left with a severely hobbled analysis of meaning. While this is true in some literal sense, it strikes us as being a critique that misses the point because the real genius of topic models is precisely that for this specific type and level of meaningful content, it appears as though relationality trumps syntax. It turns out you *can* remove all of that other information from the analysis and still get robust results. So a better question to ask is what sort of a trade-off shall we make in terms of surrendering this more localized (syntactic and semantic) information in order to realize gains in information of the sort that topic models can afford us?

But this does point to why topic models will be good for some kinds of meaning measurement projects, but be a poor choice for others. So, for example, as scholars from a variety of disciplines have now demonstrated, narratives can be very usefully modeled as tie-based networks— see for example Bearman and Stovel, (2000) for a network analysis of Nazi life-stories, Franco Moretti's (2013) network analysis of Shakespearian plays or even the work of David Herman (2004) who develops a logic model for narratives. All of this suggests that topic models, since they have discarded this type of localized relational information, would be much less useful for studying narrativity. [11]

To us, a more worrisome concern that has also been expressed is the utter simplicity of topic models as a textual analysis method (Grimmer and Stewart, 2013; Schmidt 2012). One might be forgiven for imagining that one needs nothing more than a text to analyze and a copy of a software program like Mallet to produce brilliant cultural research. In truth, as with any scholarly pursuit, the quality of the knowledge about the case and the clarity of thinking about the phenomena determine the utility and the richness of the analysis regardless of the sophistication of the methods employed. With topic models the researcher is responsible for knowing enough about the phenomena under investigation to be able to

---

[11] Then again, we might have also predicted the same would be true of a genre such as poetry but Rhody (2012) finds differently.

understand what the discourse field is about. They must pick a corpus that has substantively meaningful content within the field under investigation and be familiar enough with that corpus to have a good sense of how the text reads and how its contents will address the analytic problem at hand. Moreover, researchers need to be able to make sense of the topic word clusters that are produced by the algorithm and to be able to recognize when a set of topics are worthless or misleading (because, perhaps, there are no well organized topics in the corpus or because the number of topics asked for by the researcher doesn't match the actual number of topics in the corpus, etc.) and when the topics are indeed capturing word clusters that makes good sense to a well informed observer (a subject-area specialist) who understands the discursive context of the corpus.[12] Of course there are also technical requirements—the analyst needs to prepare the text (this might include, for example, removing stop-words, etc.) and meet all the formal assumptions of the model.[13] Researchers must also interpret the topic model output, probably iteratively, so that a best fit can be found between the number of topics and an overall level of interpretability. And finally all of this new topic model data must be fitted into a well-informed, explanatory or substantively meaningful analysis of the social phenomena under investigation.

Seen in this light it is useful to think about topic models not as providing an automatic text analysis program but rather as providing a lens that allows researchers working on a problem to view a relevant textual corpus in a different light and at a different scale. DiMaggio, Nag and Blei use this metaphor and, as they note, it is a frame with its roots deep in the analytic process itself, "(f)inding the right lens is different than evaluating a statistical model based on a population sample. The point is not to estimate population parameters correctly, but to identify the lens through which one can see the data most clearly" (pp. xx this issue).

---

[12] Some progress is being made on developing more formal decision rules for goodness of fit of different levels of topic. Emily Marshall presents some ideas about this in her contribution to this special issue. See also the suggestions here by Bonilla and Grimmer.

[13] Another assumption is that documents within the corpus are independent, i.e. each text is generated by the author without the knowledge/reference or temporal dependencies to the rest of the documents in the corpus. While addressing such assumptions is essential in certain analysis endeavors, they also keep the models simple and general and do not require expert/human knowledge or additional information.

One implication is that well informed interpretive work — hermeneutic work — is still required in order to read and to interpret the meanings that operate within a textual corpus even when one is peering through the lens of a topic model. It is not the need for a deep understanding of one's textual corpus that has changed, it's the place where this style of knowledge comes into play. We began this section with a quote from Harold Lasswell and his colleagues (1952) about the importance of obtaining a deep historical and contextual understanding of one's corpus before beginning to count. With topic models, this is inverted. One counts and then one begins to interpret. In this sense, what topic models and other types of automated text analysis tools do for cultural researchers is to shift the locus of subjectivity within the methodological program — interpretation is still required, but from the perspective of the actual modeling of the data, the more subjective moment of the procedure has been shifted over to the post-modeling phase of the analysis.[14]

3. Why do topic models matter?

> Content analysis will not tell us whether a given work is good literature; it will tell us whether the style is varied. It will not tell us whether a paper is subversive; it will tell us if its contents change with the party line. It will not tell us how to convince the Russians; it will tell us what are the most frequent themes of Soviet propaganda.
>
> Harold Lasswell, et. al. (1952, p. 45), *The Comparative Study of Symbols: An Introduction*

Topic models matter for a lot of reasons. Most obviously they matter because they provide a way for researchers to obtain reasonable automated content coding of large text corpora. As social and cultural scientists become increasingly engaged with what is now being called Big Data — large-scale data streams taken from the internet, social media sites,

---

[14] Topic models are certainly not the first. The social sciences have used many such types of methods in the past. Starting back with Lazarsfeld's theory of latent factor analysis, and the use of methods such as factor analysis, multi-dimensional scaling or Multiple Correspondence Analysis (MCA) as used in Europe.

or archives like Google books—having tools that scale becomes increasingly important. Thus topic models matter because they enable us to take the measure of large-scale social phenomena that we could not have previously been able to do. Whether our goal is to study attitude change in twitter feeds (Ramage, Dumais and Leibling, 2010) or genre shifts in literary fields (Jockers, 2013; Moretti, 2013), topic models matter because they enable researchers to study phenomena of the sort that can only be viewed through a macroscopic lens.

But topic models also matter because they can used for viewing small-scale text corpora. In the Mohr, Wagner-Pacifici, Breiger and Bogdanov article (this issue), topic models are one of three text analysis methodologies that are combined to study a relatively small corpus (of a half million words) that is well within reach of a traditional 'close reading' by experts in hermeneutics and relevant subject areas. Here formalization supplements (rather than displaces) a close reading of the corpus. So, again, topic models matter because they provide new lenses for new projects.

As we have sought to highlight here, topic models also matter because they facilitate a fundamental shift in what we have called the locus of methodological subjectivity—from pre-counting to post-counting. This is another major reason why topic models matter and just to emphasize this quality we turn to one last set of examples that can usefully illustrate this contribution. This comes from a study of a corpus of 20,000 newspaper editorials sampled from ten major newspapers (in five countries) over a sixty-year period. Researchers identify a number of topics but we will focus here on just two. The first is "International Violence." It is defined by the terms—war, combat, battle, weapons, enemy, front, trench, foxhole, prisoners, soldiers. The next is labeled "Domestic Violence" and it includes the terms—riot, murder, strike, disorder, pickets, suicide, prison, jail, lynching, gangs (Lasswell, et. al. p. 68).

The example comes from the research done by Harold Lasswell and his colleagues at Stanford in the years just after World War II. The project had the goal of identifying "trends in the key symbols of modern politics" between the years of 1890 and 1945 (Lasswell, et. al. p.iii).[15] No computers were used to identify these topics. Instead, using methodologies that Lasswell had pioneered as director of "the experimental division for the study of wartime communications, established at the Library of Congress during World War II" (Lasswell, Lerner and Pool, 1952, p. 40), he and his colleagues assembled a team of human coders and (to insure consistent coding of the text), they created a coding protocol (they called it a rulebook) that channeled the interpretive focus of the coders down to a narrow range of explicitly pre-considered choices and clear decision rules.

Their method was highly dependent upon adequate pre-specification of the "key symbols" that were to be coded. Lasswell divided these into three types — those referring: "to persons or groups (symbols of identification), to preferences and volitions (symbols of demand), and to the assumption of facts (symbols of expectation)" (p. 15). Coders were instructed to check each editorial for "the presence of any of 416 symbols which constituted our symbol list. Of these, 206 were the names of national or similar units: countries, national minorities, continents, etc.; and 210 were key symbols of the major ideologies which have been contending in world politics during the past half-century. These included, to cite the "N's" as an example, Nationalism, Nazism, Neutrality, and Nonintervention" (1952, p. 43).

---

[15] In his classic essay "Why be quantitative?" Lasswell (1949) describes his frustrations at seeing so many otherwise interesting and important detailed analyses of texts that were nonetheless suspect precisely because "…we are left in the dark about why he quotes one paper one day or week and omits it the next time. Even if we assume that his judgment is good, it is permissible to ask if such arbitrary selection procedures create a properly balanced picture, or whether they result in special pleading based, if not on deliberate deception, then on unconscious bias" (p.44). Rogers (1994) has a useful review of Lasswell's career.

As we have seen in the passage cited earlier, Lasswell and his colleagues worried a lot about the processes, time and effort that went into establishing their lists of key symbols. They had good reason to worry. Once the coding categories were negotiated, pre-tested, written into the rulebook and the team had begun to code the corpus, there was no going back—no chance to re-code, re-compile or re-run. This meant that Lasswell and his team had to establish a deep knowledge of the case and do so well before the counting began. To do this, they crafted a six-step process that culminates in "…constructing our tentative symbol list. This we can do partly on the a priori basis of our reflections on the past and present, partly on the empirical basis of our preliminary scrutiny of the media to be analyzed" (p. 68).[16]

But to what end? After orchestrating a text analysis project of this magnitude and precision, what did Lasswell and his colleagues want? Interestingly, it seems that what they really wanted was something like a topic model. Consider the project described here. We noted that Lasswell started by having his coders track 416 key symbols. On closer inspection we find that the key symbol list begins to look a lot like processed lists of terms in which stop-words have been removed, stemming has occurred, and synonymic terms have been collapsed together.[17] And once the corpus had been coded according to this scheme, what next? On this, Lasswell and colleagues are clear. Although they lament the lack of viable theoretical models for understanding how to model idea structures, they insisted on the importance of advancing on the problem with empirical research,

> Today we have no models at all and, therefore, no basis for predicting how symbols will behave under specified conditions…We do have some models of attitude formation, propaganda effects, and ideological behavior… But we should not confuse theories about

---

[16] The six steps—"First, decide which segments of the population we wish to test for this particular change in symbolic behavior…Next, select…a representative medium of symbolic behavior…Third…estimate roughly the period to be covered…we should next set up a tentative scheme of periodization…fifth…state our hypothesis with sufficient definiteness to enable us to construct the list of symbols which would index it…With these propositions before us, we can take the sixth step of constructing our tentative symbol list" (pp. 67-68).

[17] Indeed, those techniques, matched with a named entity recognition (NER) program might be able to provide a list of key symbols that come pretty close to what Lasswell was after. The Mohr, Wagner-Pacifici, Breiger and Mohr paper (in this issue) explains NER processors in more detail.

> ideas with theories about symbols…Ideas are expressed by symbols. Their manifest form is nothing more than a conglomeration of symbols…We need models of how symbols operate to produce the configurations called ideas, attitudes, ideologies…And our knowledge of symbolic behavior can be advanced only if we learn how ideas take form out of the symbolic elements through which they are expressed
>
> (Lasswell, Lerner and Pool, 1952, pp. 64-65).

In other words, for Lasswell and his colleagues, the reason to have teams of human coders track 416 symbols across ten newspapers (from five countries across sixty years) was so that these data could be used to identify larger structures of meaning that could then be linked back to broader research agendas.

> If the list of symbols is sufficiently extensive, it will be found that groups of symbols follow common patterns. It will be possible, in other words, to apply a sort of factorial analysis to the list, which one will find that the large number of symbols occurring do not each represent an independent variable, but that groups of symbols form constellations, certain words appearing together. The independent factor is an idea to which the group of symbols refers and whose fluctuations it indexes.
>
> (Lasswell, Lerner and Pool, 1952, p. 55)[18]

Thus for Lasswell and his team members, the goal was to find ways to measure *ideas* which were latent constructs indexed by constellations of word *symbols*. They worked at developing a model for capturing this process, but they didn't succeed. "Criteria for the validity of a list may now be stated more formally, although the statistical working out of the procedure remains to be done" (p. 56). They concluded somewhat optimistically,

> "Symbolic behavior seems to be prone to factorial analysis, since a limited number of unit ideas fall into complex constellations. It seems unlikely that the probability of the appearance of symbols with respect to each other and over time could be represented by a regular surface. We hope that statisticians will address themselves to testing these hunches and resolving some of these problems.
>
> (Lasswell, Lerner and Pool, 1952, p. 57)

Too bad for Lasswell, he was born half a century too early to be able to make use of LDA models to analyze his corpus. What then did he do? They used their best guesses. "It should be

---

[18] Lasswell et al go on to sketch the basic model of a latent factor analysis model for content analysis by drawing on Lazarsfeld's ideas about latent factor analysis in survey research.

noted that, for certain studies, an a priori list remains most appropriate." And so, in fact, our last set of topic models examples —"International Violence" and "Domestic Violence" —were not models at all, but a set of best guesses about how to go about assembling a set of index measures by hand from a dataset about which the researchers already know a great deal. After having gone through six steps of preparation, Lasswell and his colleagues write, "we can quickly think up several dozen symbols which most Americans associate with violence. Here are two groups which occurred to these writers, by free association, within a few minutes" (p. 68).

What is striking to us today is just how much Lasswell, at the very beginning of the modern field of content analysis, began with the goal of assembling a set of text analysis measures that end up looking a lot like what topic models deliver. In that sense, we might say the creators of topic models have stepped up to Lasswell's challenge. But of course, topic models do a lot more than solve Lasswell's statistical problem. In fact, with topic models the entire process of creating the code lists, writing the rulebook as well as the actual coding of the corpus itself are replaced by a set of automated algorithmic procedures. One might say this puts content analysis back on an equal footing with traditional modes of scientific research, no longer must content analysis start where all other methods end (as Lasswell had warned). But, as we have also tried to emphasize here, topic models do not remove the scholarly or the hermeneutic work from the project of analyzing a textual corpus, it simply moves the bulk of this labor over to the other side of the data modeling procedure. And so one last way that topic models matter is that they—in this sense—represent something of an ending to a first chapter in the history of content analysis methodologies and the beginning of another.

5. Conclusion

> …the amount of wasted effort will be much less with adequate preparation of the sort we recommend.
>
> Harold Lasswell, et. al. (1952, p. 66), *The Comparative Study of Symbols: An Introduction*

In this essay we have described what we see as the important features of topic models for scholars in the social and cultural sciences who might want to choose to use this method. Of course there are also limitations and caveats (and we have discussed some of those here) but it is clear that to the extent that topic models prove to be an effective way of coding the meanings inside text corpora, then these are methods that can provide a way to analyze texts (including "Big Data" texts) that is substantively quicker, more efficient and more objective than traditional methods of content analysis in the social and cultural sciences.

We have gone to some lengths to trace out the parallels between the work of the founder of modern content analysis methods, Harold Lasswell, and new developments in this emerging field of contemporary topic modeling. While the two projects may have initially seemed quite far removed from one another, we have sought to demonstrate that they are in fact perfect bookends to one period in the history of modern content analysis methodologies, a period that got its start (as so many other modern social scientific methodological programs) in the crucible of applied social science during World War II.[19]

What Lasswell and his colleagues initially invented as a set of procedures for human beings has now been fully supplanted by a set of algorithms. And though this really does—in a profound sense—change everything, many of Lasswell's precautions and concerns remain with us still. We still need to have learned well about the case. We still need to think clearly and analytically about the connections between the measure of textual content and the way in which these measures articulate into other types of social structures. And in this, notice that the ambition of content analysis researchers continues unabated. Lasswell and his colleagues saw this as the real end and ultimate goal of content analysis and they described this style of work as "interaction analysis."

---

[19] Mohr and Rawlings (2010) discuss some of the ways that other formal models of culture emerged during this historical period. See also Platt (1996) for a broader historical review.

They write, “The aim of interaction analysis is to associate the flow of symbols directly with the flow of events. In the ideal case, fluctuations with respect to a single type of event…could be correlated with fluctuations in treatment of a single type of symbol” (p. 38). They also warn us, “This is the most difficult use of content analysis” (p. 38). We agree but we also believe that ultimately, the goal of modern content analysis should be to emphasize this very kind of interaction (or duality) analysis and, in so doing, to help to rebalance the social sciences, by bringing the formal study of culture and meaning back into some form of parity with the quantitative study of social structures and material logics that have generally been ascendant since about the time that Lasswell and his colleagues were writing (Mohr, 1998).

Acknowledgements

We thank Timothy Dowd for his superb editorial support both on this essay as well as on all of the papers collected in this special issue (and for his patience). We also would like to thank the members of the Department of Sociology at the University of Barcelona where we presented an early version of this paper. Special thanks to José A. Rodríguez, José Luis C. Bosch, Liliana Arroyo and Jesús de Miguel for their useful feedback and suggestions.

References:

Bearman, P., Stovel, K., 2000. Becoming a Nazi: a model for narrative networks. Poetics 27, 69-90.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993-1022.

Blei, D.M., Lafferty, J, 2006. Dynamic topic models. In International Conference on Machine Learning, pages 113–120, New York, NY, USA, 2006. ACM.

Blei, D.M., Lafferty, J.D., 2007. A correlated topic model of *Science*. Annals of Applied Statistics 1(1), 17–35.

Blei, D.M., Lafferty, J.D., 2009. Text mining: Classification, clustering, and applications. In: Srivastava, A. and Sahami, M., (Eds), Text Mining: Classification, Clustering, and Applications. Taylor & Francis, London, UK, pp. 1-24.

Blei, D.M., 2011. Introduction to Probabilistic Topic Models, Princeton University.

Blei, D.M., 2012a. Topic modeling and Digital Humanities. Journal of Digital Humanities 2(1), 8-11.

Blei, D.M., 2012b. Probabilistic topic models. Communications of the ACM 55 (4), 77-84.

Breiger, R.L., 2000. A tool kit for practice theory. Poetics, 27(2-3), 91-115.

Brett, M.R., 2012. Topic modeling: A basic introduction, Journal of Digital Humanities 2(1), 12-16.

Chang J. and Blei D.M., 2010. Hierarchical relational models for document networks. Annals of Applied Statistics, 4(1),?.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R. 1990. Indexing by latent semantic analysis, Journal of the American Society of Information Science 41(6):391-407.

Diesner, J., Carley, K.M., 2005. Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. In: Narayanan, V.K., Armstrong, D.J. (Eds.), Causal Mapping for Research in Information Technology. Idea Group Publishing, Hershey, PA, pp. 81-108.

Diesner, J., & Carley, K. M., 2008. Conditional Random Fields for Entity Extraction and Ontological Text Coding. Journal of Computational and Mathematical Organization Theory, 14, 248 – 262.

Diesner, J., Carley, K.M., 2010. A methodology for integrating network theory and topic modeling and its application to innovation diffusion. In proceedings of Social Computing (SocialCom), 2010 IEEE Second International Conference on Social Computing, Minneapolis, Minnesota, August 20-22, 2010, Pp. 687-692.

Griffiths, T., Steyvers, M., 2004. Finding scientific topics, National Academy of Sciences, 101 (suppl. 1) 5228-5235.

Griffiths, T., Steyvers, M., Blei, D.M., and Tenenbaum, J. 2005. Integrating topics and syntax. In L. K. Saul, Y. Weiss, and L. Bottou, editors, Advances in Neural Information Processing Systems 17, pages 537–544, Cambridge, MA, 2005. MIT Press.

Grimmer, J., 2010. A Bayesian hierarchical topic model for political texts: measuring expressed agendas in senate press releases. Political Analysis 18(1), 1-35.

Grimmer, J., King, G., 2011. General purpose computer-assisted clustering and conceptualization," Proceedings of the National Academy of Sciences, 3 February (2011).

Grimmer, J., Stewart, B., 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political documents Political Analysis, 2013. 21 (3), 267-297

Herman, D., 2004. Story Logic. University of Nebraska Press, Lincoln, NB.

Hofmann, T., 1999. Probabilistic latent semantic analysis. In Uncertainty in Artificial Intelligence (UAI).

Jelisavcic, V., Furlan, B., Protic, J.; Milutinovic, V., 2012. Topic models and advanced algorithms for profiling of knowledge in scientific papers," *MIPRO, 2012 Proceedings of the 35th International Convention* , vol., no., (May) pp.1030-1035.

Jockers, M.L., 2013. Macroanalysis: Digital Methods and Literary History. University of Illinois Press, Urbana, IL.

Kaplan, S., Vakili, K. 2012. Breakthrough innovations: Using topic modeling to distinguish the cognitive from the economic. Unpublished manuscript, Rotman School, University of Toronto.

Lasswell, H.D. 1949. Why be quantitative? In: Lasswell, H.D. and Leites, N. (Eds.), Language of politics: Studies in quantitative semantics. George W. Stewart, Publisher, New York, NY, pp. 40-52.

Lasswell, H.D. and Leites, N. (Eds.) 1949. Language of politics: Studies in quantitative semantics. George W. Stewart, Publisher, New York, NY.

Lasswell, H.D., Lerner, D., Pool, I.d.S., 1952. The Comparative Study of Symbols: An Introduction. Stanford University Press, Palo Alto, CA.

McNamara, D.S., 2010. Computational methods to extract meaning from text and advance theories of human cognition, Topics in Cognitive Science 1-15.

Meeks, E. and S. Weingart, S., 2012. The Digital Humanities contribution to topic modeling. Journal of Digital Humanities 2(1), 1-6.

Mohr, J.W., 1998. Measuring meaning structures. Annual Review of Sociology 24, 345-370.

Mohr, J.W., Bogdanov, P., forthcoming. Collection and analysis of relational data from digital archives. In: Alhaji, R., Rokne, J. (Eds.), Encyclopedia of Social Networks and Mining. Springer, London.

Mohr, J. W., Rawlings, C., 2010. Formal models of culture. In: Hall, J., Grindstaff, L., Lo, M.-C. (Eds.), A Handbook of Cultural Sociology. Routledge, New York, pp. 118-128.

Moody, J. and Light, R. 2006. A view from above: The evolving sociological landscape." The American Sociologist 37(2):67-86.

Moretti, F., 2013. Distant Reading. Verso, London.

Mützel, S. 2012. "Newness and collaborative category construction from stories." Unpublished manuscript. Social Science Research Center Berlin.

Osgood, C.E., Suci, G., & Tannenbaum, P., 1957. The Measurement of Meaning. University of Illinois Press, Urbana, IL.

Newman, D., Hagedorn, K., Chemudugunta, C., Smyth, P., 2007. Subject metadata enrichment using statistical topic models. Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 366-375.

Platt, J., 1996. A History of Sociological Research Methods in America, 1920-1960, Cambridge University Press, Cambidge, UK.

Propp, V. 1958. Morphology of the Folktale, University of Texas Press, Austin, Tx.

Ramage, D., Dumais, S, Liebling, D. 2010. Characterizing microblogs with topic models. Proceedings of the Fourth International AAAI Conference.

Rhody, L. 2012. Topic modeling and figurative language. Journal of Digital Humanities 2(1), 19-35.

Rogers, E. 1994. A history of communication study: A biographical approach. The Free Press, New York, NY.

Saussure, F. 1959. Course in General Linguistics. McGraw-Hill, New York, NY.

Schmidt, B.M., 2012. Words alone: Dismantling topic models in the Humanities. Journal of Digital Humanities 2(1), 49-65.

Teh, T., Jordan, M. Beal, M., and Blei, D. 2006. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476):1566–1581.

Wallach, H. 2006. Topic modeling: Beyond bag of words. In Proceedings of the 23rd International Conference on Machine Learning.

Appendix -1

Top 8 words per topic (part 1).

	W-1	W-2	W-3	W-4	W-5	W-6	W-7	W-8
T-1 Engaging the canon	representative 0.008605	understood 0.007753	showing 0.006049	canonical 0.005197	engaging 0.004345	gennembruds 0.004345	bugge 0.003493	accepts 0.003493
T-2 Predicting economic expectations	expectations 0.006762	sufficient 0.006762	manipulations 0.005928	focuses 0.005928	constant 0.005093	overlap 0.005093	uncontroversial 0.004258	limited 0.004258
T-3 Archives & struggles	struggle 0.006889	archive 0.005188	collections 0.004337	revealed 0.004337	position 0.004337	descriptions 0.004337	captured 0.004337	feel 0.004337
T-4 Identification & extraction of nouns	identification 0.005878	remove 0.005878	extract 0.003951	nouns 0.003951	understood 0.003951	live 0.003951	skram 0.003951	health 0.002987
T-5 Computer models of language	language 0.020458	lda 0.012451	models 0.011650	field 0.010850	document 0.008848	fields 0.008848	labels 0.008447	identified 0.006446
T-6 Forgotten versions	forgotten 0.006375	versions 0.005477	closely 0.003681	half 0.003681	independence 0.003681	budgets 0.003681	interpretive 0.003681	pair 0.003681
T-7 Earlier efforts	existing 0.008785	hundreds 0.006855	efforts 0.004924	earlier 0.004924	unrest 0.003958	publication 0.003958	accepted 0.003958	slightly 0.003958
T-8 Topic models	topic 0.059637	topics 0.051214	words 0.026614	model 0.020812	analysis 0.015402	corpus 0.015178	time 0.012278	texts 0.012222
T-9 Bauditz largely deliberately missing	bauditz 0.009259	missing 0.006757	largely 0.005923	deliberately 0.005088	urban 0.005088	contents 0.004254	small 0.004254	aspects 0.004254
T-10 Anniversary result	examining 0.007472	result 0.006651	perspective 0.006651	anniversary 0.005830	ensuring 0.005009	capture 0.005009	fall 0.005009	interpretation 0.004188
T-11 themes, authors, gender	themes 0.031632	thematic 0.019601	corpus 0.014588	author 0.012583	gender 0.011580	work 0.011079	century 0.010076	theme 0.008572
T-12 Original, arbitrary and begrudgingly famous	original 0.005855	instinct 0.005855	arbitrary 0.004895	begrudgingly 0.003935	famous 0.003935	lens 0.003935	tales 0.003935	serpents 0.003935
T-13 Author turn began	author 0.011393	turn 0.004436	began 0.004436	represented 0.003566	interests 0.003566	suggest 0.003566	mechlenburg 0.003566	moretti 0.003566

Appendix -1

Top 8 words per topic (part 2).

	W-1	W-2	W-3	W-4	W-5	W-6	W-7	W-8
T-14 Frames for coverage of Art news	topic 0.048580	arts 0.027056	assigned 0.016984	times 0.015467	nea 0.011741	art 0.009672	coverage 0.009396	grants 0.008016
T-15 Great Britain, WWII & uninformative topics	frequently 0.006631	great 0.005697	wwii 0.005697	discuss 0.004763	treated 0.004763	program 0.004763	uninformative 0.004763	probability 0.004763
T-16 Studying the media effects of terror alerts	alert 0.027741	terror 0.024073	alerts 0.020404	public 0.014793	percent 0.012419	arts 0.011556	attention 0.011232	support 0.010261
T-17 Films and meanings	film 0.026118	solutions 0.011061	produced 0.010482	films 0.010482	museum 0.010482	hollywood 0.009324	meanings 0.009324	core 0.007586
T-18 Novels as bags of character names	names 0.017345	novels 0.007584	frequently 0.006833	character 0.006833	bag 0.006833	reach 0.006082	fiction 0.005331	influence 0.005331
T-19 Standard, relational, predicted, occurrences	standard 0.007113	relational 0.005357	predicted 0.005357	occurrences 0.004478	london 0.004478	contributes 0.003600	confidence 0.003600	box 0.003600
T-20 Authors' Gender	female 0.029062	male 0.027337	authors 0.026475	figure 0.023025	data 0.021588	topic 0.020726	novels 0.017276	word 0.015551
T-21 Crime, banditry, unrest & rebellion	rebellion 0.029128	crime 0.019148	unrest 0.018317	violence 0.017277	state 0.011872	records 0.009585	bandits 0.009169	major 0.008753
T-22 Communities of authors: Research on literary passages & demography journals	fertility 0.020368	literary 0.012460	demographic 0.011006	passages 0.010188	british 0.010097	breakthrough 0.009916	journal 0.009916	research 0.009825
T-23 Honor, position & conscience	honor 0.009242	societal 0.006497	corpus 0.005582	position 0.004667	conscience 0.004667	original 0.003752	precise 0.003752	tested 0.003752
T-24 Banditry as an ontological question	textual 0.012484	historical 0.011105	phenomenon 0.008346	robber 0.006966	concept 0.006277	accounts 0.005587	bandit 0.005587	ontological 0.005587
T-25 Texts, meaning & national security	text 0.019047	states 0.013456	security 0.013456	texts 0.011965	united 0.010474	scene 0.009356	national 0.009356	act 0.008238